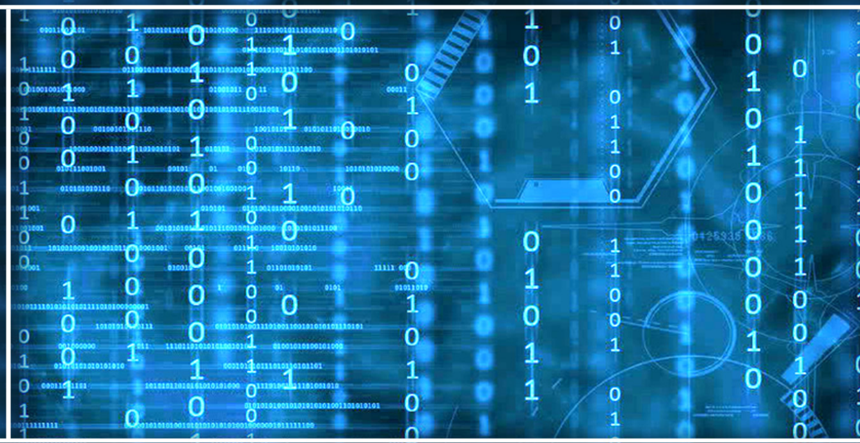
IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 9 Issue 4

April 2018



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

SAI

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 9 Issue 4 April 2018
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

CONTENTS

On Prospects of Development of Telecommunication Systems and Services based on Virtual Reality Technology

Andrey Zuev, Roman Bolbakov
Institute of Information Technologies
Moscow Technological University MIREA
Moscow, Russian Federation

***Abstract*—Virtual reality technologies are considered to be a basis and a promising development trend of telecommunication systems' and services. New opportunities and sci-tech problems that need to be solved are currently undergoing analysis. In the nearest future, the possibility of creating a new segment in the international telecommunication services market that is oriented towards providing communication services to the mass consumer and the entertainment industry, new tools for goods promotion and education content delivery, as well as toward solving experimental problems in managing state structures and social economical systems, is being justified. Within the context of studying the necessary prerequisites to implementing this opportunity the following is being done: an approach to creating and charging for according telecommunication services is being given, opportunities for using the existing national and international cellular connection infrastructure is researched; the prospects of shifting towards new cellular network standards of the fifth generation are being analyzed, employing new ways of connecting to the Internet, as well as new formats of transmitting and processing multimedia, including the provision of immersion in the virtual reality environment; safety aspects of end-user equipment and VR-based telecommunication service exploitation are being updated. As a result of consolidating the given material, the authors are working on building grounds for the rapid development in this direction of telecommunication services and options provided by the Russian national and transnational cellular network operators. The paper was created as part of the project 2.7178.2017/БЧ "Researching Cognitive Semiotics in the Multimedia Virtual Reality Environment".**

***Keywords*—Virtual reality; cellular network; mobile internet; telecommunication service; digital economy; informational society**

I. Introduction

Establishing the knowledge economy, forming the informational community, implementing digital production and developing the digital economy are what is now considered to be the actual future of technologically developed countries. Bringing these trends to life will be achieved based on the fifth and sixth technological tenors [1] and will require significant alterations in the provisionary infrastructure in all areas of societal activities. One of the key elements of this infrastructure is the area of telecommunication technologies that provides all of the telecommunication services - mechanisms and instruments of transmitting and processing the ever-growing volumes of information - cellular network, mobile data, social networks, mobile applications, etc. The composition of telecommunication services delivered by various providers within said services is currently identical and differs only by separate options of their use and pricing policies. On the one hand, this points to high levels of competition, full satisfaction of customer needs and embracing the whole telecommunication services market, and, on the other hand, raises the relevant issue of predicting the future appearance of new needs conditioned by the innovative development of society and economy, leading to the market growth and changes in the structure of its segments.

Developing the telecommunication industry cannot be limited to enhancing existing technology and equipment. Within the context of expected large-scale alterations in the system of social interactions and relations, the functioning of social institutes and economical subjects, the organization of industrial and logistical processes and so forth, it is necessary that telecommunications find the directions for rapid development, the result of which needs to be the timely appearance of new, but to-be-demanded segments of the telecommunication services market [2]. The scientific-technological progress, transition to the fifth tenor and approaching expansion of the advantages of the sixth will require not only employing the new approaches to building telecommunication systems, but will also provide the opportunities for implementing services that are based on new and innovative technologies. From the point of view of economical efficiency, offering new telecommunication services makes sense for the largest segments of the existing market, ones that correspond with the most preferred target audiences - them being internet access and mobile (cellular) connection.

At the moment, regulatory documents have been adopted and programs are being implemented which regulate the development of informational and telecommunication technologies in the RF, including "Digital Economy of the Russian Federation", "Russian Federation's Informational Community Strategy for the years of 2017-2030" and "The Strategy for Developing the Information Technology Industry in the Russian Federation in Years 2014 – 2020 with an Outlook to 2025". However, from the authors' point of view, the search for an answer to the question of what directions of rapid development are possible, needs to be conducted in scientific and engineering communities. One of such directions may be the development of virtual reality (VR) to the extent of employing it as a basis to innovative services that are shaping the segment of the corresponding market. Immersiveness will be the key improvable feature of VR environments, and not only private individuals, but also legal entities will be its target audience, and the area of VR-technology use will include not only the multimedia exchange, but also practical tasks aimed at managing state structures and social-economical systems [3].

The article provides positive results of the analysis made by the authors of the capabilities and prospects of the new segment on world telecommunication services market in VR environment, their architecture and approach to billing approaches. The authors also present the result of designing an immertion telecommunication VR service “Virtual Situational Center”.

II. Telecommunication Services based on Virtual Reality

Currently, the cellular network and mobile data are easily accessible in most of the countries of the world, and both the coverage and audience of these services are constantly growing. National and transnational mobile operators have coordinated the commutation issues between each other, as well as tarification issues regarding the users of local, regional and national infrastructures, and developed architectural and technological solutions can be distributed over the whole world in the future. Any cellular user will become identifiable by their telephone number, and additional telecommunication services based on VR technology may be provided to them through the use of the mobile services already open to them. Let us review the concept of implementing such an additional service based on a typical telecommunication system (TCS) of cellular data:

1) Modern mobile phones and smart phones are able to launch VR applications, these devices are used as screens for helmets (cardboard VR), therefore fulfilling the main function of TCS end-user devices.

2) Interaction with a launched VR application is conducted through the use of controllers similar to Samsung Gear VR, ones that complement the functionality of TCS end-user devices.

3) Send-receive device networks used to connect users to a service are supplied by the already used cellular network equipment - base stations, connection lines, commutation centers, etc.

4) In order for consumers to commute an additional service, one simply needs to add a server to a typical TCS, providing user synchronization throughout the use of a given VR application. Note that this server may also provide the resources necessary for the users who use any kind of Internet access.

Basically, the cellular network operator automatically provides their TCS and mobile data as necessary technical provision for the service in question. At that, they have no grounds for asking the user and/or operator (provider) for additional payments. As a result, the operator’s additional expenses are only connected to supporting the synchronization servers. Apart from mobile phones and smart phones as end-user devices, any specialized VR device, such as HTC Vive or Oculus Rift may be used.

With that, the task of organizing functioning servers that synchronize VR applications remains quite challenging, since it entails the necessity to exploit specialized data processing centers. Defining the requirements for the equipment and software to be used there can be a separate research problem, solving which will call for, among other things, conducting a swries of experiments modeling the stress load put on the structure elements of the telecommunication service by end-user devices.

The general view of the service structure of telecommunication services in VR environments is shown in Fig. 1.

The general view of the VR service server is shown in Fig. 2.

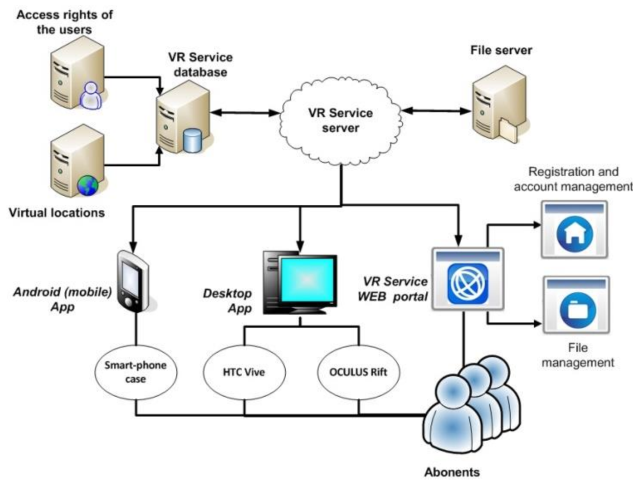


Fig. 1. General view of the provided service structure.

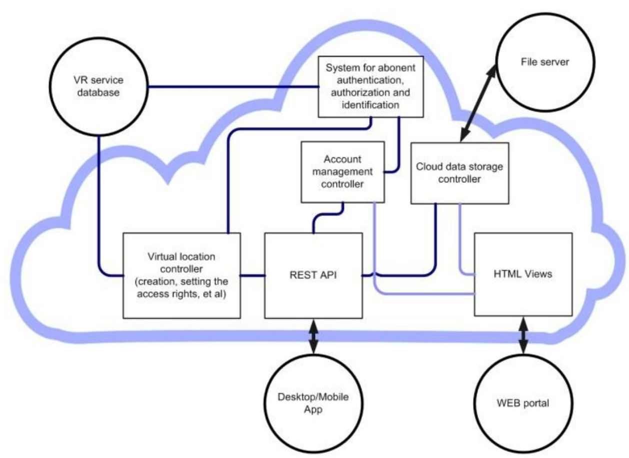


Fig. 2. General view of the VR service server structure.

This description given on the approach to VR relecommunication services implementation and architecture is, of course, cursory and requires detailed elaboration. Their future specification will be one of the studies to be conducted by the authors.

The main factors currently limiting the wide spreading of the VR applications and, accordingly, the main tasks that need to be resolved in order to implement the telecommunication services based on it, are:

1) The necessity of providing sufficient computing powers of end-user devices for the given services while achieving the minimal possible sizes of said devices for the mobility of their users.

2) The necessity of quickly transmitting and processing large amounts of multimedia information.

In the near future, these tasks may have a comprehensive solution with a consideration of development tendencies in data transmission and processing described earlier:

1) Implementation of the fifth generation of connection standards expected in 2020 suggests an increase in data transmission of up to 20 times and reduction of lag time down to 10 times compared to the 4G standards used [4].

2) One of the International Telecommunication Union's (ITU) - a United Nations facility specializing in informational telecommunication technology - directions of activity is developing a specialized MPEG Immersive Media (MPEG-I) standard by 2022, providing not only transmission efficiency, but also its immersiveness [5] - creating a participation effect in the given environment.

3) The biggest players of the telecommunication services market are actively searching for new ways of setting up Internet access:

a) Facebook's Aquila project offers using land-based stations and drones.

b) Google's Project Loon offers using land-based stations and hot-air balloons in the stratosphere.

c) SpaceX's Falcon is oriented toward developing reusable satellite placement methods that would allow, among everything else, forming groups of lesser-sized cellular satellites on lower orbits while keeping to lower costs.

III. Tarification of VR-Based Telecommunication Services

The modern practice of providing telecommunication services is based on the principles of paying beforehand for a package of services that is to be used in the future. For instance, the cellular network user pays for a set of services and options from a certain amount of traffic Gbps, instant messages, minutes, etc. for the following month. A similar principle may be used for charging for new VR-based services - the package may include the summary length of connection sessions, the size of the cloud storage, the maximum number of session participants, the volume and composition of multimedia content, options for accessing social networks and gaming servers, end so forth, Table I. At the same time, it is possible to have both the options for fixed tarification with a limited number of packages with certain sets of services, as well as introducing flexible tarification, calculating the price of the package in accordance with the parameters entered by the user and quantifiable values of specific services included in said package. The last option would define the amount of a preliminary payment depending on such parameters and the combinations chosen by the user previously

TABLE I. An Example of VR Telecommunication Services Package Tarification

The composition of a VR telecommunication service package	**Tarification plans (package service options)**				
	Plan 1	***Plan 2***	***Plan 3***	**…**	***Plan n***
Session duration (minutes)	A1	A2	A3	…	variative
Max number of users (people)	P1	P2	P3	…	variative
Max volume of traffic (Gb)	T1	T2	variative	…	variative
Cloud storage size (Gb)	V1	variative	variative	…	variative
Access to social networks (yes/no)	no	yes	yes	…	yes/no
Access to gaming servers (yes/no)	no	yes	yes	…	yes/no
Online television (yes/no)	no	no	yes/no	…	yes/no
………	………	………	………	…	………
Virtual cinema (yes/no)	no	no	yes/no	…	yes/no
Package price (currency unit)	S1	S2	S3	…	Sn

In accordance with the "Digital Economy of the Russian Federation" program, virtual and augmented reality technology is one of the nine directions depicted as one of the main directions of cross-cutting digital technology. Digital industry development directions and key institutions covered by the program cater to the development of the already existing conditions for the appearance of breakthrough and promising cross-cutting digital platforms and technology, as well as creating the conditions for launching new platforms and technology where competencies are formed for market and economy (areas of activity) growth. In this context, Russia's national and transnational network operators are gaining an opportunity to organize an innovative telecommunication service [6] that has no competition and presents itself as a new and, potentially, high-grossing segment of the telecommunication services market. Considering the absence of competition and a basically free use of the whole existing TCS infrastructure of other mobile operators, all users of the mobile and stationery internet can be considered potential users of this service on a world-wide scale. From this point of view, a corresponding project not only fully complies with the Russian Federation's Informational Community Strategy for the years of 2017-2030, but can also be considered as an innovative source of national income in terms of digital economy. However, considering the current level of VR development in Russia, an issue arises in terms of finding and choosing a platform for designing and prototyping the services described.

IV. Current Results and Scientific Technological and Practical Tasks

At the Moscow Technological University's Information Technologies Institute (IIT MIREA), a telecommunication service prototype based on VR has been designed and is currently successfully tested. Mobile devices, as well as HTC Vive and OculusRift headsets may be used as end-user devices. Users have access to an option of synchronized interaction with each other and the objects of the virtual reality, including audio and video playback, watching videos recorded by remote cameras, working with graphic images and PDF documents. An option has been designed for forming a set of personal user virtual environments, as well as using them for a cooperative and/or personalized cloud storage [7]. The project was named "The Virtual Situational Center" and was presented at the Junction 2017 hackathon as SCVR (Situation Center Virtual Reality) (first place in Microsoft challenge of Entertainment track). Solutions it includes are universal and can be used for implementing telecommunication services based on VR for various practical areas and tasks, both for mass use and solving specific tasks in social-economic process management. Let us review the examples of areas of possible application of such innovative telecommunication services:

1) A virtual situational center (center for strategic management, situation room, control station, multimedia center, etc. [3]) - a set of methodical, informational, hardware and software devices, aimed at managers and/or groups of experts. The relevance of this service on various levels of state and corporate management is partly determined by the need for raising decision-making efficiency based on a large number of information sources and streams under time pressure.

2) A virtual office (conference room) - a telecommunication service providing interactive communication with employees and contractors for higher work efficiency and less time-spending on meeting/negotiation organization, discussing and coordinating documentation, etc.

3) A virtual office for analytic - a service for organizing individual and collective expert work when analyzing multimedia information that depicts the aspects of the situation that is undergoing analysis. It offers an opportunity of visualizing evidence and factors, physical individuals and legal entities, objects and locations in a virtual environment in order to form, show and analyze the connection between them. Using this service on a mobile device allows, for instance, efficient access to visualized information and expert interaction during business trips and onsite events.

1) Virtual communication in social networks and messengers (Facebook, Twitter, Vkontakte, etc., Skype, Viber, WhatsApp, etc.) significantly improves the opportunities for exchanging media content, raising competitive performance and user attraction, provides new opportunities for posting advertising content.

2) The entertainment industry, education and goods promotion - multiplayer online games in VR environments may become one of the main development trends of the gaming industry, virtual tours and educational content can be in high demand on every educational level, and virtual advertising can become a more vivid method of visualization and goods usage.

3) Addition to mobile and video connection services - using a VR-based telecommunication service on a mobile device grants new opportunities and forms of communication for users.

During the process and according to the results of the "Virtual Situational Center" project, a series of scientific technological and practical tasks has been fulfilled, in order to effectively use VR technology for forming new telecommunication services:

1) End-user device miniaturization: minimizing mass-size parameter values and composition of equipment used. In pursuit of this, IIT MIREA is designing a construction, hardware and software of a compact VR helmet and manipulator that will allow using a mobile phone and all options of interactive communication.

2) Researching aspects of human-machine interaction and special features of information perception in VR environments. In this direction, in terms of a governmental task, IIT MIREA is implementing the 2.7178.2017/БЧ project "Researching Cognitive Semiotics in the Multimedia Virtual Reality Environment".

3) Designing and improving methods of providing immersiveness for VR environments, including the development and implementation of special standards for multimedia information processing and playback.

4) Researching the effect hardware and working patters have on the state of the user, designing recommendations and demands to protecting him from harmful factors. Special attention needs to be paid to researching the influence VR devices have on organs of sight, and researching the environments' effect on the users' psycho-emotional state.

5) Designing and developing a specialized convergent billing system that would allow user commutation in VR environments and access to multimedia content, tarification and payments to the service provider.

6) Designing a data processing center that has computing capacities big enough for telecommunication services in VR environments to be accessible simultaneously to a large number of users.

7) Researching opportunities and prospects for using new standards of mobile connections (the fifth generation), as well as new formats of transmitting and processing multimedia information, including the provision of immersiveness in VR environments.

Tasks listed can be solved by MIREA's own scientific departments, or in cooperation with industrial partners.

V. Conclusion

Virtual reality technology could be used in organizing new telecommunication services in the nearest future. Basically, the operator that is the first to offer said services, will not only be able to gain all the users of mobile and stationary internet as their clients on a world-wide scale, but also use existing TCS and already provided services of other mobile operators for forming this new segment without paying additional costs. New connection standards and forms of data transmission and processing that are getting ready for production will contribute to rapid coverage of this new telecommunication service market segment and to the connection of users to new services without consideration for national mobile operators. MIREA's Information Technologies Institute invites all interested physical persons and legal entities to participate in this direction of research and development.

References

[1] Glazyev S. The strategy of Russia advancing development in the conditions of global crisis. M.: Ekonomika, 2010. — 256 c.

[2] Zuev A.S., Bolbakov R.G. On prospects of development of telecommunication systems and services based on virtual reality technology. M.: Russian technological journal. — Vol. 5. — №6. P. 3 – 10.

[3] A.S. Zuev, I.S. Fadeev Virtual Command Centers — New Management Tool for Socio-Economic Systems // Information technologies. — 2016. Vol. 3. — P. 229—232.

[4] View on 5G Architecture (Version 2.0). 5G PPP Architecture Working Group. — 18.07.2017. — 113 P.

[5] Turban L., Urban F., Guillotel P. Extrafoveal Video Extension for an Immersive Viewing Experience. IEEE: Transactions on Visualization and Computer Graphics. 2017. №5. P. 1520-1533.

[6] Zuev A. On the Direction of the Rapid Development of Telecommunication Systems and Markets. Interactive Systems: Problems of Human – Computer Interaction. – Collection of scientific papers. – Ulyanovsk: USTU, 2017. – 290 p.

[7] Fischbach M., Latoschik M.E., Wiebusch D. Semantic Entity-Component State Management Techniques to Enhance Software Quality for Multimodal VR-systems. IEEE: Transactions on Visualization and Computer Graphics. 2017. №4. P. 1407-1416.

Unifying Modeling Language-Merise Integration Approach for Software Design

Issar Arab[1]
Department of Informatics, Technical University of Munich,
Garching/Munich, Germany

Safae Bourhnane[2], Fatiha Kafou[3]
School of Science and Engineering
Al Akhawayn University in Ifrane, Ifrane, Morocco

***Abstract*—Software design is the most crucial step in the software development process that is why it must be given a good care. Software designers must go through many modeling steps to end up with a good design that will allow for a smooth development process later. For this, designers usually have to choose between two main modeling methodologies: Merise and UML. Both methodologies are widely used; however, each one has its own advantages and disadvantages. This paper combines both techniques and merges their advantages to come up with an approach that would help software designers make the best of both methodologies. This integration mainly targets the software design step in general but can be specifically applied to database design. It presents the weaknesses and strengths of each one of UML and Merise as two techniques used in database modeling and design. Later in this paper, a comparing of UML and Merise diagrams is lead and based on it, the decision on which of the two is the best at each step of the modeling process.**

Keywords*—*UML; Merise; modeling; design; databases

I. Introduction

Database is what all software developers are concerned with in the first place. If you have a well-designed database, you can be sure that the entire development process will go as smoothly. For the purpose of designing the best database, tools and frameworks like Merise and UML can be used for data modeling. However, none of the existing frameworks is perfect. That is why it is thought that presenting a new approach based on UML and Merise would help having a good database design just by applying little effort and avoiding the drawbacks of each technique.

Although UML is the methodology that is widely used, it, definitely, has some disadvantages that make its usage tedious to some extent. UML is very complex with more than 13 diagrams and more than 100 types of classes [1]. This makes it hard to adopt and even harder to master. UML is also time consuming. It takes a lot of time to manage and maintain UML diagrams [2]. On top of that, software developers do not benefit from UML diagrams as much as you would hope, because they work with code and programs rather than pictures and diagrams. UML is rather beneficial for project managers that are concerned with the way the software tool would work [3].

Merise, on the other hand, is not as widely used. It also has a set of advantages and several disadvantages. Merise does a great job with the modeling and the conception of small databases. But, when designing large databases, it may not be the best methodology to opt for. Also, it is limited to the 3rd normal form [4]. In addition to that, it is best suited to work with modeling sequential tasks and does not deliver a good result when dealing with distributed ones. It is not meant to model semantic data.

The new approach comes to circumvent the disadvantages mentioned above for both methodologies through creating a new process to model the system in general and databases in particular. This integration of both techniques has less limitation than each methodology when applied by its own.

The rest of the paper is organized as follows: Section II presents the work that has been previously done in the same field. Section III consists of a comparison of UML and Merise. The description of the work done is presented in Section IV. Finally, Section V describes the suggested final process to follow.

II. Related Work

Knowing that MERISE is a methodology that is mainly used in France and that is being adopted in European engineering community more than other communities, the author tries in [5] to make the methodology more suitable for English speaking users. However, this work does a perfect job in trying to spread MERISE in English speaking community, by somehow translating the existing elements of the methodology, but doesn't in any way try to hide the disadvantages and limitations of the methodology itself.

In [6], the author presents the different concepts of UML as an object-oriented modeling language. These concepts definitely have many problems and limitations, which actually don't exist in the first methodology. But UML does have advantages that, in contrast, don't exist in MERISE. Hence, comes the idea of combining the two approaches by integrating the diagrams from each to satisfy the user's need in different scenarios. In the next sections, the paper will present how the integration is to be devised.

III. Unified Modeling Language vs Merise

Before getting on with the integrated approach, it is necessary to look at the comparisons between UML and Merise that have been done in the literature.

UML and Merise are not completely similar. Each one has a different concept. UML, for example, takes care of the object-oriented modeling, while Merise works best for relational databases. Even though UML is more widely used than Merise, both methodologies can nonetheless be used in modeling and conceiving databases.

On the one hand, Merise is said to constitute a real methodology that respects the standards. Earlier, in 2003, Merise was divided into three main pillars: steps to follow, formalism, and organization. However, some of these aspects did not survive in front of the advancement of technology and needs of the recent applications. The "steps to follow" for example, is no longer needed in order to have a good methodology, while the importance of "formalism" persists.

On the other hand, methodologists claim that UML presents a very good formalism with a high level of standardization, but it is lacking the process to follow in addition to the organization to be a real methodology. Besides, Merise works best with organizational information systems while UML is designed for object-oriented based information systems. That is why the two methods actually complement each other and can be used at the same time.

The purpose of this paper is to combine these two methodologies and to prove that they can together be leveraged in the modeling of the same project.

IV. Description of the Work Done

It goes without saying that the order of the UML diagrams to be used is not fixed as it depends on the type of the application and the style of the designer or developer.

In this paper, an attempt to unify the process of software design is made, by making all the steps standardized and clear.

In the first place, the classes that constitute the system are identified then the actors of the application are looked at. Right after that, the exchanged messages between the actors of the system are studied, their sequence as well as the order in which these messages appear. Then, further light is shed on the set of activities that are performed within the application.

At each step, a comparison of the diagrams used in each methodology is lead, and then the assessment.

The figure below (Fig. 1) summarizes the process as described above.

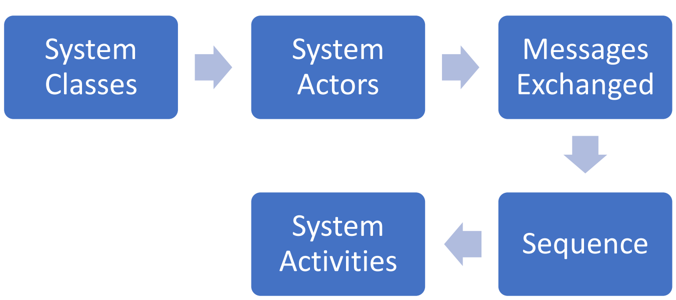


Fig. 1. Software design process.

A. *UML Class Diagram vs Conceptual Model of Merise*

1) *UML Class Diagram*

a) *Definition*

It provides a general overview of the final system by describing the classes involved in the system and by explaining the relationships between them. It allows the users to go from domain specific data structures to a detailed design of the final product. The main components of the class diagram are [7]:

Class: grouping of objects with the same characteristics

Method: part of a class that shows the behaviors of a certain object of that class

Attribute: part of a class that represents the static properties of an object of the same class

Multiplicity: indicates that one of the related classes refers to the other and it can take many values.

Relationship: represent the logical relationship between classes. There are many types of relationships in the class diagram of UML.

Object: instances of a specific class

Access Level: data privacy is determined by assigning an access level to it

b) *Example*

Fig. 2 shows the different components mentioned in the previous section.

Fig. 2. UML class diagram.

As shown in the figure above, it represents a simple class diagram that consists of two classes: Person, and Address. Each class has attributes (e.g. Name, Address, …) and operations or methods (in this case only the class Person has a method that is called BuyCar()). The multiplicities in this example mean that an address is associated to one person maximum, while a person does necessarily have one address.

2) *Merise Conceptual Model*

a) *Definition*

At this level of the modeling process, the *entity/relationship schema* is used in Merise. A typical entity/relationship diagram would contain two main components as its name suggests: *entity* and *relationship*.

An entity, short for entity type, can be compared to a class in the context of UML, but it only contains *properties (attributes)*. In general, an entity can be defined independently of the rest of the data and corresponds to one row in a database table [8], [9].

In an E-R diagram, the entity needs to have a unique identifier which could be one or a set of properties (e.g. orderid + date to characteristic of an order).

The *relationship* (or association) links together one or more entities and can itself contain additional attributes.

b) Example

Fig. 3 shows a graphical representation of the E-R diagram.

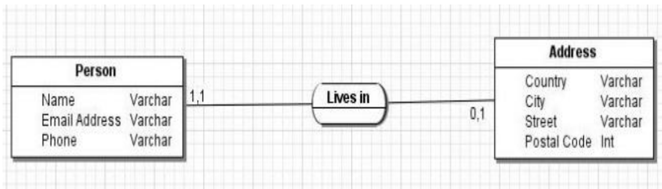


Fig. 3. Merise conceptual diagram.

In the example above, the same classes as in the previous UML class diagram are kept to show the differences that exist between the two modeling techniques when dealing with system classes. The first difference that pops up is in the multiplicities. They still mean the same in both diagrams, but they inversed. Another difference resides in the fact that associations in Merise can have attributes, thing which does not exist in UML.

3) What to use and why

If a translation from the conceptual model to the class diagram of Merise was to be done, nothing much would be done: in fact, each relationship will be transformed to an association, each entity will be a class and relationships with attributes will be transformed into a class association with the same attributes.

Although the differences are not that big, it is recommended to use the UML class diagram for the following reasons:

- The multiplicities in the class diagram are more intuitive and make more sense to the designers that are new to the domain. It is easier to understand that a tutor has a program rather than a program is owned by a tutor.
- The class diagram gives a better illustration and overview of the system because it presents not only the attributes of the objects but also their data types in addition to behaviors and their return data types.
- UML class diagram is closer to the implementation as it lets you think about the code and the things to be implemented in the coding phase. This saves a huge amount of time in the implementation.
- The UML class diagram is more for object-oriented languages (java, visual basic, .net …), and the object-oriented paradigm is gaining a lot of popularity among programmers these days.

B. UML use Case Diagram vs Conceptual Model for Communication of Merise

1) UML Use Case Diagram

a) Definition

Use case diagrams give a general overview of the usage requirements of the final system [10]. They are mainly used to represent the stakeholders of the entire project. It is also helpful in the deployment phase as programmers find it easy to go from actual use cases from the diagram to functions in the system. The Use Case diagram consists of the following components:

Use cases: they are horizontal ellipses that represent the sequence of actions that are done by a user and that would be of additional value to them.

Actors: The main users of the system, they can be humans or external entities (operating systems).

Associations: They represent the relationship between the actors and the use cases, between use cases (include, extend), or even between users (inheritance).

System Boundary: Represented by a rectangle drawn around the use cases. Their main goal is to delimit the scope of the project.

Packages: Packages are totally optional. They are used to group use cases of the same type together allowing for a better organization of the entire diagram [11].

b) Example

Fig. 4 is a simple use case diagram that shows the different actions performed by the student and professor in a university.

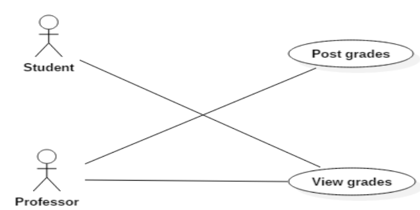


Fig. 4. UML use case diagram.

Concerning the example in the previous figure, it presents two main actions that are performed within a university by the professor and the student. It can be inferred from the diagram that the professor posts and views the grades while the student can only view the grades.

2) Conceptual Model for Communication

a) Definition

This model is complementary to what is called “Context Diagram”. The “Context Diagram” shows the external entities that interact with the system to be designed [12]. The Conceptual Model for Communication completes this diagram in the sense that it decomposes the system into many internal actors who exchanges messages between them.

Graphically, the actor is represented by an ellipse whereas the messages are represented by arrows [13].

b) Example

In Fig. 5, the organization is composed of 2 internal actors who are the professor and the student, and they are interacting with the system through performing two main actions that are Post Grades and View Grades. The actions done by each actor can be inferred just like in the previous diagram.

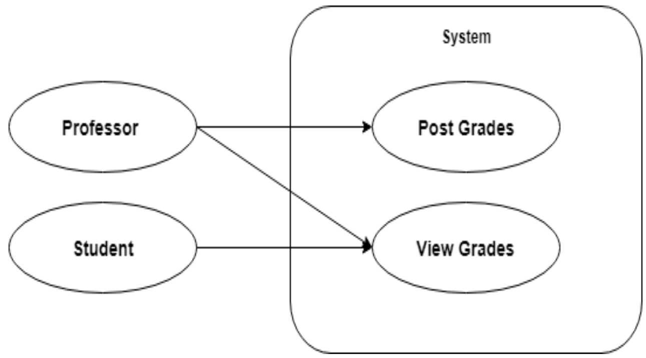


Fig. 5. Conceptual model for communication.

3) What to use and why

Both the use case diagram and the conceptual model for communication show the set of internal actors that exist in the system and the actions that are performed by those.

It is recommended to use the conceptual model for communication if there is an extensive interaction not only between the internal actors, but also between those and other external entities. The conceptual model for communication explicitly shows the interaction between the entities. This implies that even if the system in question has a significant number of entities that are communicating with each other, it can easily be represented in a nice and readable diagram.

However, if the focus needs to be done on each individual actor (to limit the privileges and describe them), then it would be more suitable to use the use case diagram. This latter focuses on the user rather than the actions done. It takes a better care of the privileges given to each actor which consequently affects the actions to be performed by that actor.

C. UML Sequence Diagram vs Merise Data Flow Diagram

1) UML Sequence Diagram

a) Definition

Sequence Diagram is a high-level interaction diagram that shows how operations are carried out between the different parts that exist in the system [14]. Graphically, the messages exchanged during the interactions are ordered vertically in an increasing chronological order. The vertical line that represents time is called the *lifeline*. It extends as long as the life of the actor in question within the system. The horizontal axis shows the different objects involved in the interactions the diagram shows. Each of those objects is called a participant and has its own lifetime [15].

b) Example

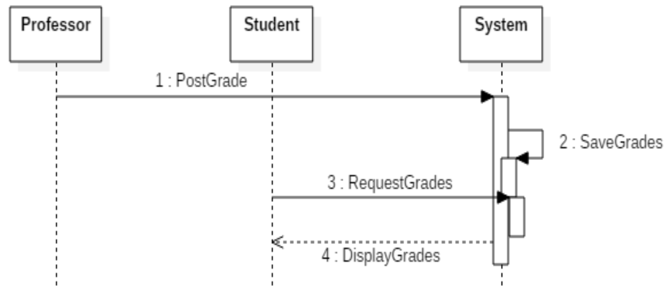


Fig. 6. UML sequence diagram.

The example above (Fig. 6) has the same actors as the use case diagram. The sequence diagram shows the messages that are exchanged between the two actors and the system. The professor posts the students' grades to the system which saving them later to the database. To view their grades, the students request the grades from the database which then replies by displaying the grades.

The messages shown in the diagram follow a chronological order, meaning that the first message sent is displayed in the top of the diagram and has the ID number 1 and so on.

2) Merise DataFlow Diagram

a) Definition

This diagram shows which activities are related to each other and how they are involved in solving the problem stated [16]. At this stage, this diagram is done without taking into consideration the actual behavior of the system (scheduling, synchronization …). It shows the activities and relationships between them in a non-sequenced manner [17].

b) Example

Fig. 7 shows an example of a Merise Data Flow Diagram.

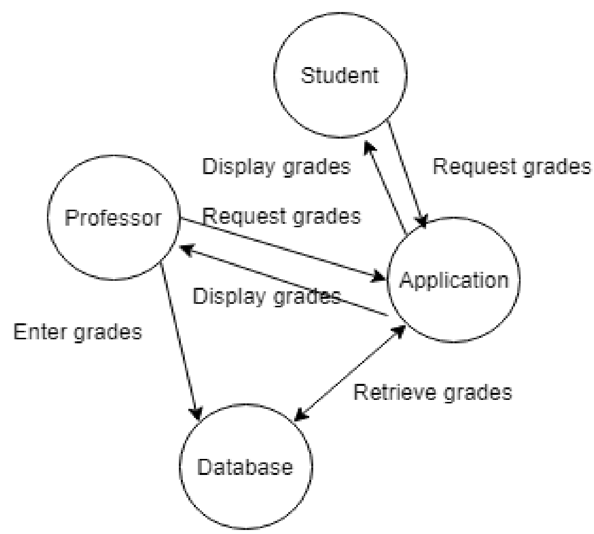


Fig. 7. Merise data flow diagram.

3) What to use and why

It is clear that both diagrams are about exchanging messages. The main and obvious difference would be that UML's sequence diagram looks more structured and organized because it takes into consideration the time and it shows the activities in a chronological order.

However, it can be observed that the use of the data flow diagram brought by Merise will do a better job in giving a general view about the communication of the objects within the system:

- The sequence diagram somehow gives an idealistic representation of the messages exchanged between instances, while the view given by the data flow diagram is more realistic.
- In the data flow diagram, there is no need to follow a specific order the thing that allows for a certain level of flexibility. This way, the user will be able to see different scenarios and choose a specific instance to initiate the scenario.
- The data flow diagram is easy to master with few symbols and notations compared to the complex UML sequence diagram. Plus, it is more intuitive and easy to explain to project managers or clients who, not necessarily have a computer science background.

D. UML Collaboration Diagram vs Merise Dataflow Diagram of Merise

1) UML Collaboration Diagram

a) Definition

The collaboration diagram is similar to the sequence diagram. The difference is that the collaboration diagram is object-centered whereas the sequence diagram is time-oriented [18], [19].

b) Example

Fig. 8 shows a simple example of a UML collaboration diagram.

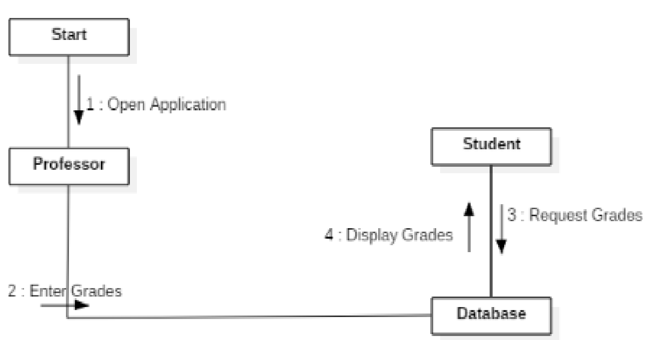


Fig. 8. UML collaboration diagram.

2) What to use and why?

In this step of the modeling, no Merise diagram is introduced. However, one can opt for an intermediate solution in this case. It is recommended to use the UML collaboration diagram, and if not applicable (for the specifications of the application in question), simply replace the objects by the usual entities used in Merise.

That is because:

- The collaboration diagram shows more details about the messages between objects/entities.
- There might be a chronological order introduced to the diagram.
- Actors can be included in the diagram
- It gives a clear and structured overview of the system in a later step of the design

E. UML Activity Diagram vs Merise MCT

1) UML Activity Diagram

a) Definition

The activity diagram consists of activities, states and transitions between those. It shows how activities coordinate to achieve and provide certain services and defines the main events of the system needed to make a given service, and how those events relate to each other.

It is an advanced flowchart that combines other details such as the actors, the starting point, and the finishing point of the system.

In addition to that, it captures the dynamic flow of the system [20], [21].

b) Example

Fig. 9 illustrates an example of UML's activity diagram.

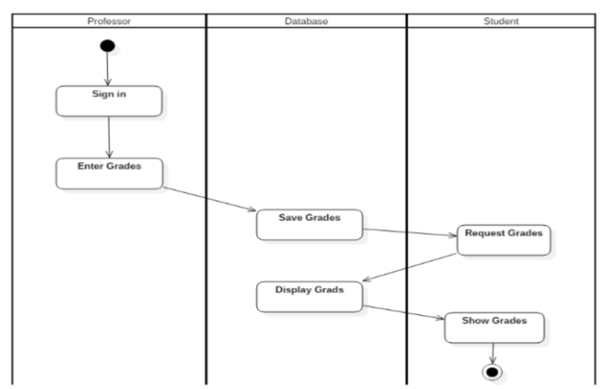


Fig. 9. UML activity diagram.

2) Merise MCT

a) Definition

The conceptual model of treatment is one of the most famous diagrams in Merise. It allows for the treatment of the dynamics of the information system meaning the event-driven operations that are carried out within the system.

This diagram helps then to model the activities of the system using clear schemas. It simply defines what should be done without giving any idea about how, when or where.

The components below describe the MCT diagram [22]:

Process: A subset of the enterprise activities. This means that the entity uses many processes within the same activity.

Operation: Is a set of actions executed after an event or a conjunction of events.

Event: An event represents the change in the external universe of the information system or in information system itself.

b) Example

The example showed in Fig. 10 presents a Merise conceptual treatment model.

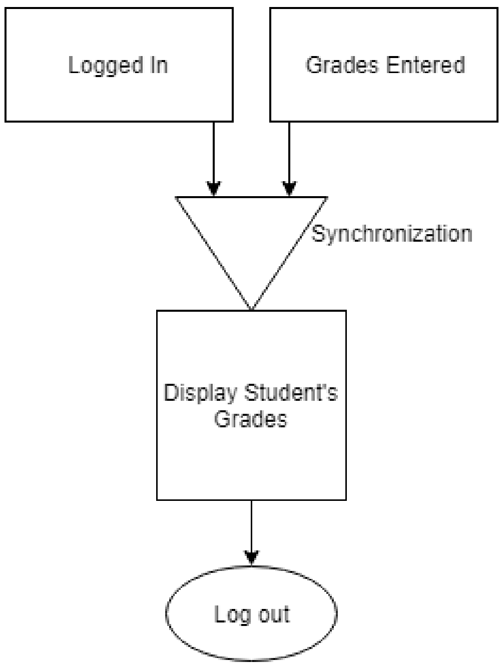


Fig. 10. Merise MCT.

3) What to use and why?

Both the activity diagram of UML and MCT of Merise can be used in the same stage of the conception and design phase. They both show the flow of activities in the system to be conceived. But, UML's activity diagram tends to be more powerful for the following reason:

- The separation between the actors of the system will be of a great help in trying to really understand the system. It makes it clear for the user which activity is done by which actor.
- The MCT provides the use of some rules that when added may result in increasing the complexity of the diagram.
- The MCT does not clearly identify the initial and final events while it is really important to state when the flow of activities starts and when it ends.

V. Entire Process to Follow and its Limitations

Now that all the recommendations regarding the usage of UML and Merise diagrams were given, it is time to discuss the entire process to follow.

This is done through providing the steps of the entire process that is suggested in this paper. Fig. 11 summarizes the suggested process.

As mentioned previously in this paper and as shown in the diagram, It is recommended to start either with UML use case diagram or the Merise conceptual model for communication. Then, go for UML class diagram. For the following step, it is suggested to opt for Merise dataflow diagram. As for the forth step, UML activity diagram is said to do a better job, then end up with an integrated approach that combines collaboration and dataflow diagrams.

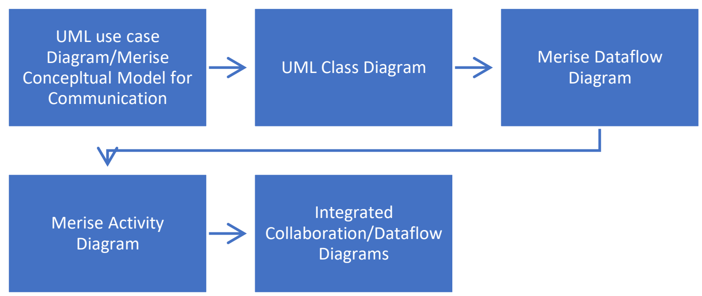


Fig. 11. Entire process to follow.

In the process of comparing Merise and UML diagrams, many challenges were faced. The first one was to find common aspects and features in both methodologies. This required closely looking at the applications of each one seperatly. The second challenge was mainly about keeping the required functionalities at each stage of the design process. The last challenge resides in keeping the process as efficient as it initially is while modifying the diagrams used.

The limitations of this new approach reside in the fact that it has not been based on experiments. It is based on studying each step of the design process along with the diagrams associated with it. Besides, this new approach may not suit all types of projects and all communities. People used to Merise digrams will have hard time merging it with another approach, and the same goes for software designers who are more of UML users.

VI. Conclusion and Future Work

UML and Merise are two methodologies that are used by software designers. Unlike what many practitioners in the domain think, these two modeling methods are not very different from each other. UML and Merise complement each other in a way allowing for their integration which means that they can be used at the same time with the same application that is why software designers always have problems choosing the framework to use.

This paper came up with an approach to unify UML and MERISE approaches in a way that reduces the drawbacks and takes advantages of each one of them.

This new approach aims at helping software designers by simplifying the design process and making it smooth.

As future work, this new approach needs to be implemented in a real-world project and tested in terms of performance. A good way of doing that could be by testing each approach, MERISE and UML, and comparing it to the one suggested in this paper. These tests based are to be performed on a real-life project that is complex to enough in order to push these methodologies along with the new approach to the limit, and hence be a proof of concept. An interesting metric to measure this performance would be the number of iterations done in each methodology before getting the appropriate model to implement. Besides, the feedback of the software designers will also be valuable in assessing the performance of each technique.

References

[1] D. E. Avison. (1991). MERISE: A European methodology for developing information systems. Available online at: https://link.springer.com/article/10.1057/ejis.1991.33

[2] Bernd Bruegge and Allen H. Dutoit (1999). Object-Oriented Software Engineering Using UML, Patterns and Java. Available online at: http://dbmanagement.info/Books/MIX/POO_Software_Engineering_Using_UML_Patterns_and_Java_3rd_Edition.pdf

[3] Keng Siau, Qing Cao (2003). How complex is the Unified Modeling Language? Available online at: https://dl.acm.org/citation.cfm?id=960145

[4] Tom Mens, Ranghild Van Der Straeten, and Jocelyn Simmonds. Maintaining Consistency between UML Models with Description Logic Tools. Available online at: http://www.cs.toronto.edu/~jsimmond/docs/mcc/TomMensEtAl.pdf

[5] Murray Cantor (1998). Object-Oriented Project Management with UML. Available online at: https://www.wiley.com/en-us/Object+Oriented+Project+Management+with+UML-p-9780471253037

[6] Third Normal Form (3NF). Available at: https://ww.1keydata.com/database-normalization/third-normal-form-3nf.php

[7] The Class Diagram. Available at: https://www.ibm.com/developerworks/rational/library/content/RationalEdge/se04/bell/

[8] UML 2 Class Diagrams: An Agile Introduction (2014). Available at: http://ww.agilemodeling.com/artifacts/classDiagram.htm

[9] Yann Thierry-Mieg (2007). Database Design & Modeling: Entity / Relationship. Available at: https://pages.lip6.fr/Yann.Thierry-Mieg/old/EFREI-DBMS/07-Design-E-R.pdf

[10] Margaret Rouse (2014). Use Case Diagram. Available at: http://whatis.techtarget.com/definition/use-case-diagram

[11] Edwin Obenauf (2017). UML use case diagram example. Available ate: http://studiootb.com/uml-use-case-diagram/uml-use-case-diagram-uml-example-adorable-photo-what/

[12] Chris Adams (2016). What is a system diagram and what are the benefits of creating one. Available at: http://www.modernanalyst.com/Careers/InterviewQuestions/tabid/128/ID/1433/What-is-a-Context-Diagram-and-what-are-the-benefits-of-creating-one.aspx

[13] Sabah Al-Fedaghi, Ala's Alsaqa, Zahra'a Fadel (2009). Conceptual Model for Communication. Available online at: https://arxiv.org/ftp/arxiv/papers/0912/0912.0599.pdf

[14] Donald Bell (2004). The Sequence Diagram. Available at: https://www.ibm.com/developerworks/rational/library/3101.html

[15] Sequence Diagram. Available at: https://en.wikipedia.org/wiki/Sequence_diagram

[16] What is a Data Flow Diagram? Available at: https://www.lucidchart.com/pages/data-flow-diagram

[17] Data flow diagram. Available at: https://www.smartdraw.com/data-flow-diagram/

[18] Margaret Rouse. Collaboration Diagram. Available at: http://searchsoftwarequality.techtarget.com/definition/collaboration-diagram

[19] UML 2 Communication Diagramming Guidelines. Available at: http://agilemodeling.com/style/collaborationDiagram.htm

[20] UML – Activity Diagrams. Available at: https://www.tutorialspoint.com/uml/uml_activity_diagram.htm

[21] Activity Diagram for Inventory Management System (UML). Available at: https://www.lucidchart.com/pages/activity-diagram-for-inventory-management-system-UML

[22] Merise (data-processing). Available at: http://wikipedia.qwika.com/fr2en/Merise_(informatique)

Linear Quadratic Regulator Design for Position Control of an Inverted Pendulum by Grey Wolf Optimizer

Hüseyin Oktay ERKOL
Department of Mechatronics Engineering
Faculty of Technology, University of Karabuk
Karabuk, Turkey

***Abstract*—In this study, a linear quadratic regulator (LQR) based position controller is designed and optimized for an inverted pendulum system. Two parameters, vertical pendulum angle and horizontal cart position, must be controlled together to move a pendulum to desired position. PID controllers are conventionally used for this purpose and two different PID controllers must be used to move the pendulum. LQR is an alternative method. Angle and position of inverted pendulum can be controlled using only one LQR. Determination of Q and R matrices is the main problem when designing an LQR and they must be minimized a defined performance index. Determination of the Q and R matrices is generally made by trial and error method but finding the optimum parameters using this method is difficult and not guaranty. An optimization algorithm can be used for this purpose and in this way; it is possible to obtain optimum controller parameters and high performance. That's why an optimization method, grey wolf optimizer, is used to tune controller parameters in this study.**

Keywords—Grey wolf optimizer; inverted pendulum; position controller; linear quadratic regulator; optimized controller design

I. INTRODUCTION

Linear Quadratic Regulator (LQR) is one of the optimum control methods and it is successfully applied to many systems. Selection of the controller parameters is the main problem when designing an LQR controller. The selected parameters must minimize a performance index. The selection process is conventionally made by trial and error method and it makes the process difficult, not guarantees finding the optimum parameters and may take long time. Optimization algorithms help designers to overcome these problems and guarantee finding one of the optimum solutions.

One of the basic systems for control theory is DC motor and LQR controller is one of the methods to control its speed and position. Ruderman et al. designed an LQR based speed controller for a DC motor [1]. Abut compared the PID controlled DC motor and the LQR controlled DC motor under disturbance and the results showed the LQR based system has better performance than PID based one [2]. Haron deigned speed and position controllers for a DC motor in his study. In the study, PID and LQR controllers was used and made a performance comparison. The results show again the LQR controller has better performance than PID controller [3].

Another popular system for control theory applications is inverted pendulum. Kumar et al. designed an LQR based controller for balance and trajectory tracking problem of a Self-Erecting Single Inverted Pendulum [4]. They reported that LQR based system had faster and smooth stabilizing process compared to Full State Feedback controller designed by pole placement. Prasad et al. made a study to analyze and compare the PID and LQR controlled system under disturbance [5]. The results was justified that the advantages of the LQR controller.

Trial and error method is widely used method to determine the elements of the Q and R matrices of an LQR controller [6]. However there are many study shows the optimization algorithms help to determine the optimum parameters for the controller. Ata et al. designed an LQR based controller for an inverted pendulum on a cart. In the study, elements of the Q and the R matrices of the controller were selected by Artificial Bee Colony Algorithm to achieve the optimum performance. Optimization process was made on a nonlinear model and the results showed the ABC optimized system had good performance [7].

In another study, an unmanned rotorcraft pendulum was controlled using LQR optimized by ABC and Particle Swarm Optimization algorithm [8]. The designed system was also tested under disturbance and the results showed that the ABC optimized system had better performance than the PSO optimized system. Çatalbaş et al. was designed an LQR controller for a Boeing 707 flight model and the unstable model was controlled successfully by the LQR controller [9].

In this study, an inverted pendulum system is modeled and controlled by LQR. Q and R matrices of the LQR are optimized by Grey Wolf Optimizer (GWO). All the study is made by simulation using Matlab program. Two different objective functions are used for the optimization process: firstly performance index of the LQR is used and then an improved objective function obtained adding settling time and total absolute error to the performance index is used. The controller is successfully optimized using both of the objective functions.

II. LINEAR QUADRATIC REGULATOR

LQR is one of the optimal control methods and widely used in the optimal control problems. The LQR method is used to control of complex systems that needs high performance. A system described by linear differential equations can be shown in steady-state form given in (1) and (2). *A* is system matrix, *B* is input matrix, *C* is output matrix and *D* is feed forward matrix. "*x*" is the state vector, "*y*" is the output vector and "*v*" is the input vector. A conventional LQR problem is to find the *Q* an *R* matrix which minimizes the cost function (performance index) based on the input "*v*" [10]. Performance index "*J*" is defined as given in (3). The control energy is represented by $v(t)^T Rv(t)$, while the transient energy is expressed as $x(t)^T Qx(t)$ [11]. *Q* is symmetric positive semi definite matrix and *R* is symmetric positive definite matrix.

$$\dot{x}(t) = Ax(t) + Bv(t) \tag{1}$$

$$y(t) = Cx(t) + Dv(t) \tag{2}$$

$$J = \frac{1}{2}\int_0^{\infty}[x^T(t)Qx(t) + v^T(t)Rv(t)]\,dt \tag{3}$$

Designing a LQR controller consists of the following steps:

Step 1: *Q* and *R* matrix, minimizing *J*, must be chosen.

Step 2: Then the algebraic Riccati equation, given in (4), is solved to obtain *P* using *Q* and *R*.

Step 3: Optimum feedback gain matrix "*K*" is calculated using (5).

Step 4: System response is checked. If the system response is not met the required specifications, repeat all steps again.

$$AP + PA^T - PB^T R^{-1} BP + Q = 0 \tag{4}$$

$$K = -R^{-1}B^T P \tag{5}$$

A pre-compensation factor must be used when the system has a bigger steady-state error than expected. Pre-compensation factor calculation can be made by the equation given in (6).

$$N = -(C(A - BK)^{-1}B)^{-1} \tag{6}$$

As seen as, the system must be well modeled to design an LQR controller. The system must be linearized if the system is not linear. All states of the system must also be measurable or observable. Therefore, LQR design has a complex procedure but it has an important advantage. Controlling the all system states is possible with one LQR controller. In this study, pendulum position and vertical angle of the pendulum are controlled by one LQR controller.

III. INVERTED PENDULUM ON A CART

Inverted pendulum is a popular system, which is naturally nonlinear and unstable, in control theory. Inverted pendulum balance research is classically based on an inverted pendulum on a cart and the aim is balancing the pendulum by moving the cart [12]–[14]. The basic system is given in Fig. 1. Invers pendulum is fixed on the cart by a rotating joint. "*θ*" angle changes when an enough amount of force applied to the cart. The aim of the system is balancing the pendulum on vertical axis. The position control of the cart is also possible using an extra controller.

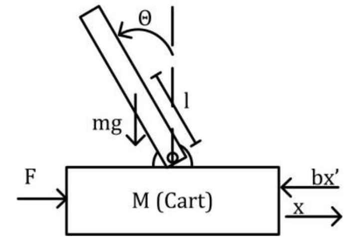


Fig. 1. Structure of inverted pendulum on a cart.

The differential equations of the system can be derived using Euler-Lagrange method. The equations of the system given in Fig. 1 are given in (7) and (8) [15], [16]. "*I*" is the moment of inertia of the pendulum, "*m*" is the mass of the pendulum, "*M*" is the mass of the cart, "*l*" is the length of the pendulum, "*x*" is the cart position, "*θ*" is the angle between the pendulum and the vertical axis, "*F*" is the input force.

$$(I + ml^2)\ddot{\theta} + mglsin\theta + ml\ddot{x}cos\theta = 0 \tag{7}$$

$$(M + m)\ddot{x} + b\dot{x} + mgl\ddot{\theta}cos\theta - ml\dot{\theta}^2 sin\theta = F \tag{8}$$

When the "*θ*" is enough small, the equations can be linearized and steady-space equations (given in (9)) of the system can be obtained. "*a*" used in (9) is given in (10). The system parameters are given in Table I.

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & \frac{m^2gl^2}{a} & \frac{-b(I+ml^2)}{a} & 0 \\ 0 & \frac{mgl(M+m)}{a} & \frac{-mlb}{a} & 0 \end{bmatrix} \begin{bmatrix} x \\ \theta \\ \dot{x} \\ \dot{\theta} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{I+ml^2}{a} \\ \frac{ml}{a} \end{bmatrix} F = \begin{bmatrix} \dot{x} \\ \dot{\theta} \\ \ddot{x} \\ \ddot{\theta} \end{bmatrix} \tag{9}$$

$$a = I(M + m) + Mml^2 \tag{10}$$

TABLE I. SYSTEM PARAMETERS

Symbol	Value
M	0.5Kg
m	0.2Kg
l	0.3m
g	$9.81m/s^2$
I	$0.006kg.m^2$
b	$0.1N/m.s^{-1}$
F	- N
θ	- °

IV. GREY WOLF OPTIMIZER

An optimization algorithm minimizes (or maximizes) a function called objective function. The objective function is a special function defined for a system (or for a problem). It is affected by the parameters of the system. The optimization algorithms minimize the objective function by changing the function variables in a special way. This special way is inspired from the creatures in nature in some algorithms like Artificial Bee Colony Algorithm, Particle Swarm Algorithm or Grey Wolf Optimizer.

Grey Wolf Optimizer algorithm is inspired from the Grey Wolfs in nature. They are social animals and live in groups which size is generally 5-12. There are four levels in a group called as alpha, beta, delta and omega. The group leaders called alpha make important decisions like about hunting, sleeping and etc. The alphas are the most dominant wolves in the group. The alphas may not be the most powerful member of the group but they are best in managing. Beta wolves help alpha wolves for everything. When the alphas get away, ill or very old, betas do coordination and decision making processes for the group. They are under control of the alpha wolves but they can command the other wolves in the group. They also give feedback to alphas about the other wolves and works. There are omega wolves at the end of the hierarchy. They always do what the dominant wolves want. They are the last wolves allowed to eat. It seems like omega wolves do not have an important role in the group but it is observed that the group has some problems like internal fighting in the absence of omegas [17], [18].

The delta wolves are another type and they are responsible of hunting, scouting, sentineling, and some of them may be caretakers or elders. Hunters help the alphas and betas. Sentinels protect the group, scouts watches around and warns the group if there is any danger. Caretakers care the weak or ill wolves.

They have also a special hunting strategy. They track and approach the prey. Then they encircle, pursue and harass the prey until it stops moving. Finally they attack the prey. Detailed information of the mathematical model of the algorithm can be found in [17].

V. Experimental Study

In this study, an inverted pendulum model is designed, and controlled by an LQR controller. All study is made by simulations using Matlab program. Q and R matrices of the LQR controller are optimized by GWO algorithm. General block diagram of the LQR controlled system is given in Fig. 2. A, B and C are system matrices; K is feedback gain matrix and N is pre-compensation factor.

Solution of the differential equation in the simulation is made by the four steps Runge-Kutta method and the used time step is 0.001s. Total simulation time is 10s. Number of Search agents (individuals in the group) is selected as 30 and the iteration number is selected as 50 for the GWO algorithm. Q and R matrices are defined as diagonal matrices and the range of the each element is 1.10^{-4}-1.10^{10}.

An objective function is needed to tune the Q and R matrices when used an optimization algorithm. The main objective function is the performance index J, given in (3), for LQR design. Optimum controller design is possible when J is used as an objective function.

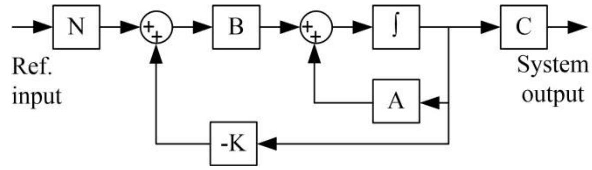


Fig. 2. Structure of LQR controlled system.

The system outputs are given in Fig. 3 when only J is used as objective function. The settling time of the position output is 14.18s with 2% tolerance. The settling time of the θ output is 19.36s. Maximum error of θ is 0.04° and performance index J is 0.141.

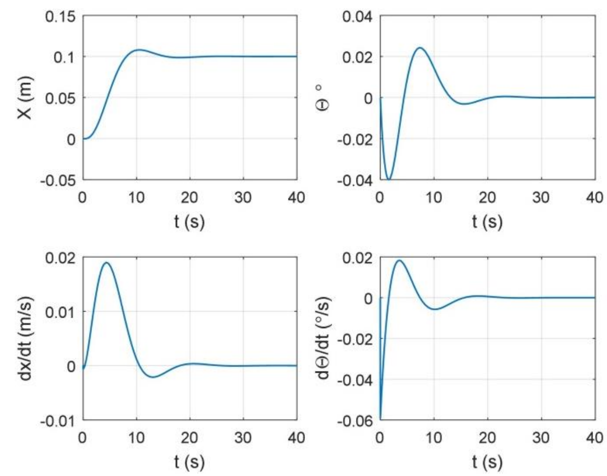


Fig. 3. System outputs 1 (J is used as objective function).

As seen as, the controller works good but the settling time is very long. That means, the results may not meet design requirements. In this case, an improved objective function is needed. The used objective function to meet the design requirements is given in (11). ST denotes the settling time and $Z_{1,2}$ is a coefficient to increase the effect of ST and integral of absolute error on objective function. $Z_{1,2}$ is selected as $1\text{x}10^8$.

$$o.f. = J + z_1 ST + z_2 \int abs(e) \tag{11}$$

At the end of the optimization process, Q and R matrices optimized as given in equation 12 and equation 13. The value of the performance index J is $3.195\text{x}10^5$ for the given Q and R matrices. Pre-compensation factor, N is calculated as -19.884. The system outputs are given in Fig. 4. The settling time of the system for position control is 1.26s with 2% tolerance and it is 2.06s for θ control. Maximum error of the θ angle is 1.77°.

$$Q = \begin{bmatrix} 3.471x10^6 & 0 & 0 & 0 \\ 0 & 4.91x10^3 & 0 & 0 \\ 0 & 0 & 3.54x10^{-4} & 0 \\ 0 & 0 & 0 & 1.11x10^4 \end{bmatrix} \tag{12}$$

$$R = 8779.759 \tag{13}$$

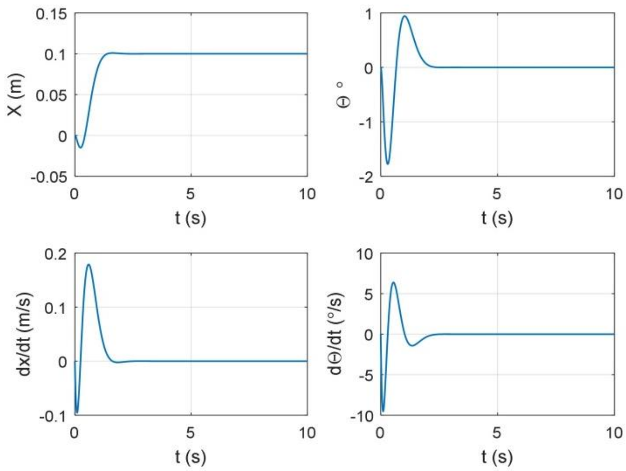


Fig. 4. Sytem outputs 2 (equation 11 is used as objective function).

As seen as, the settling time is shorted using the improved objective function but error of θ angle is increased. As a result, both of (3) and (11) are successful with GWO and selection of the objective function is depended on the design requirements.

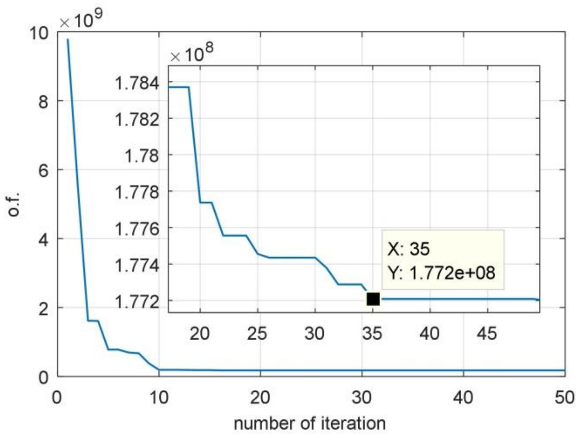


Fig. 5. Objective function value vs. Iteration number.

The speed of the algorithm is another important parameter. Objective function output vs. number of iteration graph is given in Fig. 5. GWO reaches the best solution at 35^{th} iteration.

VI. Conclusion

In this study, an LQR based position controller is designed using GWO algorithm. Determination of Q and R matrices, minimizing the performance index, is the main problem when designing an LQR controller. Minimizing the performance index using Q and R matrices is an optimization problem and GWO is successfully used to obtain optimum Q and R matrices.

Using only performance index J helps to design optimum controller but it may not meet the design requirements like settling time or maximum overshoot. In this case, the objective function must be improved using the effect of the system outputs which must be meet design requirements. Settling time and integral absolute error may be added to the objective function to obtain shorter settling time.

As a result, GWO algorithm successfully optimizes the LQR controller. The settling time of the position controller is 1.26s and maximum error of θ angle is 1.77°. GWO reaches the optimum results at the 35^{th} iteration.

References

[1] M. Ruderman, J. Krettek, F. Hoffmann, and T. Bertram, “Optimal State Space Control of DC Motor,” Proc. 17th World Congr. Int. Fed. Autom. Control, vol. 17, no. 1, pp. 5796–5801, 2008.

[2] T. Abut, “Modeling and Optimal Control of a DC Motor,” Int. J. Eng. Trends Technol., vol. 32, no. 3, pp. 146–150, 2016.

[3] U. Tun and H. Onn, “Linear Quadratic Regulator (Lqr) Controller Design for Dc Servo Motor,” no. July, 2013.

[4] E. Vinodh Kumar and J. Jerome, “Robust LQR controller design for stabilizing and trajectory tracking of inverted pendulum,” Procedia Eng., vol. 64, pp. 169–178, 2013.

[5] L. B. Prasad, B. Tyagi, and H. O. Gupta, “Optimal control of nonlinear inverted pendulum system using PID controller and LQR: Performance analysis without and with disturbance input,” Int. J. Autom. Comput., vol. 11, no. 6, pp. 661–670, 2014.

[6] H. Wang, H. Dong, L. He, Y. Shi, and Y. Zhang, “Design and simulation of LQR controller with the linear inverted pendulum,” Proc. - Int. Conf. Electr. Control Eng. ICECE 2010, pp. 699–702, 2010.

[7] B. Ata and R. Çoban, “Linear Quadratic Optimal Control of an Inverted Pendulum on a Cart using Artificial Bee Colony Algorithm : An Experimental Study Yapay Arı Kolonisi Algoritması ile Bir Arabalı Ters Sarkacın Lineer Kuadratik Kontrolü : Deneysel Bir Çalışma,” vol. 32, no. June, pp. 109–123, 2017.

[8] S. Changhao and H. Duan, “Artificial bee colony optimized controller for unmanned rotorcraft pendulum,” Aircr. Eng. Aerosp. Technol., vol. 85, no. 2, pp. 104–114, 2013.

[9] M. C. Çatalbaş, Z. U. Bayrak, and A. Gülten, “Hava Ta şı tlar ı n ın Uçuş Kontrolü i ç in Do ğ rusal Kuadratik Regülator Yöntemi ile Optimal Kontrol ör Tasar ı m ı OPTIMAL CONTROLLER DESIGN VIA LINEER QUADRATIC,” vol. 27, no. 2, pp. 1–9, 2015.

[10] A. Salehahmadi and A. Farmanbordar, “Modeling Linear Quadratic Regulator LQR / LQT / LQGT for,” vol. 5, pp. 1–11, 2015.

[11] A. Al-mahturi and H. Wahid, “Optimal Tuning of Linear Quadratic Regulator Controller Using a Particle Swarm Optimization for Two-Rotor Aerodynamical System,” vol. 11, no. 2, pp. 184–190, 2017.

[12] H. O. Erkol, “Ters Sarkaç Sisteminin Yapay Arı Kolonisi Algoritması ile Optimizasyonu,” J. Polytech., vol. 900, no. 4, pp. 863–868, 2017.

[13] S. K. Mishra and D. Chandra, “Stabilization and Tracking Control of Inverted Pendulum Using Fractional Order PID Controllers,” J. Eng., vol. 2014, pp. 1–9, 2014.

[14] H. Niemann and J. K. Poulsen, “Analysis and design of controllers for a double inverted pendulum,” Proc. 2003 Am. Control Conf. 2003., vol. 4, pp. 2803–2808, 2003.

[15] A. Poorhossein and A. Vahidian-Kamyad, “Design and implementation of Sugeno controller for inverted pendulum on a cart system,” SIISY 2010 - 8th IEEE Int. Symp. Intell. Syst. Informatics, pp. 641–646, 2010.

[16] A. Patnaik and L. Behera, “Evolutionary Multiobjective Optimization Based Control Strategies For An Inverted Pendulum On A Cart,” IEEE World Congr. Comput. Intell., pp. 3141–3147, 2008.

[17] S. Mirjalili, S. M. Mirjalili, and A. Lewis, “Grey Wolf Optimizer,” Adv. Eng. Softw., vol. 69, pp. 46–61, Mar. 2014.

[18] S. K. Verma, S. Yadav, and S. K. Nagar, “Optimization of Fractional Order PID Controller Using Grey Wolf Optimizer,” J. Control. Autom. Electr. Syst., vol. 28, no. 3, pp. 314–322, 2017.

Evaluation of Photo Contents of Conversation Support System with Protocol Analysis Method

Zhou Xiaochun, Miyuki Iwamoto, Noriaki Kuwahara
Graduate School of Engineering and Science
Kyoto Institute of Technology
Kyoto, Japan

***Abstract*—With the deepening of aging and low birth rate in China, the solitary elderly or old couple living alone is becoming more and more, who has a higher risk of senile dementia caused by disuse of cognitive function because of loneliness without communication. Due to the shortage of care workers, the young volunteer is expected becoming communication partner for them. But it is difficult for the young volunteer without the experience of communication with the elderly, and for two generations to find common topics. However, Conversation Support System was proposed so that the elderly and the young volunteer can talk smoothly with common photo contents. In order to evaluate the utility of photo contents of the system in China, we did the conversation experiment by photos in China, to analyze the expression and stress of subjects during the conversation. As a result, the photos which made the elderly and the young volunteer feel easy and difficult for the conversation were found. Then we analyzed utterance data of subjects with protocol analysis method to discuss the common features of these photos.**

Keywords—Protocol analysis method; the elderly; the young volunteer; photo contents; conversation

I. Introduction

A. Situation of Aging and Low Birth Rate in China

In China, the number of the elderly aged 65 and over was 158.31 million in 2017, and the percentage of the elderly rises to 11.4%. Meanwhile, China is in the face of low birth rate as the result of "One Child Policy". In 2017, China's population increased by 172.3 million, a year-on-year decrease of 630,000, and the birth rate dropped from 1.295 percent in 2016 to 1.243 percent in 2017 [1].

B. Shortage of Nursing Care Staff in China

The National Medium and Long-Term Development Plan for Civilian Talents (2010-2020) proposed that the number of nursing care staff should grow from 30,000 in 2010 to 6 million in 2020 [2]. However, by the end of 2016, the total number of skillful nursing care staff was only 12,144 [3].

C. Related Works

With the deepening of aging and low birth rate in China, the solitary elderly or old couple living alone is more and more, who has a higher risk of senile dementia caused by disuse of cognitive function because of loneliness without communication. However, due to the severe shortage of nursing care staff, the young volunteer is expected becoming communication partner for the elderly. But it is difficult for the young volunteer without any care and communication experience. Co-imagination as a communication support method for prevention of dementia was proposed by Otake in 2006, which should support daily interactive communication between people without care experience and the elderly with images [4]. However, Iwamoto Miyuki investigated the Conversation Support System, which was expected that the young volunteer would feel less stress when talking with the elderly by common contents of photographs and videos [5]. Iwamoto decided to use photos as the contents of this system, because from the experimental results so far, although the video made the subjects feel less pressure, the photo made the elderly and young volunteers talk longer [5].

D. Photo Contents of System

Referring to related works on photo contents for Intergenerational Conversation Support System in Japan, the photo categories of "Food" and "Events" are common to any generation [6]. Meanwhile, from the results of the preliminary experiment in China, Chinese elders were interested in photo categories of "School" and "Commodity" [7]. Therefore, we selected photo categories of "Food", "Events", "School" and "Commodity" as contents of the system in this study, and for each category, we used ten photos in China for the conversation.

E. Protocol Analysis

Protocol analysis is a method of extracting the problem of the system by analyzing the utterance data (protocol) of the user while using the system [8]. In this study, the subjects are the elderly and young beginners who speak slowly when looking at the photo not seen before. Protocol Analysis is an effective method because utterance data of subjects is easy to be obtained.

F. Research Objectives

The goal of this study is to construct a system making the Chinese elderly and young volunteer talk smoothly by photo contents. In order to find the common features of optimal photos for the system, by which the elderly should talk smoothly, and the young volunteer should feel no stress in conversation, we evaluated each photo for each category with protocol analysis method.

II. Experiment

A. Summary

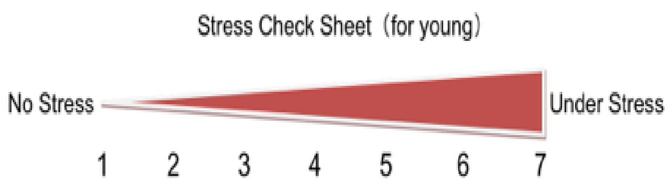
In this experiment, the elderly and the young volunteer talked each other while looking at the common photos. We examined the degree of stress for the young volunteer, and the expression of the elderly depending on each photo in conversation.

The purpose of this experiment is to evaluate the photo categories which make the conversation between two generations easy or difficult.

B. Evaluation Item

1) Stress Check for Young Volunteer for Photo

The young indicated the degree of stress on the stress check sheet every minute in conversation. The stress check sheet represented a 1-7-scale. 1 means the least stress and 7 means the most stress (Fig. 1).



Category Photo	Stress(1~7)
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Category Photo	Stress(1~7)
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Fig. 1. Stress check sheet.

2) Questionnaire for Category

The elderly and the young answered ten questions by 5-stage subjective evaluation each time after each category was finished.

3) Expression Analysis for the Elderly for Photo

We analyzed the expressions of the elderly from the video recordings. We used the degree of happiness or unhappiness as a result of each photo.

C. Subject

The conversation participants were seven young female caregivers of Suzhou Social Welfare Home, and four senior women and two senior men without dementia living there. They don't know each other before.

D. Environment of Experiment

Because the function of eyes is often kept better than ears for the elderly, we brought a photo into close-up and displayed on the screen by the projector. The elderly and the young talked to each other while looking at the screen. The layout of the experiment was shown in Fig. 2 and 3.

Fig. 2. Experimental environment.

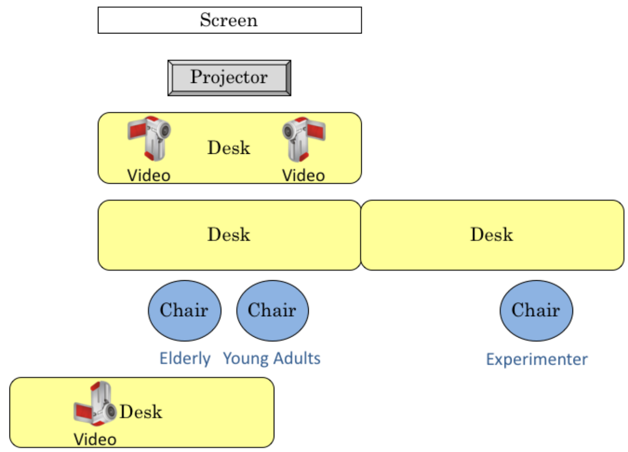


Fig. 3. The layout of the experimental environment.

- We used a meeting room, in which we placed desks and chairs side by side.
- We used a MacBook in which photos were uploaded for a 10-minutes conversation session.
- We used a camera to capture the expression of the elderly throughout the sessions.

E. Materials

This experiment used photo categories of "Food", "Events", "School", and "Commodity". Each category was with ten photos.

F. Methods

- We prepared four photo categories of "Food", "Events", "School" and "Commodity", and ten photos for each category by Keynote.
- One elderly people and one young volunteer sat down side by side and faced to the screen. Then we displayed each photo on the screen for 1 minute. They were asked to talk about each photo for 1 minute.
- The young volunteer indicated the degree of stress on the stress check sheet every minute for each photo.
- Both the elderly and young volunteer answered the questionnaire for category each time after the conversation was over.
- A camera was used to capture the expression of the elderly throughout the sessions.

- The expression of the elderly was analyzed from the video recordings. We used the major literature "Expression Analysis" techniques to know which photo made the subject feel happy [9]. We framed the expression images of each category every second. Frame images for the number of conversation hours were classified by their respective pictures and analyzed by using the frame images. By making it into a frame, it is possible to delete facial expressions of parts other than the conversation in the video, and it can be thought that picture and emotion can be related more. We evaluated the degree of the smile on a frame-by-frame basis. We define 0% as a state of expressionlessness, and 100% as the state of the highest degree of smile. We used the highest degree of the smile as a result of each photo.

III. Protocol Analysis Method

Table I shows an example of protocol analysis method.

TABLE I. EXAMPLE OF PROTOCOL ANALYSIS

Speaker	Utterance Data		
	Utterance Contents	*Utterance Amounts*	*Speaking Time*
Young	It's a blackboard.	1	1. 3"
Elderly	Is it...?	1	0. 9"
Young	Lei Feng	1	0. 9"
Elderly	En, Lei Feng	2	1. 4"
Young	Is it that Learn-from-Lei Feng a few years ago?	1	3. 6"
Elderly	Yes yes.	1	1"
Young	We had to Learn from Lei Feng when we were children.	1	2"
Elderly	Yes yes.	1	1. 3"

- Transcribed all utterances of subjects from the video recording, while "Speaker", "Utterance Contents", "Utterance Amounts", "Speaking Time" are recorded for each photo. The speaking time used the time code of video recording.
- Extracted out photos that were supposed to be easy and difficult for the conversation on the results of stress check, expression analysis, and speaking time.
- Extracted out the key topic of photos above from the transcribed utterance protocol on the results of utterance amounts, which was classified to estimate their common features.

IV. Result

1) Questionnaire for Category

From the results of the questionnaire, we knew that all photo categories of "Food", "Events", "School", and "Commodity" got the less stress for subjects under stage 3, which can help them talk easily [7].

2) Stress Check & Expression Analysis for Photo

Fig. 4 shows the relationship between the stress of young volunteer and expression of the elderly for each photo. The photos with both of no stress and happy are "Children's Day", "Duanwu Festival", "Spring Festival Evening", "Popcorn", "Soy Milk", "Fried Rice", and "Television". On the contrary, the extreme unhappy photo is "Dumpling".

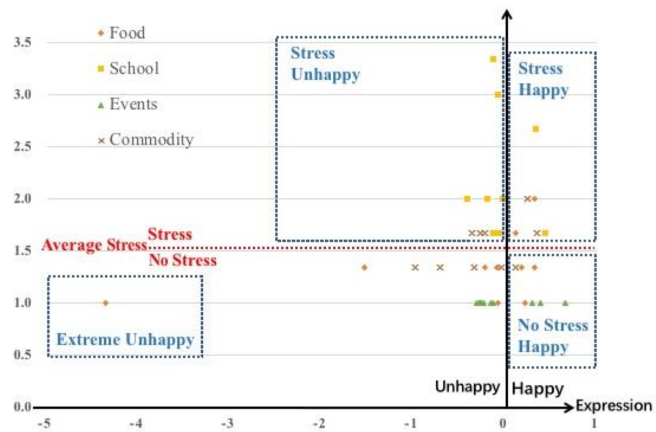


Fig. 4. Relationship between stress and expression.

3) Speaking Time for Photo

Fig. 5 shows the percentage of speaking time of the elderly for "Food" category. Because each conversation unit for each photo is the same 60 seconds, and the average speaking time of the elderly for each photo is 20.3 seconds. The percentage is over 34% (20.3 seconds) which is supposed to be good photo making conversation easy. From this figure, it is known that all photos in "Food" category are good.

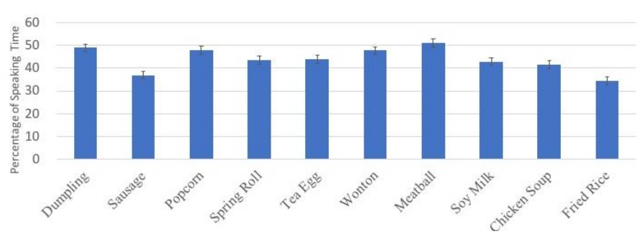


Fig. 5. Percentage of speaking time of the elderly for food.

Fig. 6 shows the percentage of speaking time of the elderly for "Events" category. From this figure, it is known that only photo of "Spring Festival Evening" is good.

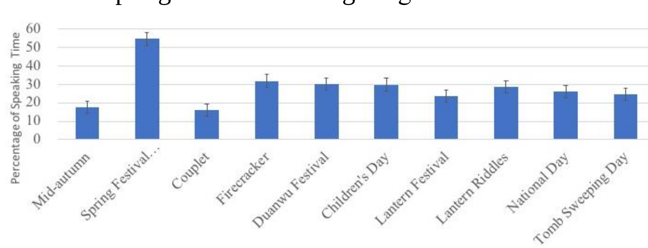


Fig. 6. Percentage of speaking time of the elderly for events.

Fig. 7 shows the percentage of speaking time of the elderly for "School" category. From this figure, it is known that only photo of "School Bag" is good. On the contrary, the extreme short speaking time is "Table-tennis".

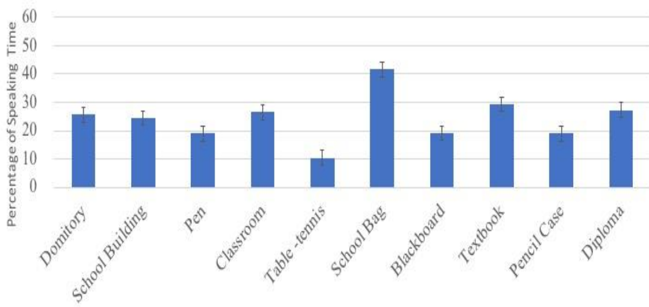


Fig. 7. Percentage of speaking time of the elderly for school.

Fig. 8 shows the percentage of speaking time of the elderly for "Commodity" category. From this figure, it is known that photos of "Sewing Machine", "Lunch Box", "Bicycle", "Kerosene Lamp", "Cold Cream", "Quit", "Hot Pot", "Small Coal" are good.

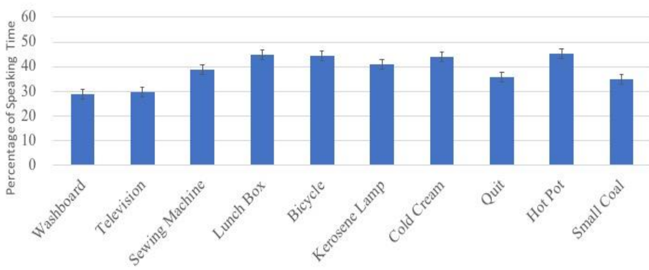


Fig. 8. Percentage of speaking time of the elderly for commodity.

Even if all categories are easy for conversation, the result of each photo is very different, from which we get the best photos with no stress, happy, and speaking time above 34%, that is "Spring Festival Evening" (Fig. 9), "Popcorn" (Fig. 10), "Soy Milk" (Fig. 11), and "Fried Rice" (Fig. 12). On the contrary, we also get the extreme unhappy photo of "Dumpling" (Fig. 13), and extreme short speaking time of "Table-tennis" (Fig. 14).

Fig. 9. Spring festival evening.

Fig. 10. Popcorn.

Fig. 11. Soy milk.

Fig. 12. Fried rice.

Fig. 13. Dumpling.

Fig. 14. Table-tennis.

4) Utterance Amounts for Photo

We analyzed the utterance contents of photos above and extracted out key topics on the result of utterance amounts.

Fig. 15 shows the percentage of utterance amounts of key topic for each photo.

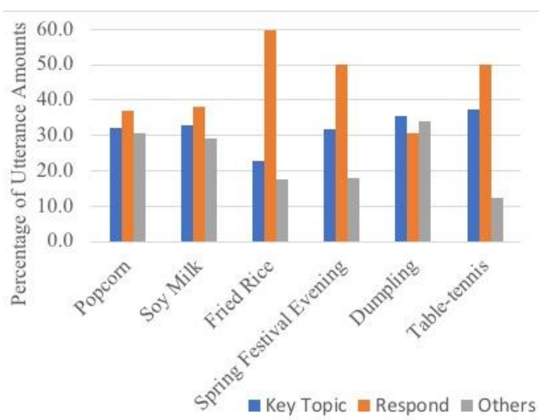


Fig. 15. Percentage of utterance amounts of key topic.

a) Spring Festival Evening: The most utterance amounts of the elderly are responding to the young partner. Except it, the key topic with the most utterance amounts concerns “Presenters of Spring Festival Evening”.

b) Popcorn: The most utterance amounts of the elderly are responding to the young partner. Except it, the key topic with the most utterance amounts concerns “How to Make Popcorn”.

c) Soy Milk: The most utterance amounts of the elderly are responding to the young partner. Except it, the key topic with the most utterance amounts concerns “How to Make Soy Milk”.

d) Fried Rice: The most utterance amounts of the elderly are responding to the young partner. Except it, the key topic with the most utterance amounts concerns “How to Make Fried Rice”.

e) Dumpling: The most utterance amounts of the elderly are key topic concerning “Northerners Like Dumpling” and “Southerners Don’t Like Dumpling”.

f) Table-tennis: The most utterance amounts of the elderly are responding to the young partner. Except it, the key topic with the most utterance amounts concerns “Countryside” and “Shabby”.

V. Discussion

We classified the key topics of photos above:

A. Easy for the Conversation

1) Making Food Method

For the photo of “Popcorn”, “Soy Milk”, and “Fried Rice”, the key topic with the most utterance amounts of the elderly is that how to make the food and materials for making.

2) Recent Famous People

For the photo of “Spring Festival Evening”, the key topic with the most utterance amounts of the elderly is presenters of Spring Festival Evening, who are recent famous people in China.

From the result that the most utterance amounts of the elderly are responding to the young partner, we knew that these topics also made the most interaction between the elderly and the young volunteer.

B. Difficult for the Conversation

1) Different Area

For the extreme unhappy photo of “Dumpling”, the key topic with the most utterance amounts of the elderly is that northerners like the dumpling, not southerners. In China, different area has wholly different customs.

2) Shabby Things

For the photo of “Table-tennis”, the elderly talked few, only about what it looks like the countryside, and everything is shabby. Although photos of 1960’s are nostalgic, the elderly don’t like shabby things of that time.

VI. Conclusion

From this experiment, we got the common features of photos making the elderly feel easy to talk with the young volunteer as follows:

- Recent famous people.
- Diet can be done oneself.

This conclusion is similar to the related work on photo contents in Japan that regardless of age, what is familiar in the current environment can become a category of common interest for the young people and the elderly. For example, when the young people see food they enjoy, the elderly can teach how to make it. This is a good topic that makes both the elderly and the young people feel happy [6].

Meanwhile, we have to avoid the features making conversation difficult:

- Different area.
- Shabby things.

This conclusion is also similar to the related work in Japan that in conversations, the burden of the young people will increase when there is a category with less commonality between the young people and the elderly [6].

For examining the common features of photo contents mentioned above, we should use more photos and do more experiments in China.

VII. Future Works

In this study, we knew that the “Food”, “Events”, “School”, “Commodity” were suitable categories for conversation for the elderly in China, in which only four photos got the results with no stress, happy, and long speaking time. We should prepare more photos for the next experiment to examine more optimal photos and common features of them.

Because measure device is difficult for the elderly, in future research, we expect to evaluate the elderly and young volunteer’s emotional reaction to photos with sentiment analysis.

The young volunteers of this study were all the caregivers in this study. For the next experiment we should let the young people without any care experience become the conversation partners.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 15H01698.

References

[1] Statistical Communiqué of the People's Republic of China on the 2017 National Economic and Social Development, National Bureau of Statistics of China, 2018.

[2] The National Medium and Long-Term Development Plan for Civilian Talents (2010-2020), Ministry of Civil Affairs of the People’s Republic of China, 2011.

[3] China Civil Affairs’ Statistical Yearbook (Statistics of China Social Services), Ministry of Civil Affairs of the People’s Republic of China, 2017, p.134.

[4] M. Otake, M. Kato, T. Takagi, and H. Asama, “Coimagination Method: Communication Support System with Collected Images and its Evaluation via Memory Task”, Universal Access in Human-Computer Interaction Constantine Stephanidis (Ed.), LNCS 5614, Springer-Verlag, 2009.

[5] M. Iwamoto, N. Kuwahara, K. Morimoto, “Comparison of Burden on Youth in Communicating with Elderly using Images Versus Photographs”, vol.6, No.10, IJACSA, 2015.

[6] M. Iwamoto, N., Kuwahara, K., Morimoto, “A Study of Conversation Support System between the Elderly Person and Young Adults by Using Facial Expression Analysis”, Springer-Verlag, 2011.

[7] X. Zhou, M. Iwamoto, N. Kuwahara, K. Morimoto, “A Study of Photographs as Communication Content for Intergenerational Conversation Support System,” Lecture Notes in Computer Science, vol 10287, Springer, Cham, 2017.

[8] H. Kaioho, E. Harada, Introduction to Protocol Analysis. Shin-Yo-Sha, Japan, 1993.

[9] T. Ota, M. Tamura, M. Arita, N. Kiso, Y. Saeki, “Expression Analysis-comparison of the Characteristics of the Facial Expression”, Ekman, 2010.

Mutual Coupling Reduction of MIMO Antenna for Satellite Services and Radio Altimeter Applications

Saad Hassan Kiani
Electrical Engineering Department
Iqra National University
Peshawar, Pakistan

Ahsan Altaf
School of Electrical and Electronics Engineering
Istanbul Medipol University
Istanbul, Turkey

Khalid Mahmood
Electrical Engineering Department
University of Technology
Nowshehra, Pakistan

Alex J. Cole
School of Engineering and Digital Arts
University of Kent
Canterbury, United Kingdom

***Abstract*—Ground irregularities also known as defected ground structures (DGS) is a freshly presented innovatory way in designing of patch antennas to boost up the performance of antenna constraints. This study presents a novel proposal of ground irregularities or defected ground structure is proposed for suppression of mutual coupling effects among 2x1 multiple input multiple output patch array designed on Roggers Duroid 5880. The two adjacent M shape structures surrounding Dumbbell Shaped structure and sandwiched between Dumbbell shape patterns showed the significant level of surface wave suppression up to -42dB while maintaining the gain of 4.7dB and 5.6dBi of directivity. The patch array operates at 4 to 4.3GHz for Fixed and Radio satellite services (FSS) and (RSS) and radio altimeter application systems.**

Keywords—Multiple input multiple output (MIMO); mutual coupling; defected ground structures (DGS); fixed satellite services (FSS); radio satellite services (RSS); radio altimeters

I. Introduction

With the rapid advancement of technology, the communications industry has seen significant growth in order to fulfill the criteria of higher consumer data rate demand. Antenna configuration has seen a very large amount of interest among antenna designers and researchers. In particular, the microstrip patch antenna has been the subject of much research because of its unique offering features as light weight, low fabrication cost and conformal geometry. Antennas offer higher performance and higher compatibility in array topologies, to which microstrip patches adapt well. Multiple Input Multiple Output (MIMO) technology has gained significant attention in the design of state-of-the-art wireless communication schemes due to its ability to increase channel capacity whilst maintaining bandwidth. MIMO antenna systems require very good element isolation in order to avoid mutual coupling of surface waves, but simultaneously necessitate a reduced size for their incorporation into portable, handheld devices. Coupling amongst the patch structures is typically significant in MIMO structures. Mutual coupling between probe elements or patches is an undesirable spectacle that alters the conductance of radiating elements as each patch element radiates over the air or through ground plane surface currents conduction. Furthermore, in applications such as Imaging Radar System and aircraft radio altimeter applications, coupling phenomena requirements is to be as minimum as possible [1]. In order to control surface waves, Electromagnetic band gap (EBG) structures and Defected ground structures and line resonators have been most common techniques [2]-[5]. Defected ground structures are used to control size of antenna significantly and producing multiband response [6] as their shape aims to control the propagation of waves from ground to radiating element through substrate. Their ability to control electromagnetic suppression just like of Electromagnetic band gap structures gives them edge as their implementation is more easy as compared to others [7], [8]. In [9] more nearly-20dB of isolation among elements is presented. Increasing higher layers of EBG structures causes abnormal response [10]. In [11] dumbbell with spiral rings showed better response in isolation of array with higher efficiency. In this study a novel design of DGS is presented providing up to -42dB of isolation and antenna performance parameters. CST microwave studio 2014 is used for designing the antenna. This paper is presented as under.

Section I covers introduction area. Section II covers single element and array configuration and structure of novel design. Section III covers the discussed results and last conclusion. Ease of Use

II. Antenna Design

A. Single Element Design

Roggers Duroid 5880 is taken as substrate with relative permittivity of 2.3 due to its behavior noted in [12]. The patch size is for fundamental frequency of 4.1GHz. Following formulas are used to sum patch dimensions [13].

$$W = \frac{c}{2f0\sqrt{\frac{\varepsilon r+1}{2}}} \tag{1}$$

Where εr = relative constant and f_0 =functioning frequency

$$L = L(eff) - 2\Delta L \tag{2}$$

Where

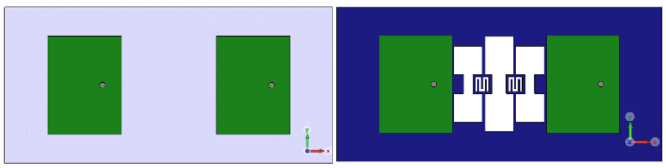
$$L(eff) = \frac{c}{2f0\sqrt{\varepsilon(reff)}} \quad (3)$$

And

$$\varepsilon_{(reff)} = \frac{\varepsilon r+1}{2} + \frac{\varepsilon r-1}{4}\left(1+\frac{12h}{W}\right)^{-1/2} \quad (4)$$

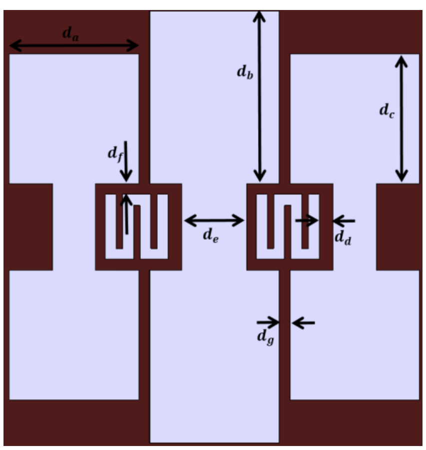
B. Array Design

The array is composed of two identical patches separated by half wavelength distance of 34mm. The overall dimension of array is 7776mm^2. With and without isolating structure array is shown in Fig. 1. The substrate thickness is taken 2mm and height of both ground plane and patches are kept 0.787mm.

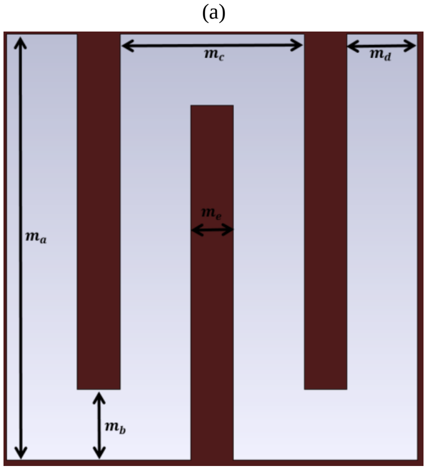


(a) (b)

Fig. 1. Conventional antenna (b) Proposed antenna.



(a)



(b)

Fig. 2. (a) Whole isolating structures; (b) M Shape isolating structure.

The isolating structure is composed of three dumb bells with two M shaped structures. The central dumb bell is surrounded by M shaped structures followed by smaller dumb bells. The distance between all dumb bells is kept 2.5mm and in that distance M shape is introduced. Both isolating structures are shown in Fig. 2(a) and (b) respectively. The edge to edge distance is kept 20mm and center to center is kept 34.5mm. A 1.9mm whole is also made in ground plane to connect SMA connector inner probe to the patches.

The matching impedance is set to be 50ohm. The overall dimensions of dumb bell and M shaped isolating unit are given in Tables I and II.

TABLE I. DIMENSIONS OF DUMB BELL STRUCTURE

Parameters	***da***	***db***	***dc***	***dd***	***de***	***df***	***dg***
Value (mm)	6	8	6	0.025	3	0.025	0.025

TABLE II. DIMENSIONS OF M STRUCTURE

Parameters	***ma***	***mb***	***mc***	***md***	***me***
Value (mm)	3	0.5	1.2	0.5	0.2

ma, mb, mc, md, me are the dimensions of M shape isolating structure, whereas *da, dc* are the dimensions of central dumbbell while *db, de* are the dimensions of side dumbbells respectively. *df, dd, dg* are the separating distance among isolating elements. Patterns are introduced in efforts to change the electromagnetic conduction properties of structures. These structures relative to size of wavelength acts usually as block filers for certain number of frequencies and can be considered as LC network. Surface wave's interference reduction was seen with insertion of isolating structure.

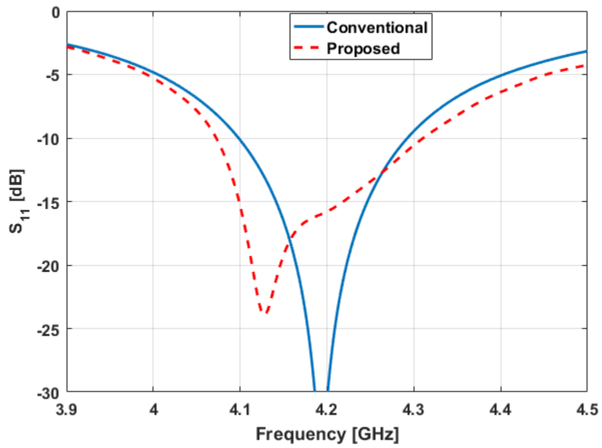
III. Results and Discussion

The proposed structure as stated was designed in Computer Simulation Technology 2014 version. Performance parameters like return loss, gain, directivity, bandwidth, and VSWR, E and H fields are discussed with and without isolating structure. The detail results comparison are shown in Table III.

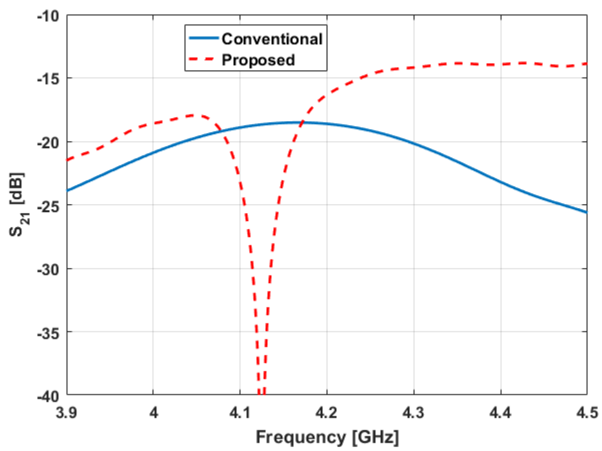
TABLE III. PERFORMANCE PARAMETERS

Parameters	**Conventional**	**Proposed**
Return Loss,S11	-30.25dB	-28.0dB
Coupling, S22	-18.00dB	-42.05dB
Gain	5.18dB	4.82dB
Directivity	6.15dBi	6.8dBi
VSWR	1.05	1.25
Bandwidth	180MHz	200MHz

The return loss plot or reflection coefficient was slightly decreased in proposed antenna with gain but overall directivity patterns have increased. The return loss of both designs is shown in Fig. 3(a) and isolation enhancement with and without proposed defected ground structure is shown in Fig. 3(b). Isolation enhancement is taken in terms of S12 as only one of the patch elements has been excited. It is clearly visible that up to -28dB of isolation is increased summing up to -42dB which is clearly much more them [14], [15].



(a)



(b)

Fig. 3. S parameters of conventional and proposed antenna.

The proposed antenna has shown good bandwidth results with addition of 20MHz. Also antenna has shown satisfactory results and has shown minimum miss match losses.

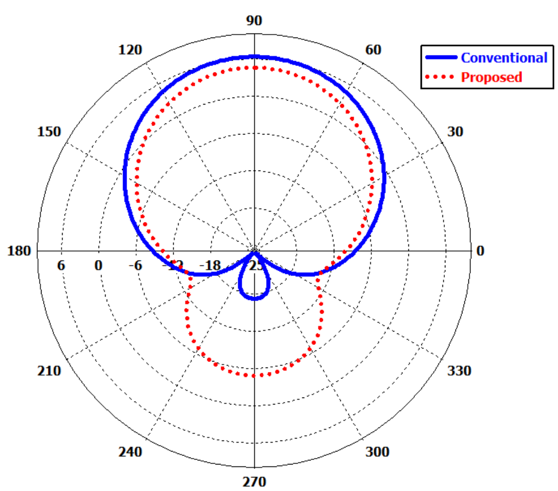


Fig. 4. E field pattern.

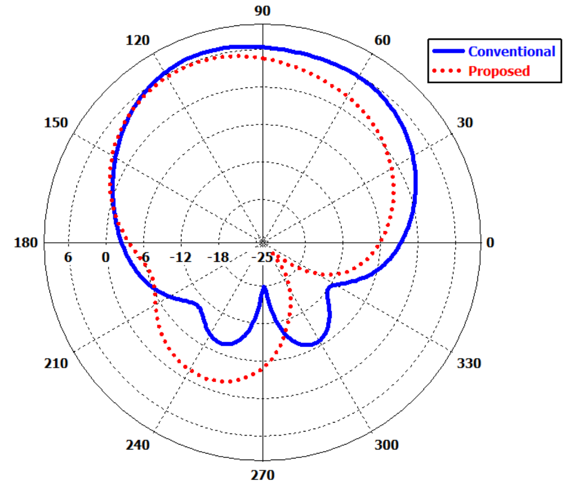


Fig. 5. H field pattern.

The standardized Gain pattern in E and H planes of conventional and proposed structures is shown in Fig. 4 and 5 respectively. The H plane patterns have been tilted a little bit because of the proposed structure and so the direction of current of excited patch is perpendicular to the passive patch.

From figures it is cleared that existence of DGS have not altered radiation patterns. Only the patterns are marginally propped since as of asymmetry in the plane from which it is evident that not only proposed structure successfully blocked the surface waves but also mends the patterns.

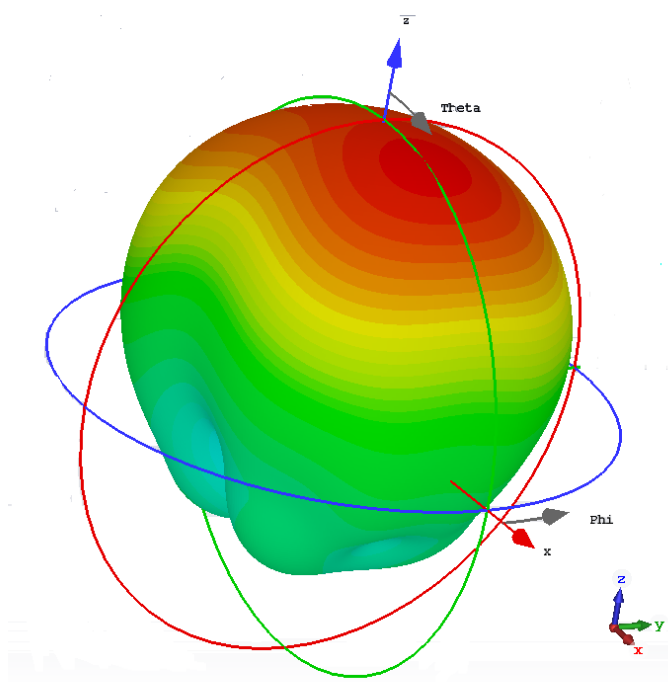


Fig. 6. 3D view of far field.

Fig. 6 shows the 3D directivity view of proposed antenna. The proposed antenna shows high directivity results of 6.8dBi with main lobe direction of 16 degrees.

IV. Conclusion

A fresh DGS design is presented in this study to improve isolation enhancement between MIMO antennas. A very simple structure is designed comprising of two shapes M and Dumb bells using Computer simulation technology software. By insertion of proposed patterns reduction of mutual coupling was achieved up to more than 20dB and was reduced to sum of -42dB also improving 20MHz of bandwidth. The antenna showed good performance parameters results such as gain, directivity, miss match losses with minor alteration in radiation patterns. The isolated embedded structure antenna can be used for satellite services and for radio altimeter applications.

V. Future Work

The proposed antenna work in future can be extended for circular polarized antennas that are suitable for avionics, aerospace and satellite applications. Also this work can be extended to enhance isolation among MIMO elements by increasing DGS elements.

References

[1] Mekala, C., Saranya, P., & Narayanan, V. S. (2014). Survey on Mutual Coupling Reduction Techniques for Imaging Radar Application. i-Manager's Journal on Communication Engineering and Systems, 3(3), 28.

[2] Soltani, Saber, and Ross D. Murch. "Designing planar MIMO antennas using a novel canonical dual port antenna." *Antennas and Propagation Society International Symposium (APSURSI), 2014 IEEE*. IEEE, 2014.

[3] Altaf, Ahsan, M. A. Alsunaidi, and Ercument Arvas. "A novel EBG structure to improve isolation in MIMO antenna." *USNC-URSI Radio Science Meeting (Joint with AP-S Symposium), 2017*. IEEE, 2017.

[4] S. D. Assimonis, T. V. Yioultsis, C. S. Antonopoulos, "Design and Optimization of Uniplanar EBG Structures for Low Profile Antenna Applications and Mutual Coupling Reduction", Antennas and Propagation IEEE Transactions on, vol. 60, pp. 4944-4949, 2012

[5] Peng, L., Sun, K., Xie, J. Y., Qiu, Y. J., & Jiang, X. (2016, October). Coupling reduction of closely packed antennas by stringing EBG structure. In Antennas, Propagation and EM Theory (ISAPE), 2016 11th International Symposium on (pp. 716-719). IEEE.

[6] Saad Hassan Kiani, Shahryar Shafique Qureshi, Khalid Mahmood, Mehr-e- Munir and Sajid Nawaz Khan, "Tri-Band Fractal Patch Antenna for GSM and Satellite Communication Systems" International Journal of Advanced Computer Science and Applications(IJACSA), 7(10), 2016.

[7] Karthik, R., Manjunath, R. K., & Kumaraswamy, H. V. (2015). Rectangular slotted ground structure to reduce Mutual coupling in Microstrip Patch antenna array. International Journal of Advanced Research in Electronics and communication Engineering (IJARECE), 4(5).

[8] Altaf, Ahsan, M. A. Alsunaidi, and Ercument Arvas. "A novel EBG structure to improve isolation in MIMO antenna." USNC-URSI Radio Science Meeting (Joint with AP-S Symposium), 2017. IEEE, 2017.

[9] Ahmed, M. I., Sebak, A., Abdallah, E. A., & Elhennawy, H. (2012, June). Mutual coupling reduction using defected ground structure (DGS) for array applications. In Antenna Technology and Applied Electromagnetics (ANTEM), 2012 15th International Symposium on (pp. 1-5). IEEE.

[10] Li, Q., Feresidis, A. P., Mavridou, M., & Hall, P. S. (2015). Miniaturized double-layer EBG structures for broadband mutual coupling reduction between UWB monopoles. IEEE Transactions on Antennas and Propagation, 63(3), 1168-1171.

[11] Zulkifli, F. Y., Rahardjo, E. T., & Hartanto, D. (2010). Mutual coupling reduction using dumbbell defected ground structure for multiband microstrip antenna array. Progress In Electromagnetics Research Letters, 13, 29-40.

[12] Kiani, Saad Hassan, Khalid Mahmood, Umar Farooq Khattak, Burhan-Ud-Din, and Mehre Munir. "U Patch Antenna using Variable Substrates for Wireless Communication Systems." International Journal of Advanced Computer Science and Applications 7, no. 12 (2016): 286-291.

[13] Balanis, Constantine A. Antenna theory: analysis and design. John Wiley & Sons, 2016.

[14] Tu, D. T. T., Van Hoc, N., Quan, H., & Van Yem, V. (2016, July). Compact MIMO antenna with low mutual coupling using defected ground structure. In Communications and Electronics (ICCE), 2016 IEEE Sixth International Conference on (pp. 242-247). IEEE.

[15] Veeramani. A, Afsane Saee Arezomand, Vijayakrishnan. J, Ferdows B. Zarrabi, "Compact S-shaped EBG Structures for Reduction of Mutual Coupling," 2015 Fifth International Conference on Advanced Computing & Communication Technologies, 2015.

Detection of Soft Atheroscelarotic Plaques in Cardiac Computed Tomography Angiography

Muhammad Moazzam Jawaid
Department of Elec. & Electronics Engg.
City University London, EC1V0HB

Sajjad Ali Memon, Imran Ali Qureshi, Nasrullah Pirzada
Department of Telecommunication Engineering
Mehran University of Engineering & Technology, Jamshoro 76062

***Abstract*—Computed tomography angiography (CTA) has turned non-invasive diagnosis of cardiovascular anomalies into a reality as state-of-the-art imaging equipment is capable of recording sub-millimeter details. Based on high intensity, the calcified plaques are easily identified in cardiac CTA; however, low density based non-calcified plaque detection has been a challenging problem in recent years. We propose an efficient method in this work for automated detection of the non-calcified plaques using discrete radial profiles. The plaque detection is accomplished using support vector machine to differentiate abnormal coronary segments. We investigated a total of 32 CTA volumes and the detection mean accuracy of 84.6% was achieved, which is in-line with the reported literature.**

***Keywords*—Coronary tree segmentation; support vector machines; non-calcified plaque detection; mean radial profiles; Rotterdam CTA dataset**

I. Introduction

Coronary heart disease (CHD) refers to the deposition of materials (also termed as coronary plaques) inside coronary arteries. The growth of plaque results in a reduced blood flow towards heart muscles. Consequently, the heart muscles become oxygen starved resulting in fatal cardiac consequences including angina, heart failure and arrhythmias. In context of the flow of this paper, we present relevant literature in Section I-B, which is followed by the clinical data description. In the subsequent section, the proposed model is explained, followed with results of Section III. The last section presents some limitations and the future extension for this work.

A. Clinical Motivation

According to the fact sheet of the World Health Organization [1], CHD was the leading cause of death globally in 2013, with 8.14 million deaths (16.8%) compared to 5.74 million deaths (12%) in 1990. Moreover, the recent statistics of the National Health Services, United Kingdom [2] reveals that over 2.3 million people in the United Kingdom suffer from CHD where the annual death toll is approximately 73,000 (an average of one death every seven minutes). The huge levels of growing morbidity and mortality have led to a increased interest abnormality detection methods. From a clinical point of view, the detection and quantification of arterial plaque can help physicians avoid or at least delay the worst cardiac events by addressing behavioural risk factors [3]. State of the art developments [4] in non-invasive imaging technology have revolutionized the clinical diagnosis methods in recent years. For instance, sub-millimeter based acquisition of the internal organs has made CTA a feasible alternative to cardiac cauterization for detecting coronary obstruction [5]; however, the composition of the coronary plaques pose a difficult challenge in the effective diagnosis. High intensity calcified plaques can be detected easily in CTA imagery [6]–[9]; however, the detection of the non-calcified plaques has been a challenging problem in clinical practice due to close proximity with blood voxels.

Clinically, the non-calcified plaques have been established as the most important indicator of acute coronary syndromes due to their fragile nature [10]. Moreover, unexpected rupture has made soft plaques much threatening, i.e. for many individuals, sudden death becomes the first sign of soft plaque in contrast to the calcified plaques which often lead to disease symptoms at early stages. In addition, the positive remodeling associated with soft plaques further amplifies the detection challenge as the radial stenosis detection based methods often miss the non-calcified plaques [6]–[9], [11]. Consequently, the intense focus of the current research is an early detection of soft plaques to predict and avoid worst cardiac events [12].

B. Related Work

Based on the fact that soft plaque detection is a complex phenomena, there exists a little literature [13]–[17] addressing automatic detection of soft plaques in CTA imagery. Out of the reported work, the majority of the research have been clinical pilot studies or generic anomaly detection techniques. One framework in this context was proposed by Clouse *et al.* [13]; however the main focus was the quantification of manually identified soft plaques. Accordingly, a total of 49 coronary segments (41 normal, 8 abnormal) were chosen for investigation from a dataset of 40 CTA volumes, to validate the proposed quantification method. For precise quantification of soft plaques, the authors established correspondence between two normal cross sections at the terminal sites of the plaque region to approximate the outer boundary of the vessel. In the subsequent step, all voxels having intensity equal of the lumen were subtracted and those left over were identified as soft plaque. Accordingly, the research illustrated that the soft plaques can be quantified in CTA; however, the results were based on certain manual inputs i.e. a pre-selective set of the segments was used in investigation with a manual selection of plaque terminal points. An extension of this work further validated the quantification correlation among two imaging modalities (i.e. CTA and intravenous ultrasound (IVUS) plaque quantifications [18]. Similar to the base study, the selection of the coronary segment was made in a pre-processing stage for optimal results. Accordingly, 20 soft plaque effected segments were chosen from a set of 12 CTAs. Despite of the successful correlation , this method does not fulfil the automated spirit as it was based on manual selection of the plaque positions.

Machine learning based soft plaque detection was first reported by Wei *et al.* [14], in which authors employed a linear discriminant analysis (LDA) to minimize the false positives for a set of 120 pre-selected plaque candidates. According, the efficiency of the LDA classifier was based on NCP candidate selection criteria, and the machine learning was used to maximize the performance by suppressing false candidates. Starting with a manually corrected coronary centreline, the vessel radius along the length of the centreline was obtained in the first stage. In a subsequent stage, the obtained radius was used to identify seed points for 2 mm long plaque candidate regions. From a set of 83 CTA volumes, 120 plaque candidate regions were used in detection process, for which the reported sensitivity was 92.5%. Another use of learning method was reported by Tessman [17], in which coronary stenosis effected cross-sections were detected. In the first step, the pre-extracted coronary centreline was used to map the vessel segment with a series of multi-scale overlapping cylinders to identify the sampling points inside the segment. Subsequently, image based features like intensity, gradient and the first-second order derivatives were extracted at the sampled points to identify high intensity calcifications. Moreover, global features including image mean, entropy and variance were used in combination with Haar-like features to detect the low intensity soft plaques. According to the reported results, the plaque detection accuracies were 94% and 79%, respectively for two classes of plaques i.e. calcified and non-calcified. It should be noted that the low accuracy for non-calcified plaques illustrate that soft plaque detection demands a more sophisticated system i.e. beyond stenosis based computations to efficiently address vessel remodelling.

An important method focusing on cross-section based vascular abnormality detection was proposed by Zuluaga *et al.* [19]. Based on the "density level detection" technique of Steinwart [20], authors employed an unsupervised learning approach in this work for detecting abnormal cross-section. In this method, the vascular cross-sectional images were discretely sampled around centreline to derive the feature set for suppressing outliers Subsequently, they used an SVM model trained on normal cross sections to label the outliers (i.e. the cross sections which violate the intensity pattern of normal class) as abnormal. According to the reported results, a good detection rate of 79.62%, was reported for 9 clinical CTA datasets; however, the selection of anomaly concentration parameter ρ plays an important role in overall results. In addition, a large number of normal cross-sections having similar intensity pattern are required for good training of SVM due to one-class nature of supervisor.

Similarly Renard and Yang [15], Lankton *et al.* [16] and Li *et al.* [21] proposed different approaches for the plaque detection in CTA; however, these method were validated for small datasets and require certain manual inputs from the user, which preclude the automated solutions. Likewise, a number of plaque quantification algorithms [22]–[25] have been proposed in recent years with a motive of correlating CTA with intra-vascular ultrasound (IVUS) measurements; however, these methods again employ manual input in terms of the plaque position and length in respective coronary vasculature.

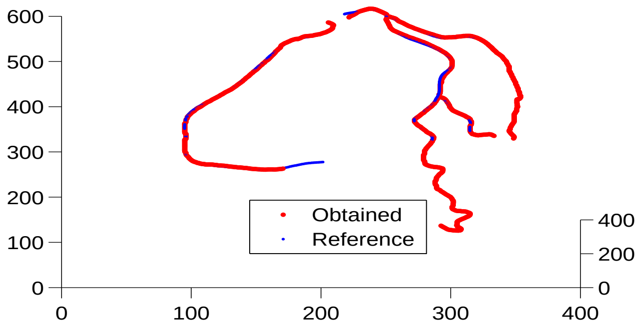
In contrast to manual input based quantification, we addressed the problem of automated detection of plaque in this work. Accordingly, our contribution is an efficient method for the detection of the non-calcified plaque in coronary vasculature. We employed a machine learning technique (SVM) for identifying non-calcified plaque affected coronary segments. It should be noted that the proposed method differs from LDA based Wei *et al.* [14] classifier and anomaly detection methods of [17], [19] in the sense that we segment coronary tree in CTA using hybrid energy formulation. Accordingly, the segment radial information based on the segmented tree is employed in classification to handle both positive and negative remodeling associated with the soft plaques. Experimental results demonstrate that the proposed method achieves a good agreement (detection accuracy of 88.4% with respect to manual annotations), and in-line with anomaly detection methods of [17], [19].

C. CTA Dataset Acquisition

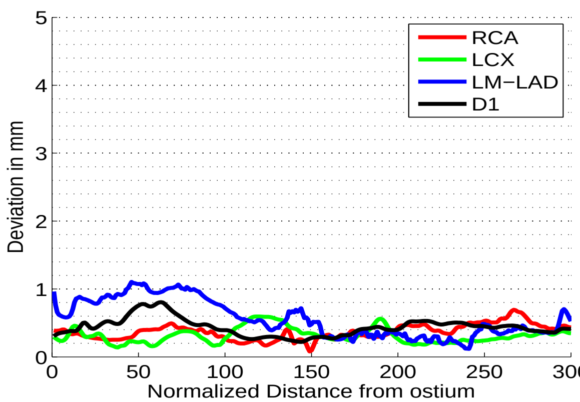
In this study, we have investigated three CTA datasets in context of the soft plaque detection. A first dataset comes from public database of Rotterdam Coronary Artery Evaluation framework. This dataset contains 18 CTA volumes coming from different scanners and different institutes as explained in [26], [27]. The second dataset consisting of 12 clinical CTA volumes was obtained from Guys & St. Thomas's Hospital London. In addition, a third dataset consisting of only two CTA volumes was obtained from Semmelweis University, Budapest Hungary. It should be noted that the multi-vendor data increases the complexity of the plaque detection problem; however, the reproducibility of the method can be validated successfully. Moreover, all the CTA data provides complete ground truth in terms of segment nature (normal or plaque effected), type and position of the plaque in abnormal segments and stenosis information (if any) for the vessel boundary respectively.

II. Methods

The first step in the plaque detection process is the segmentation of the coronary tree in CTA using a hybrid energy formulation as proposed in Jawaid *et al.* [28]. After tree extraction, we performed the skeletonization using fast marching implementation of the thinning algorithm of Van *et al.* [29]. As the plaque detection method heavily relies upon the centreline accuracy, we evaluated the deviation error with respect to manual reference ground truth as presented in Fig. 1. The visual comparison for complete coronary vasculature is presented in Fig. 1a, whereas the deviation error for individual segments (RCA, LCX, LAD and D1) is shown in Fig. 1b. After generating the tree skeleton, we used 17-segment model of American Heart Association (AHA) [30] to label the individual segment present in respective coronary tree. In the subsequent step, we employed segment-wise centerlines to extract cylindrical volume using interpolation in 3D space. In the final step, the cross-section based cylindrical volume is used in a support vector machine framework to identify abnormalities in segment respectively. For a mathematical interpretation of the paper, let I defines a 3D CTA image and $\mathbf{x}$ represents a spatial



(a) over-riding with reference



(b) Deviation error

Fig. 1. Centreline accuracy with respect to reference centreline [27]. (a, c) shows obtained centreline overlain with reference in 3D space, whereas (b, d) represents mean deviation for major segments in millimeters. It can be observed that mean deviation with respect to the reference is less than or around 1mm.

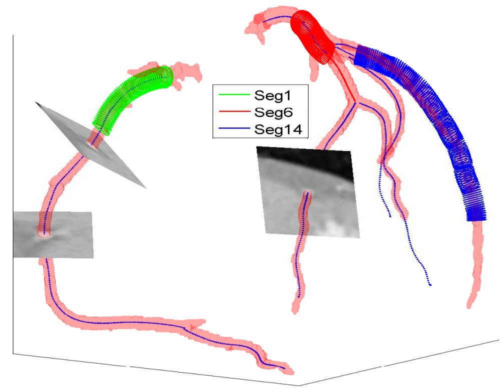
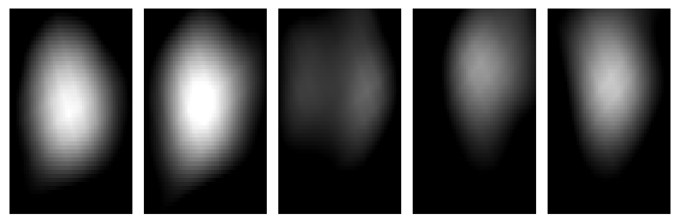
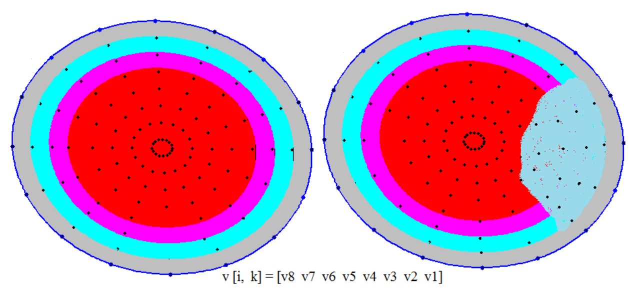


Fig. 2. Cylindrical model for coronary segments using 6mm circle. It can be observed that 3D surface is overlaid with centreline (black) along with oblique planes in 3D space. Moreover, blue, red and green contours represent the curved cylindrical approximations for three segments of AHA model.



(a) normal (b) normal (c) normal (d) normal (e) normal



(f) Intensity composition for normal and abnormal coronary cross sections.

Fig. 3. Cross-section based visualization of coronary segment. (a)-(e) represents grey scale visualization for sequential cross sections, whereas (f) shows colour interpretation in context of concentric rings (dots). It can be observed from (f) that left represents a normal cross section with adequate flow of blood, whereas right shows a plaque leading to blood obstruction.

location in domain Ω. Moreover, it is important to mention that the high intensity based calcifications are regularized in a pre-processing stage to optimize the non-calcified detection as reported in [13]–[16], [18], [21], [31], [32]. Accordingly, the high intensity plaques were assigned lumen intensity value to minimize to work with-in the scope of this research.

A. Cylindrical Modelling of Coronary Segments

In order to identify the intensity in-homogeneity along the coronary tree, we employed the ideal of the mean radial profile. However, in contrast to conventional 2D image based profiles, we used an extended version to detect intensity abnormality in 3D vessel structures. Accordingly, we extracted oblique cross sections along the length of segment by substituting $\mathbf{n}_{xyz} = [n_x, n_y, n_z]^T$ (normal of the plane) and $\mathbf{c}_{xyz}$ (centreline point at respective location) in by (1).

$$\mathbf{n}_{xyz} \cdot (\mathbf{x} - \mathbf{c}_{xyz}) = 0 \tag{1}$$

The normal of the plane is computed using consecutive points of the centreline to precisely follow the vessel orientation.

To effectively represent the coronary segments, we used the diameter for cylindrical model to be 6 mm,as it represents the maximum possible expansion of the normal coronary structure in CTA [11], [33]. This 6mm modelling is illustrated in Fig. 2 where it can be observed that the circumference of the cylindrical model serves interface between lumen and the background. In the subsequent step, we computed the customized radial profile for the coronary segment using discrete approximations as expressed by (2).

$$v[i,k] = \frac{1}{L}\sum_{t=1}^{L} I(r_i, \theta_t, q_k)\ \forall i,k,\ i = 1,...,8,\ \ k = 1,...K \tag{2}$$

where q_k represents the k^{th} cross sectional of coronary segment and K defines the total number of points along the length of the segment. L denotes the total number of projected rays, which is set equal to 16 in our model and the respective projection angle is computed as $\theta_t = t\left(\frac{\pi}{8}\right)$. Moreover, i denotes the concentric ring formed at radius $r_i = 0.4(9-i)$mm. It is important to mention that the discretization parameters are selected to achieve a balance between profile accuracy and processing load. Accordingly, this sampling interval used for radial and cylindrical axis represents 0.4mm (isotropic voxel size), 22.5° angular interval projects 16 rays on the sampling plane for estimation of the radial profile on respective plane. Moreover, the formulation for radius r_i parameter reflects that concentric rings are numbered in an inward fashion, i.e. the outer ring is labeled as v_1 with inner most ring named as v_8 as shown in visual illustration of cross sections in Fig. 3f. It can be observed that in general inner rings (v_5 to v_8) define the blood filled lumen and outer rings (v_1 to v_4) define the interface between lumen and the CTA background.

The concentric ring based labeling phenomena is further demonstrated in Fig. 4, where the intensity response is presented along the length of the coronary segment. It can be observed that four external rings defining external interface assumes low intensity values and remain stable irrespective of the normal or abnormal cross-section, whereas the internal four rings reflect the contrast filled blood in terms of high intensity. Moreover, it can be observed from the figure that normal cross-sections lead to stable response for the inner rings, whereas the presence of low intensity material results in significant concavities for inner four rings (see Fig. 4b).

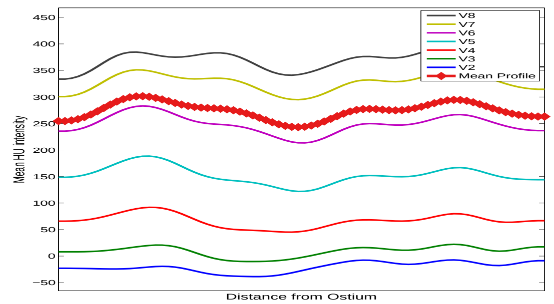
Accordingly, we start with the assumption that this concavity property of the diseases segments can be effectively used in support vector machine based classification. In the subsequent step, we computed mathematical representation (s) of coronary segment using intensity response of four inner rings as follows:

$$s[k] = \frac{1}{4}\sum_{i=5}^{8} v[i,k],\ \forall k,\ k = 1...K \tag{3}$$

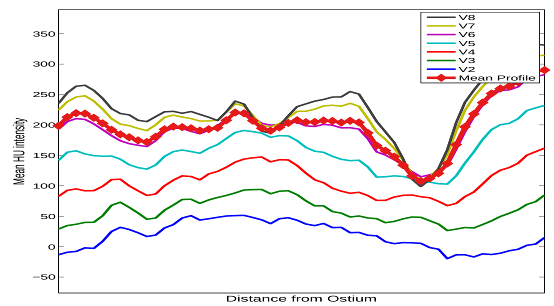
The mean segment representation of the coronary segment often undergoes short term transitions, which are smoothed with the help of moving average operation. Accordingly, we compute the smoothed statistical representation of the segment in terms of moving mean and moving standard deviation using a [1] by [3] moving window as expressed in (4). Moreover, it can be visually observed from Fig. 3f that the different coronary segments have variable lengths. Accordingly, this length variation is apportioned in this step with the help of spline-based interpolation to construct fixed length characteristic functions $\mu_s^{'}$, $\sigma_s^{'}$ (each having 100 samples) for respective coronary segments.

$$\sigma_s[k] = \sqrt{\frac{1}{(2n+1)-1}\sum_{i=-n}^{n}(s[k+i]-\mu[k])^2},$$

$$\mu_s[k] = \frac{1}{2n+1}\sum_{i=-n}^{n} s[k+i],\ \forall k,\ k = 1,...,K \tag{4}$$



(a) CTA Volume 6 Segment 1 (normal).



(b) CTA Volume 5 Segment 2 (abnormal)

Fig. 4. Intensity plot for 8 concentric rings (v_1 to v_8) for two segments (normal and abnormal). It can be observed that central ring (v_8) exhibits HU intensity and outer ring (v_1) assumes lower intensity value. Moreover, the mean representation of the coronary segment is computed by averaging the four inner rings (v_8) - (v_5).

B. SVM Based Segment Classification

1) Feature Based Representation for Coronary Segments: The computation of fixed length characteristic function is followed with SVM based differentiation of the plaque effected coronary segments. The performance of the SVM classifier heavily relies on the selected features, as distinctive features helps classifier in optimal performance, whereas ambiguous features lead to poor accuracy. In context of the non-calcified plaque detector SVM, the intensity plays most important role as it is the only indicator of the non-calcified plaques. Accordingly, we derived hand-crafted features capable to project intensity variations before application of the SVM classifier. Accordingly, we extracted the features by splitting the segment characteristic functions $\mu_s^{'}$ and $\sigma_s^{'}$ into m windows as expressed in (5).

$$f_\mu[m] = \sum_{n=1}^{5}\mu_s^{'}[n+5(m-1)], \forall m = 1,2,...,20$$

$$f_\sigma[m] = \sum_{n=1}^{5}\sigma_s^{'}[n+5(m-1)], \forall m = 1,2,...,20 \tag{5}$$

The idea of m windows is used to exploit the relative variations in intensity along the length of the segment. However, selection of appropriate number of windows is a challenging task as it rationalizes the feature vector dimension at the cost of approximation error. Relationship between subsets and the approximation error is presented in Fig. 5, where it can be observed that the quantization error is inversely proportional to the number of subsets, i.e. approximation improves as the number of windows is increased. In order to maintain a balance between the accuracy and the feature vector size, we defined number of subsets m equal to 20, as the quantization error becomes steady for $m = 20$ as illustrated in the Fig. 5. Accordingly, the discriminative capability of subset based extracted features (f_μ and f_σ) to distinguish the intensity patterns in to two classes is illustrated in Fig. 6 (see Fig. 6a - reffig:fig8f).

Furthermore, an additional parameter namely mid-lumen intensity f_{mid} is added to improve the performance of the SVM classifier. Mathematical formulation for the mid lumen intensity is expressed by (6), i.e. mid-lumen response is acquired by modelling the intensity of the central concentric ring v_8.

$$f_{mid}[m] = \frac{1}{5}\sum_{n=1}^{5} v_8[n + 5(m-1)], \forall m = 1, 2, ..., 20 \quad (6)$$

The visual justification for the additional feature mid lumen intensity is presented in Fig. 6c and 6f. Apparently f_{mid} replicates the distribution pattern of f_μ; however, this feature encodes the concentration of contrast medium in the lumen centre along the length of the segment. It is important to mention that a non-calcified plaque located at the start of the coronary segment results in lower intensity in the mid of lumen; hence, the segment must be labelled as abnormal. However, due to the stable mean and variance along the segment, classifier may erroneously identify segment as normal. Accordingly, the mid lumen feature f_{mid} ensures that the classifier takes into account not only the intensity variations but the mid-lumen response of segment for efficient classification. Next, we concatenate three feature sets f_μ, f_σ and f_{mid} to obtain a feature based representation $\mathbf{Fx}_i$ for respective coronary segment with dimensions [1 x 60].

2) SVM Classification Framework: For a support vector machine based classifier, the input data consists of a feature space along with training labels, i.e. N feature vectors of the form $\mathbf{X}_n$ and the associated binary labels Y_n defining the class of feature vector as normal or diseased as expressed by (7). Here d represent the dimensions of feature vector, i.e. defined equal to 60 in this work and N represents total number of samples in the classifier test.

$$D = \left\{ (\mathbf{X}_n, Y_n) \,|\mathbf{X}_n \subseteq R^d, Y_n \subseteq \{-1, 1\} \right\}_{n=1}^{N} \quad (7)$$

In context of the binary classification problem, the support vector machine computes an optimal hyperplane by minimizing the norm of weights for ideal segregation; however, a slack variable is often integrated to relax the constraints for a feasible solution as expressed in (8).

$$min \left|\mathbf{w}\right|^2 + P\sum_{i=1}^{n} \varepsilon_i \quad (8)$$

$$subject\,to: Y_n\left(\mathbf{w}^T\mathbf{X_n} + b\right) \geq 1 - \varepsilon_i,\ \varepsilon_i \geq 0, for\ i = 1, 2, ...n$$

where $P = 10^0$ defines the penalty cost, i.e. it is responsible for regularizing the influence of individual support vectors in the classification. A small value of P leads to quick and inaccurate classification, i.e. having frequent violations, whereas high value results in slow and accurate classification using hard margin in classification. For mapping data into higher space, we employed a non-linear radial basis Gaussian kernel with σ set equal to 1.

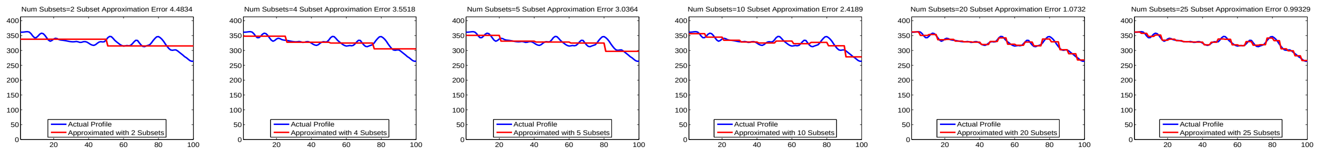
III. Results

A. Results for SVM Classification

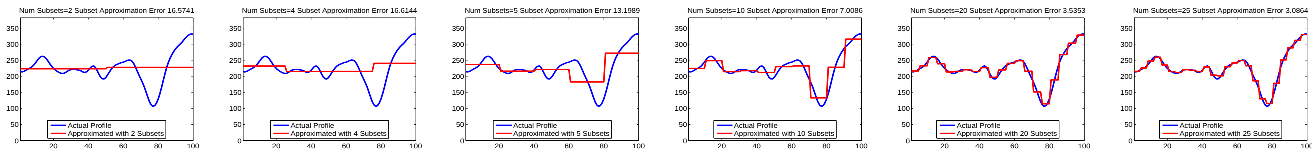
The first step towards verification of the results is the formulation of ground truth reference. For this research work, the ground truth comes along with the CTA image data, i.e. all CTA images accompany manual expert-based segment-wise labels. The corresponding labels indicate the status of the coronary segment in terms normal or plaque affected, and for plaque affected segments the ground truth further reveals the potential position. As the scope of this work is detection of plaque in coronary vasculature using SVM classifier, we therefore employed ground truth in context of normal versus plaque affected coronary segments.

Accordingly, we evaluated the plaque detection performance of the SVM classifier by extracting a total of 344 (200 normal, 144 abnormal) segments from 32 CTA volumes. The statistical validation for detection performance has been performed using Leave One Out (LOO) cross validation as shown in Fig. 7. It is important to mention that for N samples, LOO validation method employs $N-1$ samples in the training and One sample in testing. From computational point of view, LOO validation consumes extra time in comparison with $K - fold$ validation; however, it reveals the true efficiency of the SVM model, as every sample is evaluated individually. It can be observed that a promising sensitivity rate of 92% is achieved. Moreover, positive predictive value for SVM classifier is 81.4%, negative predictive value is 86.9% and overall accuracy of soft plaque detection is equal to 84.6%. A relatively high value for these metrics reveal that the automated detection methods is capable of detecting non-calcified plaques with a good agreement with human expert, which is ultimate theme of any computer assisted application.

Next, we evaluated the performance of the SVM model on three data sets individually to validate the generalization of our model. In this evaluation, we extracted test segments individually from three datasets (66 from Rotterdam, 76 from St. Thomas and 36 from Semmelweis) and SVM classification results are presented in Fig. 8. In the subsequent step, we used the trained SVM classifier to investigate the impact of feature vector dimensions on the classifier efficiency. In this test we used 122 coronary segments extracted from 3 datasets (70 normal and 52 abnormal according to the manual ground truth) and compared the classifier performance in terms of accuracy and processing time. It has been observed that the windowed mean and deviation based 40 features lead to a classification accuracy of 76.8%, where the addition of mid-lumen f_{mid} and features improved the classifier accuracy by approximately 8%. Moreover, the comparative analysis demonstrates that the further increase in the feature space dimensions show only a marginal improvement in the classifier accuracy, while the computational time increases significantly. These results lead to the conclusion that the classifier performance becomes

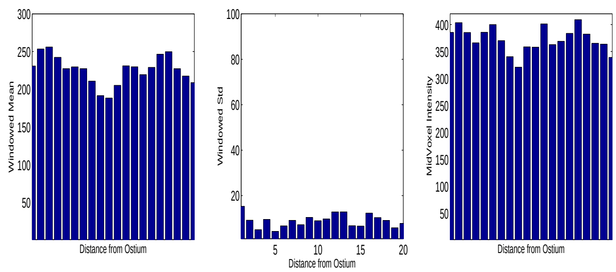


(a) 2 subsets (b) 4 subsets (c) 5 subsets (d) 10 subsets (e) 20 subsets (f) 25 subsets

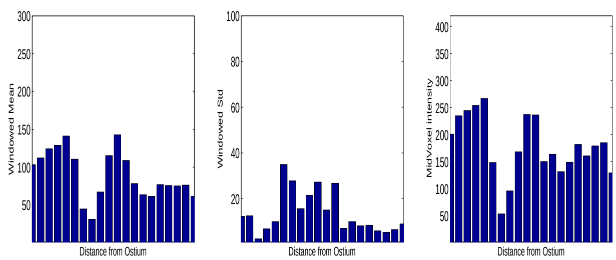


(g) 2 subsets (h) 4 subsets (i) 5 subsets (j) 10 subsets (k) 20 subsets (l) 25 subsets

Fig. 5. Subset based signal representation to reduce the dimensions of the feature vector. It can be observed from a pairwise comparison that both normal (top) and abnormal (below) segments can be adequately represented using 20-25 subsets. The top row represents normal segment and bottom row represents abnormal segments.



(a) Normal windowed mean. (b) Normal windowed std. (c) Normal mid-voxel intensity.



(d) Abnormal windowed mean. (e) Abnormal windowed std. (f) Abnormal mid-voxel intensity.

Fig. 6. illustration for the segment descriptor features. (a, c) shows stable values for moving mean and deviation for a normal segment, whereas (b, d) shows unexpected jumps for a soft plaque effected segment. Moreover (e, f) reflects the discriminatory power of mid lumen intensity i.e normal segment (e) assumes higher HU in contrast to abnormal segment (f) effected with low density soft plaque.

resistant to the feature vector dimensions at a certain point due to the redundancy of features.

IV. Discussion

It can be observed from the performance graph (Fig. 7) that the overall performance for proposed SVM detection model is 84.6% with respect to the clinician based manual detected ground truth, with a sensitivity of 92% and specificity of 80.3%. Moreover, it can be observed from Fig. 8 that plaque detection performance remain consistent around 82% for three individual datasets. The detection performance for three individual datasets can be further explained based on the fact that a "significant dip" in the segment profile ensures greater accuracy for the classifier. Accordingly, a large number of severe plaque instances in the Rotterdam data results in higher sensitivity, whereas the relative low accuracy for Semmelweis CTA data indicates the absence of severe plaque instances. In addition to the manual ground truth based validation, the efficiency of the proposed model is compared with state-of-

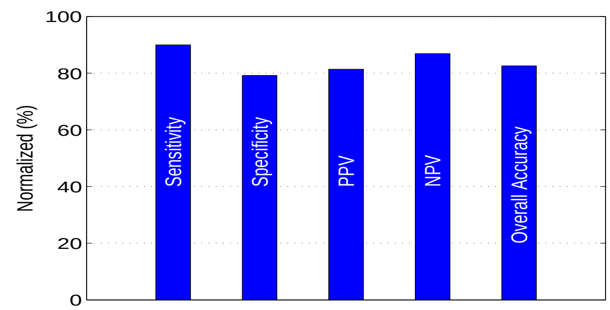


Fig. 7. Plaque detection performance of the SVM classifier. Leave one OUT based cross correlation shows an overall detection accuracy around 84.6%, with reasonable sensitivity, specificity, PPV and NPV rates.

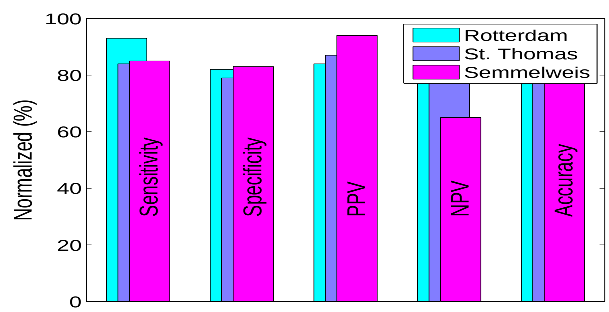


Fig. 8. SVM classifier performance for three individual datasets. The overall accuracy is centred at 80% with a consistent sensitivity and specificity ratio for three datasets.

the-art plaque detection methods of Wei *et al.* [14] (sensitivity of 93%), Lankton *et al.* [16] (sensitivity of 88%) and Tessmann *et al.* [17] (sensitivity of 79%) to establish a correlation with the reported literature. The future work aims to extend this work for a quantitative analysis of plaque in abnormal marked coronary segments. One possible extension is the use of deep learning framework to avoid the computation of hand crafted features for SVM model. Accordingly, we believe that the automated detection of non-calcified plaque can significantly increase the diagnostic of clinical experts to reduce cardiac fatalities.

V. Conclusion

We proposed an efficient method for automated detection of non-calcified plaques in cardiac CTA imagery. The innovation of this work is statistical representation of coronary segments, and the support vector machine based 2-class interpretation of respective segments. Accordingly, the proposed model delivers a very good detection rate for non-calcified plaques with respect to manual expert detections. In context of the future expansion of this work, there exists many potential extensions. An important aspect is to use the detected plaques for precise voxel-wise quantification of total plaque volume, which is the most important indicator of coronary heart disease. Another possible extension is to employ the deep learning in plaque detection process. This can allow auto feature extraction and minimize the user burden by eliminating the need of hand crafted features. In addition, one potential extension is to obtain the ground truth from multiple human experts and evaluate the plaque performance of the proposed method with respect of independent observers in context of inter-observer agreement.

Acknowledgment

We acknowledge the support of Dr. Ronak Rajani, consultant cardiologist, Guys & St. Thomas Hospital, London and Dr. Pal Maurovich Horvat of Semmelweis University Hungary for the Clinical CTA data and ground truth references.

References

[1] S. Waxman, F. Ishibashi, and J. E. Muller, "Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the global burden of disease study 2013," *Lancet*, vol. 385, no. 9963, pp. 117–171, 2015.

[2] U. K. NHS. Coronary Heart Disease, statistics for united kingdom. Available at http://www.nhs.uk/Conditions/Coronary-heart-disease/Pages/Introduction.aspx(2016/11/11).

[3] W. H. Organization. Cardiovascular diseases CVDs, the global statistics. Available at http://www.who.int/mediacentre/factsheets/fs317/en/(2016/11/11).

[4] T. Flohr and B. Ohnesorge, "Multi-slice ct technology," in *Multi-slice and Dual-source CT in Cardiac Imaging*. Springer, 2007, pp. 41–69.

[5] M. S. David C. Levin and D. Fischman. Coronary CTA, a cost-effective alternative to cardiac catheterization for the evaluation of cad, study suggests. Available at https://www.sciencedaily.com/releases/2010/04/100421162617.htm (2016/12/07).

[6] S. C. Saur, H. Alkadhi, L. Desbiolles, G. Székely, and P. C. Cattin, "Automatic detection of calcified coronary plaques in computed tomography data sets," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2008, pp. 170–177.

[7] G. Brunner, U. Kurkure, D. R. Chittajallu, R. P. Yalamanchili, and I. A. Kakadiaris, "Toward unsupervised classification of calcified arterial lesions," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2008, pp. 144–152.

[8] I. Išgum, A. Rutten, M. Prokop, and B. van Ginneken, "Detection of coronary calcifications from computed tomography scans for automated risk assessment of coronary artery disease," *Medical physics*, vol. 34, no. 4, pp. 1450–1461, 2007.

[9] B. Mohr, S. Masood, and C. Plakas, "Accurate lumen segmentation and stenosis detection and quantification in coronary cta," in *Proceedings of 3D Cardiovascular Imaging: a MICCAI segmentation challenge workshop*, 2012.

[10] R. Virmani, A. P. Burke, A. Farb, and F. D. Kolodgie, "Pathology of the vulnerable plaque," *Journal of the American College of Cardiology*, vol. 47, no. 8s1, pp. C13–C18, 2006.

[11] R. Shahzad, H. Kirisli, C. Metz, H. Tang, M. Schaap, L. van Vliet, W. Niessen, and T. van Walsum, "Automatic segmentation, detection and quantification of coronary artery stenoses on cta," *The International Journal of Cardiovascular Imaging*, vol. 29, no. 8, pp. 1847–1859, 2013.

[12] S. Waxman, F. Ishibashi, and J. E. Muller, "Detection and treatment of vulnerable plaques and vulnerable patients novel approaches to prevention of coronary events," *Circulation*, vol. 114, no. 22, pp. 2390–2411, 2006.

[13] M. E. Clouse, A. Sabir, C.-S. Yam, N. Yoshimura, S. Lin, F. Welty, P. Martinez-Clark, and V. Raptopoulos, "Measuring noncalcified coronary atherosclerotic plaque using voxel analysis with mdct angiography: a pilot clinical study," *American Journal of Roentgenology*, vol. 190, no. 6, pp. 1553–1560, 2008.

[14] J. Wei, C. Zhou, H.-P. Chan, A. Chughtai, P. Agarwal, J. Kuriakose, L. Hadjiiski, S. Patel, and E. Kazerooni, "Computerized detection of noncalcified plaques in coronary ct angiography: Evaluation of topological soft gradient prescreening method and luminal analysis," *Medical Physics*, vol. 41, no. 8, p. 081901, 2014.

[15] F. Renard and Y. Yang, "Image analysis for detection of coronary artery soft plaques in mdct images," in *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2008, pp. 25–28.

[16] S. Lankton, A. Stillman, P. Raggi, and A. Tannenbaum, "Soft plaque detection and automatic vessel segmentation," in *12th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer Berlin Heidelberg, 2009, pp. 25–33.

[17] M. Tessmann, F. Vega-Higuera, D. Fritz, M. Scheuering, and G. Greiner, "Multi-scale feature extraction for learning-based classification of coronary artery stenosis," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2009, pp. 726 002–726 002.

[18] H. Brodoefel, C. Burgstahler, A. Sabir, C.-S. Yam, F. Khosa, C. D. Claussen, and M. E. Clouse, "Coronary plaque quantification by voxel analysis: dual-source mdct angiography versus intravascular sonography," *AJR. American Journal of Roentgenology*, vol. 192, no. 3, p. W84, 2009.

[19] M. A. Zuluaga, I. E. Magnin, M. H. Hoyos, E. J. D. Leyton, F. Lozano, and M. Orkisz, "Automatic detection of abnormal vascular cross-sections based on density level detection and support vector machines," *International Journal of Computer Assisted Radiology and Surgery*, vol. 6, no. 2, pp. 163–174, 2011.

[20] I. Steinwart, D. R. Hush, and C. Scovel, "Density level detection is classification." in *NIPS*, 2004, pp. 1337–1344.

[21] Y. Li, W. Chen, K. Liu, Y. Wu, Y. Chen, C. Chu, B. Fang, L. Tan, and S. Zhang, "A voxel-map quantitative analysis approach for atherosclerotic noncalcified plaques of the coronary artery tree," *Computational and Mathematical Methods in Medicine*, vol. 2013, 2013.

[22] S. Achenbach, F. Moselewski, D. Ropers, M. Ferencik, U. Hoffmann, B. MacNeill, K. Pohle, U. Baum, K. Anders, I.-k. Jang *et al.*, "Detection of calcified and noncalcified coronary atherosclerotic plaque by contrast-enhanced, submillimeter multidetector spiral computed tomography," *Circulation*, vol. 109, no. 1, pp. 14–17, 2004.

[23] A. W. Leber, A. Becker, A. Knez, F. von Ziegler, M. Sirol, K. Nikolaou, B. Ohnesorge, Z. A. Fayad, C. R. Becker, M. Reiser *et al.*, "Accuracy of 64-slice computed tomography to classify and quantify plaque volumes

in the proximal coronary system," *Journal of the American College of Cardiology*, vol. 47, no. 3, pp. 672–677, 2006.

[24] T. Schepis, M. Marwan, T. Pflederer, M. Seltmann, D. Ropers, W. G. Daniel, and S. Achenbach, "Quantification of non-calcified coronary atherosclerotic plaques with dual-source computed tomography: comparison with intravascular ultrasound," *Heart*, vol. 96, no. 8, pp. 610–615, 2010.

[25] M. Otsuka, N. Bruining, N. C. Van Pelt, N. R. Mollet, J. M. Ligthart, E. Vourvouri, R. Hamers, P. De Jaegere, W. Wijns, R. T. Van Domburg *et al.*, "Quantification of coronary plaque by 64-slice computed tomography: a comparison with quantitative intracoronary ultrasound," *Investigative radiology*, vol. 43, no. 5, pp. 314–321, 2008.

[26] W. Theo. The Great Challenge, coronary artery stenoses detection and quantification evaluation framework. Available at http://coronary.bigr.nl/stenoses/(2016/11/11).

[27] H. Kirişli, M. Schaap, C. Metz, A. Dharampal, W. B. Meijboom, S. Papadopoulou, A. Dedic, K. Nieman, M. De Graaf, M. Meijs *et al.*, "Standardized evaluation framework for evaluating coronary artery stenosis detection, stenosis quantification and lumen segmentation algorithms in computed tomography angiography," *Medical image analysis*, vol. 17, no. 8, pp. 859–876, 2013.

[28] M. M. Jawaid, R. Rajani, P. Liatsis, C. C. Reyes-Aldasoro, and G. Slabaugh, "A hybrid energy model for region based curve evolution-application to cta coronary segmentation," *Computer Methods and Programs in Biomedicine*, vol. 144C, pp. 189–202, 2017.

[29] R. Van Uitert and I. Bitter, "Subvoxel precise skeletons of volumetric data based on fast marching methods," *Medical Physics*, vol. 34, no. 2, pp. 627–638, 2007.

[30] G. L. Raff, A. Abidov, S. Achenbach, D. S. Berman, L. M. Boxt, M. J. Budoff, V. Cheng, T. DeFrance, J. C. Hellinger, R. P. Karlsberg *et al.*, "Scct guidelines for the interpretation and reporting of coronary computed tomographic angiography," *Journal of Cardiovascular Computed Tomography*, vol. 3, no. 2, pp. 122–136, 2009.

[31] F. Renard and Y. Yang, "Image segmentation for detection of soft plaques in multidetector ct images," in *2008 IEEE Southwest Symposium on Image Analysis and Interpretation*, March 2008, pp. 121–124.

[32] A. Ponnapan. Automatic Soft Plaque Detection from CTA, georgia institute of technology. Available at https://pdfs.semanticscholar.org/5ad6/81ce27f6720ee2c9c00298226dcef3 d01e7b.pdf(2016/11/11).

[33] J. Dodge, B. G. Brown, E. L. Bolson, and H. T. Dodge, "Lumen diameter of normal human coronary arteries. influence of age, sex, anatomic variation, and left ventricular hypertrophy or dilation." *Circulation*, vol. 86, no. 1, pp. 232–246, 1992.

Ethical Issues and Related Considerations Involved with Artificial Intelligence and Autonomous Systems

Saud S. Alotaibi
Department of Information Systems
College of Computer and Information Systems
Umm Al-Qura University, Makkah, Saudi Arabia

Abstract—**The applications of artificial intelligence (AI) and automated systems (AS) demonstrate excellent outcomes in various sectors of industrial units to replace the human from jobs. However, the competitive world with evolutions, advancement of technologies and thrive for success by industrial gains by the managements are leaving the interests and benefits of larger number of human beings in the society. In this paper, various ethical issues related with the implementation of AI/AS are demonstrated with different perspectives. The ongoing developments in the area of AI/AS are critically evaluated and related advantages and serious concerns of the society are discussed. Various global initiatives and legal amendments across the globe to limit the excessive usage of AI/AS are being examined with critical assessments.**

Keywords—Ethical issues; artificial intelligence; automated systems; human values; evolutions

I. Introduction

The complete IT sector in the modern days of innovation and science is racing towards improving the quality of life and living standards as compared to early life. The internal desire of human being for a safe, comfortable and smooth life created a scenario where the man-made machines are being compared with the abilities of the human beings itself. Some of the authors like Holmes, in the past even questioned whether a digital brain [14] can replace a human mind? The mind of a human thinks, feels, exercises, makes perceptions, judges, reflects, etc. All these characters will be changing based on the time, situation and environment. However, in recent times the performance of human beings is being calculated with respect to the accuracy of automated systems (AS) designed by mankind itself. Two biggest technological titans of Silicon Valley are seen debating the potential of artificial intelligence (AI) at various platforms widely in recent times. The role of AI in designing various automated systems such as health industry, automobile industry, etc. were highlighted optimistically by Facebook founder, CEO, *Mark Zuckerberg* and was condemned fearfully by *Elon Musk* in an open debate. Supporting the statements of Elon Musk, the other business tycoon of Alibaba's *Jack Ma* stated that, AI may lead to world war III, but at the end humans will win according to CNBC News in April. The role of AI for the next generations were quoted to be pushing the human being towards decades of pain and any kind of repetitive job that has no emotional connection is finished [8]. Innovation at the cost of human being is not accepted by many people around the world and of course the advantages of AI also cannot be ignored but must be limited to the areas where it is necessary. The role of various governments across the globe to define the conditions and boundaries for any kind of automated activity in the near future to avoid the situation where the human jobs are replaced completely by involving the automated machines. The machines may be smarter, faster and stronger and may be intelligent but cannot be wise enough to take decisions as in the case of human beings. A human being is added with lots of emotions and feelings which a machine does not have. They may have ample knowledge and ability to work faster and efficiently using AI, but they do not have wisdom to understand various influential factors such as changes in environment, biological factors, emotions, feelings, etc. and therefore the decision making process by an automated system differs with the mankind.

A. Background of Artificial Intelligence

Early invention of artificial intelligence was mostly reported in the fictions, philosophy, imaginations, etc. and later was reported to take a good progress in the area of electronics and engineering [19]. The turning machine in the year 1936 was the milestone in the area of AI [4]. Various intellectual issues related to AI were discussed by Newell in his book [20]. The replacement of labors with robotic machines created a serious revolution by the labor unions. There after most of the robotic applications were limited for security agencies and purposes [12]. However most of the industrial units continued the usage of robotic systems based on AI to deliver the huge market units with greater accuracy. Especially in the area of automobile industry the usage of robotic machines and AI based systems established a strong hold to increase the supply and benefits. This in turn gained lot of attraction from most of the investors and industrialists. For example, the strategy of Korea for industrial development at different stages is shown in Fig. 1 includes the role of AI and IOT as a key for their success [15].

Machine learning (ML) strategies are very important to generate AI was used from a long time [6]. For example, genetic algorithms were used to find the solutions for the intractable computing problems such as scheduling. Similarly, neural networks were used to understand the human learning and also for industrial controls, monitoring activities and to classify various dynamic parameters. As a whole the applications of AI helped to create a faster computing and helped to develop solutions for the problems with complex nature. They provided easy solutions for various real-time industrial problems where the entry of humans is always a concern.

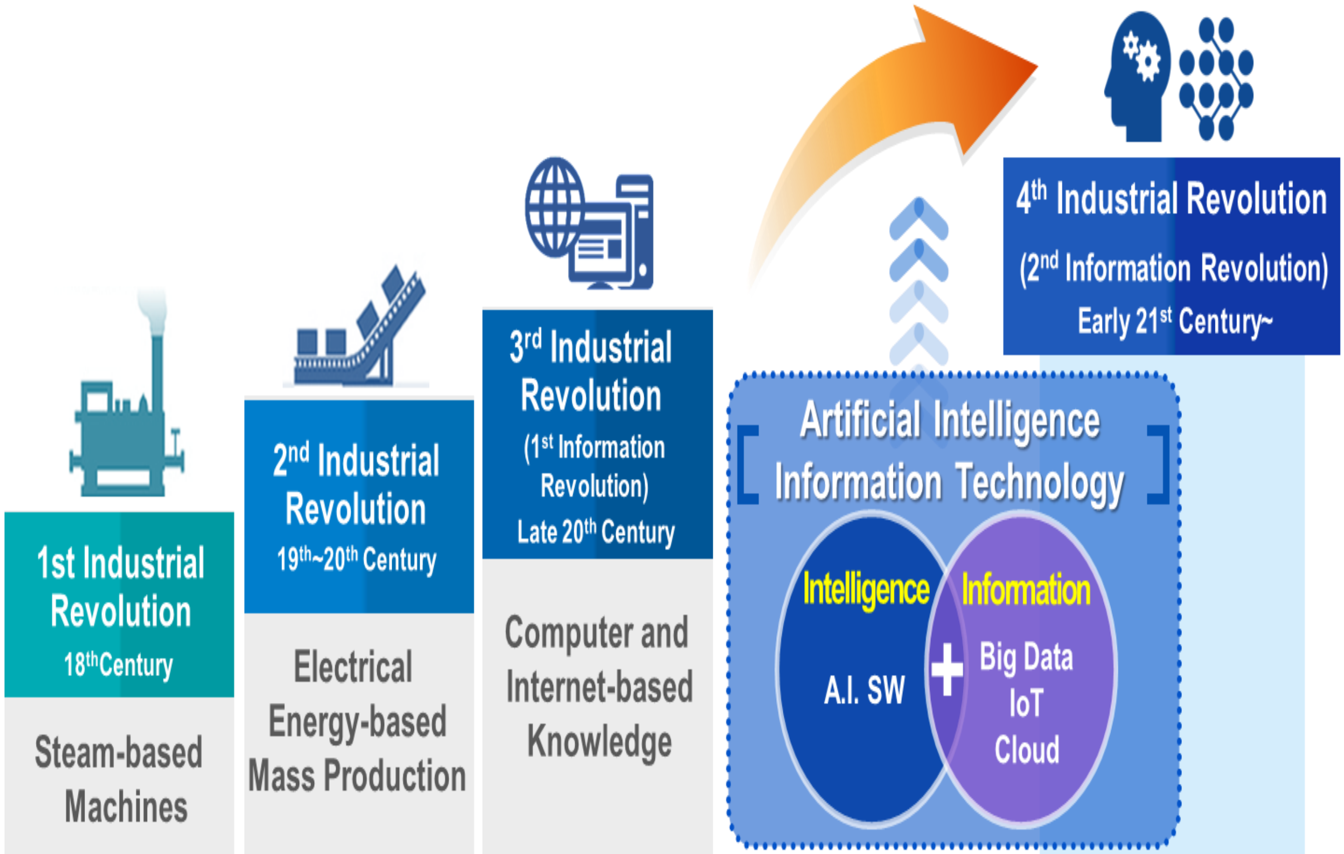


Fig. 1. Role of AI and IOT in 21st century for Korea's development program.

B. Problem Statement

AI being a young filed to the market since the introduction in 1940s by Alan Turning, it has gained the attention of many people due to wide scope of dealing things and activities of the market. It is imbibed with wider spectrum of ideas, viewpoints, logical activities, techniques, working actions, etc. [13]. Due to the complex nature of dealing with things from different disciplines, people have plenty of viewpoints towards the credibility of AI and its applications. Some of the questions include: why people are depending more on mechanically developed products or why the society need human models or why business people want to replace human with machines, etc. Similarly, there are many doubts and concerns raised by various societies and organizations include: will have no heart or feelings or instant decision making or does not know the humanity, etc. Some argue that AI can be very much useful to deal with serious accidents, can identify people with ill health, in troubles at the time of absence of family members, etc. Therefore according to Shahriari and Shahriari [22] the definition of AI is "… the science and engineering of making intelligent machines, especially intelligent computer programs" (pp. 197). It attempts to understand intelligent entities and provides healthy solutions for the human kind, which are almost next to impossible to predict by a human as faster as an AI based system. Therefore, the intelligence of the machine can learn, reason, solve, percept, rationale, think and deliver natural language process. It means the machine is mimicking the human and therefore can have huge influence and impact on the human life and may change the complete life style of the human beings [22]. It means that machine can control everything in the absence of human and deliver the results. However, if any system failed in the whole process of delivering the end results means a lot of damage to the society. For example, recent accident in Delhi, due to driverless train rams into a wall could have been resulted to a serious damage and many would have lost their life and loved ones [25].

There are so many ambiguous, contradictory and concerning issues were stated in introduction by various tech giants and business leaders on AI and some of the books and authors from different sections of society raised ethical and moral perspectives. These concerns definitely have different set of answers from different viewpoints and guidelines from the legislations of different state and international governments. Therefore it is very clear that understanding and define the dimensions of problem statement is quite complex in nature and proposing a solution may not be suitable for all types of industrial units.

II. Leading Ethical Issues Related with AI

There are many ethical issues related with AI and some of them are listed here from various sources. All the issues may not have similar importance and significance by using AI. For example, the usage of AI in an industrial robot helping the equipment in a production unit may differ with the robots working in healthcare and nuclear plants in terms of stringency, environment and human safety [17]. Sometimes it is also necessary to consider the cultural diversity and the opinion of people towards defining the norms, rules and regulations as autonomous intelligent systems are based on algorithmic biases and which may prove to be disadvantageous to a particular group or section of people. For example, the Muslim community prefers to eat halal food from the slaughter houses. They may feel offended with the regular slaughter houses we found in most of the western countries. Therefore the influence of certain communities, religions and habits of people must be considered carefully. Now in the following sections, various ethical issues related with AI are discussed [5].

A. Unemployment

The biggest affected area due to excessive automation is labor. More complex roles are being carried out by the machines at various production units creates a cognitive labor which creates a serious gap. For example, consider the transportation and travelling departments. If all the drivers are replaced by the self-driving transporting facilities means millions of employees will get unemployed and naturally there will be a serious imbalance in the society.

B. Inequality

It will be very difficult to assess the income sharing if the machines are playing a vital role in earning for a company. Most of the companies will be in a dilemma to decide the wages as it can depend more on the machines as compared with human work force. In other words the company providing AI based facilities will earn more money as compared to human beings.

C. Humanity

It is very difficult to obtain the sensitive nature of human beings in manmade machines. They cannot be made to respond intelligently towards building relationships and in case of human beings they are limited in terms of attention and kindness. Also it is very important to note the usage of software's and automated machines are under the usage effectively. However when they are handled by the wrong person they proved to be serious threat for entire humankind.

D. Artificial Stupidity

It is very difficult to guard against the mistakes made by the machines. Both human and systems learn to understand and detect various elements. But if something missing in the meantime means the human can act sensibly and in case of a machine it will be fooled. Especially if the role of AI is considered in the area of security, finance and healthcare industries it will be having a serious impact on the human life and society.

E. Racist Robots

It is evident that most of the AI based robots are capable of performing the tasks with high speed and capacity as compared to humans, however, it is always not the same and fair to trust them blindly. For example, Google uses AI to identify photos, people and objects in GOOGLE's Photos services and always it will not be able to produce the correct results. Also if any camera misses mark on racial sensitivity then the predicted criminals will be always the black people. Therefore it is very much evident that AI based systems are biased and judgmental.

F. Security

There are many countries in the world using powerful weapons and monitoring devices based on automated mechanisms and AI. Many countries like USA, Japan, China, Iran, Israel, etc. are having automated drones and in the battle field they are very much destructive and can win the wars without the human presence. The magnitude of destruction they can cause to human beings is unimaginable [24]. The possibility of such weapons in the hands terror organizations is always a serious concern. In the recent times, most of the countries at United Nations Summit appealed the ban on the automated war equipment and animations.

G. Evil Genies

So far the discussion was about external inputs and wrong usage of AI, but if the AI based systems turned against us as depicted in some of the Hollywood movies and etc. can damage the situation to a terrible unforeseen consequences. In an interesting scenario, the AI systems suggested to kill everyone on the planet to end the cancer. Such a goal of no more cancer could have been achieved by the computers very efficiently as compared to humans.

H. Singularity

Due to the ingenuity and intelligence the human dominance is seen everywhere on the earth and a better platform can be built in a bigger, faster and stronger manner as compared to the animals. This is due to the tool development and controlling skills of the human kind. These tools may include cages, weapons, etc. and with their help human kind can dominate the rest of the living entities. However, if AI does a better machine to anticipate the moves to defend it, human beings will be no longer an intelligent being on the earth.

I. Robot Rights

There are many research programs across the world to understand and unlock the secrets of basic mechanism of the brain and neuroscientists are still working in this area. Scientists are trying to build the mechanism of reward and aversion for the AI based systems as well. If suppose the widely available genetic algorithms are in a successful positions to survive and are able to form next generations to improve various instances, then how these machines will be trained towards ethical considerations and values of human life.

So there should be some kind of control over all these activities using AI systems to save the human existence on the earth. Any form of misinterpretation using AI could lead to a serious damage to the society, environment and humans too.

III. Need for Standardizing the Ethical Design for AI and AS

The fear and excitement are part and parcel of the AI systems and therefore it is essential to have a structured system whereby the negative side of it could be under control. The rules applied to one group, state and country may not be suitable or applicable at times to all others. In such scenarios, it is very important to have a standard procedure to implement various conditions in a systematic manner. Here some of the concerns of AI towards domestic and commercial aspects were discussed by Bryson and Winfield [6] raised some important questions as follows: does AI can compete with human beings; does AI undermine the societal stability; and does AI will be able to harm privacy, personal liberty and autonomy of the people. These questions raise serious concerns in the society towards the safety of humankind and the technology has developed to a level where they're in a position to replace the human entity in most of the areas including healthcare industry to conduct surgical operations to make some of the conclusive decisions while conducting the treatment for a patient. In simple words, it is proved that the machines are competing with the labors, experts and even with the human brains as well in most of the professional jobs. Now, considering the adverse part of AI's applications in the hands of anti-social elements always a serious threat to the society. Therefore it is very much essential to monitor the developments taking place in the area of AI. In this direction, some of the organizations linked with technology and scientific developments such as IEEE taken various initiatives to inculcate various ethical considerations in AI based systems. The overall motive of such initiative is to "… ensure every technologist is educated, trained, and empowered to prioritize ethical considerations" [6, p. 118].

A. Various IEEE Global Initiatives

To ensure a safe society as quoted by many tech giants and industrial experts in the area of design, IT, manufacturing and AI/AS, it is very important that the people involved in industries, universities and corporations must be well aware of what is good and what is bad for the society. Various people related with AI, law and ethics, philosophy, etc. provided collected inputs for various initiatives to be implemented as a move to educate the masses of people to consider ethics and values at the time of their work and research.

In general, the standards are based on various consensuses and agreed on different ways to do things in a particular format to reach a common goal. This is a good way that a system is providing all the guidelines to do the things and a kind of uniformity is maintained to establish coordination and confidence within the community of same interests. Already various standards such as British Standards (BS) 8611:2016, Robots and Robotic Devices explicitly addresses the guidelines for a Robotic designer and gives the platform to assess ethical risks involved in their designs. Similarly, IEEE conducted a program where multiple voices in AI/AS communities expressed their opinion, concerns and helped to share the information to empower people involved in AI to prioritize the ethical consideration in various design and development programs they initiate at their respective places.

In order to give a clear idea and guidelines, IEEE published a document *Ethically Aligned Design* (EAD) in December 2016. This Version-1 of the EAD is included with eight sections for covering the AI/AS related with ethics and moral values at the time of design and development stages includes: a) general principles, b) values to be added to the autonomous intelligence systems (AIS), c) guidelines to be followed for ethical designs, d) various safety and benefits of AI, e) security for personal data and access control while working with AI, f) reframing the autonomous weapon systems (AWS), g) impact on economy, h) humanitarian issues and laws. There were about 50-60 drafts and recommendations in EAD, covering issues and respective solutions to establish new standards. Presently according to the EAD presented four standards addressing the ethical concerns include the following [6], [7], [9], [10], [16].

- P7000-Model Process for Addressing Ethical Concerns during System Design
- P7001-Transparence of Autonomous Systems
- P7002-Data Privacy Process
- P7003-Algorithemic Bias Considerations

The first standard (P7000) is related with values to be involved in the system design and methodology. The second standard (P7001) is related to transparency, clarity and appreciation between people, when two or more are involved in a project. The third standard (P7002) is related to methods and approaches to be followed for maintaining the privacy issues. Finally, the fourth standard (P7003) is related to avoid and eliminate the negative bias in the algorithms based programs. Apart from these there are three more standards developed by the IEEE standards committee focused on the data of child, employer and personal data as listed below [23]:

- P7004-Standard for Child and Data Governance
- P7005-Stndard for Transparent Employer Data Governance
- P7006-Standard for Personal Data Artificial Intelligence Agent

The last three standards P7004-P7006 are very important when it comes to the information of overall society, people, their habits, cultures, and privacy. It cannot be assumed that the AI systems will be safer at all times since they are trained at the initial stages. However, if the same system is under the influence of negative elements of the society, the problems for everyone are unimaginable. Entire society can be under the scanner with a threat to life and as it was mentioned earlier that for cancer free earth the automated systems may successfully complete the task by killing everyone. Therefore serious discussions of opinions on such scenarios have been included to draft the last three standards.

B. Impact on the Society from IEEE Initiatives

There are quite sensible impacts seen across the globe between various communities, states and nations towards maintaining the trust and harmony by improving the understanding between each other. Main objective of building the trust between two personalities, communities, etc. was achieved by following such streamlined standards. The systems developed by professionals will be exposed against safety standards so that the care must be taken to ensure that AS safety certification is awarded to a smooth launch of the product or machine in the market. The overall global spending towards safety measures on AI based machine intelligence systems is increasing rapidly as shown in Fig. 2.

In case of accidents a transparent investigation must be carried out by these autonomous systems, so that actual faults must be traceable. The transparency must be included even with the people informing the evidence and with the lawyers or experts who are assessing the failures or accidents. Most of the disruptive technologies such as driverless trucks, cars, trains, etc. need the transparency with various development activities and testing processes to ensure the safety of people and to establish the confidence in technology.

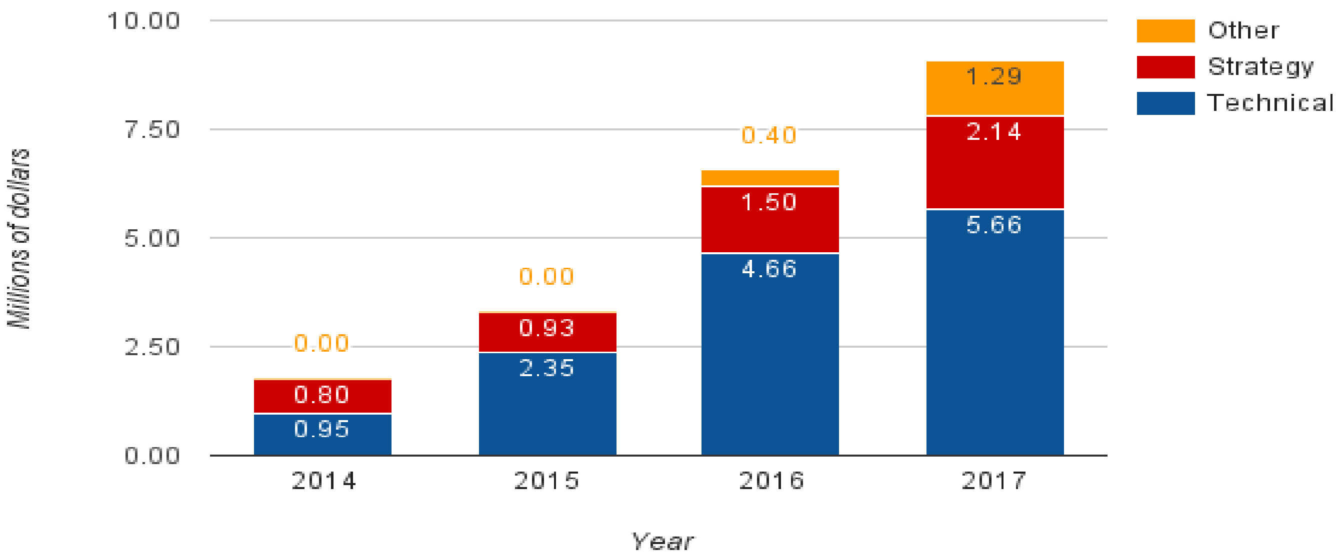


Fig. 2. Increasing global expenditure on AI to establish safety measures.

Source: Russell and P. Norvig (2016) [1]

C. *Role of AI/AS Systems in Gulf Countries*

The role of AI based systems in the Gulf countries helped various petroleum companies to ensure a systematic approach to maintain standard oil prices at different places. The artificial neural network (ANN) based tools helped these companies to solve complex problems involved in the oil companies towards pricing, measurements, assessments, forecasting, etc. were solved in effectively to ensure a stability among the competitors [18]. A kind of realistic approach for predicting the commercial oil prices was possible with such AI based systems. On the other hand, some of the programs on solar radiations are under progress by using AI based systems using MATLAB tool for predicting global solar radiation (GSR) in AI Ain city of UAE [2], [3]. In an attempt to predict the water quality parameters of Karoon River in Iran, scientists are using AI based models and analyzed the results in the real-time environments [11].

However, various cultural aspects were discussed and highlighted while using the automated systems and Riek and Robinson [21] opined that culture plays a substantial role towards interaction between people and even between people and machines. Therefore it is very important for machines to learn the social and cultural norms that impact overall people in the society.

IV. Conclusion

The changes observed in real-life due to excessive usage of AI/AS based systems raised serious concerns towards a negative impact on the job opportunities, inequalities, humanity and security. These concerns were discussed at international forums and platforms to build ethics and values for the professionals and engineers involved in design and development programs of AI systems. The global initiatives taken by various organizations such as IEEE, IEE, BS, etc. are optimistic to ensure that right steps are being followed so that the AI/AS based systems are useful for the humans, rather creating the problems. The global expenditure towards safety measures were seen increasing rapidly to ensure a safe usage of AI applications.

References

[1] S. J. Russell and P. Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[2] M. H. Al Shamisi, A.H. Assi and H.A. Hejase. “Using MATLAB to develop artificial neural network models for predicting global solar radiation in Al Ain City–UAE,”. In Engineering education and research using MATLAB. InTech, 2011.

[3] M. H. Al Shamisi, A.H. Assi and H.A. Hejase. “Artificial neural networks for predicting global solar radiation in al ain city-uae,”. International journal of green energy, 10(5), pp.443-456, 2013.

[4] A. Benko and C. S. Lányi. “History of artificial intelligence. In Encyclopedia of Information Science and Technology,” Second Edition (pp. 1759-1762). IGI Global, 2009.

[5] A. Yapo and J. Weiss. "Ethical Implications of Bias in Machine Learning." In *Proceedings of the 51st Hawaii International Conference on System Sciences*. 2018.

[6] Bryson, J. and Winfield, A., “Standardizing ethical design for artificial intelligence and autonomous systems,”. Computer, 50(5), pp.116-119, 2017.

[7] Calo, R., “Artificial Intelligence Policy: A Roadmap,” 2017.

[8] CNBC, “Jack Ma’s Artificial Intelligence and Automation Warning,” 2017. [Video]. https://youtu.be/cgAPG3Lh3gQ?t=70/., [Accessed on November 12, 2017].

[9] De Spiegeleire, S., Maas, M. and Sweijs, T., “Artificial Intelligence and the Future of Defense: Strategic Implications For Small-and Medium-Sized Force Providers,”. The Hague Centre for Strategic Studies, 2017.

[10] Diebel-Fischer, H., “Research Ethics in the Digital Age: Fundamentals and Problems,”. In Research Ethics in the Digital Age (pp. 7-21). Springer VS, Wiesbaden, 2018.

[11] Emamgholizadeh, S., Kashi, H., Marofpoor, I. and Zalaghi, E.,”Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models,” International Journal of Environmental Science and Technology, 11(3), pp.645-656, 2014.

[12] Gates, B.,”A robot in every home,”. Scientific American, 296(1), pp.58-65, 2007.

[13] Hall, P., 3. edition,”Stuart J. Russell and Peter Norvig,“. Artificial Intelligence: a Modern Approach. Upper Saddle River, New Jersey: Prentice Hall, 2009.

[14] Holmes, N.,”Would a digital brain have a mind?,”. Computer, 35(5), pp.112-111, 2002.

[15] V. Devedžić. "Web intelligence and artificial intelligence in education." *Educational technology & society* 7, no. 4 (2004): 29-39.

[16] Maia Alexandre, F.,”The Legal Status of Artificially Intelligent Robots: Personhood, Taxation and Control,” 2017.

[17] R. Sparrow. "Killer robots." *Journal of applied philosophy*24, no. 1 (2007): 62-77.

[18] Movagharnejad, K., Mehdizadeh, B., Banihashemi, M. and Kordkheili, M.S., “Forecasting the differences between various commercial oil prices in the Persian Gulf region by neural network. Energy,” 36(7), pp.3979-3984, 2011.

[19] Negnevitsky, M.,”Artificial intelligence: a guide to intelligent systems,” Pearson Education, 2005.

[20] Newell, A.,”Intellectual issues in the history of artificial intelligence,” (No. CMU-CS-82-142), 1982.

[21] Riek, L.D. and Robinson, P.,”Challenges and opportunities in building socially intelligent machines,” [social sciences]. IEEE Signal Processing Magazine, 28(3), pp.146-149, 2011.

[22] Russell, S. J. and Norvig, P. “Artificial Intelligence: A Modern Approach (2 ed.),” Pearson Education.2003.

[23] Shahriari, K. and Shahriari, M.,”IEEE standard review—Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems,” In Humanitarian Technology Conference (IHTC), 2017 IEEE Canada International (pp. 197-201). IEEE, 2017.

[24] Springer, P.J., ”Military robots and drones,” a reference handbook. ABC-CLIO, 2013.

[25] Times Now Bureau, “Delhi Metro Magenta Line driverless train rams into wall, DMRC Says accident due to ‘Human Error’, 2017. [News Report] [Accessed on November 23, 2017].

An Adaptive Cruise Control Model based on PDLCA for Efficient Lane Selection and Collision Avoidance

Khawaja Iftekhar Rashid
Department of Computer Science & Technology
Dalian University of Technology, (DUT)
Dalian, Liaoning, Post (116024), P.R. China
https://orcid.org/0000-0001-5003-7969

Muhammad Tahir
School of Software Technology
Dalian University of Technology, (DUT)
Dalian, Liaoning, Post (116621), P.R. China

Ding Nan
Department of Computer Science & Technology
Dalian University of Technology, (DUT)
Dalian, Liaoning, Post (116024), P.R. China

Anil Ahmed
Department of Computer Science & Technology
Dalian University of Technology, (DUT)
Dalian, Liaoning, Post (116024), P.R. China

***Abstract*—Todays, safe and sans collision free traveling is essential in the present World. Because of the particular needs required by different usages of Vehicular Ad Hoc Networks (VANETs), mainly in evolving lane, outlining a constant Collision Avoidance (CA) and securely lane changing has turned into the foundation of controlling the vehicle in the dense environment. In our system, we proposed efficient lane selection and collision-free vehicle transportation system in the light of patience, speed and the distance between all types of vehicles. To maximize the efficiency of the existing transportation infrastructure and reduce collision, it is necessary to improve safety and communication in all vehicles. In this paper, we first define the patience based method and simulate model for safety and efficiency of traffic flow using the game theoretical technique. Then we have improved a new algorithm for lane selection and collision control using distance and speed of all vehicles gather from the vehicular communication. After two different methodologies for lane change and collision-free driving are compared *like;* random patience based method and car following method, based on different traffic density. In the part of contribution as an outcome, our experimental simulation setup is executed on highway road traffic scenarios to demonstrate the capacity of the all type of vehicles to select best-fit lane, identify collision, and explore around them to maintain the safe distance. Finally, it is concluded that our proposed algorithm outperforms in lane selection and collision avoidance to satisfy safety and efficiency in the Vehicular Ad Hoc Networks**.

Keywords*—*Automatic vehicle; collision avoidance; lane selection; patience; threshold; traffic density

I. Introduction

Vehicle-to-Vehicle communications (V2V) provide the opportunity to establish the connection between vehicles through the recent advancement in communication technologies. The experiments in planning a Collision Avoidance (CA) and the framework are adjusting the viability of maintaining a strategic distance from collisions versus the danger of false alerts. The assignment of a CA is the framework to track objects of potential collision hazard and to maintain the threshold for avoiding the collision. Vehicle applications specifically display few difficulties; dense traffic causing complex situations with many moving articles; minimum distance sensors and computational units must be utilized [1], [2]. According to author's focus on how to make decisions based on uncertain estimates and in the presence of multiple obstacles. Timely detection of other vehicles in the region of outrageous significance to help avoid an accident and potential loss of human life, and traffic jams. Driver distractedness, weakness, and juvenile conduct are the fundamental variables causing street mishaps [3]. As indicated by the National Highway Traffic Safety Administration (NHTSA), that almost *25%* of police-announced accidents embroil some sort of driver distractedness the driver may be diverted, exhausted, snoozing or somewhere out in dreamland [4].

Acceleration and lane-change decisions are drivers' main operational and tactical decisions. Lane change is considered as a standout amongst the most difficult driving moves to comprehend and to foresee, and with related driving choices are frequently observed as a noteworthy wellspring of collision [5]. Among current examinations, a large portion of lane change models in light of vehicle kinematics estimate that when the neighbor vehicle changes lane, the accompanying vehicle of target lane keeps uniform movement [6], [7]. Adaptive Cruise Control (ACC) is typical and surely understood driving assistant system. By depending on a frontal radar or a '*LIDAR sensor'*, it can automatically adjust speed to keep up a protected progress distance between the vehicles in a similar lane. A few automakers have just presented highlights, for example, the ACC, in their top-of-the-line cars and some are seeking after research to incorporate the ACC, framework and collision warning or collision avoidance frameworks in different -situations. These frameworks utilize the basic plan that announces the target vehicle as the nearest one right now in subject's lane. Significant disadvantages coming about because of these fundamental methods are the failure to identifying vehicle on other lane and performing sufficient response ahead of time and the restricted execution in high arch location [8], [9].

A few works have been done in the field of ACC, for instance in the field of automatic braking, parking, forthcoming collision control, automatic crossing point assistance, automatic lane change and lane keeping framework that take after the speed of the lead vehicle. Apart from these social or legitimate issues while managing ACC, with VANET, obviously there is an achievement for enhancing wellbeing if the frameworks can act quickly and successfully. However, it should be notice that the highly complex quality associated with these situations, these frameworks request a high measure of vehicles data and the environment [10]. Congestion arises in the network when traffic increase and stuck in the lane instead of changing their lanes. ACC, assistances to enhance driver support and increase highway limit. *Likewise;* by applying automatic vehicle following framework, the collision between vehicles can be sufficiently decreased. Regardless, the greater part of them is executed in straight forwarding situations. Broad research has been done on the longitudinal speed control of vehicles. *In this research paper;* we compared Random Patience based Lane Change Algorithm (RPLCA), and Lane Change Model based on Car-Following Behavior (LCMCFB), that help the autonomous vehicle to change lanes in different situations and avoid the collision. Based on this concept we present our new idea of lane change by setting a threshold based on the distance between vehicles and different values of patience. Moreover, we use the game theoretical techniques based on zero-sum non-cooperative game theory under incomplete information. The type of game represents lane changing decisions that are made when the driver has partial information about other driver's decision. Rest of our paper deals as follows: In Section **II**, we are providing the overview of some work related to our experimental design which is based on collision avoidance methodology as well as patience based approach. Section **(III)** provides simulation scenario and mentions proposed method briefly. In Sections **(IV)** and **(V)**, we discuss the experimental environment and discuss the result respectively. In the end, we conclude our remarks in Section **(VI)**.

II. Related Work

VANETs, apply numerous channels, includes Control Channels (CCH), and Service Channels (SCH), to give open street wellbeing service and enhanced comfort and productivity of driving. 75MHz of DSRC, the range at 5.9GHz extraordinarily assigned by FCC that is uniquely intended for vehicular correspondence is separated into seven equivalent channels called CCH and SCH [11]. Subsequently, the collision may occur when a vehicle changes its lane without having proper information related neighbor vehicles. ACC is a procedure used to control the speed of a vehicle to maintain safe distance from the leading vehicles. The essential point of this strategy is to give a reasonable alleviation to the driver. Locally available sensors help to control speed and maintain distance with the help of data received from Road Side Infrastructure (RSI), satellites and messages broadcast from surrounding vehicles. Some of the application installed in vehicles to detect leading vehicle which are mention here. A laser-based framework couldn't recognize leading vehicle in foggy climate or when the vehicle is muddy. Other techniques including single radar-based framework and binocular computer vision framework in view of the front camera are used in vehicle detection [12], [13].

A. Collision Avoidance Methods

Several methods have been introduced in VANET to avoid collision between vehicles to protect lives and efficient driving. Here we are discussing some techniques related collision avoidance by different researchers. Alessandro Colombo et al. consider the issue of impact control at the intersection point. The author considers it as an NP-hard issue and is around illuminated in this paper, by receiving concentrated approach in light of supervisor having provable error bound [14]. The supervisor depends on a hybrid approach that utilizes a dynamic model of the vehicles and occasionally tackles a scheduling issue. The supervisor act as a channel between a coveted inputs forms from driver side and the physical framework. This structure is effortlessly combined with different controllers, acting between the driver input and the supervisor, to seek after optional execution destinations inside the arrangement of safe control activities permitted by the supervisor.

An effort related collision avoidance and detection of independently moving vehicles are settled by receiving multi-operator frameworks is discussed in [15]. Vehicles can commonly illuminate themselves about their objectives and goals for future development and subsequently can comprehend a conceivable collision significantly more productively than in the non-collaborative case. The proposed technique combine with a mind stunning conduct enabling it to consult with the workstation operators about items transportation, explore through the system to get to a specific workstation as quick as could be allowed, recognize and avoid collisions with different Automated Guided Vehicles (AGVs), and even powerfully identify and maintain a safe distance from hurdles on road.

Marwah M. Almasri et al. presents another solution for lane following and collision avoidance in the mobile robotic frameworks. The proposed system depends on the utilization of minimal effort infrared sensors and includes the sensible level of estimations, with the goal that it can be effectively utilized as a part of ongoing control applications. This technique demonstrates that the robot has been effectively followed extremely congested curves and encounter any obstacle on its way [16]. Above work is firmly identified with our simulation related collision recognition and lane change control. In our collision avoiding scenario we consider above work in term of the vehicular environment. Where we get the location from beacon messages and compute the distance between the vehicles.

B. Patience based Method

In our simulation, we dynamically choose the lane that is suitable for safe and efficient transport on road by collecting data of close-by vehicles, their speed, distance and number of vehicles in various lanes. These features help us to take the -correct decision. Dissimilar to [17], we better utilize patience level to change lane (moving from the congested lane to the one having low traffic). The factor of patience has been discussed by few researchers. We are reviewing some of them in this section. Patience parameter was presented in [18]. While going astray from the favored speed vector in protocol constrained autonomous collision avoidance algorithms permitting a weighted balance of a lane selection and speed change. The patience parameter and suggested measurements of research permit autonomous collision avoidance with surrounding vehicles to maintain a safe distance from the collision. The patience parameter is presented for autonomous vehicles in sea scenario utilizing multi-target advancement with a non-parallel collision avoidance space involving distance factor. The experiment has performed in light of various values of the patience from *1-99*. In our simulation, we utilize patience parameter and set the value to maximum for all the vehicles, and calculate distance and speed to select lane and speed. We will prove that the speed of vehicles and change lane incredibly influence the collision and lessen blockage and help the system to work efficiently. We compare two algorithms and based on above mention factor presents our algorithm "*Adaptive Cruise Control Patience and Distance-based Lane Change and Collision Avoidance Algorithm (ACCPDLCA)"* to improve traffic efficiency and avoid the collision. ACC can just reduce the speed of the vehicle and help drivers to avoid the collision by maintaining the safe distance from the leading vehicle and help him to change speed.

III. Proposed Work

A. Traffic Scenario

The substance of our exploration in view of Cooperative awareness messages - According to [19], the cooperative awareness well known as beacon message contains the information related current status, geolocation and service announcements to/from those vehicles. This information related geolocation is further used in our research in calculating the distance between the vehicles. For safety driving, it is essential to have the location of surrounding vehicles to calculate the distance and make a decision whether to apply break, slow down speed, speed up or change the lane.

In aforementioned work based on patience level, the lane change factor of a vehicle depends upon the level of patience parameter, when the vehicle applies break each time its patience values reduce by 1 and speed reduced with the factor of deceleration. We have simulated different scenarios on the basis of different variables involved likewise acceleration, deceleration, patience value, and vehicles density. When it reaches to 0, the vehicle changes its lane and shift to one of the closest lanes. In a high-density traffic environment when a vehicle changes its lane only based on patience level, chance of collision increase. For example, an engine failure or a mishap involving two or more vehicles can lead to a traffic jam. For this scenario, the objective is to build the vehicle efficiency for singular vehicles. Moreover, contingent upon various factors, for example, the significance of the accidental area, the vehicle framework may figure and recommend elective lanes to an expansive arrangement of vehicles considering a more extensive perspective of the movement requests keeping in mind the end goal to reduce the effect of this accident to areas not near to that instance. For this scenario, the objective is to expand the general transport productivity. Another issue is that when a vehicle changed its lane the value of patience is again set to maximum to avoid frequent lane change. Traffic scenario for lane selection and change is illustrated in Fig. 1.

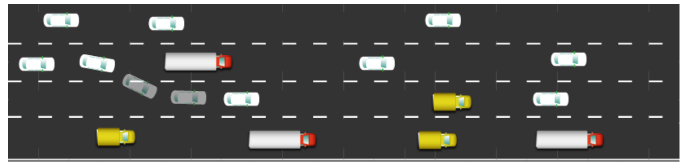


Fig. 1. The Road Traffic Scenario based on all types of vehicles when a high-speed vehicle stuck in the lane and change the lane from the best available lane.

B. Proposed Simulation Model

In our proposed solution we set the different threshold to change lane with proper information and having the best available solution that we achieved by using Bayes probability algorithm in our simulation. First, we check the traffic density of each lane and the speed of neighbor vehicles before lane change. Second, we start to make the decision of lane change based on distance and speed of the vehicle ahead. The third one we used Bayes probability theorem to decide whether to change lane or remain intact with the speed of the vehicle ahead. Increase in the density of traffic reduces the overall speed of vehicles as well as increase the chance of collision. To achieve maximum speed on-road vehicles choose a lane having low traffic with maximum speed. During this process, several collisions occur. To tackle the problem-related lane change and increase the overall speed of the vehicles we proposed an algorithm based on patience value and distance between the vehicles.

Another push to control blockage and increase the speed of the vehicles to run smoothly and efficiently we used the patience factor. However, the preference and threshold of patience are different in our simulation than the one used in the ocean based technique as we discuss in Section II. The equation to detect threshold for patience is given below.

$$p_{threshold} = \frac{p_{Max} * v_{xy}}{\mathrm{L}} \tag{1}$$

Where p is the patience, v_{xy} is the speed of the vehicle and L represents the number of lanes. After calculation of the thresholds, after that, we decide to change the lane of the vehicles along considering other factors. When the lane of the vehicle changed the patience value is again set to max. An equation to compute the safety threshold of distance is as under.

$$d_{threshold} = \alpha * (sizec_{ahead} + sizec_{self}) \tag{2}$$

α should lie in between [0.5, 1]. If $sizec_{ahead} > sizec_{itself}$ then the value of the constant should be equal to 1 for the lane change to avoid the collision. If $sizec_{ahead} \leq sizec_{itself}$ then the constant value should be set as 0.5. If the value of constant is less than 0.5 then chance of collision increased. Size_{ahead} and

$size_{itself}$ both are variables, used to store values of the size the vehicles.

Before settling on any choice on the premise of information and earlier outcomes we initially check the possibility of collision. On the premise of those probabilities, we choose the proper lane and distinctive limits to diminish the danger of collision. Utilizing these probability estimations we plan structure for performing simulations. To stay away from collision we should have former data of the vehicular network. On the premise of information obtained from the vehicular network we characterize the probabilities and change the lane and speed of the vehicle to avoid collision. In Bayesian probability hypothesis, one of these "instances" is the assumption, H, and the other is information, D, and we wish to judge the relative truth of the theory given the information.

In our research, we have an arrangement of variable information D = (D1, D2, D3, ... , Dn) the reasons for collision β and an arrangement of feasible lane μ. As indicated by Bayesian control we utilize following recipe to choose the following move of the vehicle about the lane change.

$$P(\mu|D) = \frac{P(D|\mu)P(\mu)}{P(D)} \quad (3)$$

Since the data related lane P(μ) is earlier access to the vehicle when it joins the network. P(D) is likewise obtained by summing up P(D|μ)(P|μ). The term P(D|μ) is the probability capacity and it evaluates the likelihood of the examined information acquired from the data received from various sources for our scenario beacon messages, GPS and so forth. P(μ|D) is posterior. To maintain a strategic distance from the collision between vehicles one must realize what caused the collision. From (4) we induce the best arrangement of causes to clarify a given bit of information, as a rule, includes boosting the posterior over sources.

$$\hat{\beta} = \frac{\arg\max_{\beta} P(\beta|D,\mu)P(\beta|\mu)}{P(\beta|D)} \quad (4)$$

In (4) we utilize Nash Equilibrium in light of the fact that in a Nash balance, every vehicle's decision of activity is the best reaction to the moves really made by his adversaries. Subsequent to ascertaining the qualities and deciphering the accessible information it is important to choose the best lane to limit the hazard. Particularly in the dense network, it is hard to change lane. Utilizing Bayes likelihood hypothesis we assess the best reasonable lane from all the access lanes. Our activity is to choose the best lane μ from the information interpreted. To benefit the best lane we need to choose the lane having the best estimation of μ. Mathematically it is expressed as

$$\hat{\mu} = arg\max_{\mu} P(\mu|D) \quad (5)$$

C. Game Theoretical Method

Adaptive Cruise Control (ACC) is a discretionary voyage control framework for roadside vehicles that naturally alters the vehicle speed to maintain safe distance from the leading vehicles. Control depends on sensor data from locally available sensors. ACC innovation has broadly viewed as a key segment of any future ages of intelligent autos. They affect driver well-being and accommodation and also expand the limit of roads by maintaining safe distance among the vehicles and diminishing driver's mistakes. The intelligent control of vehicles is one of the present difficulties of intelligent transport frameworks. The utilization of Artificial Intelligence strategies to vehicle frameworks empowers driving help frameworks to interface with nature. Previous work-related ACC can help to avoid front-end collision based on data interpreted by the sensors onboard.

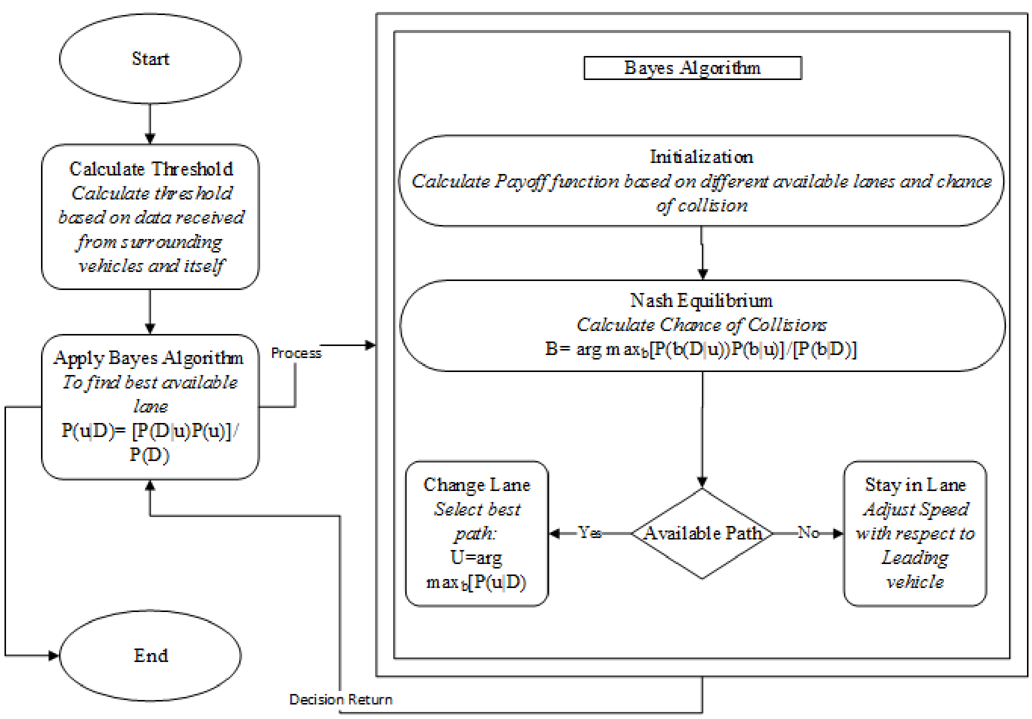


Fig. 2. A flow diagram of lane selection process for ACCPDLCA.

Fig. 3. Proposed simulation setup executed on low-density traffic for high-way road.

In our research we are dealing with lane change scenario, having distinctive extra highlights moreover vehicle ad-hoc correspondence with the surrounding vehicle to consider their speed, area, and other essential messages to guarantee well-being and effectiveness on road. VANET, correspondence helps us to gather relative information from different vehicles in the system utilizing reference point messages. With the assistance of these messages, we compute the general proficiency of the system. Area and speed of encompassing vehicle is to execute choice made by the vehicle. The framework of our experimental simulation is shown in above Fig. 2.

Let a vehicle c_i moves with a speed of v_i having patience value p_i at time t in lane L_i following vehicle c_j with a distance d. Utilizing ad-hoc correspondence between vehicles it will likewise assemble related data communicate much of the time from the encompassing vehicles. All these information stored in a tabular form maintained by each node. These beacon messages enable them to know current movement and speed of its surrounding vehicles. It persistently computes the distance between the vehicles ahead and furthermore has the data about their speed. At the point when the vehicle approaches the base distance criteria $d_{threshold}$ it will begin decelerating with the factor of deceleration. The patience esteem diminishes much of the time with deceleration. Meanwhile, when patience esteem p_i approaches $p_{threshold}$ esteem it will then execute next step. At the point when the distance is way to deal with least limit and the patience likewise fulfill its condition then the estimation in light of the accessible information will perform through game theory approach. At that progression, we ascertain the most effective way and reasonable scenario for the vehicle. We simulate our experiment on different values of patience and traffic density and conclude the results in Section V.

IV. Experimental Setup

The experiment is performed in NetLogo software version v 6.0.2 [20], designed by Uri Wilensky. It shows programming ideas utilizing specialists as vehicles, patches, links and the spectator. We perform our experiments in highway scenario, by considering all types of the traffic density. We perform simulations in low density, medium density, and high-density traffic condition. The number of vehicles will be chosen progressively. Additionally, acceleration and deceleration are likewise factor and are also accessible at User Interface (UI), for dynamic determination. The patience level is also dynamic. We have performed several experiments with different level of patience from one to hundred and study the behavior of vehicles with respect to speed and collision avoidance. Simulation environment in *Net Logo* with low-density traffic is shown in Fig. 3.

Here we present another idea of ACC, Patience and Distance-based Lane Change Algorithm (ACCPDLCA), based on the game theory approach. It takes the position of the leading vehicles to set the speed and distance of every single close-by vehicle to measure collision probability. ACCPDLCA, based on a game theory proposed a few principles as per which every vehicle act. To accomplish our destinations we infer any forecasts whatsoever in various circumstance by describing how vehicles carry on, what their targets are and how guidelines of the diversion they attempt. Simulation results that vehicle knows the procedures accessible to every vehicle have finished and reliable preferences over conceivable results and they know about those preferences. Besides, they can decide the best lane for themselves and faultlessly execute it. On the off chance that a result isn't a Nash equilibrium, at that point no less than one of the vehicles isn't best reacting, and at some point or another a vehicle in that part will happen to arrive on a higher activity which will then be received by the vehicles a while later. This helps vehicles to choose lane having maximum speed. The decision is based on the minimum threshold value of patience and distance between the vehicles.

The number of accessible vehicles in a lane is consistently refreshed and was shown in a table. While in the dual mood we use two different algorithms run at a time on different vehicles to compare the effect of communication on the network. We can select the vehicle dynamically and observe its behavior in different frames. Here we apply ACCPDLCA on the selected vehicle while all the other vehicles in the network work base on ACC technology. In given simulation minimum speed of the vehicle is 0, i.e., complete break while we set start speed equal to 0.5. After that speed increased randomly up to 1.2.

A. Simulation Parameters

We performed the simulation in NetLogo default setup with four number of lanes. Each of the underneath simulations keep running for 20 and 40 minutes. Therefore, we ponder the informational index of this renovation made on each step. Results are placed in the spreadsheet and also we obtain in tabular form. For each activity condition, we have an informational index of 648,540. General we interpreted 2,594,160 of informational collections during simulation. Result generated based on simulation data is discussed in Section V. In Table I, we represent out simulation parameters.

TABLE I. REQUIREMENTS OF SIMULATION PARAMETERS

Simulation Tool	NetLogo
Traffic Environment	Low, Medium, High, Extreme Dense
Traffic Density (vehicle/lane)	variable
Average speed (at time of joining)	0.5
Max Speed	1.2
Patience	10, 20, 30, 40, 50, 60, 70, 80, 90, 100
Acceleration	0.002, 0.006, 0.01
Deceleration	0.02, 0.06, 0.1
Simulation Time (sec)	1200, 2400
Data Set	At each step

V. FINAL RESULTS AND DISCUSSION

We simulated several experiments to prove that our algorithm help vehicles to select the better lane and enhance the speed by maintaining a safe distance from traffic congestion and collision between vehicles. We compared our algorithm with Random Patience based Lane Change Algorithm (RPLCA), and with Lane Change Model based on Car-Following Behavior focusing on the kinematic behavior of the lane-changing vehicle in the process of accelerated lane change (LCMCFB). Table II has generated from low traffic condition. The simulation of our algorithm also included various patience values from 10-100. In low-density traffic lane change occur infrequently, however as indicated by the qualities we measure that if the lane change marvels happen just on the bases of patience with the random incentive at the joining of the system then it decrease the general speed of vehicles on road.

TABLE II. VALUES OBTAINED AT LOW-DENSITY TRAFFIC

Patience	Average Collision in			Average speed with		
	RPLCA	ACCPDLCA	LCMCFB	RPLCA	ACCPDLCA	LCMCFB
10	3	0	0	0.70	0.80	0.79
20	3	0	0	0.72	0.78	0.79
30	3	0	0	0.71	0.78	0.79
40	3	0	0	0.73	0.80	0.80
50	3	0	0	0.73	0.80	0.79
60	3	0	0	0.73	0.78	0.80
70	3	0	0	0.72	0.79	0.80
80	3	0	0	0.75	0.80	0.80
90	3	0	0	0.72	0.80	0.78
100	2	0	0	0.74	0.78	0.79

Unnecessary lane change condition can cause the collision with surrounding vehicles that travel at high speed in the destination lane. Fig. 4 represents the number of collisions on various level of patience. After performing several experiments with different level of patience ACCPDLCA performs better than RPLCA. When we increase the traffic on road up to medium level the odds of collision between the traffic increase uncommonly when the driver changes the lane without having any data related traffic scenario in different lanes and in addition to the speed of vehicles. In this circumstance, it is important to settle on the better choice. Values obtained in Table III validate the effect of lane change among medium level traffic. In this scenario when the number of lane change increases unnecessarily the general speed of the vehicles decrease. In such scenario, when a vehicle change lane the following vehicles have to applied break to maintain the safe distance to avoid the collisions and the patience of the following car decreased.

Unnecessary lane change decreases the overall speed of the vehicles. Comparison of these algorithms with respect to the number of collisions have illustrated in Fig. 5. ACC in view of PDLCA performs better in this scenario. In low and medium level density condition we see that general speed of vehicle remain close to each other in all algorithms. Also, collision avoidance in ACCPDLCA and LCMCFB are apparently equal. However, when the traffic increase, it is hard to oversee speed of the vehicles particularly when each vehicle has different patience level subsequent after joining VANET.

TABLE III. VALUES OBTAINED AT MEDIUM DENSITY TRAFFIC

Patience	Average Collision in			Average speed with		
	RPLCA	ACCPDLCA	LCMCFB	RPLCA	ACCPDLCA	LCMCFB
10	33	3	4	0.36	0.69	0.50
20	43	3	3	0.39	0.70	0.49
30	36	3	4	0.41	0.68	0.68
40	38	2	4	0.42	0.69	0.69
50	38	4	3	0.43	0.68	0.68
60	39	3	5	0.44	0.68	0.68
70	40	3	4	0.43	0.68	0.68
80	39	4	4	0.45	0.69	0.69
90	40	3	5	0.45	0.68	0.68
100	38	4	3	0.47	0.69	0.69

TABLE IV. VALUES OBTAINED AT HIGH-DENSITY TRAFFIC

Patience	Average Collision in			Average speed with		
	RPLCA	ACCPDLCA	LCMCFB	RPLCA	ACCPDLCA	LCMCFB
10	164	20	30	0.20	0.55	0.33
20	196	24	28	0.22	0.54	0.33
30	213	23	31	0.23	0.54	0.33
40	219	23	28	0.24	0.54	0.33
50	226	21	32	0.26	0.54	0.33
60	224	20	31	0.25	0.55	0.33
70	229	22	28	0.26	0.54	0.33
80	228	24	34	0.26	0.55	0.33
90	227	24	29	0.26	0.55	0.33
100	233	22	30	0.26	0.55	0.33

TABLE V. VALUES OBTAINED AT EXTREME HIGH-DENSITY TRAFFIC

Patience	Average Collision in			Average speed with		
	RPL CA	ACCPDL CA	LCMC FB	RPL CA	ACCPDL CA	LCMC FB
10	690	289	678	0.11	0.30	0.19
20	879	323	839	0.13	0.31	0.19
30	981	184	850	0.14	0.31	0.18
40	1008	294	781	0.14	0.31	0.19
50	1070	265	769	0.15	0.31	0.19
60	1128	259	774	0.15	0.31	0.19
70	1163	302	639	0.16	0.31	0.19
80	1151	278	536	0.16	0.31	0.19
90	1167	276	666	0.16	0.31	0.19
100	1173	240	734	0.16	0.30	0.19

In Section III, we as of now specify that how factor estimation of patience impact the general speed of the vehicle and cause traffic block. In a high-density condition for the most- part patience value decreased to 0 of approximately every vehicle, since when we apply break or diminish speed with respect to the speed and conduct of the leading vehicle, subsequently patience value decreased to 0. ACC in view of PDLCA help vehicles to sit tight for the others appropriate time to start lane change as portraying earlier. However, the performance of LCMCFB decreases as the traffic increase. Safety of vehicle during lane change scenario cannot only depend upon surrounding vehicle. A side collision occurs during lane change when we transfer to a lane in high-density traffic with proper information. Our proposed technique helps vehicles to remain in lane and continue proceeding onward with steady speed unless system generates safety message for transferring lane. These messages are generated as discussed in Section III above. Tables IV and V have been generated based on high-density and extreme high-density traffic from RPLCA, PDLCA, LCMCFB, on different patience esteem, respectively. Graphically represented in Fig. 6 and 7, respectively.

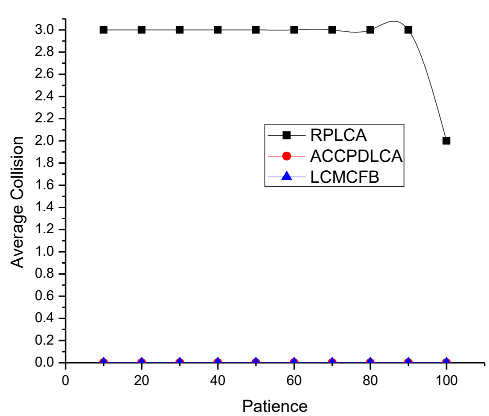


Fig. 4. Collision in low-density traffic environment.

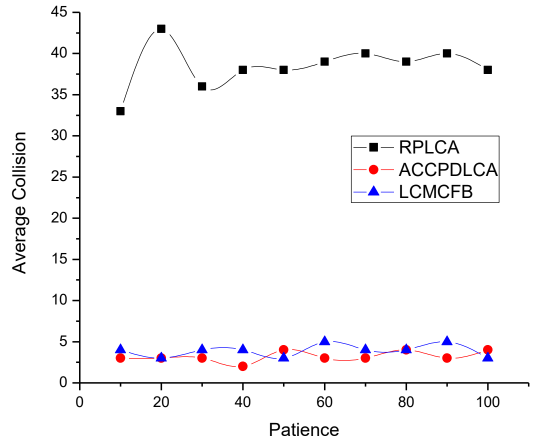


Fig. 5. Collision in medium density traffic environment.

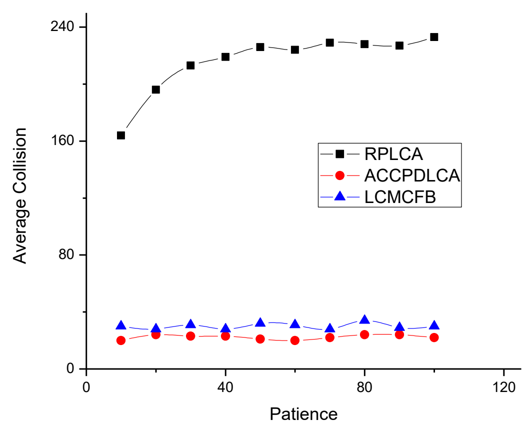


Fig. 6. Collision in high-density traffic environment.

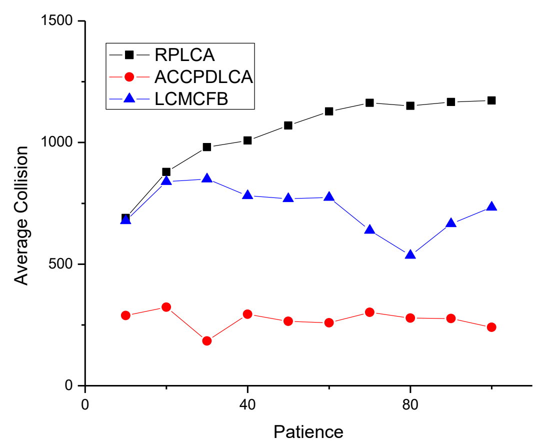


Fig. 7. Collision in extreme high-density traffic environment.

In Fig. 8, we compare overall improvement in speed of vehicles in different density networks, using ACCPDLCA, LCMCFB, and RPLCA. The minimum speed of the vehicle can approach 0 while we set the starting speed 0.50. The maximum speed touches 1.2. From the graph, it is dissipating that overall performance in term of speed in ACC with PDLCA is better than ACC without PDLCA (RPLCA). However, performance in term of efficiency (speed) of LCMCFB has also improved in low and medium density network. But in high and extreme dense environment LCMCFB shows lack of efficiency as well as safety.

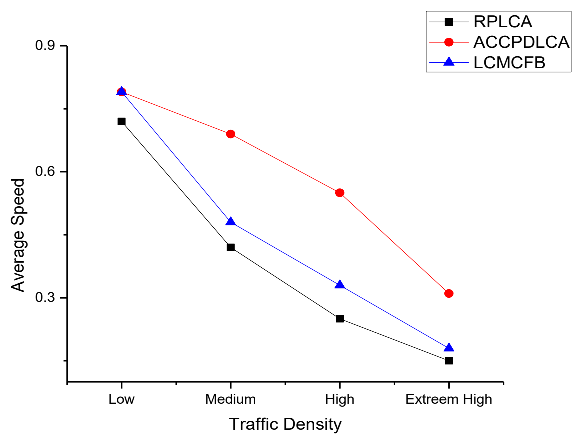


Fig. 8. Comparison of overall performance of vehicles in term of speed.

VI. Conclusion

The strategy to enhance traffic efficiency in terms of speed and collision avoidance is discussed in this paper. The results obtained from our simulation verify that factor for patience with the coordinated effort of distance among vehicle performs better to enhance general execution of the system to keep away from a traffic jam in the dense environment. Our proposed solution Adaptive Cruise Control model based on Patience and Distance-based Lane Change and Collision Avoidance Algorithm (ACCPDLCA), performs better than Random Patience based Lane Change Algorithm (RPLCA), and with Lane Change Model based on Car-Following Behavior (LCMCFB).

The results shows that vehicles following ACCPDLCA have higher speed while RPLCA, decreases the speed and the efficiency of the traffic. Also, unnecessary lane change behavior in vehicles has been overcome by using Bayes algorithm and fixing the patience level to maximum when a vehicle joins the network in the specific area.

Maximum speed in high-density traffic are *0.55, 0.33*, and *0.26* in *ACCPDLCA*, *LCMCFB*, and *RPLCA*, respectively. However, in extreme high-density environment where the distance between vehicles is lower than the threshold, our algorithm performs quite well as compared to others. The number of collision with respect to other two algorithms is also decreasing in number. Additionally, research about dense vehicular network will likewise be led.

VII. Future Scope

Safe and sound collision free traveling with high speed is the basic demand of the present World. The purpose of the algorithm is to suggest efficient lane selection and improve speed on the highway. Our algorithm is assisting drivers, by selecting the best suitable lane for safe and efficient driving. In the future, we will implement our proposed system in city traffic and multi-lane systems.

Acknowledgment

This work was supported by the National Natural Science Foundation of China No. 61471084, No. U1608253, and the open program of State Key Laboratory of Software Architecture No. SKLSA2016B-02.

References

[1] J. Jansson, "Collision Avoidance Theory with Application to Automotive Collision Mitigation," Electrical Engineering, Link¨oping University, 2005.

[2] A. K. G. G. W. T. Batra, "Collision detection system for vehicles in hilly and dense fog affected area to generate collision alerts," presented at the Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, India, 2014.

[3] P. S. S. L. Sagar D. Devagire, "Collision Avoidance Systems: A Review of Emerge Technology, and its Evolution," International Journal of Innovative Research in Science, Engineering, and Technology, vol. 5, 2016.

[4] A. M. L. X. T. B. Tang, "Vehicle Detection Techniques for Collision Avoidance Systems: A Review," IEEE Intelligent Transportation Systems Society, vol. 16, 24 March 2015.

[5] H. S. M. Alireza Talebpour, Samer H. Hamdar "Modeling Lane-Changing Behavior in a Connected Environment: A Game Theory Approach," Transportation Research Procedia, vol. 21st International Symposium on Transportation and Traffic Theory 2015.

[6] F. R. PENG Jinshuan, SHI Leilei, ZHANG Qiong "Research of Driver s Lane Change Decision-making Mechanism," Journal of Wuhan University of technology, vol. 33, pp. 46-50, 2011.

[7] Y. H. WU Xiaorui, "A Lane Change Model with the Consideration of Car-Following Behavior," Procedia - Social and Behavioral Sciences, pp. 2354 – 2361, 2013.

[8] D. K. S. M. J. P. H. J. K. K. Yi, "Design of an Adaptive Cruise Control / Collision Avoidance with lane change support for vehicle autonomous driving," presented at the ICCAS-SICE, Fukuoka, Japan, 2009.

[9] D. K. S. M. J. P. H. J. K. K. Yi, "The Development of a Cooperative Heavy-Duty Vehicle for the GCDC 2011: Team Scoop," Intelligent Transportation Systems IEEE Transactions on, vol. 13, pp. 1033-1049, 2012.

[10] J. E. N. Felipe Jiménez, José Javier Anaya , Fernando García, Aurelio Ponz , José María Armingol "Advanced Driver Assistance System for road environments to improve safety and efficiency," Transportation Research Procedia, vol. 14, pp. 2245-2254, 2016.

[11] H. J. Reumerman, Z. Yupeng, G. R. Hiertz, G. S. Max, and L. Stibor, "Method for multi-channel resource reservation in a wireless mesh network," ed: Google Patents, 2012.

[12] D. A. F. Armin Askari, Alex A. Kurzhanskiy, Pravin Varaiya, "Measuring Impact of Adaptive and Cooperative Adaptive Cruise Control on Throughput of Signalized Intersections," 2016.

[13] G. W. Ziran Wang, Matthew J. Barth, "Developing a Platoon-Wide Eco-Cooperative Adaptive Cruise Control (CACC) System," presented at the IEEE Intelligent Vehicle (IV) Symposium, Redondo Beach, CA, 2017.

[14] A. Colombo and D. D. Vecchio, "Efficient algorithms for collision avoidance at intersections," presented at the Proceedings of the 15th ACM international conference on Hybrid Systems: Computation and Control, Beijing, China, 2012.

[15] R. C. L. C.-W. Kuo, "Intelligent Seven-DoF Robot With Dynamic Obstacle Avoidance and 3-D Object Recognition for Industrial Cyber-Physical Systems in Manufacturing Automation," Proceedings of the IEEE vol. 104, 24 March 2016.

[16] M. M. A. A. M. A. K. M. Elleithy, "Trajectory Planning and Collision Avoidance Algorithm for Mobile Robotics System," IEEE Sensors Journal vol. 16, June 15, 2016.

[17] S. Miyata, T. Nakagami, S. Kobayashi, T. Izumi, H. Naito, A. Yanou, et al., "Improvement of adaptive cruise control performance," EURASIP Journal on Advances in Signal Processing, vol. 2010, p. 295016, 2010.

[18] K. Woerner and M. Benjamin, "Autonomous Collision Avoidance Tradespace Analysis for High-Speed Vessels," in 13th International Conference on Fast Sea Transportation. Society of Naval Architects and Marine Engineers, 2015.

[19] X. W. Wenyang Guan, Ye Jin, "Channel congestion performance analysis for DSRC vehicle ad hoc network," presented at the 10th International Conference on Communications and Networking in China (ChinaCom), China, 2015.

[20] U. Wilensky. (1999). NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

AUTHORS' PROFILE

Khawaja Iftekhar Rashid received the B.Sc. and MCS degree from University of Punjab, Lahore, Pakistan and University of Azad Jammu and Kashmir, Muzaffarabad, Pakistan in 2012 and 2014 respectively. Currently pursuing his Master degree in Computer Science and Technology from Dalian University of Technology, Dalian, P.R. China. His research interests include vehicular networks, network information theory, wireless sensor networks, etc.

Nan Ding received the B.S., M.S., and Ph.D. degrees in Computer Science and Engineering from Dalian University of Technology, Dalian, P.R. China, in 1999, 2002, and 2006, respectively. He was a visiting scholar with the Electrical and Computer Engineering of University of New York University, U.S., from Sep 2013 to Sep 2014. He has been an Associate Professor in Computer Science and Engineering from Dalian University of Technology, Dalian, P. R. China since 2008. His research interests include vehicular networks, wireless security, wireless sensor networks, cache management, and distributed fault-tolerant computing.

Muhammad Tahir received the BS (Hon's) degree in Software Engineering from University of Sindh, Jamshoro Pakistan in 2008 and Master of Engineering Degree in Software Engineering from CHONGQING University P.R. China in (2014). Currently, He is a Ph.D. Research Scholar in the School of Software Technology, Dalian University of Technology, P.R. China. His research interest includes Web-based Software defects prediction models, vehicular networks, Web-based tuning, IOT, Cloud Computing, SDN & 5G.

Anil Ahmed received the B.S. degree in Computer Science from Mirpur University of Science & Technology, Mirpur, Pakistan in 2012. Currently, He is pursuing Master Degree in Computer Science and Technology from Dalian University of Technology, Dalian, P. R. China. His research interest includes Vehicular Networks, Deep learning, and CNN.

Object Contour in Low Quality Medical Images in Curvelet Domain

Vo Thi Hong Tuyet
Faculty of Information Technology
Ho Chi Minh City Open University
Ho Chi Minh City, VietNam

Nguyen Thanh Binh*
Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology, VNU-HCM
Ho Chi Minh City, Vietnam

***Abstract*—The diagnosis and treatment are very important for extending the life of patients. The small abnormalities may also be manifestations of the diseases. One of the abnormalities is the contour of each object in medical images. Therefore, the contour is very important and special to low quality medical images. In this paper, we propose a new method to detect the object contour of low quality medical images based on the self affine snake, one of the types of active contour models. The method includes two periods. Firstly, we use augmented lagrangian method to remove noise and detect edges in low quality medical images in curvelet domain. Finally, the active contour model is improved to show the contour of objects. After comparing the appearance of the contour and the time processing with other algorithms, we confirm that the proposed method is better.**

Keywords—Object contour; curvelet transform; augmented lagrangian

I. Introduction

Nowadays, many diseases cannot be detected by the naked eye, such as the diseases of bone, liver, cancer, etc. If dangerous diseases are detected early, the diagnosis and the treatment will be easier. Therefore, the medical image is an important tool for specialists. They contain much important information inside human body, and the contour of each object is very necessary. The location of boundary is sometimes to refer to as fracture, enlargement or tumor. Most of the important task in medical image is edge detection. The result of edge detection depends on the quality of the image [2], [5]. However, the quality of medical images depends on many factors such as: technical skills, machines, printer, etc. because these reasons make medical images have noise or blur details, and the quality of them is reduced. So, this is a big challenge.

Medical images are abundant about types. So, the boundary of each agency is difficult for detection because of the intensity of colorful or connection between them. In the last time, there were many methods for object contour in natural image, such as Sobel [4], B-spline [6]-[10] or the generation types of wavelet transform as [1], [6], [11], [12]. Brigger and his partners used B-spline snakes to detect the boundaries of objects in [7], multiscale of them, and to develop it as a flexible tool in [10]. Then, [23] is a method which also uses snake for detection and segmentation in medical images. The combination between curvelet with self-affine mapping is proposed in [25] for the medical images which have low quality. So, the relation between transform and snake can be useful for contour detection. If the results from edge detection step are good, the contour will be more accurate and clearer.

Although multi-scale or space domain gives positive results, the wavelet transform is one of the methods that is space domain. Discrete wavelets transform (DWT) [13] and complex wavelet [12] is continued. However, DWT has three disadvantages [14]: poor directionality, shift sensitivity, lack of phase information. So, the emergence of more optimal methods restored, such as: contourlet transform [15], nonsubsampled contourlet transform [16], [17], ridgelet transform or curvelet transform [18] and shearlet transform in [1].

Medical images contain one or more objects. They have noised or blurred pixels or both, and called as weak object. A concept of strong and weak objects in contour is defined in [12]. Therefore, the low quality is the big challenge for object contour detection. In the recent years, the applied of curvelet threshold in medical image is very powerful such as denoising [26], detection [27] and segmentation [28]. However, the time processing for using curvelet coefficients is very high. Yet, the slow time can be accepted in medicine as long as the result is good. However, the time can be smaller, the life can be longer. With this problem, augmented lagrangian method [3] is a good resolution.

Based on the previous knowledge, we use augmented lagrangian method (ALM) and selft-affine snake to detect the boundary of medical images. However, we use the space domain, curvelet domain, instead of the frequency domain. The contributions of this paper are: 1) to design a competent curvelet domain adaptive threshold, capable to increment the quality of medical image. 2) The active contour model is explained. 3) Authors proposed a new method for object contour in low quality medical image in curvelet domain. The rest of the paper is organized as follows: In Section 2, we present the basic concepts which relates to the proposed method in this paper, the Section 3 present the proposed method for object contour, the Section 4 present the experiments and results. The final section is conclusions.

II. Brief Knowledge

A. Curvelet Transform

Wavelet transform is a popular method for image restoration [14]. Developing it, the new generation wavelet is showed; and the curvelet is one of them. The first generation of curvelet transform is ridgelet transform [18], [19]. In ridgelet domain, the hard threshold and soft threshold are applied. If we call T_{hard} to be the hard threshold and T_{soft} to be the soft threshold, $\lambda \geq 0$ is parameter wavelet, I is normal parameter value. We have equations:

$$T_{hard}\left(\hat{d}_{jk},\lambda\right)=\hat{d}_{jk}\, I\left(\left|\hat{d}_{jk}\right|>\lambda\right) \quad (1)$$

and

$$T_{soft}\left(\hat{d}_{jk},\lambda\right)=\mathrm{sign}(\hat{d}_{jk})\max\left(0,\left|\hat{d}_{jk}\right|-\lambda\right) \quad (2)$$

Ridgelet transform can be adapted to represent objects with curved edges that are almost straight [21]. Two concepts are monoscale and multiscale, which represent the theory of ridgelets [25] based on the ridge functions to develop ridgelet system. The orthonormal ridgelets are indexed using as (3):

$$\lambda=\left(s,l,a,al,x\right) \quad (3)$$

where, s is the ridge scale, l is the ridge location, a is the angular scale, al is the angular location and x is a gender token. The ridgelet transformation can be showed as follows [20]:

- Smoothing step based on the Fourier transform.
- Calculating bivariate ridgelet and depended on the radon transform that is the important process for ridgelet values.
- Performing the scale wavelet transform with ridgelet coefficients to define the result images.

The curvelet transform is a non-adaptive technique. It is the representation of multi-scale of objects. The curvelet is superior method to keep information and to improve the quality of images which have curves between edges. Similar to wavelets, curvelet transform can be dilated in two dimensions and translated. And like ridgelets, curvelets can be applied for all scales, locations and orientations. Each subband of curvelet is a curve with width ≈ length^2 and is analyzed by a local ridgelets. The subbands of curvelets have the nonstandard form $[2^{2s}, 2^{2s+2}]$. Authors in [18], [19] give the process of curvelet transform to include four steps:

Firstly, the subband is decomposed. The image f is decomposed into subbands by [18], [19]:

$$f \mapsto (P_0 f,\Delta_1 f,\Delta_2 f,...) \quad (4)$$

Each subband is the presentation of the value which needs to adapt to reconstruct coefficients.

Secondly, smooth partitioning. Each subband is applied with scale which has the sidelength $\sim 2^{-s}$, as follows by (5):

$$\Delta_s f \mapsto (w_Q \Delta_s f)_{Q\in Q_S} \quad (5)$$

where, w_Q is a collection of smooth window localized around dyadic squares:

$$Q=[k_1/2^s,(k_1+1)/2^s]\times[k_2/2^s,(k_2+1)/2^s] \quad (6)$$

Thirdly, the renormalization. Each resulting square is renormalized to unit scale

$$g_Q=(T_Q)^{-1}(w_Q\Delta_S f), \qquad Q\in Q_S \quad (7)$$

Finally, ridgelet analysis. Each square is analyzed via the discrete ridgelet transform.

While restore the images, besides of the curvelet coefficients, we can use the threshold and filter [14], [21] to support or combine between them [20]-[22]. But the execution time is very long. The time is slow because the adapting of the image to many filters or thresholds of transforms.

B. Augmented Lagrangian Method

When restoring the image restoration, any method must recover the sharpness of the image. The idea to minimize a total variation optimization problem for spatial temporal data was proposed [3]. It is Augmented Lagrangian Method (ALM) to solve the constrained problem. When we call a vector denoting the unknown (potentially sharp) image which has size M x N is called f ($f \in R^{MN\times 1}$), where f is an ingredient of equation to find the observed image g, $g \in R^{MN\times 1}$. It is a linear shift invariant imaging system to calculate as: $g = Hf + \eta$, where f is a vector denoting the unknown (potentially sharp) image of size M x N, g is a vector denoting the observed image, $\eta \in R^{MN\times 1}$ is a vector denoting the noise/blur, and the matrix $H \in R^{MN\times MN}$ is a linear transformation representing convolution operation. This algorithm includes two types: TV/L1 minimization (for denoising image) and TV/L2 minimization (for deblurring image). They were defined as [3]:

$$\underset{f}{\text{minimize}} \quad \frac{\mu}{2}\left\|Hf-g\right\|^2+\left\|f\right\|_{TV} \quad (8)$$

and

$$\underset{f}{\text{minimize}} \quad \mu\left\|Hf-g\right\|_1+\left\|f\right\|_{TV} \quad (9)$$

where, μ is the regularization parameter. The authors were to find a saddle point of L(f, u, y). Then, they used the alternating direction method (ADM) to solve f-subproblem, u-subproblem with TV/L2 and f-subproblem, u-subproblem and r-subproblem with TV/L1. The equation as [3]:

$$\underset{f,u}{\text{minimize}}\frac{\mu}{2}\left\|Hf-g\right\|^2+\left\|u\right\|_1 \quad (10)$$

and
$$\underset{f,r,u}{\text{minimize}}\ \mu\|r\|_1 + \|u\|_1 \tag{11}$$

Subject to r = Hf - g and u = Df. Augmented lagrangian method can be summarized as follows [3]:

- Input: the information of vector denoting, convolution matrix, regularization parameter, the isotropic total variation.
- Set parameter with value default. This step depends on other types of denoising or deblurring.
- Initialize for the first value of each problem.
- Compute the matrices of displacement in directions, such as: horizontal, vertical and temporal.
- Check this value is coverage or not. If true, ALM will remove values (noise or blur) and update parameters. Then the checking process is continued.
- When having any value to be not coverage, the algorithm will stop to check.

C. Active Contour Model

Active contour model, snake, is a method to detect boundaries of object in images. The position of a snake parametrically is known by v(s) = (x(s), y(s)) and energy functional as [7], [10]:

$$E^*_{snake} = \int_0^1 E_{snake}(v(s))ds = \int_0^1 E_{int}(v(s)) + E_{image}(v(s)) + E_{con}(v(s))ds \tag{12}$$

where, E_{int} represents the internal energy of spline due to bending is:

$$E_{int} = (\alpha(s)|v_s(s)|^2 + \beta(s)|v_{ss}(s)|^2)/2 \tag{13}$$

E_{image} gives rise to the image forces, and E_{con} gives rise to the external constraint forces to calculate by [24]:

$$E_{con} = \frac{1}{2}(\alpha(s)|v_s(s)|^2) \tag{14}$$

where, $\alpha(s)$ and $\beta(s)$ are user-defined weights. The $\alpha(s)$ is a large weight for the continuity term penalizes. The distance between points in the contour is changed by $\alpha(s)$. In addition, $\beta(s)$ is for the smoothing term of the contour. The features of images are the showing of image energy. These features are calculated by [24]:

$$E_{image} = w_{line}E_{line} + w_{edge}E_{edge} + w_{term}E_{term} \tag{15}$$

where, w_{line}, w_{edge}, w_{term} are weights of these salient features. The process of snake includes computing force and then combining them and converging snake model. With self-affine snake includes a step in the first of process which is extracting self-affine maps. However, the active contour model is only to apply when we have the object detection.

III. Object Contours in Curvelet Domain

Energy reducing is a popular direction for presenting the contour methods. In [22], the authors use energy reducing, self-affine snake in wavelet domain for contour detection. Edge detection is almost difficult because of the quality of images. The result of the contour object detection depends on the results of the edge detection step. Special with medical images, the quality of them is bad because of many reasons: skill of technical, machine, etc. … Edge detection is hard work because it is easy to lose information in medical images. Curvelet is one of the best choices for denoising medical images. In [24], [25], the authors use curvelet domain and Snake for contour detection and segmentation which have positive results. So, we can develop from it.

Based on these ideas, we propose a new method for object contour. The process method is presented clearly in Fig. 1. In Fig. 1, the process method includes two periods. Firstly, object detection based on ALM for smoothing. Finally, the self-affine Snake for active contour. Both period 1 and 2 are also in curvelet domain.

A. Object Detection based on ALM for Smoothing

Firstly, the input of the method is low quality medical images. These images have blur or noise values. So, objects in them are weak. Then, the decomposition which level is 5 and direction is 32 to be applied. The curvelet's process is as follows: decompose input images into curvelet domain with scales depending on the level and direction. Then, the loop statement will begin in 1. Each repeat, calls block size, applies digital ridgelet transform.

Secondly, the smoothing image: this step applies TV/L1 process of ALM, which follows [3]:

$$minimize_{f,r,u}\ \mu\|r\|_1 + \|u\|_1 \tag{16}$$

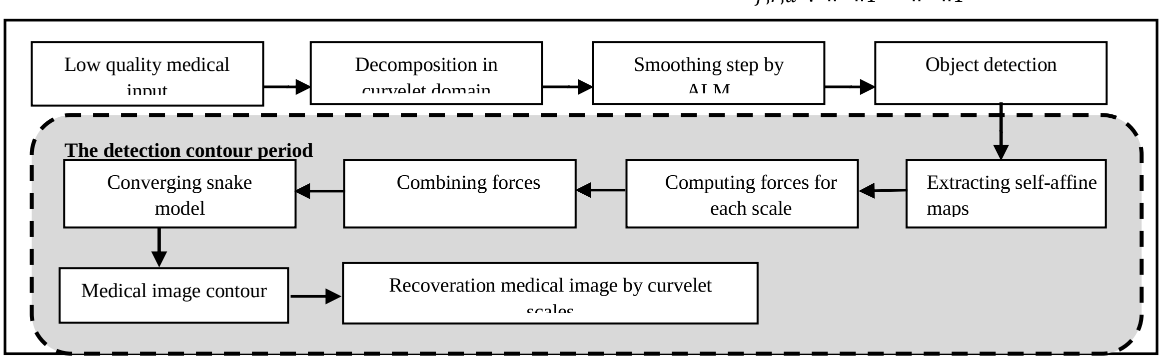


Fig. 1. The proposed method for object contour.

Equation (16) is solved as follows:

INPUT: vector denoting the observed image (g) and convolution matrix (H), regularization parameter μ , the isotropic total variation β_x , β_y , β_t.

OUTPUT: the medical image, which is denoised and deblurred. The object is to apply smoothing.

BEGIN:

- Set parameter with value default for $\rho_r = 2, \rho_0 = 100$ where $\alpha_0 = 0.7$ and ρ_r: regularization parameter.
- Initialize $f_0 = g$, $u_0 = Df_0$, $y_0 = 0$, $r_0 = Hf_0 - g$, $z_0 = 0$
- Compute the matrices of the first-order forward finite difference operators along the horizontal, vertical and temporal directions.
- ***While (not coverage) do:***

(i) Solve the f-subproblem is:

$$minimize_f = \frac{\rho_0}{2}\|r - Hf + g\|^2 + \frac{\rho_r}{2}\|u - Df\|^2 + z^T Hf + y^T D \quad (17)$$

where, $D = \begin{bmatrix} D_x^T & D_y^T & D_t^T \end{bmatrix}^T$ with: D_x, is the order forward finite-difference operators along the horizontal direction, D_y is vertical direction and D_t is temporal direction.

f-subprobem is improved by equation:

$$f = \mathcal{F}^{-1}\left[\frac{\mathcal{F}[\rho_0 H^T g + H^T(\rho_0 r - z) + D^T(\rho_r u - y)]}{\rho_0|\mathcal{F}[H]|^2 + \rho_r\left(|\mathcal{F}[D_x]|^2 + |\mathcal{F}[D_y]|^2 + |\mathcal{F}[D_t]|^2\right)}\right] \quad (18)$$

where, $\mathcal{F}$ denotes the 3D Fourier Transform operator.

(ii) Solve u-subproblem:

$$u_{k+1} = \arg\min_u \|u\|_1 - y_k^T(u - Df_{k+1}) + \frac{\rho_r}{2}\|u - Df_{k+1}\|^2 \quad (19)$$

where, $$u_x = \max\left\{|v_x| - \frac{1}{\rho_r}, 0\right\} * sign(v_x) \quad (20)$$

(iii) Solve r-subproblem:

$$minimize_r \quad \mu\|r\|_1 - z^T r + \frac{\rho_0}{2}\|r - Hf + g\|^2 \quad (21)$$

by equation: $$r = max\left\{\left|Hf - g + \frac{1}{\rho_0}z\right| - \frac{\mu}{\rho_0}, 0\right\} * sign\left(Hf - g + \frac{1}{\rho_0}z\right) \quad (22)$$

(iv) Update the Lagrange multiplier y and z:

$$y_{k+1} = y_k - \rho_r(u_{k+1} - Df_{k+1}) \quad (23)$$

$$z_{k+1} = z_k - \rho_0(r_{k+1} - Hf_{k+1} + g) \quad (24)$$

(v) Update:

$$\rho_r = \begin{cases} \gamma\rho_r, if \|u_{k+1} - Df_{k+1}\|_2 \geq \alpha\|u_k - Df_k\|_2 \\ \rho_r, otherwise \end{cases} \quad (25)$$

(vi) Check convergence: if $\| f_{k+1} - f_k \|_2 / \| f_k \|_2 \leq tol$ then break

End while.

END

Thirdly, the object detection. In contour detection, after smoothing step, methods begin to calculate the strengths between edges together known to be the determining gradients. We use the equation of Eucidean distance as (26):

$$E = \left(\sum_{i=1}^{n}|x_i - y_i|^2\right)^{1/2} \quad (26)$$

where, (x, y) is coordinates of image pixels. E will be used to define edges to be shown. Each pixel in the gradient image, 45^0 for direction and the connected neighborhoods help for the calculating process. The next problem is to show the relationships between pixels by direction, it is:

$$\theta = \arctan\left(\frac{|G_y|}{|G_x|}\right) \quad (27)$$

where, G_x and G_y are the gradients in the x and y directions respectively. G_x and G_y are as (26).

The edges are connected when the strength is greater than 8-connected neighborhoods, and it will be removed. The boundaries of medical images are given, but they are not clear and powerful. We continue with them in the next period.

B. Contour Detection

In this period, we show the boundaries from the edges to be detected of the first period. The processing of this period, we present [24] as: extracting the self-affine maps, combining forces, converging snake model. In here, we don't use the coefficients of curvelet as [24]. We improve the number of edges in the part A of this section. The matching code C is also evaluated by (28):

$$C = \sum_{x \in M_i} |g(x) = g(m_i(x))| \quad (28)$$

where, g(x) is the intensity values. This is a first step of our methods. The second step is decomposition in Gaussian steps and the sum calculating (13) and (14). From this, the energy of medical images has given, similar to (15). This is the grounds for forces.

Then, the coefficients for showing the boundaries of objects must fit with:

$$2^n \times E_{image} > E \quad \text{and} \quad n > \log_2 (E / E_{image}) \qquad (29)$$

where, E is the values of equation (15) and n is the coefficients of smoothing step (by the comparison the smaller value between ALM and curvelet). If any value fits, it will be used for restoration and appearance. If not, it will not be a member of object contour. The difference of our method with [24] is we don't only depend on coefficients of denoising images and improve this result of contour. Because of the denoising in space domain can lack information of medical images. We want to keep edges and connect between weak edges. They have very little gray intensity.

IV. Experiments and Results

The contour of objects in medical images has been very important because the size and location of them is manifestation of a disease. In Section 3, we present our method for object detection and improve the quality of boundaries in medical images. In the proposed method, we do not only use the features of curvelet transform for denoising medical images as other methods [24], [25] but also apply ALM for smoothing images. That is the preparatory step for the object detection in curvelet domain. Then, the self-affine snake in curvelet domain show the contour of objects. But we have another improvement to be the comparative values of forces.

The results of our proposed method are compared with the other methods such as self-affine snake [23] and curvelet-based geodesic snake method [25]. We have more edges and clear contour than other methods. We collect dataset from many hospitals with many sizes in more than 100 types of medical images. We test and show the contour of objects about the clarity and accuracy. Besides, we also compare the time processing between methods.

Fig. 2 is the results of the proposed method and other methods with low quality medical images. Fig. 2(a) is the low quality medical images, Fig. 2(b) is the result of self- affine snake [23], Fig .2(c) is the result of curvelet-based geodesic snake method [25] and the proposed method result is the Fig. 2(d).

Fig. 3 is another example about testing. Similar to Fig. 2, Fig. 3(a) is the low quality medical images. Fig. 3(b) is the result of self-affine snake [23]. Fig. 3(c) is the result of curvelet-based geodesic snake method [25] and the proposed method result is Fig. 3(d).

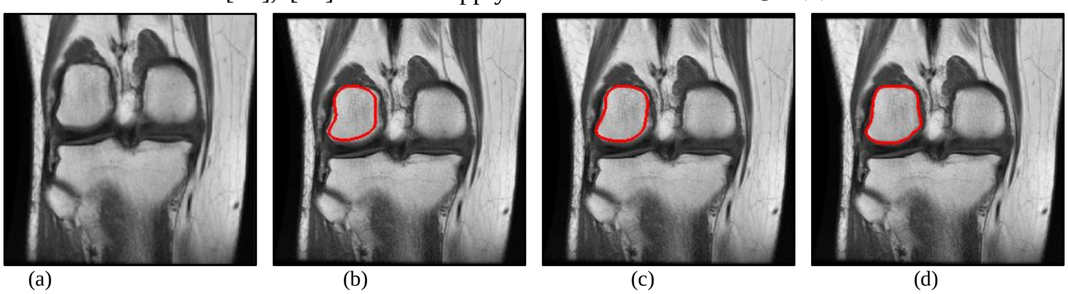


Fig. 2. The results of contour detection by other methods with weak object: (a) The weak object in a noised and blurred medical image. (b) The result of detection by self-affine snake [23] with the time processing ~ 3.152 seconds. (c) The result of detection by curvelet-based geodesic snake method [25] with the time processing ~ 4.523 seconds. (d) The result of detection by the proposed method with the time processing ~ 4.047 seconds.

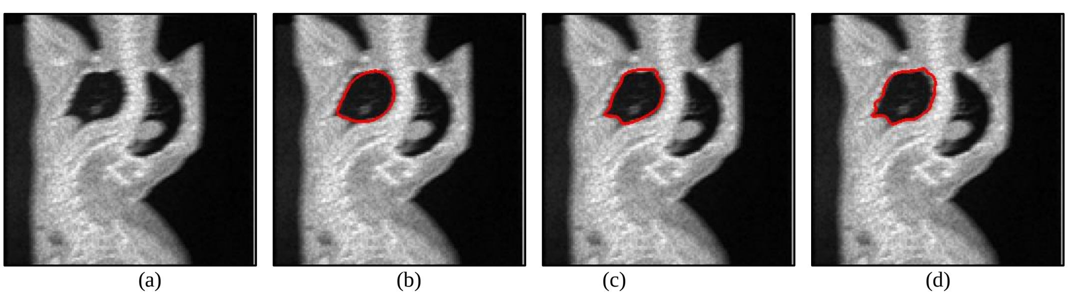


Fig. 3. The results of contour detection by other methods with weak object. (a) The weak object in a noised and blurred medical image. (b) The result of detection by self-affine snake [23] with the time processing ~ 3.941 seconds. (c) The result of detection by curvelet-based geodesic snake method [25] with the time processing ~ 5.067 seconds. (d) The result of detection by the proposed method with the time processing ~ 4.797 seconds.

The images in Fig. 2(a) and 3(a) also have no powerful objects. Because of the addition of 0.0005 of noise value and 0.001 of blur value in them. These numbers are popular in images.

From Fig. 2 and 3, we can see the accuracy of contour detection by our method is better. The time processing is smaller than curvelet-based geodesic snake method [25]. Because the proposed method uses ALM to improve the smoothing step to shorten the time processing. The time processing of self-affine snake [23] is the shortest, but the accuracy of contour is not good. Our results are better than other methods because of two reasons:

Firstly, on one hand, that is improving of the smoothing step by ALM and using curvelet coefficients to active contour.

In edge detection step, the time processing is shortened to base on the ALM that removes noise from input images, because it does not depend on any threshold or filter. This is different from curvelet-based geodesic snake method [25]. The calculating of curvelet coefficients is applied in the next step, after denoising in smoothing step. That is more accurate and faster than other method. With medical images, the small details are very useful. We choose ALM to improve the smoothing step because of this reason. We do not want to lose the any pixels or weak values from input images. Because any pixel has a difference from others, it may be an early express.

Secondly, on the other hand, the curvelet coefficient is very useful for active contour. The combining forces need the wavelet scales. In curvelet domain, these coefficients are very fit in this step. Moreover, the scales and the number of direction are in decomposition. Each subband of curvelet is a curve, to medical images that is an advantage. The previous step, edge detection, has good results and keeps much information of input images; the object contour is an advantage to be showed. And specially, we still use the results from the comparision with ALM coefficients to show forces.

V. Conclusions and Future Works

Object contour is an important task for segmentation, but the edge detection decides the results of contour. So, the improving result in the first step is necessary. When smoothing by transform, the result is very good but the processing time is very slow. We use the ALM to shorten the time but still keep information and applying self-affine snake to active contour. All of the steps are in curvelet domain, but we don't only base on the curvelet coefficients. The authors compare the results with self-affine snake [23] and curvelet-based geodesic snake method [25]. The comparison is not only the sharp contour, but also the execution time of algorithms in low quality medical images. In the future works, we extend to strong low quality medical images. We calculate and compare the information in medical images before and after applying our method.

Acknowledgment

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number C2017-20-20. The authors thank the anonymous reviewers for their constructive comments, which improved the quality of the paper.

References

[1] G. Easley; D. Labate; W.Q. Lim. "Sparse directional image representations using the discrete shearlet transform". Journal of Applied and Computational Harmonic Analysis, Vol 25, pp. 25-46, 2008.

[2] L. Zhai, S. Dong; H. Ma. "Recent methods and applications on image edge detection". International Workshop on Geoscience and Remote Sensing, pp. 332-335, 2008.

[3] Stanley H. chan; Ramsin Khoshabeh; Kristofor B. Gibson; Philip E. Gill and Truong Q. Nguyen. "An Augmented Lagrangian Method for Total Variation Video Restoration". IEEE Trans. Image Process, Vol 20 (11), pp. 3097-3111, 2011.

[4] O. R. Vincent; O. Folorunso. "A Descriptive Algorithm for Sobel Image Edge Detection". Proceedings of Informing Science & IT Education Conference, pp. 97-107, 2009.

[5] Bryan S. Morse. "Edge Detection. Brigham Young University". http://homepages.inf.ed.ac.uk, (last accessed: 28 Nov 2017)

[6] Wang Yuping; Cai Yuanlong. "Multiscale B-spline wavelet for edge detection". Science in China (Series A), Vol. 38(4), pp. 499-512, 1995.

[7] Patrick Brigger; Michael Unser. "Multi-scale B-spline Snakes for General Contour Detection". Wavelet Applications in Signal and Image Processing VI, SPIE, Vol 3458, pp. 92-102, 1998.

[8] A.D. Bhatt; R.V. Warkhedkar. "Reverse engineering of human body: a B-Spline based heterogeneous modeling approach". Computer-Aided Design and Applications, Vol 5, pp. 194-208, 2008.

[9] A.D. Bhatt; R.V. Warkhedkar. "Material-solid modeling of human body: a heterogeneous B-Spline based approach". Computer-Aided Design, Vol. 41, pp. 586-597, 2009.

[10] P. Brigger; J. Hoeg; M. Unser. "B-Spline snakes: a flexible tool for parametric contour detection". IEEE Transactions on Image Processing, Vol. 9, No. 9, pp. 1484-1496, 2000.

[11] Lei Zhang; Paul Bao. "Edge detection by scale multiplication in wavelet domain". Pattern Recognition Letters 23, pp.1771-1784, 2002.

[12] Nguyen Thanh Binh. "Image contour based on context aware in complex wavelet domain". Journal of Human-centric Computing and Information Sciences, Vol 5(14), pp. 1-17, 2015.

[13] Marcin Kociolek; Andrzej Materka; Michal Strzelecki; Piotr Szczypínski. "Discrete Wavelet transform – derived features for digital image texture analysis". Proc. of International Conference on Signals and Electronic Systems, pp. 163-168, 2001.

[14] Nguyen Thanh Binh; Ashish Khare. "Image Denoising, Deblurring and Object Tracking, A new Generation wavelet based approach". LAP LAMBERT Academic Publishing , 2013.

[15] Minh N. Do; Martin Vetterli. "The contourlet transform: an efficient directional multiresolution image representation". IEEE Trans. Img. Processing, pp. 2091-2106, 2005.

[16] Arthur L. da Cunha; Jianping Zhou; Minh N. Do. "Nonsubsampled Contourlet Transform: Theory, Design, and Applications". IEEE Trans. Img. Proc, pp. 3089-3101, 2005.

[17] Arthur L. da Cunha; J. Zhou; Minh N. Do. "Nonsubsampled Contourlet Transform: Filter design and applications in denoising". Proceedings of the 12th IEEE International Conference on Image Processing, pp. 749–752, 2005.

[18] Zhang; J. M. Fadili; J. L. Starck. "Wavelets, ridgelets and curvelets for poisson noise removal". IEEE Transactions on Image Processing, pp.1093-1108, 2008.

[19] Starck J L; Candès E J; Donoho D L. "The curvelet transform for image denoising". IEEE Trans. Image Processing, Vol. 11(6), pp. 670-684, 2002.

[20] Nguyen Thanh Binh, Ashish Khare. "Multilevel threshold based image denoising in curvelet domain". Journal of computer science and technology, Vol 25, pp. 632-640, 2010.

[21] Nguyen Thanh Binh, Vo Thi Hong Tuyet. "The combination of adaptive filters to improve the quality of medical images in new wavelet domain", chapter 3 in book: Classification and Clustering in Biomedical Signal Processing, IGI Global Publication, pp 46-76, 2016.

[22] M. Saadatmand-Tarzjan; H. Ghassemian. "Self-affine snake: a new parametric active contour". IEEE Int. Conf. Signal Process. Commun, pp. 492-495, 2007.

[23] M. Saadatmand-Tarzjan; H. Ghassemian. "Self-affine snake for medical image segmentation". Pattern Recognition. Letters 59, pp. 1-10, 2015.

[24] Vo Thi Hong Tuyet, "Active contour based on curvelet domain in medical images", Proceedings of The International Conference on Nature of Computation and Communication, Vol 168, pp 325-333, 2016.

[25] Hao Shan; Jianwei Ma. "Curvelet-based geodesic snakes for image segmentation with multiple objects". Journal Pattern Recognition Letters, Vol. 31, pp. 355-360, 2010.

[26] Amir Moslemi; Amir Movafeghi; Shahab Moradi. CT Medical "Images Denoising Based on New Wavelet Thresholding Compared with Curvelet and Contourlet". International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol 9(10), pp. 2174-2179, 2015.

[27] Jiange Liu; K. Andrea Scott; Ahmed Gawish; Paul Fieguth. "Automatic Detection of the Ice Edge in SAR Imagery Using Curvelet Transform and Active Contour", Journal of Remote Sens, Vol 8, pp. 480-489, 2016.

[28] Latha Subbiah; Dhanalakshmi Samiappan; P. Muthu. "Automated denoised ultrasound carotid artery image segmentation using curvelet threshold decomposition". IEEE International Conference on Wireless Communications, Signal Processing and Networking, pp. 2261-2266, 2017.

Analytical Study of Algorithms for Solving Inverse Kinematic Problems in Robot Motion Control Systems

Dr. Osama Ahmad Salim Safarini
Department of Computer Science
The University College of Taymaa
University of Tabuk,
Tabuk 71491, Saudi Arabia

***Abstract*—The given article covers the general formulations of inverse kinematic problems for robot motion control systems. We have discussed the difficulties how to solve such problems using analytical and numerical methods. We have also analyzed the convergence of iterative algorithms with the regularization on the trajectory with the points outside of the gripper reachability. The example of an iterative calculation of joint trajectories for a 3-link robot using the recursive algorithm for the Jacobi matrix calculation has been presented.**

***Keywords*—Robot motion control systems; inverse kinematic problems; iterative methods; algorithm convergence; regularization**

I. Introduction

When modeling manipulative robot motion control systems (MCS) it is necessary to solve with the help of a computer the inverse kinematic problems (IKP) for their executive mechanisms (EM) using analytical or iterative methods. Algorithms for solving such problems constitute the mathematical basis for development (MCS) of robots.

A robot manipulator is composed of a serial chain of rigid links connected to each other by revolute or prismatic joints. A revolute joint rotates about a motion axis and a prismatic joint slide along a motion axis. Each robot joint location is usually defined relative to neighboring joint. The relation between successive joints is described by 4x4 homogeneous transformation matrices that have orientation and position data of robots. The number of those transformation matrices determines the degrees of freedom of robots. The product of these transformation matrices produces final orientation and position data of a n degrees of freedom robot manipulator. Robot control actions are executed in the joint coordinates while robot motions are specified in the Cartesian coordinates. Conversion of the position and orientation of a robot manipulator end-effector from Cartesian space to joint space is called as inverse kinematics problem, which is of fundamental importance in calculating desired joint angles for robot manipulator design and control [1].

For a manipulator with n degree of freedom, at any instant of time joint variables is denoted by ($q_i = q(t)$, $i = 1;2;3;....n$) and position variables ($x_j = x(t)$, $j = 1;2;3;….m$). The relations between the end-effector position x(t) and joint angle q (t) can be represented by forward kinematic equation, $x(t) = f(q(t))$ where f is a nonlinear, continuous and differentiable function. On the other hand, with the given desired end effector position, the problem of finding the values of the joint variables is inverse kinematics, which can be solved by, $q(t) = f(x(t))$ Solution of (q(t)) is not unique due to nonlinear, uncertain and time varying nature of the governing equations [2].

The different techniques used for solving inverse kinematics can be reviewed with some articles where, Wu et.al. [3], a new analytic inverse kinematics (IK) solver is proposed which is suitable for multiple constrained 12-DOF human limbs. By decomposing human skeleton into five parts one head chain, two arm chains and two leg chains, a multi-constrained human skeleton can be solved analytically.

Drzevitzky [4] introduced Inverse Kinematics problems for anthropomorphic limbs and have shown how to solve those analytically in order to obtain symbolic solutions. The symbolic solutions can be modified and re-computed to match, for example, other input values that serve as constraints when solving the according Inverse Kinematics problem.

In the theoretical robotics solutions for IKP often use algorithms based on analytical expressions that require calculating inverse trigonometric and transcendental functions. Such algorithms are obtained directly from geometric kinematics models of EM, or by vector-matrix models in the representation of Denavita-Hartenberg, describing the kinematics of EM in homogeneous coordinates [7]-[13].

However, for robots with complex kinematics, analytical solutions of the IKP on a given trajectory of grasping may turn out to be erroneous in specific configurations of EM (link positions), as well as at the boundary and outside the reachable zone of the grasp (gripper) because of the degeneracy of the Jacobi matrix due to the lowering of its rank. In such cases, only approximate solutions of IKP can be obtained by iterative methods. But when using such methods, it is necessary to study the convergence of their algorithms.

The article discusses the difficulties of the analytical solution of the IKP using the example of a three-link robot with rotational links, draws attention to the need for regularization of iterative algorithms to ensure their convergence, describes the recursive algorithm for calculating the Jacobi matrix, and analyzes the operability of iterative algorithms on the trajectories of the gripper containing areas with special configurations of EM.

II. General Formulations of the IKP

In systems of positional, high-speed and force-torque control movement of robots, different inverse problems of kinematics are solved.

A. IKP about Positions of Links

Given a 6 × 1 vector of linear position coordinates and angular coordinates of the orientation of the gripper

$$s_C = (x_c, y_c, z_c, \varphi_c, \theta_c, \psi_c)^T$$

N × 1 vector of generalized coordinates of links is calculated

$$q = (q_1, q_2, \dots, q_N)^T = \Phi^{-1}(S_C), \qquad (1)$$

Where Φ^{-1} is the N × 1 vector-valued function inverse to the 6 × 1 vector-valued function Φ (q) corresponding to the kinematic scheme of the robot's EM.

The problem (1) is the most complicated from the computational point of view, since it requires the solution of a system of nonlinear algebraic equations of the form.

$$\Phi(q) - s_c = F(q) = 0$$

Where, 0 is the zero 6 × 1 vector.

IKP about link speeds

For given vectors of linear and angular velocities of grasping (of the gripper):

$$V_C = (v_x, v_y, v_\tau)^T;\ \Omega_c = (\omega_\varphi, \omega_\theta, \omega_\psi)^T$$

the N × 1 vector of generalized link speeds is calculated.

$$\dot{q} = (\dot{q_1}, \dot{q_2}, \dots, \dot{q_N})^T = J^{-1}(q).\binom{V_C}{\Omega_c} \qquad (2)$$

Where $J^{-1}(.)$ is the inverse (or pseudoinverse) matrix of the Jacobi matrix J(q) of the vector-valued function Φ(q).

The problem (2) is a solution of a system of linear algebraic equations of the form:

$$J(q).\dot{q} = .\binom{V_C}{\Omega_c}$$

B. IKP about the Forces and Moments in the Joints (Hinges) of the Links

By the given vectors of projection of force and moment in the gripper:

$$F_C = (F_x, F_y, F_\tau)^T;\ M_c = (M_\varphi, M_\theta, M_\psi)^T$$

The vector of generalized forces in the hinges is determined

$$Q = (Q_1, Q_2, \dots, Q_N)^T = J^T(q).\binom{F_C}{M_c}. \qquad (3)$$

Expression (3), which requires the calculation and transposition of the Jacobi matrix, is valid only for the case of an ideal EM that does not have energy losses in the joints of the links.

The Jacobi matrix in problems (1) - (3), depending on the number of links N of a robot can be square or have a rectangular form:

$$J(q) = \frac{\partial\Phi(q)}{\partial q} = \frac{\partial F(q)}{\partial q} = \left\{\frac{\partial F_i(q)}{\partial q_j},\ i = \overline{1,6},\quad j = \overline{1,N}\right\}$$

In a more general formulation, problem (1) can be formulated as the problem of minimizing the square of the norm of the discrepancy vector:

$$\|F(q)\|^2 \to \min_{q \in R^N}, \qquad (4)$$

and solved by iterative methods without computing the Jacobi matrix or gradient methods using its numerical approximations. However, the convergence of algorithms of such methods is slower and when they are used, more steps of the iterative process are required [6].

In [8], an example of an iterative solution of the IKP in the formulation (4) for a 6-link robot is given, on the program trajectory of grasping which the constant orientation of the grasping is given not by the Euler angles, but by the vector of the direction cosines.

If the values of the vector of generalized coordinates are bounded by the admissible domain $q \in D^N$, then problems of the form (4) should be solved by methods of conditional minimization. In this case, the exact solution of the IKP may in principle be absent.

C. Analytical Solution of the IKP

Consider a 3-link robot of the BBB type operating in an angular coordinate system for which problem (1) is solved ambiguously and depends on the sign of the angular position of the third link (the lower or upper arm configuration):

$$q_1 = \operatorname{arctg}\left({}^{y_c}/_{x_c}\right);$$
$$q_3 = \pm\arccos(C);$$
$$q_2 = \operatorname{arctg}\left({(z_c - L_1)}/{R_{xy}}\right) \pm \arccos(D), \qquad (5a)$$

Where L_1, L_2, L_3 are the lengths of the links;

$$R_{xyz} = \sqrt{x_c^2 + y_c^2 + \left((z_c - L_1)\right)^2};$$
$$C = (x_c^2 + y_c^2 + \left((z_c - L_1)\right)^2 - L_2^2 - L_3^2) / (2L_2L_3);$$
$$D = (x_c^2 + y_c^2 + \left((z_c - L_1)\right)^2 + L_2^2 - L_3^2) / (2L_2R_{xyz});$$

The solution of the IKP can also be obtained from other expressions [9]:

$$q_1 = arctg\left({}^{y_c}/_{x_c}\right);\;\; E = \sqrt{1 - C^2};$$

$$q_3 = \pm \text{arctg}\,(E/C);$$

$$q_2 = \text{arctg}\left({}^{(z_c - L_1)}/_{R_{xy}} \right) - \\ - \text{arctg}\left(\frac{L_3 \sin q_3}{L_2 + L_3 \cos q_3}\right) \quad (5b)$$

However, the solutions of the IKP with MATLAB functions by the expressions (5a) on the trajectory of grasping with points outside its reachability zone turn out to be complex numbers that have no practical meaning. To obtain real solutions of the IKP, the values of the variables C and D, which are cosines, should be limited to one. The solutions of the IKP by the expressions (5b) are obtained real at points both inside and outside the reach zone of the gripper, so there is no need to limit the values of the variables C and E.

III. ALGORITHMS OF ITERATIVE METHODS

We consider algorithms for the numerical solution of problem (1), in which the Jacobi matrix is calculated in explicit form.

A. Algorithm of Newton's Method

Theoretically, in the case of a square Jacobi matrix, for example, when N = 3 or 6, the following algorithm can be used:

$$q^{k+1} = q^k - a_k J^{-1}(q^k) \cdot F(q^k), \quad (6)$$

Where $a_k \leq 1$ is the scalar value of a constant or variable step.

Wherein At each step of the iterative process, it is required to calculate the inverse matrix corresponding to the Jacobi matrix. However, in special configurations of EM, if the Jacobi matrix is poorly conditioned or completely degenerate due to a decrease in its rank, the operability of algorithm (6) is lost,

Simplified algorithm of Newton's method

If we replace the inverse matrix J^{-1} in the algorithm (6) with the transposed Jacobi matrix

$$q^{k+1} = q^k - a_k J^T(q^k) \cdot F(q^k), \quad (7)$$

Coarse (Rough) solutions of problem (1) can be obtained, but in special configurations of EM, the convergence of algorithm (7) is also not guaranteed.

B. Algorithm of the Gauss-Newton Method

When solving problems with a rectangular Jacobi matrix, when the number of links is redundant (N > 6) or insufficient (N < 6) in the algorithm (6), instead of the inverse matrix J^{-1}, it is necessary to use the left

$J^+ = (J^T \cdot J)^{-1} J^T$ or the right $J^+ = J^T (J \cdot J^T)^{-1}$ pseudoinverse matrices [5]. We obtain the following algorithm:

$$q^{k+1} = q^k - a_k J^+(q^k) \cdot F(q^k) \quad (8)$$

C. Algorithms of the Levenberg-Marquardt Method

In the algorithm (8), in calculating pseudo inverse matrices, because the matrices $(J^T \cdot J)$ and $(J \cdot J^T)$ can turn out to be poorly conditioned or degenerate, to ensure the convergence of the iterative processes, it is necessary to carry out factorization based on the matrix decomposition or to use simple regularization:

$$q^{k+1} = q^k - a_k\,[(J^T J + a_{2k} E)^{-1} J^T] \cdot F(q^k) \quad (9)$$

$$q^{k+1} = q^k - a_k\,[J^T (J J^T + a_{2k} E)^{-1}] \cdot F(q^k) \quad (10)$$

Where E is the identity matrix; a_{2k} is a regularizing scalar parameter.

When applying any of the algorithms (6) - (10), the matrix ***J (q)*** can be calculated either by the direct derivation of the vector function $\Phi(q)$by analytic expressions, which is rather cumbersome for N $\geqslant$ 3, or by a more efficient recursive algorithm [13], using the transformation matrices of homogeneous coordinates and the intersection operations (vector products) of the columns of the rotation matrix of grasping (*n*, *o*, *a*) - normal vectors, orientation and approach.

A recursive algorithm for computing the Jacobi matrix

In the base coordinate system of the robot, the Jacobi matrix performs the transformation of the vector of generalized link speeds:

$$\begin{pmatrix} V_{C,0} \\ \Omega_{c,0} \end{pmatrix} = J_0(q)\,.\,\dot{q}$$

Where $V_{C,0}$ and $\Omega_{c,0}$ are the vectors of the projections of the linear and angular velocity of grasping.

Here the Jacobi matrix is defined as the product

$$J_0(q) = J_V(q) \cdot J_N(q)$$

Where $J_N(q)$ – $6 \times N$ Jacobi matrix of velocity transformation in the coordinate system of the gripper

$$\begin{pmatrix} V_{C,N} \\ \Omega_{C,N} \end{pmatrix} = J_N(q)\,.\,\dot{q}$$

$J_V(q)$ - 6 × 6 matrix transformation of linear and angular velocities of grasping from the *N*-th to the base coordinate system.

$$\begin{pmatrix} V_{C,0} \\ \Omega_{C,0} \end{pmatrix} = J_V(q)\,.\begin{pmatrix} V_{C,N} \\ \Omega_{C,N} \end{pmatrix}$$

To calculate the matrix $J_N(q)$, 4 × 4 matrices are required that determine the position and orientation of the gripper in the *j*-th coordinate system, i.e., with respect to the (*j*-1) -th link:

$$U_j(q) = A_j(q_1) A_{j+1}(q_2) \ldots A_N(q_N) = \\ = \begin{pmatrix} R_{N,j} & P_{N,j} \\ 0\;0\;0 & 1 \end{pmatrix},$$

Where $R_{N,j} = (n_{N,j}, o_{N,j}, a_{N,j})^T$ - rotation matrix with orientation vectors of the grip coordinate system; $P_{N,j} = (X_{N,j}, Y_{N,j}, Z_{N,j})^T$ vector of the position of the center of the coordinate system of the grasp.

Columns of the matrix $J_N(q) = (J_{1,N}\ J_{2,N} \dots\dots J_{6,N})$ are calculated by the reverse recursion (j = 6, 5, ….1):

$$J_{j,N} = \begin{cases} \begin{pmatrix} d_j \\ \delta_j \end{pmatrix} & \text{For j link B-type} \\ \begin{pmatrix} \delta_j \\ 0_{3\times 1} \end{pmatrix} & \text{For j link p-type,} \end{cases}$$

Where d_j - 3 × 1 vector, composed of Z-components of vector products.

$$d_j = \begin{bmatrix} (P_{N,j} \times n_{N,j})_\tau \\ (P_{N,j} \times o_{N,j})_\tau \\ (P_{N,j} \times a_{N,j})_\tau \end{bmatrix} = \begin{bmatrix} P_{x,j}\, n_{y,j} - P_{y,j}\, n_{x,j} \\ P_{x,j}\, o_{y,j} - P_{y,j}\, o_{x,j} \\ P_{x,j}\, a_{y,j} - P_{y,j}\, a_{x,j} \end{bmatrix};$$

δ_j- a 3 × 1 unit vector directed along the axis of motion of the *j*-th link

$$\delta_j = R_{N,j}^T \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} n_{\tau,j} \\ o_{\tau,j} \\ a_{\tau,j} \end{bmatrix};$$

$0_{3\times 1}$ is the zero 3 × 1 vector.

The matrix $J_V(q)$ in the block representation has the following form:

$$J_V(q) = \begin{bmatrix} R_{N,0} & 0_{3\times 3} \\ 0_{3\times 3} & R_{N,0} \end{bmatrix}$$

Where $0_{3\times 3}$ is the zero matrix; $R_{N,0}(q) - 3 \times 3$ rotation matrix with the column vectors of the normal, orientation and approach, calculated in the base coordinate system.

A more detailed recursive algorithm for computing the Jacobi matrix is described in [9].

Analysis of the convergence of iterative algorithms

Let us analyze the processes of the iterative IKP solution using algorithms (6) - (10) with the Jacobi matrix calculation using the above recursive algorithm on two BBB type robot gripper trajectories (Fig. 1).

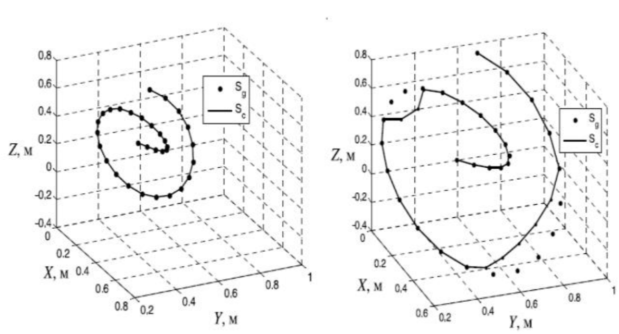


Fig. 1. Given trajectories of grasp.

The trajectories are given by the base points S_g ={X, Y, Z}, the first is in the reachable zone of grasp, the second - with some points outside its boundary.

Fig. 2 and 3 show the trajectory of the links obtained for the upper configuration of the robot by analytical methods (5a) and (5b), and calculated by these trajectories by solving the direct problem of kinematics corresponding trajectories of the gripper S_c = {X_c, Y_c, Z_c} are shown in Fig. 1 by solid lines.

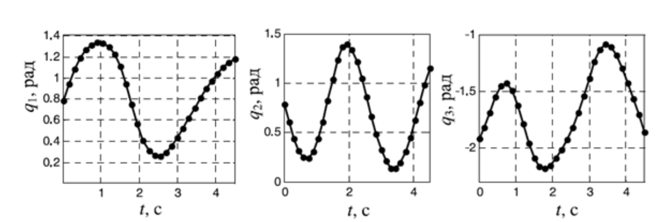


Fig. 2. Results of the solution of the IKP on the first trajectory of grasp.

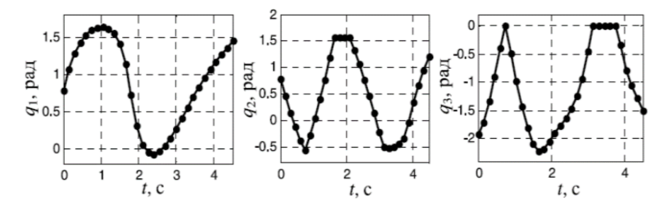


Fig. 3. Results of the solution of the IKP on the second trajectory of grasp.

In Fig. 4 are presented the graphs of the modules of the Jacobi matrix determinant of the BBB-type robot, calculated from the recursive algorithm for the two variants of the link trajectories (see Fig. 2 and 3).

Exactly the same graphs are obtained when calculating the elements of the Jacobi matrix by analytic expressions. It is seen from the graphs that for the second variant of the trajectories at the points on the boundary of the reachability zone of the grasp the determinant of the degenerate Jacobi matrix takes the value zero.

On the first trajectory of the grasp, the solutions obtained by all the compared algorithms (6) - (10) coincide with the trajectories of the links (see Fig. 2), calculated by analytical methods. At the same time, for each of the algorithms, the values of the constant parameters a_k and a_{2k} were established, which ensure the convergence of the IKP solution processes for the number of iterations $k_{max} \leq 100$.

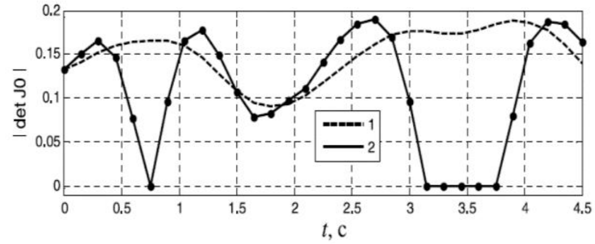


Fig. 4. Graphs of modules of the Jacobi matrix determinant.

On the second trajectory of grasping, because of the degeneracy of the Jacobi matrix at special points, the convergence of algorithms (6) and (8) is lost. Algorithms (9) and (10), with regularized left and right pseudo-inverse matrices, are the most stable. At values of $a_k = 2$, $a_{2k} = 0{,}1$ they give the same solutions as in Fig. 3, for the number of iterations $k_{max} \leq 30$ .The algorithm (7), which uses the transposed Jacobi matrix, converges, but gives very rough solutions.

IV. Conclusion

Thus, based on the results of the study, the following practical recommendations can be made.

When modeling and developing control systems for the movement (motion) of manipulative robots, it is advisable to use iterative algorithms for solving the IKP. Analytical solutions by expressions containing inverse trigonometric functions in special configurations of EM can turn out to be incorrect.

The most accurate solutions of the IKP multi-link robots can be obtained by algorithms (9) and (10) with regularization, in which the Jacobi matrix is recommended to be calculated without numerical differentiation using a recursive algorithm that uses kinematic models in homogeneous coordinates.

When programming (planning) the grasp paths (trajectories) in robot motion control systems, care should be taken to avoid the situations shown in Fig. 1 and 3.

The base points should be set in the working area of the robot, determined by the overall dimensions of the links and the Permissible variation ranges of the generalized coordinates.

References

[1] P. Baerlocher, “Iverase Kinematic Techniques of the Interactive Posture Control of Articulated Figures”. Thesis2001.

[2] S Alavandar, M.J. Nigam. “Inverse Kinematics Solution of 3DOF Planar Robot using ANFIS”, Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844. Vol. III (2008), Suppl. issue: Proceedings of ICCCC 2008, pp. 150-155.

[3] X. Wu, L. Ma, Z. Chen, Y. Gao, “A 12-DOF Analytic Inverse Kinematics Solver for Human Motion Control ”, Journal of Information & Computational Science, PP 137-141, 30-August 2004.

[4] S. Drzevitzky, “Symbolic Solutions for Inverse Kinematics Problems for Anthropomorphic Limbs”, Diplomarbeit Thesis, Germany, March, 2008.

[5] BEKLEMISHEV D.V. DOPOLNITELNYYE GLAVY LINEYNOY ALGEBRY. MOSCOW: NAUKA PUBL., 1983, 336 P. (RUS).

[6] Dennis D., Shnabel R. Chislennyye metody bezuslovnoy optimizatsii i resheniya nelineynykh uravneniy. Moscow: Mir Publ., 1988, 440 p. (rus)

[7] ZENKEVICH S.L., YUSHCHENKO A.S. UPRAVLENIYE ROBOTAMI. OSNOVY UPRAVLENIYA MANIPULYATSIONNYMI ROBOTAMI. MOSCOW: MGTU IM. N.E. BAUMANA PUBL., 2000, 400 P. (RUS).

[8] Ignatova Ye.I., Rostov N.V. Kinematicheskiye Zadachi v Matematicheskom Modelirovanii Robotov, Nauchno-tekhnicheskiye Vedomosti SPbGPU. Informatika. Telekommunikatsii. Upravleniye. St. Petersburg: SPbGPU Publ, 2010, No. 2(97), Pp. 159–164. (RUS).

[9] Ignatova Ye.I., Lopota A.V., Rostov N.V. Sistemy Upravleniya Dvizheniyem Robotov. Kompyuternoye Proyektirovaniye. St. Petersburg: SPbGPU Publ, 2014, 302 p. (RUS).

[10] Shakhinpur M. Kurs Robototekhniki, Per. S Angl. Pod Red. S.L. Zenkevicha. Moscow: Mir Publ., 1990, 527 p. (RUS).

[11] Fu K., Gonsales R., Li K. Robototekhnika, Per. S Angl. Pod Red. V.G. Gradetskogo. Moscow: Mir Publ., 1989, 624 p. (RUS).

[12] Yurevich Ye.I. Osnovy Robototekhniki. St. Petersburg: BKhV-Petersburg Publ., 2005, 416 p. (RUS).

[13] Corke P.I. Robotics Toolbox 9.9 for MATLAB R9. 2014, 286 p.

Terman-Merril Application for Intelligence Measurement

Alicia Valdez[1], Griselda Cortes[2], Laura Vazquez[3], Andrea de la Pena[4]
Research Center
Autonomous University of Coahuila
Coahuila, Mexico

Blanca Montano[5]
Human Resources Department
Faculty of Accounting and Administration
Autonomous University of Coahuila
Coahuila, Mexico

***Abstract*—The computational applications support different processes in organizations, one of these processes are those related to human resources, where one of the activities is the hiring of new personnel; the evaluation of human talent to be integrated into a company can be measured through different tests, one of which is the Terman-Merril intelligence test, which measures the intellectual quotient of candidates with a series of sub-tests. In this project, the waterfall method has been used for the development of a web computational application for the Terman-Merril intelligence test, as well as the management of the users and the results obtained to be visualized in spreadsheets for its subsequent analysis and graphing. The ASP.Net programming language and the SQL Server 2014 database have been used for the programming and the storage of information. As a result, it has been applied successfully in some companies, obtaining measurable and evaluable results on the candidates. The application has also been installed in a computer lab for students enrolled in the Bachelor of Human Resources at Faculty of Accounting and Administration at Coahuila, Mexico.**

Keywords—Human resources; Terman-Merril test; intelligence test; ASP.Net

I. Introduction

The intelligence research is an important area of knowledge for researchers and practicing psychologists, organizations require tests to evaluate new staff and select those that get the best scores, testing in the selection process is a significant element that has aim to select the high quality personnel in the organization. Special features in the selection process are represented through psychological tests and intelligence tests [1].

The assessment of intelligence required good measurement instruments, consists of standardized questions and tools for assessing an individual potential [2].

At the present day there are various psychometric tools for intelligence measurement, one of these is the "Standard Progressive Matrices Test" (SPMT), which was developed by J.C. Raven (1939) and distributed by US Psychological Corporation, consists of abstract reasoning; puzzle solving, problem solving, learning and patterns recognized [3].

Alternative tools for measuring intellectual abilities were proposed by M.A. Kholodnaya as a development of an onthological approach to intelligence studies, these tests aim to measure categorical and conceptual abilities that underlie intellectual productivity, implies the necessity to derive a new conceptual knowledge by combining three words from completely different semantic contexts into one meaningful sentence [4].

The Terman-Merril Test (TMT) was created in 1960, by L. Terman and M. A. Merril, based on the research work of A. Binet, who was a teacher at Stanford University; for the evaluation of high intellectual capacities, where intelligence is measured as a general capacity starting from a chronological age, obtaining intellectual quotient (IQ) values significant and progressive higher than those obtained with the currently used factor scales, whose primary objective is to determine the intellectual quotient of people.

It is a set of ten structured tests, requires a maximum execution to obtain the best performance of the examinee.

The objectives of the test measure are: Common sense to appreciate social situations, develop the ability to understand concepts expressed in words, knowledge of language, obtain ability to summarize, relate and abstract essential ideas, develop the ability to concentrate and work under pressure, learn to anticipate situations to foresee the future and mentally imagine the solution to a problem. Each series has a time limit and the total test requires 27 minutes, and can be administered individually or collectively [5], [6].

In this project was developed a computational application based in the TMT for intelligence measurement, considering all the variables of the ten tests that integrate it.

The software was used as a tool in a computer laboratory for the human resources students at the Faculty of Accounting and Administration at Autonomous University of Coahuila, else for examining people to be hired by the companies.

The program was created using Active Server Pages (ASP).Net [7] programming language and the SQL Server 2014 database for data storage [8].

Basically, this study has four sections.

In Section I, the introduction was shown. In Section II, the fundamental concepts are described. Also, the Terman Merril structure was stated. Thus, Section III describes the methodology, and Section IV describes the principal findings of the project.

II. Fundamental Concepts

A. Terman-Merril Structure

It is composed of ten sets of tests that measure different cognitive skills and abilities of people, each of which has a measurement in minutes to be applied, in total the full test can be achieved in 27 minutes. The particular characteristics are described below, the tests are passed when the subject meets the criterion of step for each of them; minimum performance sufficient for the proposed task, being the result dichotomous (overcome / not exceeded). Tests are assessed and scored at more than one mental age level with demanding criteria. The IQ is obtained from the corresponding norms, which relate the chronological age at the time of performing the test with the mental age obtained. As the chronological age increases, the mean and dispersion of the mental ages is also increasing [9].

- Information or knowledge, measure culture and general knowledge, long-term memory, answer in a time limit 2 minutes.
- Understanding, measures the understanding and management of reality.
- Verbal meanings, measures the ability to analyze and synthesize concepts.
- Logical selection, measures the ability to deduce and logically abstract concepts. 2 minutes time.
- Arithmetic, measures reasoning and quantitative concepts. It is answered within a 3 minutes time limit.
- Practical judgment, measure common sense. It is answered in a time limit of 5 minutes.
- Analogies, ability to reason, abstract, generalize and think in an organized way.
- Sentence ordering, planning, organization and understanding of concepts. Attention to detail. Maximum response time 3 minutes.
- Classification, measures the logical discrimination of concepts.
- Seriation, measures the deduction capacity. It is answered in a maximum of 4 minutes. Table I summarizes of test series, skill and time.

TABLE I. Terman-Merril Test Series

Serie	Skill	Time
I	Information	2 minutes
II	Understanding	2 minutes
III	Verbal meanings	2 minutes
IV	Logical selection	3 minutes
V	Arithmetic	5 minutes
VI	Practical judgment	2 minutes
VII	Analogies	2 minutes
VIII	Sentence ordering	3 minutes
IX	Classification	2 minutes
X	Seriation	4 minutes
Total		27 minutes

Once the methodology comprising the TM intelligence test is understood, the software development component is carried out.

B. Software Engineering

Software development is based on software engineering, which is an engineering discipline that covers all aspects of software production. The goal is the cost-effective development of software systems where there are no physical limitations on the software's potential. software, which can sometimes be complex and difficult to understand [10]. Another concept about software engineering that could be defined as the establishment and application of engineering principles to obtain software. Taking into account factors as important as the economic cost, the reliability of the system and an efficient operation that meets the needs of the user [11].

C. Waterfall Model

The life cycle initially proposed by Royce in 1970 [12], shown in Fig. 1, was adapted for the software from the life cycles of other branches of engineering. It is the first of the proposed and most widely followed by organizations (it is estimated that 90% of the systems have been developed under this method) [11].

It works on the basis of documents, that is, the entry and exit of each phase is a specific document type deliverable. Ideally, each phase could be done by a different team thanks to the documentation generated between the phases. The documents are: Analysis: take as input a description in natural language of what the client wants. Produces the Software Requirements Document (SRD). Design: The entry is the SRD, produce the Software Design Document (SDD). Coding: From the SDD, produces modules. In this phase, unit tests are also carried out. Tests: The integration and testing of the entire system is carried out from the approved modules. The result of the tests is the final product ready to deliver [13].

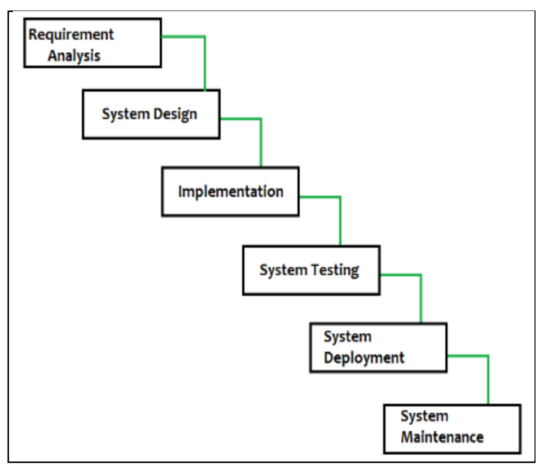


Fig. 1. Waterfall model.

This model has some advantages as the planning is simple, the quality of the resulting product is high and it allows working with low-qualified personnel. By the other hand the disadvantages are the need to have all the requirements at the beginning. Typically, the client does not have perfectly defined system specifications, or unexpected needs may arise, if mistakes have been made in one phase it is difficult to go back, you do not have the product until the end, if an error is made in the analysis phase it is reflected until the delivery, with the consequent waste of resources. The client will not see results until the end.

D. *Software used for the Development of the Application*

Microsoft SQL Server: SQL Server is a Microsoft Relational Database Management System (RDBMS) that was designed for the business environment. SQL Server runs on T-SQL (Transact-SQL), a set of Sybase and Microsoft programming extensions that add several features to standard SQL, including transaction control, exception and error handling, row processing, as well as declared variables. It also supports the management of capabilities for business intelligence and data mining [14].

Visual Studio 2015: It is an Integrated Development Environment (IDE) for Windows operating systems. Supports multiple programming languages, such as C ++, C #, Visual Basic .NET, F #, Java, Python, Ruby and PHP, as well as web development environments, such as ASP.NET MVC, Django, and others [15].

III. Methodology

The methodology phases were: Analysis of TMT components, requirements for the analysis, software specifications for application development, design and codification, testing and maintenance, and implementations and results. Fig. 2 displays the methodology phases.

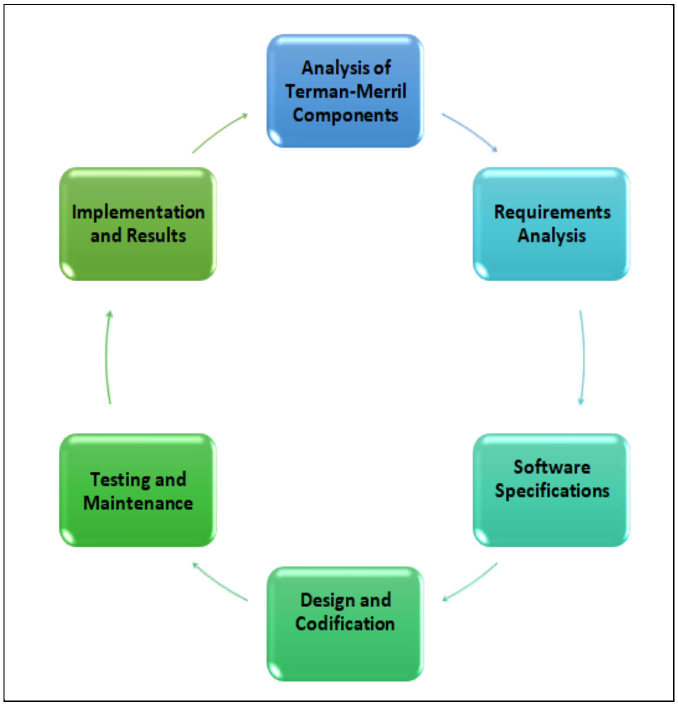


Fig. 2. Methodology phases.

Regarding the methodology to perform the research, first a detailed analysis of the requirements for the specifications of the ten series that make up the TMT was carried out, the application was developed complying with all the specifications regarding required data; how to evaluate each series, time needed, and results obtained.

The biggest challenge was the application development, since each series requires a different evaluation; the evaluation was based on a template with the correct answers and the final obtaining of the IQ; which is the data that the evaluator requests from the application.

Once the application was developed, it was applied in a sample of thirty students of the Bachelor's degree in Human Resources and in a company of the public transport branch to evaluate the applicants, obtaining successful results compared against other psychometric test such as Cleaver test.

Finally, it has been implemented in the human resources laboratory of the Faculty of Accounting and Administration and in the firm.

A. *Analysis Diagrams*

Some diagrams produced during analysis phase were displayed in Fig. 3 and 4. In Fig. 3 the application processes were shown.

The main requirements for the development in the first phase, users registry, a program where the general users data for registry in the application are entered, and with that, users can create their own account and password. After users registry phase was done, the instructions about the test were shown; before beginning the test series, samples of the questions were displayed. This process was done in the ten series of TMT. The responses were saved in the database; Fig. 4 shows part of the database diagram.

The result processing was activated through a buttom in the application once the series were responded. The TMT score obtained was calculated in the next way, first the series questions were saved in a table with the correct response; in other table related, the user answers were saved and a score was generated based in the answers and the correct answer. In this way, all the answers of the complete series were calculated and stored in a table named seriesUsers.

The user can see all the scores obtained in the ten series and visualize a result. The scores obtained place the user at a level from deficient, inferior, medium low, medium, medium high, superior to the outstanding level, which is the highest.

A part of the design of the database is shown in Fig. 4, where the main entities that are represented are: Users, SeriesUsers, Series, Questions, Answers, User Answers, among others.

With this scheme it is intended to solve the storage of the necessary data required by the TMT web application in all its processes.

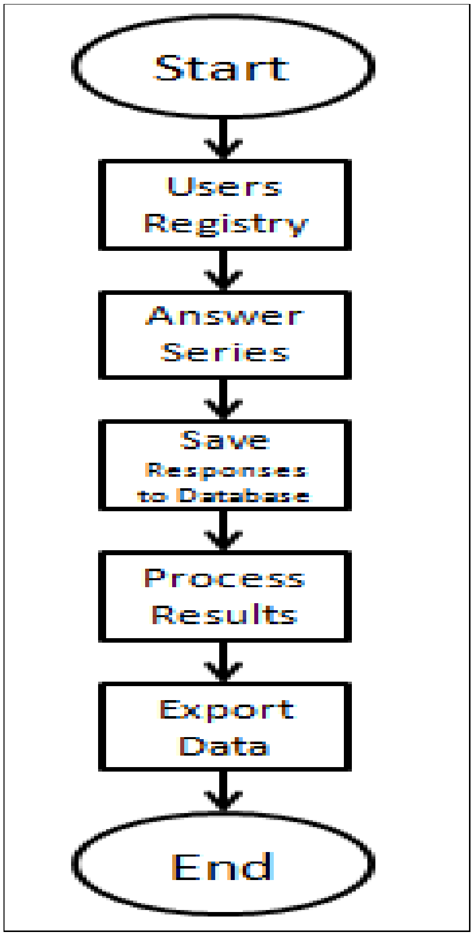


Fig. 3. Data flow processes.

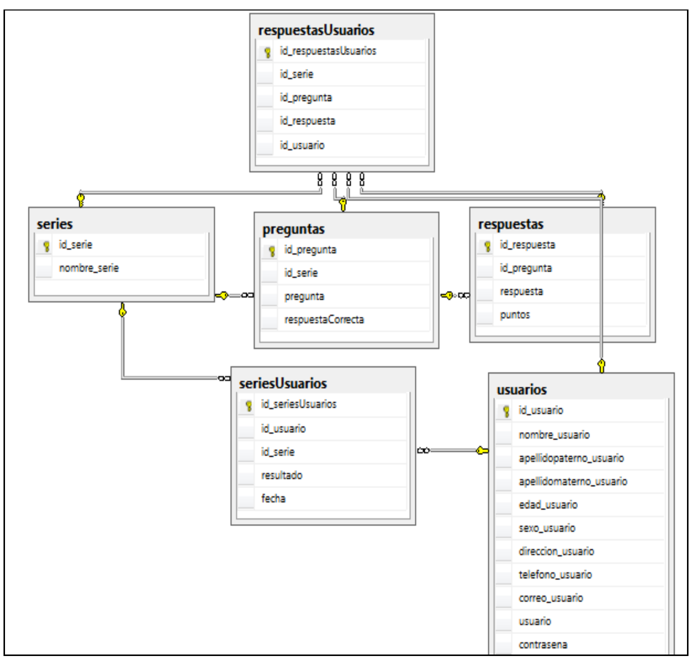


Fig. 4. Data model.

The User Registration Process is shown in Fig. 5, where the user's general data are recorded.

Fig. 5. Users registry.

A part of the program code that contains the user's registry is shown in Fig. 6.

Each of the series has its own programming since they have different questions, answers and way of evaluating.

The application was implemented in a university Windows server, the database recorded the responses to each of the series and this information can be exported to spreadsheets for been evaluated by human resources personnel.

```
1. using System; 2. using System.Collections.Generic; 3. using System.Data;
4. using System.Data.SqlClient;
5. using System.Linq; 6. using System.Web;
7. using System.Web.UI;
8. using System.Web.UI.WebControls;
9. public partial class TermanMerrilLogin : System.Web.UI.Page
10. {
11. protected void Page_Load(object sender, EventArgs e)
12. {
13. }
14. protected void Button1_Click(object sender, EventArgs e)
15. {
16. using (SqlConnection cnn = new SqlConnection(@"Data
source=VAIO\SQLEXPRESS;initial catalog=TermanMerril;Integrated
Security=True;"))
17. {
18. cnn.Open();
19. using (SqlCommand cmd = new SqlCommand("SELECT * FROM usuarios WHERE
usuario ='" + txtUsuario.Text + "' AND contrasena ='" + txtContraseña.Text
+"'", cnn))
20. {
21. //Linea de ejemplo para agregar parametros al query
22. using (SqlDataReader rdr =
cmd.ExecuteReader(CommandBehavior.CloseConnection))
23. {
24. if (rdr.Read())
25. {
26. Session["usuario"] = txtUsuario.Text;
27. Session["contrasena"] = txtContraseña.Text;
28. Session["id_usuario"] = rdr["id_usuario"];
29. //'Mandar a la página deseada despues de hacer el login
Response.Redirect("Series/Serie1Inicio.aspx");
30. }
31. else
32. {
33. Response.Write("<script language=javascript>alert('USUARIO O CONTRASEÑA NO
VALIDOS');</script>");
34. }
35. rdr.Close();
36. } }
37. cnn.Close();}}}
```

Fig. 6. Code for users registry.

IV. RESULTS

Once the application has been implemented in the faculty repository, tests have been carried out with users-students of the Accounting and Administration faculty who are studying the Bachelor's degree in Human Resources, to whom the program is useful as a laboratory.

It has also been used in the evaluation of a transport company where fifteen people were evaluated for hiring, reflecting a high degree of acceptance among the company's human resources managers. The application has a simple and effective design at the same time; since it provides the pertinent information about the people who answer it, so it can be implemented in any company or organization to evaluate the intellectual capacities of the users.

Fig. 7 shows the results of the application of the test in the company, where the data of ten people were shown, with age, the results of the ten tests and the IQ obtained.

The results obtained from the application of the test in the group of students, who answered it in the established time, where the system records the responses of each person and the intellectual coefficient obtained for each user.

These data are plotted and visualized by the test applicator.

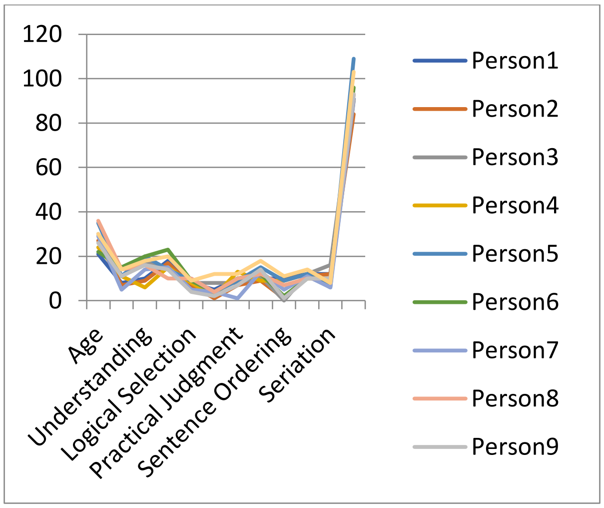


Fig. 7. Results of the Terman-Merril intelligence test application.

The application export the results to Excel spreadsheet for the graphics purposes, the application provides sufficient basis to make a decision to hire or not hire staff, other factors may influence the final decision of human resources personnel.

V. CONCLUSION

In this project, a web computational application has been designed and implemented for solving the needs of the human resources department at the Faculty of Accounting and Administration at Autonomous University of Coahuila, Mexico. Especially in the hiring of new personnel and else as an application that can be used for students in a computing laboratory, so can serve for various purposes as practices for the students and for uses in companies for evaluating new personnel to be hired.

The software has some modules: users' registry, answer series, process results, and export data, between others.

It should be noted that in the development of the TM application now plays an important role with students and teachers from Accounting Faculty and can be applied at human resources companies for helping. Some phases were defined in order to fulfillment the project; these were: Analysis of the TM components, requirements analysis, software specifications, design and codification, testing and maintenance, and implementation and results; using the waterfall model as a methodology.

The main results obtained were the process automation of the new personnel evaluation, made through a computational application, this implies better time response on the process, have the electronic information of the candidates for a position, less margin of error when applying calculations, and have a historical record of the candidates.

The implementation of the solution allowed appreciate how the use of IT, can support the processes of management and decisions in contexts such as university and companies, and contribute to strengthening initiatives aimed at improving decision-making process supported with the values obtained from the application.

The methodology used facilitated all the application development process.

REFERENCES

[1] O. Shcherbakova, D. Makarova, and E. Nikiforova, "Concept syhthesis test for intelligence research:qualitative analysis and conceptual commentaries". Bulletin of St. Petersburg State University. Psychology and Pedagogy, 2017, 2(7): p. 128-136..

[2] S. Ismat and J. Sagir, "A study of intelligence measure using Raven Standard Progressive Matrices Test items by principal components analysis" . Fuuast Journal of Biology, 2015. 5(1): p. 169-173.

[3] K. Grew and D. Flanagan, The intelligence test desk reference1998, U.S.A.: Boston7 Allyn and Bacon, pp. 65-90.

[4] M. Kholodnaya, and E. Volkova, "Conceptual Structures, Conceptual Abilities and Productivity of Cognitive Functioning: The Ontological Approach". Procedia-Social and Behavioral Sciences, 2016, 217(1): p. 914-922.

[5] N. Lemann, "The IQ Meritocracy". Time International, 1999. 153(12): p. 83-93.

[6] J. Ballesteros, "How the Terman scale measures intelligence. General psychometric model. Revised rules and updated results". Journal of Educational Psychology, 2011. 17(2): p. 179-193.Meyer, J., Le controle de gestion1969, Paris, France: Presses Universitarires.

[7] M. Amundsen and P. Litwin, Creating web sites with ASP.NET2007, España: Prentice Hall, pp. 45-67.

[8] A. Brust and S. Forte, Programación avanzada con Microsoft SQL Server 2005,2007, México, D.F.: McGraw-Hill, pp. 34-56.

[9] J. Ballesteros, "A critical review of the Terman scale. Why we should not use the Third Edition Stanford-Binet Form L-M". Psicología Educativa, 2010. 16(1): p. 23-30.

[10] I. Sommerville, Ingeniería del software.Séptima edición, ed. P. Educación2005, Madrid, España: Pearson educación. 712.

[11] R. Pressman, Ingeniería del software. Un enfoque práctico, ed. McGraw-Hill. Vol. 7a. Edición. 2010, México, D.F.: McGraw-Hill.

[12] W. Jedzej, "An approach to analysis and implementation from the waterfall model to the two-segmental model of information systems lifecycle". Business Informatics, 2012. **3**(25): p. 195-206.

[13] R. Stephens, Beginning software engineering, 2015, Canada: Wiley Publishers, pp. 67-89.

[14] J. Gabillaud, Sql server 2016; aprender a administrar una base de datos transaccional con Sql server managament studio, 1a. ed., 2017, Madrid, 2017, Editorial ENI, pp. 215-234.

[15] S. Putier, Vb.Net y Visual studio 2015 los fundamentos del lenguaje, 1a. ed., 2017, Madrid :Editorial ENI, pp. 154-170.

Software Cost Estimation using Enhanced Artificial Bee Colony Algorithm

Sevgi Yigit-Sert, Pinar Kullu
Computer Engineering Department
Ankara University
Ankara, Turkey

***Abstract*—Cost estimation is very important in software development progress so that resource and time planning can be successfully performed. Accurate estimation of cost is directly related to the decision making mechanism in the software development process. The underestimated cost might lead to fewer resources and budget problems; in contrast, customer satisfaction might diminish due to waste of resources. This study represents an estimation model for the effort required for the development of software projects using a variant of artificial bee colony (ABC) algorithm. The proposed model is performed over a dataset consisting of NASA software projects and has better performance than the previous studies.**

Keywords—Software cost estimation; enhanced artificial bee colony algorithm; NASA software

I. Introduction

As technological advances have risen, the competition in the software industry has been increased. Thus the prediction of features that affect software processes such as performance, cost, and reliability becomes essential for software companies. The cost is the most fundamental element to make a software project manageable. With the precise cost estimation, companies might evaluate the project progress and analyze what is effective for the project and get better decisions during software life cycle. It provides a reliable budget and delivery of the software product within the promised period of time.

In the literature, there are a number of models for estimating the software cost. These models are developed using algorithmic approaches, expert judgments, analogy, top-down and bottom-up strategies and machine learning techniques. Algorithmic models, one of the successful and simple software cost estimation techniques, use mathematical formulas that are based on the effort in terms of person-months at various life cycles of the software for a project [1]. In this study, we implement enhanced artificial bee colony method proposed in [2] to establish an algorithmic model for estimating software cost.

The organization of this paper is as follows: Section II explains some algorithmic models, the methodology that has been used is detailed in Section III, in Section IV, the evaluation and results of our model are shown, Section V discusses the conclusions.

II. Background

There are various major models to estimate software cost in the literature [3]-[5]. These models are based on a large number of software projects and applications completed in various organizations. In the formulations of these models, the software size is based on KLOC (Kilo Lines Of Code) that is the number of lines of source codes. The structure of the formulation is as follows:

$$\mathrm{Effort} = a(\mathrm{KLOC})^{b} \qquad (1)$$

It is thought that the software effort can be found by multiplying a power of the number of lines of the software code by a coefficient. Values a and b are constants obtained experimentally. The proposed models in the literature aim to find these constants, a and b, mathematically by using some software projects.

Some of these models are:

- Walston-Felix Model [6]: This model was proposed by Waltson and Felix in 1977. The model was developed from a database of 60 different projects and provides a correlation between effort and source code lines. This formulation can be expressed as:

$$\mathrm{Effort} = 5.2 * (\mathrm{KLOC})^{0.91} \qquad (2)$$

- Doty Model: In the same year with Walston-Felix Model, Doty presents the model called with his name [7], to estimate the effort for the number of lines of code as in the following equation:

$$\mathrm{Effort} = 5.288 * (\mathrm{KLOC})^{1.047} \qquad (3)$$

- Halstead Model [8]: This model, which predicts the error rate, was proposed in 1977. In this model, in-depth analysis is not required for the programming structure. It provides to recommend the code length and volume metrics of the software. The formulation of this model is:

$$\mathrm{Effort} = 5.2 * (\mathrm{KLOC})^{1.5} \qquad (4)$$

- Bailey-Basili Model [9]: Bailey and Basili proposed a meta-model in 1981 to calculate the software effort prediction equations that best fits the given development environment. The resulting model is based on the collection of data such as differences between projects and their environmental factors. This equation is expressed as follows:

$$\text{Effort} = 0.73 * (\text{KLOC})^{1.16} + 3.5 \quad (5)$$

- Artificial neural networks (ANN) are designed to model the way in which human brain performs. ANNs consist of simple interconnected units and are usually organized as layers. Similar to the information processing method of the brain, after a learning process, ANN has the ability to collect information and the ability to store and generalize this information with the weight of connections between cells. The learning process involves learning algorithms that enable the renewal of ANN weights to achieve the desired goal. Finnie et al. [10] used ANN method to learn parameters in effort estimation of software development in 1977. The prediction model was based on 50 sample cases.

- Genetic algorithm (GA): Genetic algorithm that is inspired by the evolutionary process, is a method used to solve optimization problems. Authors in [11] and [12] use genetic algorithms to achieve improvement on effort estimation.

- ABC algorithm model: This algorithm was developed by Karaboga et al. [13] inspired by the behaviors of biological swarm intelligence. The intelligence in foraging behaviors of honeybees is leveraged to find the optimal solutions for problems. The principal of the algorithm is based on the behaviors of honeybees on finding nectar-rich food sources and information sharing about food sources [14]. The goal of foraging process is to minimize the consumed energy and time while finding nectar-rich sources. The ABC algorithm is very simple and robust optimization algorithm. For cost estimation of a software project, it aims to learn parameters of (1).

- H-ABC [15]: The ABC algorithm can be initialized with a distribution that is either uniform or non-uniform according to the location of the optimum. In this model, Halton points are chosen among the initial points that are generated by ABC algorithm and this model is applied to optimize the parameters given in (1).

III. Methodology

The advantages of ABC algorithm are simplicity, be applicable to many different areas and its outstanding performance. On the other hand, it suffers from slow convergence and tendency to local optima [16]. Researchers have been proposed to overcome these drawbacks [2], [16]-[20]. Abro et al. [2] proposed a novel variant of basic ABC algorithm which is called Enhanced-ABC algorithm (E-ABC). The E-ABC algorithm uses multiple global-best possible-solutions (GBPS) rather than single global-best possible-solution in exploration and exploitation phase. It also has a novel procedure in the scout bee phase to improve the performance.

The bee swarm in basic ABC algorithm consists of three groups to accomplish different tasks: employed bees, onlookers, and scouts. The employed and onlooker bees indicate the number of possible solutions in the population. A scout bee is formed with the transformation of the employed bee when a food source exhausted totally [13]. A scout flies around the hive and produces a food source randomly. In the modeling of the optimization problem, the search space is represented by the hive, each possible solution is considered as a food source and the solution quality is shown by the nectar amount of these food sources.

An employed bee investigates near food sources around the hive. After she finds a food source, she tries to explore the multiple best-found locations of search space instead of the neighborhood of an assigned food source like in basic ABC algorithm.

$$s_{kl} = f_{kl} + \theta_{kl}(f_{kl} - f_{best\,/\,secondBest/\ldots/n,l}) \quad (6)$$

where s_{kl} is the produced new source from f_{kl} which represents the old food source, θ_{kl} is randomly selected number between (-1,1) [2]. $f_{best\,/\,secondBest/\ldots/n,l}$ represents one of the high-nectar-content food sources.

The fitness value of new-found food source (s_{kl}) is evaluated using the following equation:

$$fit_k = \begin{cases} \frac{1}{1+\text{o}_k}, & \text{o}_k \geq 0 \\ 1 + abs(o_k), & \text{o}_k < 0 \end{cases} \quad (7)$$

where o_k represents objective function value of k_th food source.

If the nectar amount of newly found solution is higher, then the bee learns the new one and forgets the previous solution. Otherwise, she continues with the previous position.

All population is divided into different groups, and the number of groups is determined by a user-defined value. The Global best possible solutions (GBPSs) are selected to assign each one of them to a group, regardless of its existence [2]. The aim of this task is to improve the fitness of possible solutions.

In exploitation process, employed bees share information about food sources by dancing. Each onlooker bee chooses a high-quality food source based on the observation of the dance. The quality of the food source is decided by the probability value (p_k):

$$\text{p}_\text{k} = \frac{fit_k}{\sum_{i=1}^{SN} fit_i} \quad (8)$$

where fit_k is the fitness value (nectar amount of food source) in k_{th} position. If an onlooker bee does not find a better solution than old one, she produces a new solution using the following equation:

$$s_{kl} = f_{kl} + \theta_{kl}(f_{kl} - f_{best,l}) \quad (9)$$

This formula is intended to improve exploitation process and runs if (1) failed to generate better solution than the existing solution and Selection (S), an additional control variable of E-ABC algorithm, is larger than a randomly generated number. Authors in [2] stated that higher value of S accelerates the speed of convergence.

In E-ABC algorithm, a scout bee calculates new food sources, s_{nl}, according to the formula below:

$$s_{nl} = (f_{best,l})\beta_{nl} \tag{10}$$

where $f_{best,l}$ represents one of the high-nectar-content sources. β_{nl} is a random number ranging between 0.9 and 1.1 [2].

There are a few critical parameters to be specified in E-ABC algorithm:

1) SN: the size of the population.
2) Limit value: iteration number to quit processing on a food source.
3) MIN: Maximum iteration number.
4) S value: a decision number to search a new food source.
5) n value: the number of groups in population to be assigned GBPSs.

The values are assigned as shown in Table I.

TABLE I. THE TUNING PARAMETERS FOR THE E-ABC

Parameters	**Values**
SN	100
Limit	20
MIN	750
Selection (S)	0.5
Number of groups (n)	5

B. Results

The performance of the predicted model is measured by two mainly used metrics, such as Mean Magnitude Relative Error (MMRE) and Root Mean Square Error (RMSE).

$$MMRE = \frac{1}{n}\sum_{i=1}^{n} \frac{|Effort - Estimated\ Effort|}{Effort} \tag{11}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Effort - Estimated\ Effort)^2} \tag{12}$$

IV. RESULTS AND DISCUSSIONS

This section explains the details of the used dataset and obtained results over this dataset using some common metrics.

A. Dataset

The proposed estimation model using E-ABC method in [2] is conducted on a NASA dataset [9] that contains 18 projects consisting of Kilo Line of code (KLOC), Methodology (ME) and the measured effort. The Effort is in terms of person-months. Table II shows the details of this dataset.

To estimate model parameters, the dataset is split into training set that consists of the first 13 projects, and test set that contains the rest of the projects in the dataset. The computed efforts obtained by E-ABC are shown in Table II.

TABLE II. NASA DATA OF EFFORT ESTIMATION

Project No	**KLOC**	**Methodology**	**Measured Effort**
1	90.2	30	115.80
2	46.2	20	96.00
3	46.5	19	79.00
4	54.5	20	90.80
5	31.1	35	39.60
6	67.5	29	98.40
7	12.8	26	18.90
8	10.5	34	10.30
9	21.5	31	28.50
10	3.1	26	7.00
11	4.2	19	9.00
12	7.8	31	7.30
13	2.1	28	5.00
14	5	29	8.40
15	78.6	35	98.70
16	9.7	27	15.60
17	12.5	27	23.90
18	100.8	27	138.30

Table III shows the estimated results by E-ABC algorithm are mostly close to the measured effort except for some projects, such as project 2. Some models known as benchmarks, i.e. Halstead, Doty etc., and previous studies using NN and other evolutionary algorithms, namely GA, H-ABC, are used to compare the performance of E-ABC model in terms of MMRE and RMSE. The results of the model produced by basic ABC algorithm are also provided. As shown in Table IV, basic ABC algorithm outperforms Halstead, Walston-Felix, Doty, GA, and H-ABC in MMRE metric, but not NN. It also gives lower RMSE than all models except GA., E-ABC yields better performance than all other models. It provides 22%, 78%, 81% gains in MMRE metric, and 70%, 54%, 64% gains in RMSE metric according to NN, GA, and H-ABC, respectively.

TABLE III. ESTIMATED EFFORT CALCULATED BY E-ABC

Project No	**Measured Effort**	**E-ABC Estimated Effort**
1	115,8	115,8219
2	96	62,1928
3	79	62,568
4	90,8	72,5151
5	39,6	43,052
6	98,4	88,466
7	18,9	18,8653
8	10,3	15,6933
9	28,5	30,5485
10	7	5,05
11	9	6,6968
12	7,3	11,9051
13	5	3,5164
14	8,4	7,8749
15	98,7	101,9125
16	15,6	14,579
17	23,9	18,454
18	138,3	128,4215

TABLE IV. THE PERFORMANCE RESULTS OF EFFORT ESTIMATION

Performance Criteria	Model Used								
	NN System	**Halstead**	**Walston-Felix**	**Bailey Basili**	**Doty**	**GA Estimated Effort**	**H-ABC Estimated Effort**	**ABC Estimated Effort**	**E-ABC Estimated Effort**
MMRE	11,7896	175,6550	155,5590	20,2885	302,5020	41,9072	49,0565	23,62405	9,1961
RMSE	17,4475	308,7097	123,4570	25,0224	299,4740	11,4867	14,6136	13,50596	5,2703

V. CONCLUSION

In this study, we propose a novel model to estimate the software effort. The E-ABC algorithm is used as it is faster convergence than basic ABC algorithm. With E-ABC algorithm, an algorithmic effort estimation model based on the number of lines of software has been developed. We tested our model on the NASA software projects dataset. The proposed model is compared with the algorithmic prediction models in the literature using MMRE and RMSE metrics. Results show that our model achieves better results than the previous models. The limitation of this study is that the size of the used NASA dataset is small. As future work, ABC and E-ABC can be employed to model software cost estimation over different datasets.

REFERENCES

[1] Reddy, Ch Satyananda, and Raju KVSVN. “An optimal neural network model for software effort estimation.”, Int. J. of Software Engineering, IJSE 3.1 (2010): 63-78.

[2] Abro, Abdul Ghani, and JUNITA MOHAMAD SALEH. “Multiple-global-best guided artificial bee colony algorithm for induction motor parameter estimation.”, Turkish Journal of Electrical Engineering & Computer Sciences 22.3 (2014): 620-636.

[3] Kumari, Sweta, and Shashank Pushkar. “Comparison and analysis of different software cost estimation methods.”, IJACSA International Journal of Advanced Computer Science and application 4.1 (2013).

[4] Olu-Ajayi, Razak. “An Investigation into the Suitability of k-Nearest Neighbour (k-NN) for Software Effort Estimation.”, International Journal Of Advanced Computer Science And Applications 8.6 (2017): 227-233.

[5] PVGD, Prasad Reddy. “Application of Fuzzy Logic Approach to Software Effort Estimation. ”, IJACSA Editorial (2011).

[6] Walston, Claude E., and Charles P. Felix. “A method of programming measurement and estimation”, IBM Systems Journal 16.1 (1977): 54-73.

[7] Doty Associates, Inc., “Software Cost Estimates Study,” vol. 1, (1977): 77-220.

[8] M. H. Halstead, “Elements of Software Science,” Elsevier, New York, (1977).

[9] Bailey, John W., and Victor R. Basili. “A meta-model for software development resource expenditures.”, Proceedings of the 5th international conference on Software engineering. IEEE Press, 1981.

[10] Finnie, Gavin R., Gerhard E. Wittig, and Jean-Marc Desharnais. “A comparison of software effort estimation techniques: using function points with neural networks, case-based reasoning and regression models. ”, Journal of systems and software 39.3 (1997): 281-289.

[11] Burgess, Colin J., and Martin Lefley. “Can genetic programming improve software effort estimation? A comparative evaluation.”, Information and software technology 43.14 (2001): 863-873.

[12] Sheta, Alaa F. “Estimation of the COCOMO model parameters using genetic algorithms for NASA software projects.”, Journal of Computer Science 2.2 (2006): 118-123.

[13] Karaboga, Dervis, and Bahriye Basturk. “A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. ”, Journal of global optimization 39.3 (2007): 459-471.

[14] Yigit, Sevgi, Recep Eryigit, and Fatih V. Celebi. “Optical gain model proposed with the use of artificial neural networks optimised by artificial bee colony algorithm. ”, Optoelectron Adv Mat 5 (2011): 1026-1029.

[15] Sharma, Tarun Kumar, and Millie Pant. “Halton based initial distribution in artificial bee colony algorithm and its application in software effort estimation.”, Bio-Inspired Computing: Theories and Applications (BIC-TA), 2011 Sixth International Conference on. IEEE, 2011.

[16] Zhu, Guopu, and Sam Kwong. “Gbest-guided artificial bee colony algorithm for numerical function optimization.”, Applied mathematics and computation 217.7 (2010): 3166-3173.

[17] Karaboga, Dervis, and Beyza Gorkemli. “A quick artificial bee colony (qABC) algorithm and its performance on optimization problems.”, Applied Soft Computing 23 (2014): 227-238.

[18] Akbari, Reza, Alireza Mohammadi, and Koorush Ziarati. “A novel bee swarm optimization algorithm for numerical function optimization.”, Communications in Nonlinear Science and Numerical Simulation 15.10 (2010): 3142-3155.

[19] Gao, Wei-feng, San-yang Liu, and Ling-ling Huang. “A novel artificial bee colony algorithm based on modified search equation and orthogonal learning. ”, IEEE Transactions on Cybernetics43.3 (2013): 1011-1024.

[20] Banharnsakun, Anan, Tiranee Achalakul, and Booncharoen Sirinaovakul. “The best-so-far selection in artificial bee colony algorithm. ”, Applied Soft Computing 11.2 (2011): 2888-2901.

Using Creative Educational Drama to Enhance Self-Development Skills for the Students at University Level

Hisham Saad Zaghloul
Department of Self Development Skills, Deanship of Preparatory Year & Supportive Studies,
Northern Border University, Saudi Arabia
Department of Educational Media, Faculty of Specific Education, Mansoura University, Egypt

***Abstract*—This study has been undertaken using a creative drama in teaching in order to improve the communication and thinking skills for the students of preparatory year students at the Northern Border University. It aimed to measure the differences between the experimental and control groups in skills acquisition among students. The study was conducted on 140 students from both genders (males=70 and females=70). The students were divided into four groups: each has 35 students. The study adopted an experimental approach by observing students' behavior through affecting their communication and thinking skills with using drama. The findings confirmed that using drama in teaching significantly affected on the experimental group than the students of the control group who were taught with traditional methods as the experimental group achieved better results than the control group. Furthermore, the study stressed the possibility of benefiting from the drama in teaching other practical courses at the university level and provided several recommendations in this regard as well.**

Keywords—Creative drama; creative educational drama; self-development skills; communication skills; thinking skills

I. Introduction

Creative drama is one of the most effective tools for improving communication and thinking skills. The use of creative drama technique in the classroom is a student-centered, in which learning process results in the development of the students of any curriculum. Creative drama method of teaching helps learners develop and improve their divergent thinking skills, creativity, communication and it matures their oral and written communication skills. The use of creative drama enriches the students' imagination and willingness to act or pretend as a means of reinforcing academic, emotional, and interpersonal objectives [1]. In fact, creative drama shows the learners the way to be appreciated and understand about others' the needs which will make the student themselves able to form a value judgment. Drama activities can reflect positively on the student's behavior and his personality as well. Many previous studies have confirmed the effectiveness, creative drama as an effective method of teaching in the pre-university educational stages and several courses. The approach of creative drama leads students to discover their inner strengths of "knowing" into concrete action.

Further, according to the theorists, teachers can use creative dramatics to support social, academic, personal and interpersonal goals. This emphasizes the use of learners' imaginative skill and intelligence to help them learn through their own activity and practices because learners process information differently from the highly matured students. The concept of creative dramatics also supports the growth of their language and vocabulary while stimulating high-level cognitive processes.

Therefore, it is worth saying that using creative drama for educational purposes has been the purpose of many studies that have undertaken with interactive drama techniques to facilitate the following: curriculum, language acquisition, and skills acquisition. Lastly, most of these studies have focused on pre-college education stage. Classroom exercises based on creative drama can provide a genuine experience and also fun in the classroom and guide learners to become aware of the use of their imagination. Therefore, this study will investigate the effectiveness of using drama and its impact on the learning and teaching process of self-development courses including communication and thinking skills.

II. Literature Review

Multiple studies have been conducted investigating the use of creative drama as a teaching method. The objective of this study is to investigate the degree to which improvement in communication and thinking skills is effective when using drama pedagogically. Author in [2] classifies thinking skills as crucial self-development skills. It is through thinking that students interpret and process the information they receive. Master the skill of collecting relevant information and formulate conclusions as a basis for decision making. Author in [3] states that exposing students to creative drama provokes a change in their thought patterns, enhancing their curiosity, driving them towards new discoveries, and further learning opportunities. Moreover, Çokadar and Yılmaz [4] investigated the use of creative drama and found that the participation, interaction, and harmony which emerges as part of the creative drama process, greatly enhance science students' understanding of environmental concepts, positively direct them to develop their learning. According to [5], the creative drama is an effective teaching method for use when teaching students at all stages of formal education. Indeed, reflecting on the potential value of creative drama, Kay [6] recommended that universities increase the number of training sessions in how to use creative drama in the classroom, introducing it as a fun and student-centered method to adopt when teaching.

Bayraktar and Okvuran [7] investigated the utility of creative drama and indicated that it motivates students to develop self-reflection and creative thinking skills, which can be developed alongside their writing skills, as part of their basic education. Espousing an alternative perspective, Partab [8] suggested that the primary value of drama in education is that it provides entertainment and psychological support, reducing student's anxiety. Moreover, he claimed that creative drama can develop community awareness, providing greater comprehension of learners' problems.

Similarly examining the social aspect of using drama in the classroom, Aydeniz, and Ozcelik [9] reported that the use of drama enhances students' communication skills, benefitting group discussions and eliminating shyness. Moreover, they noted that drama creates an opportunity for students to express themselves clearly, consequently improving their level of academic achievement. Aykac's [10] study further supported the view that creative drama improves students' ability to express themselves, observing that it can enhance their self-confidence by bestowing a capacity to communicate effectively, and thus creates an inner satisfaction within individuals.

The most recent study was that conducted by [11]. This study discovered that drama activities could positively influence pupils' skills acquisition; in particular fueling self-expression and self-confidence. Moreover, creative drama can be utilized to teach languages, mathematics, social studies, and music, as well as to develop sensory and motor skills.

III. Research Questions

The present study seeks to answer the following main question:

- What is the effectiveness of an educational program based on creative drama to develop communication and thinking skills among male and female students of the preparatory year at the Northern Border University?
- The following sub-questions are derived from the main question:
    - Is there a significant effect from "gender" type of communication and thinking skills?
    - Is there a significant effect from "teaching Method (Traditional method & Creative drama)" on communication and thinking skills?
    - Is there any interaction between "gender" and "teaching Method (Traditional method & Creative drama)" and the level of communication and thinking skills?

IV. Significance of the Study

The importance of the study lies in the fact that it is an attempt to:

- It draws the attention to the effectiveness of using drama in developing student's communication and thinking skills.
- It may provide teachers with applicable teaching situations using creative drama.
- It may provide results, which may be applicable to other topics.
- It may provide a clear impression of thinking and communication skills course designed to include drama with these courses as an effective method.

V. Objectives of the Study

The study seeks to explore the following objectives:

- To Design an educational program for creative drama (trainer's guide) in the field of communication and thinking skills for students in the preparatory year at Northern Border University.
- To measure the effectiveness of the creative drama program proposed in the development of communication and thinking skills of the students.
- To discover if there is any significant difference between the communication and thinking skills between the male and female students.

VI. Dimensions of the Study

- Academic dimension

This study focused on the effectiveness of drama as an educational method compared to the traditional method of teaching both communication and thinking skills.

- Human dimension

The study is concerned with students studying in the preparatory year at Northern Borders University, Arar, Saudi Arabia.

- Time dimension

The study was conducted within the first term of the academic year for 2017 -2018.

- Place dimension

This study was carried out in the preparatory year of Northern Borders University targeting its students from both genders.

VII. Definition of Terms used in the Study

It is important to define and explain the basic terms in any study, especially when there is difficulty in understanding these terms or when they carry more than one meaning. Thus, these terms are defined below:

- *Creative Drama:* The innovative positions that the teacher of the course to participate with the learner in the classroom teaching and rely on innovative attitudes and the realization of the mind and open the door to debate and dialogue and raise questions that enrich the practical aspect of the courses, which the study codified to serve the practical objectives of communication and thinking skills.

- *Educational Creative Drama:* It is the main study instrument and teacher's guide to transforming the scientific material of the communication and thinking skills of dramatic acting situations between the course instructor and students and between students and themselves and their colleagues as well in order to transform the theoretical material into practical characteristics of education.
- *Self-development skills:* These are the skills required for students of the preparatory year in their first years of university and summed up in communication and thinking skills.
- *Communication Skills:* A general course for students of the preparatory year, aim at activating the human communication of the student inside and outside the university community to get him a good citizen who benefits himself and his family and society.
- *Thinking Skills:* A compulsory course taught to the students of the preparatory year, aims to support them with the positive thinking patterns, accept others, build ideas and to make decisions, and deal with a range of other skills through brainstorming sessions and creative thinking based on student participation in the classroom.

VIII. Methodology

A. Design of the Study

A research design aims to provide a framework for data collection and analysis procedures [12]. The choice of a research design is "based on the nature of the research problem or issue being addressed, the researchers' personal experiences, and the audiences for the study" [13]. Thus, the present study employed the quasi-experimental method to investigate the effectiveness of self-development course using drama among preparatory year male and female students on their thinking and communication skills.

B. Instrument

For the purpose of the study, a structured observation was used. The observation strategy considered as a valuable strategy because it provides evidence of what happens in the classroom [14]. Moreover, Observation can provide a researcher with a rich description of the situation under investigation [15]. Thus, observation technique used in this study in order to assess any changes may occur in the behavior of the students in the experimental group during a specified period of studying the self-development courses (communication skills and thinking skills). The observation form was divided into two main sections: communication skills and thinking skills. Each of this skill was divided into three sub-sections. The course instructor of each group assesses the changes, which may occur in the period of experiment and record that for each student of his/her group. All the assessments of the observation were recorded on a five-point Likert scale, offering five choices: 1) poor; 2) acceptable; 3) good; 4) very good; and 5) excellent (see Appendix 1).

C. Procedures

The researcher adopted the following procedures:

- Review of related literature and related previous studies.
- Several dramatic teaching scenes were conducted in the light of using drama for each lesson in both courses (communication skills and thinking skills) for about five weeks.
- Research instrument (only observation card) was adopted upon a review of the literature and related studies, which paved the way for the researcher to greatly benefits from these studies to construct observation card content to suit the study objectives.
- Validity and reliability of the instrument have been measured.
- Assigning sample of the study and apply the pre-test of using drama for both control and experimental groups.
- The control group was taught using the traditional way of teaching, whereas the experimental group was taught using drama.
- The post-test of using drama was supplied by the end of teaching the assigned unit.
- Data were statistically analyzed using the software SPSS to discover if there are any significant differences.
- Results have been discovered and discussed.
- Based on the results, the recommendations and suggestions have been given.

D. Data Analysis

The experimental study data was analyzed by using SPSS program, version 22 to answer the research questions. The researcher used mainly analytical tests in the study. The first one was a t-test in order to determine the significant differences between the study variables and the impact of creative drama on the student's communication and thinking skills. Additionally, the t-test has been used to discover the results of the paired sample test for the experimental group in the case of Pre-test and Post-test in communication skills, thinking skills and their dimensions on the students. The study referred to significant differences based on the p-value. Moreover, the researcher used two-way ANOVA in order to show interaction effects between the selected variables and the communication and thinking skills.

IX. Results and Discussion

This study proved that using creative drama in the instruction of certain developing learning skills affect the achievement level of self-department skills students in a positive way. The results were clear through changes that occurred in the student's behavior in the classroom while learning.

TABLE I. DESCRIPTIVE STATISTICS OF RESEARCH GROUP'S SCORES IN COMMUNICATION SKILLS AND THINKING SKILLS

Statistical variable					**N**	**M**	**SD**	**Variance**
Variables	**Tools**	**Respondent**	**Group**	**Test**				
Communication skills	**Observation Checklist (assessment of self-development skills using creative drama)**	*Male*	*Experimental Group*	*Pre-test*	35	23.1429	2.11636	4.479
				Post-test	35	46.8000	1.99706	3.988
			Control Group	*Pre-test*	35	21.2857	2.38342	5.681
				Post-test	35	23.4286	2.47678	6.134
		Female	*Experimental Group*	*Pre-test*	35	46.3429	1.86205	3.467
				Post-test	35	46.8000	1.89116	3.576
			Control Group	*Pre-test*	35	23.8286	1.90179	3.617
				Post-test	35	22.9714	2.51450	6.323
Thinking skills		*Male*	*Experimental Group*	*Pre-test*	35	22.9429	2.31292	5.350
				Post-test	35	46.6286	1.73351	3.005
			Control Group	*Pre-test*	35	22.9429	2.31292	5.350
				Post-test	35	23.1714	2.38236	5.676
		Female	*Experimental Group*	*Pre-test*	35	22.6000	2.30345	5.306
				Post-test	35	22.9714	2.60639	6.793
			Control Group	*Pre-test*	35	23.4000	3.77453	14.247
				Post-test	35	24.8571	2.27703	5.185

According to Table I, there is a statistically meaningful difference between the two groups: control groups and experimental groups, as was clear from the score averages for the t-test, which shows a significant difference between the two groups. Moreover, a significant difference between males and females was apparent in every group. The statistics showed higher post-test values after the experiment, between both the male and female groups; these results were found after the majority of sessions of the experiment. These exceptional values, as apparent from Table I, show the positive influence of creative drama on both communication skills and thinking, confirming the value of using creative drama sessions to improve students' self-development.

The significant values obtained included the pre and post-test scores for the experimental group and encompassed all variables. Benefits of using drama as a teaching approach were apparent for both genders; the variance value for males was (5.681) before using creative drama and (6.134) after. Similarly, improvements were noted for females; before using the dramatic method, their skills acquisition was (3.617) and after it rose to (6.323). Moreover, the study confirmed that verbal communication skills were the most significantly improved skills; although non-verbal skills, thinking skills, problem-solving skills and brainstorming skills were also enhanced.

These significant values included the pre and post-test scores for the experimental group. These values illustrated statically in all research based on variables. These significant values showed significant differences in Verbal Comm. Skills (t = 8.168; p =. 014), Non- Verbal Communication (t = 11.065; p = 0.000), Electronic Communication (t = -11.469; p =. 055), Total Communication Skills (t = 11.233; p =. 000), Critical Thinking Skills (t = 11.374; p = 0.000), Problem Solving Skills (t = 13.729; p =. 000), Brainstorming Skills (t = 9.655; p =. 000), Total thinking skills (t = 12.141; p =. 000), this result reflects the effectiveness of creative drama on all research variables, see Table II. The result of the study regarding its significant impact on the learning process and development of the student's characteristics and behaviors accorded with all studies stated in the literature review of this study. For example, Aydeniz, & Ozcelik, [9], Aykac [10], Sengul [11] which confirmed the constructive role of using suing drama in a learning process which agreed with this study results.

TABLE II. RESULTS OF PAIRED SAMPLES (T) TEST FOR EXPERIMENTAL GROUP PRE-TEST & POST-TEST FOR TESTING THE IMPACT OF CREATIVE DRAMA IN COMMUNICATION & THINKING SKILLS

Variables (Measurements)		Paired Differences					t	DF	Sig.
		Mean	SD	Std. Error	95% Confidence Lower	95% Confidence Upper			
Pair 1	**Verbal Comm. Skills Pre-test Verbal Comm. Skills Post-test**	2.77857	4.02522	.34019	2.10595	3.45119	8.168	139	.000
Pair 2	**Non- Verbal Comm. Skills Pre-test Nonverbal Comm. Skills Post-test**	3.51429	3.75794	.31760	2.88633	4.14224	11.065	139	.000
Pair 3	**Electronic Comm. Skills Pre-test Electronic Comm. Skills Post-test**	5.99286	6.18247	.52251	4.95975	7.02596	11.469	139	.000
Pair 4	**Critical Thinking Skills Pre-test Critical Thinking Skills Post-test**	3.36429	3.49966	.29578	2.77949	3.94909	11.374	139	.000
Pair 5	**Problem Solving Skills Pre-test Problem Solving Skills Post-test**	3.78571	3.26258	.27574	3.24053	4.33090	13.729	139	.000
Pair 6	**Brainstorming Skills Pre-test Brainstorming Skills Post-test**	4.63571	5.68123	.48015	3.68637	5.58506	9.655	139	.000
Pair 7	**Total communication Skills Pre-test Total communication Skills Post-test**	1.22857E1	12.9405	1.09367	10.1233	14.4481	11.233	139	.000
Pair 8	**Total thinking skills Pre-test Total thinking skills Post-test**	1.17857E1	11.4858	.97073	9.86640	13.7050	12.141	139	.000

Moreover, as in Table II, it is noteworthy that the study shed light on the impact of creative drama on the study variables: gender, teaching method (whether traditional or using creative drama), and the interaction between gender and the teaching method. Here, we used the probability value, statistically termed the p-value, to discover significant differences between communication skills and thinking skills, after using drama as a teaching method. The study demonstrated that the gender has no significant effect on communication skills, as the p-value for gender (male and female) is (0.547). Meanwhile, an effect of gender emerged in relation to thinking skills, where the p-value for gender was (0.049). However, this effect is considered small; because the partial eta squared, the value is (0.028). However, the figures for the type of teaching method were clearly different; the results confirmed that teaching method has a clear impact on both communication (sig. 000) and thinking skills (sig. 000) after applying an ANOVA analysis for this variance. Finally, the study concluded by noting a moderate interaction between gender and teaching method and thinking skills and communication skills; this was apparent from the p-values for both skills, which were (547) and (0.006) respectively. For more statistical details and p values in both situations, see Tables III and IV below.

TABLE III. RESULTS OF 2WAY ANOVA TEST FOR THE IMPACT OF CREATIVE DRAMA ON TOTAL (COMMUNICATION SKILLS) OF THE EXPERIMENTAL GROUP

Tests of Between-Subjects Effects							
Dependent Variable: Total (Communication Skills) Post-test							
Source	**Type III Sum of Squares**	**Def**	**Mean Square**	**F**	**Sig.**	**Partial Eta Squared**	**Effect size**
Corrected Model	19497.257a	3	6499.086	1.298E3	.000	.966	Large
Intercept	171500.000	1	171500.00	3.426E4	.000	.996	Large
Gender (male& female)	1.829	1	1.829	.365	.547	.003	Small
Method (Traditional method & Creative drama)	19493.600	1	19493.600	3.894E3	.000	.966	Large
Gender * method	1.829	1	1.829	.365	.547	.003	Small
Error	680.743	136	5.005				
Total	191678.000	140					
Corrected Total	20178.000	139					
a. R Squared =. 966 (Adjusted R Squared =. 966)							

TABLE IV. RESULTS OF 2WAY ANOVA TEST FOR THE IMPACT OF CREATIVE DRAMA ON TOTAL (THINKING SKILLS) OF THE EXPERIMENTAL GROUP

Tests of Between-Subjects Effects							
Dependent Variable: Total (Thinking Skills) Post-test							
Source	**Type III Sum of Squares**	**Def**	**Mean Square**	**F**	**Sig.**	**Partial Eta Squared**	**Effect size**
Corrected Model	17724.936a	3	5908.312	1.364E3	.000	.968	Large
Intercept	173958.750	1	173958.750	4.015E4	.000	.997	Large
Gender (male& female)	17.150	1	17.150	3.958	.049	.028	Small
Method (Traditional method & Creative drama)	17673.779	1	17673.779	4.079E3	.000	.968	Large
Gender * method	34.007	1	34.007	7.848	.006	.055	Small
Error	589.314	136	4.333				
Total	192273.000	140					
Corrected Total	18314.250	139					
a. R Squared =. 968 (Adjusted R Squared =. 967)							

X. CONCLUSION

The purpose of carrying out this study is to investigate the effectiveness of the drama on the students' self- development, which focused on the effect of drama on both communication and thinking skills among Saudi students studying at the preparatory level in Northern Borders University. The target population of this study was targeting both genders (males and females) in the deanship of preparatory year and supportive studies of Northern Borders University for the academic year, 2017-2018. The present study discovered a number of findings in this context. The study showed that using drama in the educational situation plays a constructive role in the learning process through improving the student's skills. Besides, the significant effect was clearer of the experimental group than the students of the control group who were taught with traditional methods. Moreover, the study also showed the effectiveness of creative drama on the self-development of the students, which were clear on the impact of communication skills and thinking skills of the students in their courses. The high values of post-test for both of male and female groups in the experimental group confirmed the fundamental role of using drama developing the learning process and students' behavior.

- Limitations of the present study

This study, like all studies, has limitations. The study conducted only on one academic institute namely, the Northern Borders University in Saudi Arabia, where the researcher works as lecturer and head of self-development department in it. Consequently, this may make the study's scope is limited and its results cannot be generalized to other students in other Saudi academic institutions. Furthermore, the study did not employ more research methods for data collection such as interviews, questionnaire, etc., because the students were not happy to participate in the study includes more research methods, due to the pressure they face it in their study. Moreover, there is no more time for the researcher for designing other research methods. Additionally, there is no easy access to the female students to make interviews with them due to the segregation systems between males and females in Saudi universities.

XI. RECOMMENDATIONS AND SUGGESTIONS

Based on the findings of the study, the following recommendations and suggestions are made:

- Creative drama courses should be taught as main part of the curriculum of self-development skills to be a more effective teaching tool.
- Creative drama method should be progressively used as a teaching method in other courses as well.
- Longitudinal studies should be conducted to identify the effect of creative drama in teaching and learning communication and thinking skills
- Other studies need to investigate the changes, which occur for students, who are taught by the teachers that have increased communication and thinking skills due to being introduced to creative drama.
- More qualitative studies are recommended to be conducted to investigate the effects of drama education in more depth as well.

ACKNOWLEDGMENT

The author gratefully acknowledges the approval and the support of this research study by the grant no. (EAR-2017-4-7-F-6931) from the Deanship of Scientific Research at Northern Border University, Arar, K.S.A.

REFERENCES

[1] Froese, V. (1996). *Whole-language: Practice and Theory* (2nd Ed.). Needham, MA: Allyn and Bacon.

[2] Bayley, R. (2002). Thinking skills in the early years. *Gifted Education International*, 16(3), 248-260. doi:10.4324/9780203410851

[3] Karakelle, S. (2009). Enhancing fluent and flexible thinking through the creative drama process. *Thinking Skills and Creativity*, 4 (2), 124-129. doi:10.1016/j.tsc.2009.05.002

[4] Çokadar, H., & Yılmaz, G. C. (2010). Teaching Ecosystems and Matter Cycles with Creative Drama Activities. *Journal of Science Education and Technology*, 19 (1), 80-89. doi:10.1007/s10956-009-9181-3

[5] Adıgüzel, H. Ö, & Timuçin, E. (2010). The effect of creative drama on student achievement in the instruction of some development and learning theories. Procedia - *Social and Behavioral Sciences*, 9, 1741-1746. doi:10.1016/j.sbspro.2010.12.393

[6] Kay, N. B. (2010). Examination on self-efficacy of the pre-service teachers of the classroom teacher and pre-school teacher according to creative drama method. Procedia - *Social and Behavioral Sciences*, 2(2), 4533-4539. doi:10.1016/j.sbspro.2010.03.726

[7] Bayraktar, A., & Okvuran, A. (2012). Improving Students Writing Through Creative Drama. Procedia - *Social and Behavioral Sciences*, 51, 662-665. doi:10.1016/j.sbspro.2012.08.220

[8] Partab, V. (2012). Creative Healing-An Examination of the Value of Creative Drama in Helping South African Children Deal with the Traumas of Their Lives. Procedia - *Social and Behavioral Sciences*, 69, 487-491. doi:10.1016/j.sbspro.2012.11.437

[9] Aydeniz, H., & Ozcelik, N. (2012). Impact of Creative Drama Method on the French Achievement of University Students. Procedia - *Social and Behavioral Sciences*, 47, 962-967. doi:10.1016/j.sbspro.2012.06.764

[10] Aykac, M. (2013). The effect of creative drama activities on the personal development of housewives. *Educational Research and Reviews*, 8 (21), 2130.

[11] Sengul, O. A. (2016). The Views of the Pre-service Teachers about the Creative Drama as a Method Used in Primary Schools. *Universal Journal of Educational Research*, 4 (5), 1244-1251. doi:10.13189/ujer.2016.040539

[12] Bryman, A. (2008). *Social research methods*. (3rd ed.). Oxford: Oxford University Press.

[13] Creswell, J. W. (2009). *Research design: qualitative, quantitative and mixed methods approach.* 3rd ed. London: Sage.

[14] Borg, S. (2006). *Teacher Cognition and Language Education: Research and Practice.* London: Continuum.

[15] Cohen, L., Manion, L. & Morrison, K. (2007). *Research Methods in Education.* 6th ed. London: Routledge.

APPENDIX 1

SELF-DEVELOPMENT SKILLS EVALUATION OBSERVATION FORM

- Name of the Rapporteur (optional): ....
- Course (Communication / Thinking Skills): ....
- Observation Group (Male / Female): ....
- Date of observation: ....

Evaluation Observation Form

① = Poor ② = Acceptable ③ = Good ④ = Very Good ⑤ = Excellent

SELF-DEVELOPMENT SKILLS USING CREATIVE DRAMA			Degree by Observing Performance	Other Notes
Communication Skills	Verbal Communication Skills	The student can speak to his colleagues successfully.	① ② ③ ④ ⑤	
		The student listens to his colleagues while speaking and accepts the other opinion.	① ② ③ ④ ⑤	
		The student is good at using letter and word exits.	① ② ③ ④ ⑤	
	Nonverbal Communication Skills	The student uses body language during lectures.	① ② ③ ④ ⑤	
		The student is good at using facial expressions.	① ② ③ ④ ⑤	
		The student uses the language of things (smells, clothes, and distances).	① ② ③ ④ ⑤	
	Electronic Communication Skills	The student is proficient in dealing with international search engines.	① ② ③ ④ ⑤	
		Student maintains copyright and scientific integrity.	① ② ③ ④ ⑤	
		Fluent student deals with the dangers of the Internet.	① ② ③ ④ ⑤	
		Creative drama contributes to improving students' communication skills.	① ② ③ ④ ⑤	

Thinking Skills	Critical Thinking Skills	The student is free from fear when he launches his innovative ideas.	① ② ③ ④ ⑤	
		The student accepts the ideas of his colleagues even if they are different from his ideas.	① ② ③ ④ ⑤	
		All students participate in the launch of innovative ideas.	① ② ③ ④ ⑤	
	Problem Solving Skills	The student senses the problem and tries to resolve it.	① ② ③ ④ ⑤	
		The student assumes several assumptions and alternatives to solve the problem.	① ② ③ ④ ⑤	
		Students come up with suggested solutions to the problem he is facing.	① ② ③ ④ ⑤	
	Brainstorming Skills	The student introduces critical new ideas.	① ② ③ ④ ⑤	
		The student is well versed in teamwork in presenting innovative ideas.	① ② ③ ④ ⑤	
		The student accepts the ideas of others and develops his idea.	① ② ③ ④ ⑤	
		Creative drama contributes to improving students' thinking skills.	① ② ③ ④ ⑤	

(IJACSA) International Journal of Advanced Computer Science and Applications,
Vol. 9, No. 4, 2018

Pakistan Sign Language Detection using PCA and KNN

Muhammad Sheraz Arshad Malik, Naila Kousar, Tahir Abdullah, Muhammad Ahmed, Faiqa Rasheed
Department of Information Technology
Government College University
Faisalabad, Pakistan

Muhammad Awais
Department of Software Engineering
Government College University
Faisalabad, Pakistan

***Abstract*—Every society has a large majority group of disable people. The technology is developing day by day but no significant developments are undertaken for the improvement of these people. Sign language is an efficient mean of information exchange with special people, such as Deaf and Dumb people, they communicate with each other through sign language, but it become difficult when they communicate to outer world so sign language is used for this purpose. Different research has been done for this in America, Indonesia and India, but not much work done in Pakistan. In this research paper, author introduce a system for recognizing Pakistan Sign Language (PSL) including the alphabet to facilitate communication between special people and normal. This system capture input through webcam without making use of any additional hardware, then using segmentation approach we separate hand from the background and extract required feature from image using Principal Component Analysis (PCA) and then finally classifies the gesture feature by utilizing K Nearest Neighbors (KNN). This research will fill the communication gap between the deaf and normal people of the Pakistan country.**

Keywords*—Deaf and dumb; hand gesture recognition; k-nearest neighbors; Pakistan sign language; principal component analysis; Urdu alphabets

I. Introduction

Every society has a large maniority group of disable people. The technology is developing day by day but no significant developments are undertaken for the improvement of these people. Sign Language is an efficient mean of information exchange with special people like Deaf and Dumb, they communicate with each other through sign language, but it become difficult when they communicate to outer world so Sign Language is used for this purpose.

Different research has been done for this in America, Indonesia and India because every country has its own sign language even it's also vary from region to region. So Pakistan also has its own sign language for communication with special people but not much work has been done in Pakistan. For PSL detection a system is required.

In this paper we introduce standard Pakistan sign language (PSL) detection System through hand gestures which is implemented in MATLAB. The main contribution of this research is to design system for standard Pakistan Sign Language for Urdu Alphabets using PCA and k-Nearest Neighbors algorithm. PCA is used to reduce data dimensions. PCA takes high dimension data and convert it into lower dimension data. It finds a new set of variables, which is smaller than the original set [7]. In this research main focus will be on accuracy then the speed.

This system capture input image through webcam without making use of any additional hardware, or gloves which is being used by earlier sign detection system, then use different image processing technique for feature extraction and finally identify the gesture and display letter against that gesture using some classification method. The complete working flow of purposed system has been shown in Fig. 1.

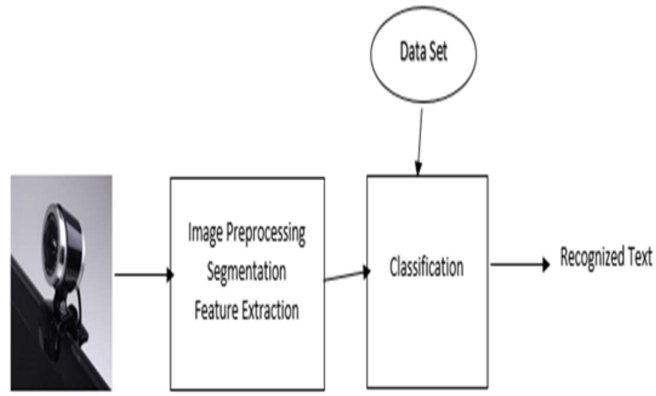


Fig. 1. Architecture of proposed system [1].

The prototype designed during this research will be introduced in "Govt. Special Education Centre, Jaranwala" having strength more then 150.

The paper is divide in the following sections. Section 2 talks about proposed model for Pakistan sign language Urdu alphabet identification based on hand gesture recognition using neural networks. Section 3 consist of the results obtained by experiment on different input. Section 4 deals with Conclusion and future.

II. Methodology

The basic idea behind the proposed model is to identify PSL Urdu alphabets depending on hand gesture captured by webcam. The work flow of proposed system is shown in Fig. 2.

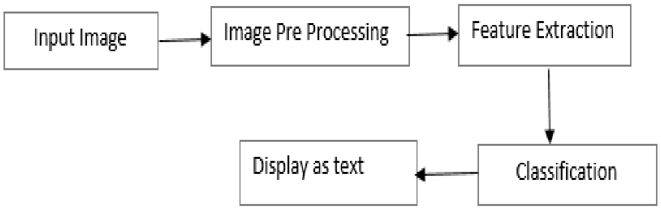


Fig. 2. Workflow of PSL detection system.

A. *Image Preprocessing*

Fig. 3 shows steps involve in image processing.In first step image is provided as input.In order to take input from the webcam a system is implemented using MATLAB. Then this image goes to image pre-processing steps in which segmentation and image conversion is performed, the RGB image is converted into grayscale image i.e. BW image for processing, because grayscale is more preferable. The function rgb2hsv is being used for conversion [2].

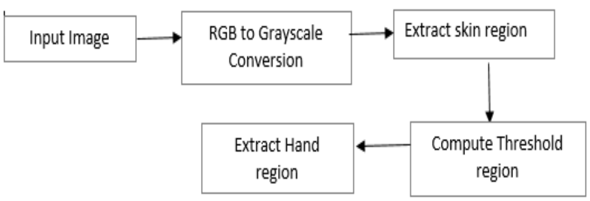


Fig. 3. Steps of image preprocessing.

Then skin region is extracted by calculating skin probability, after that thresholded region is computed. Threshold is simplest way of segmentation in which conversion of grey scale image to binary image is performed. In computer vision, segmentation or image segmentation is a process of partitioning an image into multiple segments (sets of pixels). Segmentation can also be defined as a process of separating the required region from its background. In PSL as hand gesture is used, so segmentation involve separating hand region from the background. After performing Segmentation, the hand region is extracted as white colour and other regions of image are assigned black colour. Results of skin thresholding during prepressing step has been shown in Fig. 4.

Fig. 4. Results of skin thresholding during prepressing step.

B. *Feature Extraction*

The binary images obtained after performing some processing like Skin extraction are used to extract the features using Principal Component Analysis. Feature extraction is a process in which dimension of image is reduced. As we are working on gesture the no of frames are too large to processed, so input of image pre-processing step is reduced here. Different feature extraction procedures available such as HOG [4], PCA [5], Fourier De-scripters [3], etc.

In our case we will make use of PCA algorithm for feature extraction which will return output in the form of feature weight Matrix.

1) *Principal component analysis (PCA)*

PCA is used to reduce data dimensions. PCA takes high dimension data and convert it into lower dimension data. It finds a new set of variables, which is smaller than the original set. It's maintain sample important information when extracting relevant and required information from dataset. It's useful for compression reduction and classification of data [6].

PCA algorithm steps in PSL detection:

Step1: Obtain images I1, I2, Im images for training.

Step2: subtract the mean

Ai = Γi – Ψ

Step 3: Compute Covariance matrix as

C = ATA. This is of size M x M.

Step 4: We find the best Eigen vectors from this covariance matrix R. The Eigen vectors Ui we find from this correspond to the M best Eigen vectors Vi of the C matrix. And it can be shown that these two Eigen vectors are related by Vi = A * Ui. These are the Principal Components of the Image Matrix A. Fig. 5 shows the Eigen vectors extracted using PCA of URDU alphabet.

Step 5: We find the weights of the images in this transformed space with respect to the Eigen Vectors Vi. This is calculated as:

$$\Phi_i - mean = \sum_{j=1}^{K} w_j u_j, \; (\mathrm{w_j} = u_j^T \Phi_i)$$

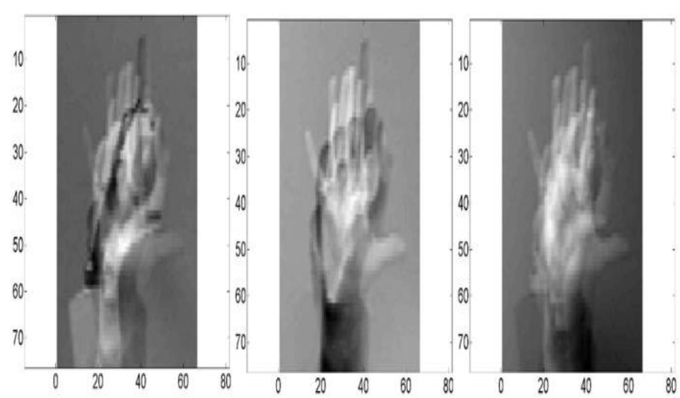


Fig. 5. Eigen vectors extracted using PCA.

C. K Nearest Neighbors (KNN) Classifier

In the last step of proposed system, we classified each input image which we get after feature extraction phase using KNN classification. Classification identify input image to a set of taining image data set. KNN is an algorithm that classifies new cases based on similarity measure i.e. distance functions. KNN is being used in pattern recognition and statistical estimation in the beginning of 1970's [7]. KNN is being used in this work because KNN is the simplest classifier used to solve classification problem.Training and Testing image have same dimensions i.e. same number of columns.

Usually Euclidean distance is used as the distance metric in KNN. In our experiment, classification is performed by finding the nearest Euclidean distance of the transformed weights of the test image. Distance using Euclidean function is calculated using following equation:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Using this distance calculation and KNN classifier Urdu alphabet for the give sign is displayed. Fig. 6 shows the output after KNN classifier.

Fig. 6. Classification using KNN.

For displaying URDU words we implemented URDU Optical character recognition system, which make use of Unicode for this purpose, such as Unicode of ب is "0628" [8] as shown in Fig. 7.

By using this Unicode and hex2dec function i.e. hex2dec(Unicode) [9], we got required Urdu Alphabet.

Fig. 7. URDU alphabets Unicode [8].

III. Implementation and Interpretation of Experimental Results

The efficiency of the PSL detection system is evaluated by testing both training and testing data set.Standard PSL have been shown in Fig. 8.

Fig. 8. PSL for Urdu alphabets.

The System has been implemented in MATLAB R2015b using PCA and KNN algorithm. The database being used in Pakistan sign language detection system is self-created because it's not available. Database of PSL alphabets is created by videos which are captured by web cam. We have used 30 image per alphabet so our database contain 38x30= 1140 images, as there are 38 Urdu alphabets. Two data set are being used one for testing and other for training.There are 140 images in Testing dataset and 1000 images in Training dataset. These images are taken from different distance and from different angle for checking system performance and its efficiency. Testing database is used to test system with possible inputs, it's also used to test to evaluate the performance of Training dataset. For the performance evaluation, the system also has been tested multiple times on testing and training data set by more than 5 user.

The proposed also system has been tested by different user using webcam having different skin type, color and gender, for test database data has been collected from people which belongs to different age.

The purposed prototype has been deployed in "Govt. Special Education Centre, Jaranwala" for testing, and it is tested nearabout 50 special children of different age , having different skin color.

After testing system fom different cases the avegare accuracy is 85% , but for some Urdu alphabet the accuracy was 100% like "الف","ب " but we get low accuracy rate when dealing with Urdu alphabet like "غ". As some of the Pakistan signs are almost similar it was very difficult to identify such as "غ" and "ف" shown in Fig. 9 below.

Fig. 9. URDU alphabets with low accuracy.

IV. Conclusion and Future Work

In this research paper, Pakistan Sign language (PSL) detection system for Urdu alphabets has been developed using MATLAB. The aim of this system is to help disabled persons living in Pakistan to communication with other. The hand gesture were recognized by using PCA and KNN classifier giving 85% accuracy rate for test data set which contain more than 1000 images and 80% when tested through webcam.

This system can further be carried out for English alphabet for single and double handed of PSL and will also improve accuracy rate and also will introduce voice fature for text to help blind one. A mobile base application for this purpose will also be designed because now a days the use of mobile phone is increasing rapidly.

Acknowledgment

The author would like to thanks department of Information Technology of Govt. College University Faisalabad, Pakistan for their support in this research work and also Pakistan Sign Language for their supporting matrial, and also Govt. Special Education Centre, Jaranwala for providing an opportunity to test the purposed system.

References

[1] Salian, S., et al. (2017). Proposed system for sign language recognition. 2017 International Conference on Computation of Power, Energy Information and Commuincation (ICCPEIC).

[2] Mathworkscom. (2018). Mathworkscom. Retrieved 2 March, 2018, from https://www.mathworks.com/help/matlab/ref/rgb2gray.html.

[3] P. C. Badhe and V. Kulkarni, "Indian sign language translator using gesture recognition algorithm," 2015 IEEE International Conference on Compuer Graphics, Vision and Information Security (CGVIS), Bhubaneswar, 2015, pp. 195-200.

[4] S. C. Agrawal, A. S. Jalal and C. Bhatnagar, "Recognition of Indian Sign Language using feature fusion," Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on, Kharagpur, 2012, pp. 1-5.

[5] Shreyashi Narayan Sawant. "Sign Language Recognition System to aid Deaf-dumb People Using PCA", International Journal of Computer Science & Engineering Technology (IJCSET), Vol. 5 No. 05, May 2014.

[6] Binjie, X. (2010). Principal component analysis for feature extraction of image sequence. 2010 International Conference on Computer and Commnication Technologies in Agriculture Engineering.

[7] Aryanie, D. and Y. Heryadi (2015). American sign language-based finger-spelling recognition using k-Nearest Neighbors classifier. 2015 3rd International Conference on Information and Communication Technology (ICoICT).

[8] Bhurgri, A.-M., Enabling Pakistani Languages through Unicode. p. 35.

[9] Mathworkscom. (2018). Mathworkscom. Retrieved 12 March, 2018, from https://www.mathworks.com/matlabcentral/fileexchange/32674-urdu-optical-character-recognition-system?focused=6999104&tab=function.

eHealth WBAN: Energy-Efficient and Priority-Based Enhanced IEEE802.15.6 CSMA/CA MAC Protocol

Wireless Body Area Network

Ansar Munir Shah[1], Abdelzahir Abdelmaboud[2], Khalid Mahmood[3],
Mahmood ul Hassan[4], Muhammad Kashif Saeed[5]
Department of Computer Science
IIC University of Technology
Phnom Penh, Cambodia

***Abstract*—This paper provided a general study of Wireless Body Area Network (WBAN) in health monitoring system as well as the study of the application of wearable and implanted Bio-Medical-Sensors (BMS) which are used to monitor the vital signs of a patient in early detection. Energy efficiency is a significant issue in WBAN which can be achieved by reducing the overhead of control packets, prioritizing sensor-nodes and sink-node selection. Moreover, uncertainty in network topologies, such as distance and link affect between sensor-nodes occurs due to the mobility of human. In this research, we propose a scheme to reduce the overhead of control packets and prioritizing the threshold values of vital signs by assigning low and high transmission power with enhanced IEEE802.15.6 CSMA/CA as well as introduce a Mobility Link Table (MLT) for selecting a sink-node to communicate with the coordinator. Compare it with existing IEEE802.15.6 CSMA/CA technique and results shows the proposed techniques regarding mean power consumption, network delay, network throughput.**

Keywords—Wireless Body Area Network (WBAN); node; IEEE802.15.6; MAC; CSMA/CA; Mobility Link Table (MLT)

I. Introduction

A WSN is a manner to organize sensors nodes to observe and control the environmental factors such as temperature, motion detection, sound, and send their data to the base station [1]. WSN consists of various applications in the fields of medical, military and supports [2]. The use of wireless sensors in the medical field is considered as most emerging and evolving wireless network to monitor the patient anytime anywhere is known as WBAN wireless body area network. A Wireless Body Area Network (WBAN) is special-purpose sensor network which is designed for medical applications and operates independently to control communication between various medical sensors, which are to be placed inside and outside the human body Patient with long-standing diseases required continuous supervision. With restrict resources, it is almost impossible to focus on every patient continuously in and outside the hospital, so it is need of the time for healthcare services to become more efficient to monitor and examine the patient anytime and anywhere [3].

Sensors used in medical monitoring are tiny, so WBAN is required many parameters such as low power consumptions, low latency, reliability, efficient bandwidth utilization and maximum throughput. The batteries used in these types of sensors are also tiny, and it is challenging to recharge or replace them, so low power consumption is considered as a significant challenge for effective WBAN. Therefore energy efficient MAC protocol is needed to overcome this challenge [4]. WBAN can be made energy efficient by updating in MAC layer as well as the power control mechanism, so reducing the overhead of control packets, prioritizing nodes and efficient sink selection using one-hop topology are concerning areas to save the energy in WBAN [5]. The patients change their postures according to their needs and posture mobility of patient body effects distance and connections variation between different sensor nodes.

In this research, we will propose two different protocol techniques. First technique will rely upon reducing the overhead of control packets by using block acknowledgment for data packets and prioritizing nodes according to the transmission power and contention window size. Second technique will relate to adaptive sink selection, which permits sink selection by using Mobility Link Table (MLT). The simulated analysis describes that purposed mechanism is succeeded to reduce mean power consumption in WBAN.

In Section II, existing media access control mechanism is described. Section III contains the proposed schemes and experimental setup and analysis are evaluated in Section IV, as well as the Section V concludes the final recommendation and future work.

II. Existing Media Access Control Mechanisms

MAC frame consists of MAC header, payload, and FCS; payload holds the informative data, and MAC header hold Control Bit frame, Recipient-identity, Sender-identity and BAN-identity [6]. Usually, B-MAC, L-MAC, S-MAC and Wise MAC protocols are used in WBAN.

B-MAC (Berkeley Mac protocol) [7], low power listening protocol, have to wake up and sleep states of the nodes which described that of B-MAC is to make nodes sleepy for a long time interval and to wake up the nodes after fixed intervals to check ongoing data communication. Timeout- timer stop the data listing if it does not find any data packet for that time, else the node will wait for entire data packets. The sender node announces a preamble to other nodes before sending the data packet B-MAC is an asynchronous protocol and fairness is not guarantee in B-MAC. CCA (Clear Channel Assessment) is a key ingredient in B-MAC and CCA mean to detect clear channel for data or control packet transmission. Data receiving node will send an Acknowledgement packet to data sending node, but it is optional, and no priority mechanism used to assign the slot to the emergency data in a critical situation and also does not considered the threshold values of vital signs simultaneously.

Lightweight Medium Access (LMAC) protocol is TDMA-based protocol [8]. Slot assignment and synchronization in an LMAC network is self-organizing. This protocol reduces energy by avoiding a collision, minimizing overheads on the physical layer, and by reducing the state-switching in the transceiver when changing the sleep-interval for nodes according to the amount of data traffic. In LMAC [9], collision is reduced by dividing time into multiple slots and assigned specific time-slot to each node for a specific time, So that each node can send its data in allocated time-slot and don’t need to compete for medium. A distributed algorithm used for time slot assignment. Both sender and receiver will turn off their transceivers after data transmission and reception respectively for saving the energy consumption. The node synchronization process is significant for LMAC, achieved by gateway control messages. Each node, on reception of gateway message, will synchronize their clocks with a gateway of their one-hop neighbor, on the other hand, it does not handle the thresh hold values of vital signs at the same time, and fixed slot allocation caused delay.

Sensor-MAC protocol considered energy-efficient media access control protocol. According to [1], SMAC reduces energy by avoiding a collision, control packet overhead, overhearing, idle listening. LMAC recognizes sleep and listens to phases for transceiver node so there should be synchronization between the network-nodes for sleeping and listen, but synchronization is not strict sometime between the nodes so they can change their sleep period as they required communication need [9]. Periodic sleeping is an essential component of SMAC which means node will change its state to sleep state in the absence of data transmission or reception. During sleep period, the node controls its wireless off, and awake it-self later according to the timer. Both contention and scheduling techniques are deployed in this protocol to reduce data collision. Fairness is considered in protocol transmission but the delay increased.

Wise-MAC is known as high energy-efficient MAC protocol presented in [10]. Wise-mac is based upon non-deterministic CSMA and became energy-efficient by using preamble sampling, and by reducing idle-listening. Idle-listening reduced by listening medium for short interval of time. The process of listening will continue until data-packet is received or medium becomes idle until finding the channel busy. Many nodes are used as relays towards the sink to avoid collision between data packets and further collision can be reduced by using medium reservation preamble in the wake-up preamble. If long data packet transmission is required, then an extra bit called more is added in front of the data packet to indicate receiver to listen, even after sending an acknowledgment. The sender sends the second packet ongoing to the received ACK-packet. CSMA/CA protocol is based on “contention-based technique”. This MAC frame structure is shown in Fig. 1.

The whole channel is divided into super-frame hierarchy [11]. In each beacon period, beacons are broadcast. In the inactive period, no beacon is transmitted. Fig. 1 represents the Beacon mode with beacon super-frame edges, Fig. 2 and 3 shows the Non-Beacon mode with super-frame edges and without super-frame edges, respectively.

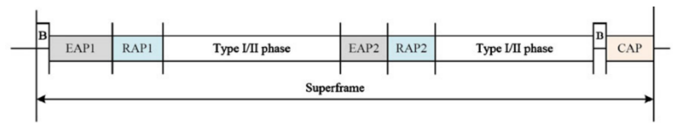


Fig. 1. Beacon mode with beacon super-frame edges.

In CSMA/CA the node detects the channel, and if find it idle, the node will broadcast the data, and if find it busy, the node will wait for random time and will try again. The node sets a back-off counter between 1 and contention window (CW) size. Node senses the random channel time, and if the channel idle, the counter decrement one at every idle and transmit the frame on the channel if the back off counter becomes zero. If the channel is found busy, the node will lock its back-off counter until the channel becomes idle. A number of failures occur are counted by another counter [12]. There are two cases possible to represent the failure solution, the Contention window (CW) size will remain unchanged if failures counter recorded odd value, the CW will be doubled if failure counter value noticed even. After data transmission completed, contention window-size is set to initial CW.

IEEE 802.15.6 based CSMA/CA [13] describes the MAC protocol with immediate acknowledgment and without acknowledgment on same contention window size (without user priority) and calculate the network delay, throughput and bandwidth efficiency. This scheme is implemented only for one sender, and one receiver node, multi-nodes, and postural mobility is not considered in the simulation. While in WBAN, we need a mobile network along with more than one receiver and transmitter.

III. Proposed Schemes

Propose technique aims to improve the overall network throughput and reduce the mean power consumption considering body mobility and posture variation factors in Intra-WBAN and evaluate the performance metrics of mobility network. Eleven nodes in the network and introduce mobility parameters to evaluate the performance parameter of the network. For energy efficiency in WBAN, we prioritize nodes by critical/emergency traffic and non-critical/normal traffic. Normal and emergency traffic are differentiated by using two techniques, one by different contention window (CW) sizes and another one is by using different transmission power for the different type of traffic. In IEEE 802.15.6 based CSMA/CA protocol, only immediate acknowledgment policy is implemented. In proposed protocol, we use block acknowledgment policy which reduces the control packet overhead so that we can get efficient power consumption in WBAN.

In the first scheme, we proposed a superframe with RAP1 and RAP2 with prioritizing the nodes by using different transmission power and different contention window size for each node. The sensor node with higher transmission power, let's call P1, will get more priority, and higher power will be given to node generating emergency traffic while node generating normal traffic will be given lower power let's say P2 which will effectively decrease the priority of that node as compared to the node with emergency traffic. Due to the high priority of emergency traffic generating node, its traffic will reach the destination without distortion by the collision which will produce less delay. Proposed technique (enhancement of IEEE 802.16.4) packet collision scenarios are defined in two states, as shown in Fig. 2, i) S1: More than one low powered packets arrival at t2 ii) S2: More than one High powered packets arrival at t5. In scenario 1, more than one low priority packets arrive which causes a collision between them. The packet collision cause discard of the packet and no packet received at the destination. Similarly, in scenario 2, there is more than one high priority packet arrived which also results in a collision.

A packet without collision scenarios are defined in three states, as shown in Fig. 3, i) S1: one high powered packet arrival, ii) S2: one low powered packet arrival, iii) S3: one high powered packet and more than one low powered packet arrival. When one high or one low priority packet arrives, it is received at the destination without collision. In scenario 3, one high powered and one low powered packet is received, as per our policy, the high powered packet will get priority and will reach the destination without collision.

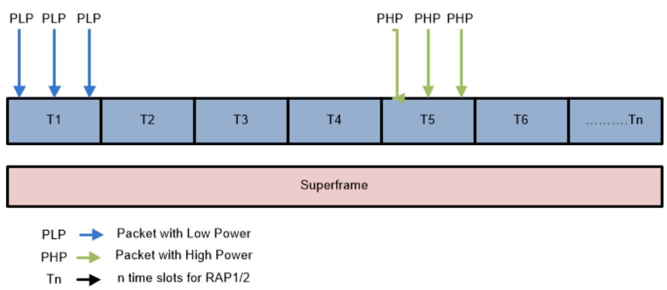


Fig. 2. Collided packet transmission with Enhanced IEEE 802.15.6.

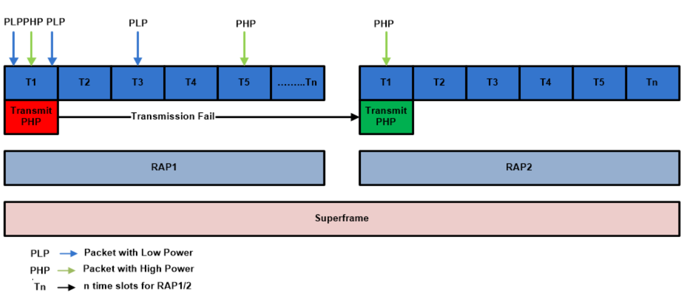


Fig. 3. Successful packet transmission in Enhanced IEEE 802.15.6.

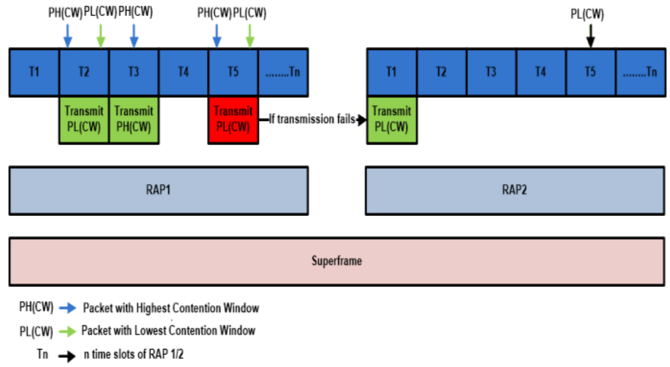


Fig. 4. Priority using contention window size.

The second method to assign user priority is by using different contention window size for each node. A node with minimum contention window size will get higher priority while a node with maximum contention window size will get lower priority. Emergency traffic will get the highest priority, so it will access the channel first due to which optimal delay and throughput are obtained. In Fig. 4 the access mechanism of the the channel is illustrated, which has three cases: i) Only low priority node transmit data ii) Only high priority node transmit data ii) High and low priority node transmit data. In case 1, at time T2 there are two packets, one with high contention window and another one with low contention window trying to access channel. The packet with low contention window will access the channel firstly, while the high window and another one with low contention window trying to access channel. The packet with low contention window will access the channel firstly, while the high contention window sized packet will wait for its turn. While in scenario 2 and 3 at time T4 and T6, one low contention window sized packet and one high contention window sized packet are trying to access the channel, so they will send their packets through the channel without collision which will ultimately cause minimum delay.

The second scheme introduced a mechanism to choose a sink node to communicate with the coordinator in Intra-WBAN. In WBAN, it is imperative to make the data transmission easy by choosing the easily accessible sink node for all other sensor nodes. In Intra-WBAN, distance and connection between different nodes vary due to the variation in posture or mobility in the human body, as shown in Fig. 5. Energy waste is a significant problem in WBAN due to improper sink selection technique and the maximized control overhead [14]. Using fix-sink or random-sink selection, using the routing table are the conventional techniques in WBAN. However, we know that information collected from every node is critical for the doctor to take any decision for patient treatment, so we need active techniques in which all nodes are accessible by the sink. Because in previous techniques other nodes may or may not access that the sink does not access particular node or other nosed.

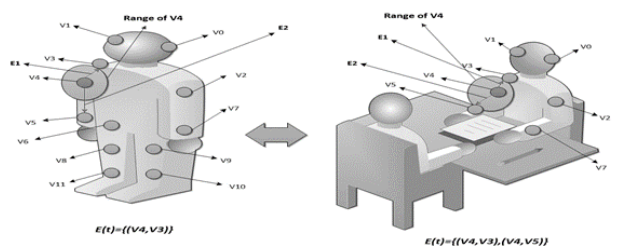


Fig. 5. Illustration of range changing of sensor node as posture changes.

As we are considering the human mobility, the MLT is a significant feature of our technique. This table contains information about no. of connections; each node has a variable called "Link" which will be incremented after each connection between two nodes is verified, through acknowledgment. MLT will use this variable to store a number of connections by each node. An MLT is distributed and maintained by each node. Complete procedure to fill this table and selecting adaptive sink by MLT described in Fig. 6.

According to Fig 6, each sensor node will broadcast control packet. All the nodes reply after getting the control packet from another node, MLT will be updated (Incrementing No. of Links) if ACK received. A node with maximum "link" value in MLT presents the sink node in selected posture. Table update before every data transmission against the current posture and node that have more links will be select as a sink node. Sink will be updated before transmission by repeating the same procedure. In this paper, we compare proposed sink selection with fixed sink, random sink selection through routing table.

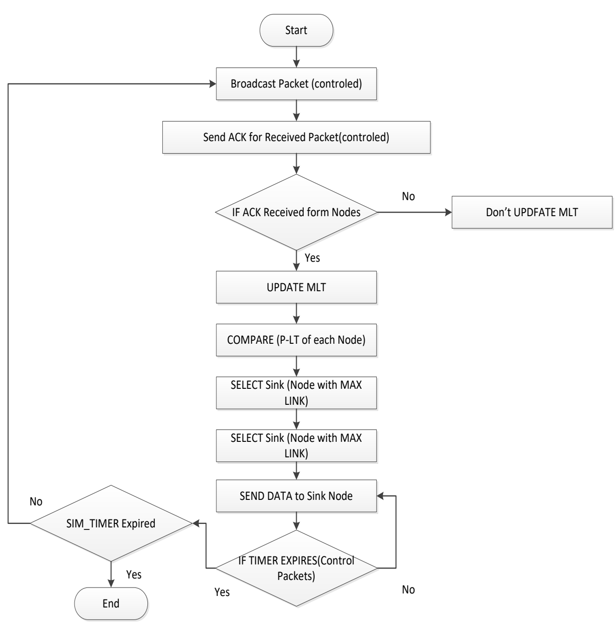
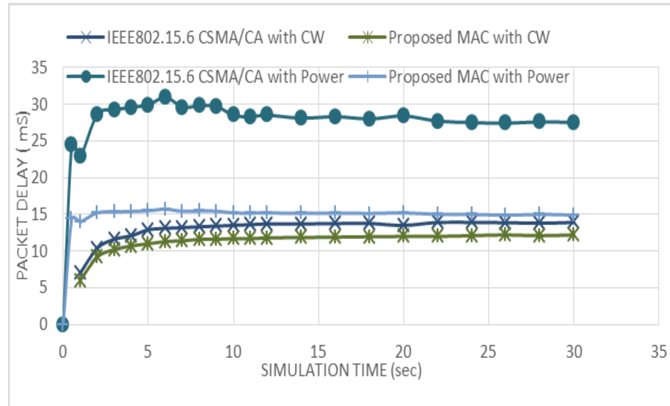


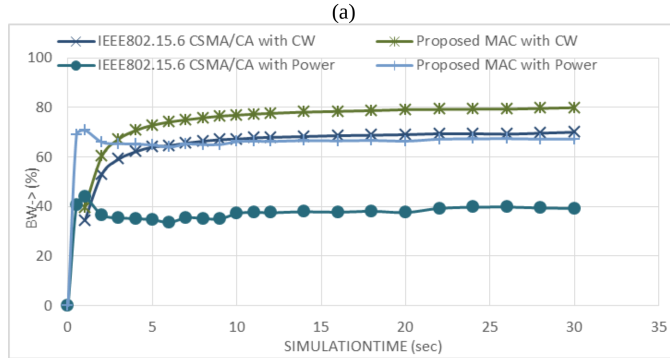
Fig. 6. Flow chart for adaptive sink selection by MLT.

IV. Experimental Setup

Proposed scheme are compared with IEEE802.15.6 CSMA/CA by simulated in OMNET++/MIXIM. Fig. 7(a) shows network delay comparison between proposed MAC and IEEE 802.15.6 based CSMA/CA by prioritizing using Contention Widow (CW) mechanism as well as the power priority mechanism. Fig. 7(b) shows network bandwidth efficiency comparison with IEEE 802.15.6 based CSMA/CA using CW and also prioritizing the nodes by the transmission power.



(a)



(b)

Fig. 7. (a) Network delay, (b) Network bandwidth.

Fig. 7(a) & (b) shows that the minimum network delay, maximum network throughput, and maximum bandwidth efficiency is also achieved for the network, in proposed MAC prioritizing by contention window (CW) due to less control packet overhead and nodal prioritization. As depicted from this figure that highest network delay, lowest throughput, and lowest bandwidth efficiency is present in case of IEEE 802.15.6 based CSMA/CA with CW. The proposed MAC with priority using contention window (CW) is showing optimum results because node with the highest priority uses the channel for a maximum time because that node has maximum data for transmission and need a channel for maximum time. Fig. 8 shows that minimum power consumption is achieved in proposed MAC by prioritizing using transmission power with block acknowledgment because when we use different transmission power and give the highest priority to highest transmission power node.

We know that nodes with normal traffic will transmit data using lower transmission power because we have assigned less priority to normal traffic. Total power consumption is the sum of transmission power, receiving power and power used in acknowledgment. Moreover, transmission power has the highest influence on power consumption, as normal traffic is mostly used, so transmission power of the whole network will be reduced which means that average transmission power is reduced so that we will get less power consumption. Moreover, block acknowledgment case further improves results by decreasing control packet overhead, which further reduces power consumption.

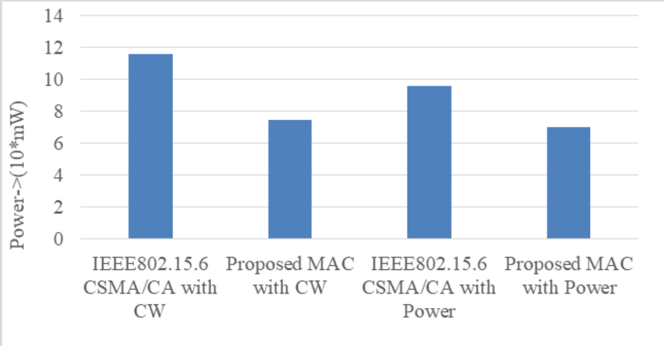


Fig. 8. Network mean power consumption.

Fig. 9 shows power consumption comparison between fixed sink, random sink selection technique, and adaptive sink selection technique through MLT, respectively for the network.

The performance analysis shows lower delay, lower power consumption, higher throughput, and higher bandwidth efficiency is achieved in adaptive sink selection technique through MLT. In adaptive sink-selection, a sink-node has maximum LINK available technique results more comfortable accessible for other nodes, which improve the performance results. On the other hand, in random sink-selection approach, selected sink-node may or may not have enough amount of connections cause more delay, high energy consumption and lower throughput.

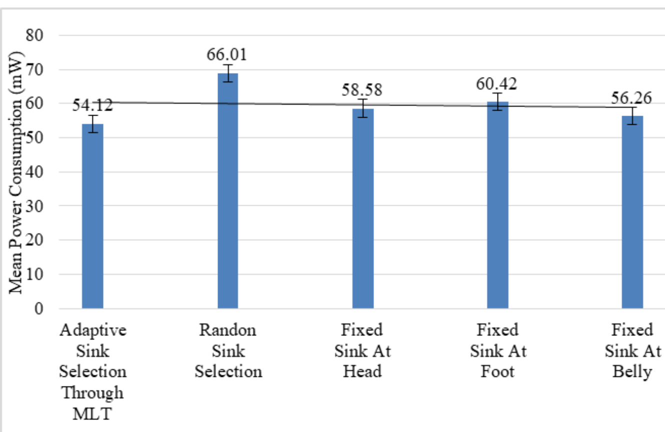


Fig. 9. Network power consumption on different sink.

V. Conclusion

Increased control packet overhead and inappropriate sink selection techniques are considered as energy wastage factor in Intra-WBAN. Two techniques proposed in this paper, the first technique helps in reducing control packet overhead, and its prioritization feature gives priority to nodes according to the power and contention window size. The second technique is based upon adaptive sink selection through MLT. These two schemes help in reducing network delay, mean power consumption, and increasing network throughput. In the first technique, control packet overhead is reduced by sending acknowledgment after 5 data packets. A second key feature which is present in this technique is two ways of assigning nodal priority according to the transmission power and contention window size, where transmission power and contention window size of every node is unique.

In the first case, the node with maximum transmission power will be assigned highest priority and priority will decrease as transmission power reduces. In the second case, priority will increase as CW size reduces and so on. In the second technique, sink node for data packet will be adaptive to the variation in a number of connections that each node holds. A number of connections for every node are present in MLT. A node with a maximum number of connections will be made sink node, and other nodes will send data to this node.

Performance analysis shows that proposed techniques outperform then IEEE 802.15.6 based CSMA/CA considering different evaluation metric as presented in performance analysis metrics such as mean power consumption, network delay, network throughput and network bandwidth efficiency. Results show that our proposed techniques are showing ten to fifteen percent improvement in IEEE 802.15.6 based CSMA/CA. We want to further enhance our work by introducing the sink node selection by links and energy power maintain by sensor nodes considering the human mobility and posture variation in Intra-WBAN. As a test - the bed is provided in OMNET++\ MIXIM that is why other researchers can also enhance our work.

Acknowledgment

We would like to express our gratitude to IIC University of Technology, the Kingdom of Cambodia for providing us excellent research environment.

References

[1] Ye, Wei, John Heidemann, and Deborah Estrin. "An energy-efficient MAC protocol for wireless sensor networks." In INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, vol. 3, pp. 1567-1576. IEEE, 2002.

[2] Park, KeeHyun. "A Ubiquitous Motion Tracking System Using Sensors in a Personal Health Device." International Journal of Distributed Sensor Networks2013 (2013).

[3] Khan, Jamil Yusuf, Mehmet R. Yuce, Garrick Bulger, and Benjamin Harding. "Wireless body area network (WBAN) design techniques and performance evaluation." Journal of medical systems 36, no. 3 (2012): 1441-1457.

[4] Cavallari, Riccardo, et al. "A survey on wireless body area networks: Technologies and design challenges." IEEE Communications Surveys & Tutorials 16.3 (2014): 1635-1657.

[5] Mehmood, Amir, and Adnan Nadeem. "Self-Organized Routing in Body Area Sensors Networks: A Review, Analysis and Simulation-based Case Study." Pakistan Journal of Engineering and Applied Sciences (2016).

[6] Kwak, Kyung Sup, Sana Ullah, and Niamat Ullah. "An overview of IEEE 802.15. 6 standard." Applied Sciences in Biomedical and Communication Technologies (ISABEL), 2010 3rd International Symposium on. IEEE, 2010.

[7] Forster, A. "Implementation of the B-MAC Protocol for WSN in MiXiM."Networking Laboratory, University of Applied Sciences of Southern Switzerland(2009).

[8] Ali, Mahmood, Annette Böhm, and Magnus Jonsson. "Wireless sensor networks for surveillance applications–a comparative survey of mac protocols." Wireless and Mobile Communications, 2008. ICWMC'08. The Fourth International Conference on. IEEE, 2008.

[9] Van Hoesel, Lodewijk FW, and Paul JM Havinga. "A lightweight medium access protocol (LMAC) for wireless sensor networks: Reducing preamble transmissions and transceiver state switches." 1st International Workshop on Networked Sensing Systems (INSS. Society of Instrument and Control Engineers (SICE), 2004.

[10] El-Hoiydi, Amre, and Jean-Dominique Decotignie. "WiseMAC: An ultra low power MAC protocol for multi-hop wireless sensor networks." Algorithmic Aspects of Wireless Sensor Networks. Springer Berlin Heidelberg, 2004. 18-31.

[11] Crosby, Garth V., et al. "Wireless body area networks for healthcare: A survey." International Journal of Ad Hoc, Sensor & Ubiquitous Computing 3.3 (2012): 1.

[12] Ullah, Sana, Min Chen, and Kyung Sup Kwak. "Throughput and delay analysis of IEEE 802.15. 6-based CSMA/CA protocol." Journal of medical systems 36.6 (2012): 3875-3891.

[13] Ait Zaouiat, C. E., and A. Latif. "Performances Comparison of IEEE 802.15. 6 and IEEE 802.15. 4." INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS 8.11 (2017): 461-467.

[14] Baig, Muhammad Rahim, et al. "AnyCasting In Dual Sink Approach (ACIDS) for WBASNs." INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS 8.3 (2017): 257-263.

Cobit 5-Based Approach for IT Project Portfolio Management: Application to a Moroccan University

Souad AHRIZ, Abir EL YAMAMI, Khalifa MANSOURI, Mohammed QBADOU
Laboratory: Signals, Distributed Systems and Artificial Intelligence (SSDIA)
ENSET Mohammedia, Hassan II University of Casablanca, Morocco

***Abstract*—Considering the problem of the management of IT project portfolios in universities, University managers face a lot of uncertainties when prioritizing projects that make up their portfolio. The alignment with their strategy becomes a major challenge and constitutes one of the essential elements of a governance approach. To overcome this challenge, the implementation of a project prioritization approach adapted to the university's strategy, vision and culture is essential. In this context, this paper aims to provide a multi-criteria approach based on a combination of AHP and TOPSIS methodologies for the selection and prioritization of IT projects in universities. The main feature of our approach is the use of COBIT 5, its principles and enablers as prioritization criteria. In order to validate our model, project portfolio managers of a Moroccan public university were involved to evaluate the criteria and to prioritize their projects. This research demonstrates that the combined use of Multi Criteria Decision Making (MCDM) methodologies proves to be suitable for the implementation of COBIT sub-process APO05.03.**

Keywords—Component; IT governance; project portfolio management; Cobit 5; AHP; TOPSIS; prioritization; university

I. INTRODUCTION

Information technology has become essential in supporting the growth and sustainability of all types of organizations including universities. The maximized investment in these technologies is forcing decision makers to implement effective IT governance mechanisms. This IT governance requires the definition and implementation of structures and processes to maximize the value from their IT investment, to better manage risks, to optimize resources and ultimately meet the strategy of the organization and its stakeholder's requests.

Portfolio governance, a part of IT governance, is the bridge between the corporate governance and the project that includes the decisions about managing projects, defines the responsible for every decision on the project and encompasses decision tasks and how these decisions should be made [1]. It aims to align the information system with the strategic priorities of the organization, to provide a global vision of all projects and allows not only to standardize management processes and rules, but also to be able to revise priorities if necessary. It ensures that not only high-value projects are added, funded and launched in a secure manner, but also executed according to stakeholders priorities and needs.

Several researchers have examined the effective management and success of project portfolios and their impact on the performance of organizations. Patanakul conducted a qualitative study for defining the attributes of portfolio effectiveness, he clarifies that project management literature in general discusses project portfolio management from management perspective and indicates that maximizing the value of the portfolio, balancing a portfolio, and aligning a project portfolio with a business strategy are three major goals for project portfolio management [2]. Unger et al. presented the positive impact that the abandonment of an ongoing project may have on the effectiveness and implementation of the strategy [3]. D. Jonas examined success factors and how they are related to the quality of management represented in "the quality of information, cooperation and resource allocation" [4]; Austin, C. et al. have conducted a study in University of Drexel in the United States in which they cited the lack of project management in higher education [5]. Indeed, the research carried out and the information collected from higher education institutions revealed the use of internal procedures for the management of IT projects, the use of experience feedback and available resources, without taking into consideration IT management good practices derived from international reference frameworks that could improve their image in a global market.

One of the major challenges that universities are facing is the large number of projects in their portfolio; they are led to optimize their resources and their investments. In fact, the selection and prioritization of projects cannot be done intuitively or based on inadequate evaluation criteria, but rather on the application of a well-defined portfolio management process.

Therefore, an effective project selection and prioritization approach is essential in order to properly balance the project portfolio and avoid selecting unprofitable projects that may have a negative impact on the performance of the process and the functioning of the institution.

This governance approach cannot succeed without the effective use of good practices frameworks and international standards. However, it is difficult to apply a common framework to all organizations that are currently demonstrating, a great interest in the adoption of these frameworks. It is therefore necessary to establish a method that is structured and adapted to the needs, strategy and culture of these institutions.

The main objective of this study is to propose a project portfolio management approach allowing IT project managers in universities to make a decision by evaluating several options in situations where no choice is perfect. The proposed approach is based on Cobit 5 framework and on AHP and TOPSIS multi-criteria decision making methods.

Cobit 5 proposes a process for portfolio management which is APO 05 "*manage portfolio*" of the domain "*Align Plan and Organize*". This process consists of six sub processes that must be satisfied (Fig. 1).

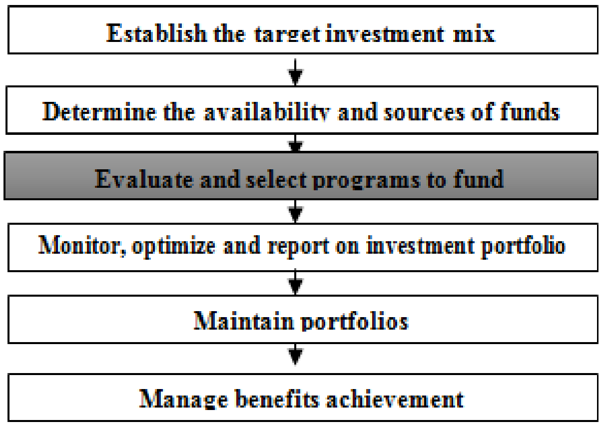


Fig. 1. Manage portfolio process.

The major limitation of Cobit framework is that it does not provide indications about the implementation of the proposed practices. Thereby, this contribution is an attempt to implement the sub process "*APO05-03 Evaluate and select programs to fund*".

This paper is organized as follows. In Section II, a theoretical background of the concepts and tools used in this study is presented. Section III describes the research methodology used and the proposed approach. A case study illustrating the implementation of the approach in universities is presented in Section IV. The results of this contribution are presented and discussed in Section V followed by a conclusion and perspectives of our research.

II. Theoretical Background

A. IT Governance

IT governance is defined as the leadership and organizational structures, processes and relational mechanisms that ensure that an organization's IT sustains and extends its strategy and objectives [6]. It is a process by which the objectives of the entity that give impact on Information technology are agreed, directed, and controlled [7]. IT Governance institute defines IT governance as "the responsibility of the Board of Directors and executive management. It is an integral part of enterprise governance and consists of the leadership and organizational structures and processes that ensure that the organization's IT sustains and extends the organization's strategy and objectives" [8].

According to ISO/IEC 38500 *"IT Governance is the system by which the current and future use of IT is directed and controlled. It involves evaluating and directing the use of IT to support the organization and monitoring this use to achieve plans and includes the strategy and policies for using IT within an organization"* [9]. Furthermore, some recent academic studies has demonstrated that the level of IT governance maturity has a significant positive impact on IT performance as well as organization performance [10], [11].

Although the problematic of IT governance has been studied by many authors, few studies can be found in the context of universities [12]-[14].

Recently, the importance of IT governance in universities has been increasingly recognized [15]. Universities have become more and more dependent on IT. To fulfill their mission and goals, they require adequate IT infrastructure and information systems which turns IT Governance into a real challenge [16]. In order to achieve their objectives, and improve their competitiveness and their effectiveness as well, they have to establish strategic objectives and make the appropriate decisions in terms of investing on IT. Universities are then showing interest in adopting the best practices and standards for IT governance because these frameworks are considered as guidelines that provide the basic structure that is flexible to apply in a certain environment. Similarly, Educause Center for Applied Research (ECAR) claims that, despite the development of ideology and procedures in terms of IT governance for business organizations, many higher education institutions have shown huge interest in implementing these ideologies to the management of IT [17].

B. Cobit 5

Control Objectives for Information and related Technology (COBIT) version 5 is a framework developed by IT Governance Institute and published on 2012 by ISACA [18]. It assists organizations in achieving their goals related to IT governance and management by providing a framework to establish the alignment of IT with the business [19], [20].

COBIT 5 enables information and related technology to be governed and managed in a holistic manner for the entire organization, taking in the full end-to-end business and functional areas of responsibility, considering the IT-related interests of internal and external stakeholders.

COBIT 5 is generic and useful for enterprises of all sizes, whether commercial, not-for-profit or in the public sector [21]. It allows the development of policies and practices for IT control throughout organizations and includes a set of 37 governance and management processes with respective metrics categorized into four domains of management and a domain of governance. It is based on 5 principles and 7 enablers that are the building blocks of the framework. COBIT 5 is an effective tool for implementing IT governance. Its latest version consists of RACI-charts to guide which stakeholders should be responsible, accountable, consulted, and informed about some activities [22].

C. Project Portfolio Management

The recognition of the strategic importance of project management is growing rapidly. This may be due to the strong belief that alignment between project management and organizational strategy can significantly increase their chance of achieving their strategic goals [23] and is positively associated with project performance [24]. Project management allows organizations to execute their strategic objectives in a structured manner and thus provides some element of control.

A portfolio refers to projects, programs, sub-portfolios, and operations managed as a group to achieve strategic objectives [25]. In a portfolio, projects must be quantifiable, classified and prioritized individually. According to Project Management Institute: Portfolio management refers to "*The centralized management of one or more portfolios, which include identifying, prioritizing, authorizing, managing and controlling projects, program and other related work to achieve specific strategic business objectives*" [26].

Blichfeldt and Eskerod define project portfolio management as the managerial activities that relate to the initial screening, selection and prioritization of project proposals, the concurrent reprioritization of projects in the portfolio, and the allocation and reallocation of resources to projects according to priority [27].

Accordingly, Project portfolio management process can be subdivided into two main phases:

- Prioritizing and selecting projects for the portfolio;
- Managing the projects within the portfolio.

The proposed metrics presented in the literature for calculating project priority have been criticized for not supporting the strategic alignment [28]. In this context, this paper aims to propose a multi criteria decision making approach for the selection and prioritization of IT projects, based on a combination of AHP and TOPSIS methodologies.

D. AHP

Thomas Saaty developed AHP as a decision-making method in the 1970s [29], it is a systematic decision making method which includes both qualitative and quantitative techniques. It is useful for obtaining single assessment value which is based on different indicators or criteria. It simplifies the process of decision making by subdividing a complex problem into a series of structured steps where each element in the hierarchy of criteria is supposed to be independent from others. The analytic network process is used when there is interdependence among criteria. AHP builds a hierarchy of decision items using comparisons between each pair of items expressed as a matrix. Paired comparisons produce weighting scores that measure how much importance items and criteria have with each other.

Decision maker examines two alternatives by considering one criteria and indicates a preference. The standard numeric scale used for AHP is 1-9 scale which lies between "equal importance" to "extreme importance", the value 9 indicates that one factor is extremely less important than the other, while value 1 indicates equal importance. At each level of the criteria hierarchy we obtain an n*n square matrix, where n is the number of elements of the level.

AHP allows building consensus among decision makers, each member can compare their judgments to those of the other members and it gives them better understanding of the impact of their priorities.

AHP decomposes the decision into the following steps [30]:

1) Define the problem and state the goal or objective.

2) Define the criteria or factors that influence the goal. Structure these factors into levels and sublevels.

3) Use paired comparisons of each factor with respect to each other that forms a comparison matrix with calculated weights, ranked eigenvalues, and consistency measures.

4) Synthesize the ranks of alternatives until the final choice is made.

E. TOPSIS

The TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) method is a multi-criteria analysis method developed by Hwang and Yoon in 1981 [31]. In this method two artificial alternatives are hypothesized: Ideal solution (IS) that presents the solution that has the best level for all attributes considered, and negative ideal solution (NIS) for the one which has the worst attribute values.

TOPSIS method performs prioritization of alternatives based on their geometric distance from the positive-ideal and negative-ideal solution.

TOPSIS decomposes the decision into the following steps [31] :

1) Establish the decision matrix,

2) Calculate a normalized decision R with coefficients Rij obtained by vector normalization.

$$R_{ij} = \frac{X_{ij}}{\sqrt{\sum_{i=1}^{m} X_{ij}}} \tag{1}$$

3) Determine the weighted decision matrix V with coefficients Vij which are calculated by multiplying each element of each column of the normalized decision matrix by the adequate weights:

$$V_{ij= W_{j*} R_{ij}} \tag{2}$$

4) Identify the positive and negative ideal solution according to the weighted decision matrix:

$$A^{+} = \{ V_1^{+}, V_2^{+}, \dots, V_n^{+} \} \text{ Where} \tag{3}$$

$$V_j^{+} = \left(maxi\left(V_{ij}\right) if\ j\ \in J\right); mini\left(V_{ij}\right) if\ j\ \in J'$$

$$A^{-} = \{ V_1^{-}, V_2^{-}, \dots, V_n^{-} \} \text{ Where} \tag{4}$$

$$V_j^{-} = \left(mini\left(V_{ij}\right) if\ j\ \in J\right); maxi\left(V_{ij}\right) if\ j\ \in J'$$

J is associated with beneficial attributes and J' with the non-beneficial attributes.

5) Calculate the separation distance of each competitive alternative from the positive ideal solution and negative ideal solution.

$$S^{+} = \sqrt{\sum_{i=1}^{n} \left(V_j^{+} - V_{ij}\right)^2} \tag{5}$$

$$S^{-} = \sqrt{\sum_{i=1}^{n}\left(V_j^{-} - V_{ij}\right)^2} \quad (6)$$

Where I = criterion index and j= alternative index.

6) Measure the relative closeness of each competitive alternative to the ideal solution,

$$C_i = \frac{S_i^{-}}{S_i^{+} + S_i^{-}} \quad 0 < C_i < 1 \quad (7)$$

7) Rank the preference in descending order; the optimum alternative is the one with the highest proximity index, and as such it represents the optimal decision, or preferred or optimal solution to the problem [32].

III. Research Methodology

To develop our approach and lead the project prioritization process, research at different levels has been done. In addition to the literature, the semi-directed interviews research method was used to collect information from experts managing IT project portfolios in universities. The target population was chosen because they have a holistic view about project portfolio management. The information collected were used to identify the project portfolio management process, to detect how decisions are made and to analyze the procedures used to select and prioritize IT projects.

The analysis of these data revealed the use of internal procedures that are based on academic needs and allocated resources and their feedback about past projects. Nevertheless, no entity has been defined for project portfolio management. Thus it calls for a huge need of developing a new approach for IT project portfolio management based on an internationally recognized framework is essential. In this context, this paper proposes a portfolio management approach to select and prioritize projects in a portfolio, taking into account the contribution of projects to the achievement of strategic objectives and their impact on the institution performance. This approach is based on COBIT 5 framework and in particular the sub-process APO05.03 "*Evaluate and select programs to fund*".

Thus, this paper proposes a set of six project selection criteria based on the five strategic axes of IT governance and the catalyst "Culture, ethics and behaviors" derived from cobit5 enablers. These criteria are intended to evaluate the quality and relevance of the projects that must be submitted to the same evaluation in order to guarantee the coherence of the portfolio.

Table I describes the criteria that will serves as a basis for project evaluation. Each criterion is associated with one or more processes in the Cobit 5 model. Table II presents the associated processes.

TABLE I. Prioritization Criteria

Criteria	**Description**
Strategic alignment	Strategic alignment is about targeting projects that are most relevant to the strategy [33] [34]. It aims to align the operation of IT with that of the organization and to ensure the value creation of IT for the organization.
Stakeholders needs	Stakeholder needs must first be identified. Then, the relationship between projects and strategy will be measured in order to select only projects that effectively meet the needs of all stakeholders.
Value delivery	This criterion consists of determining the effectiveness, material and immaterial value of each project following an individual evaluation. Only projects that generate the expected benefits will be included in the portfolio.
Resources optimization	The aim is to optimize the investment in vital IT resources (infrastructure, applications, information, and people).
Risk management	Portfolio managers must optimize risk by having a clear understanding of the institution risks and the assignment of risk management responsibilities.
Respect for the values, culture and ethics of the university	The culture, ethics and behavior of individuals and of the organization is a factor of success in governance and management activities. They must therefore be taken into consideration when selecting projects.

TABLE II. Selection Criteria and Associated Processes

Criteria	**Processes**
Strategic alignment	APO Align, Plan and Organize
Stakeholders needs	EDM 05 Ensure stakeholder transparency BAI 02 Manage requirements definition
Value delivery	EDM 02 Ensure benefits delivery
Resources optimization	EDM 04 Ensure resource optimization APO 06 Manage budget and cost APO 07 Manage Human resources
Risk management	EDM 3 Ensure risk optimization APO 12 Manage risk
Respect for the values, culture and ethics of the university	Cobit 5 Enabler 4

With the aim of constructing an approach that allows IT project portfolio selection and prioritization according to COBIT 5 practices, the study followed the procedure illustrated in Fig. 2. Each of the stages will be explained in detail below.

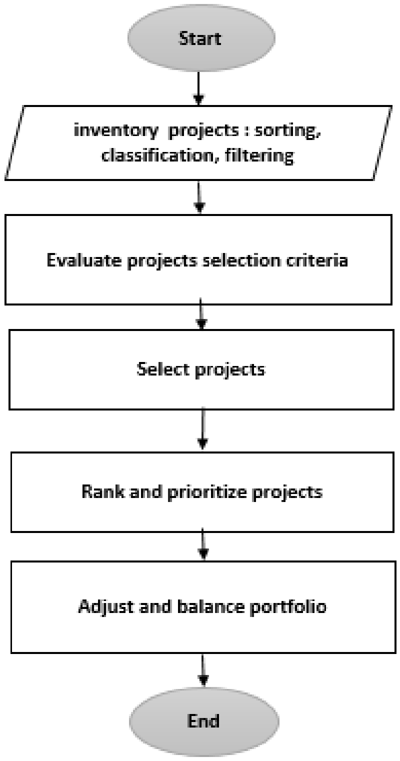


Fig. 2. Proposed approach.

- Step 1 : Projects Identification

For the majority of authors, the strategy is the starting point for identifying projects to put in a portfolio [35]. Projects must be identified according not only to the mission, vision and strategy of the university, but also to the investment budget allocated to the projects and the implanted organization structure.

- Step 2 : Criteria Evaluation

The selection of criteria depends on many different factors according to the strategic objectives of the institution. Some criteria are more important than others, therefore, for each criterion, a weighting should be assigned. Thus, we propose the use of Saaty scale (Table III) to obtain the consensus of the project team, and then the use of pairwise comparison.

TABLE III. SAATY SCALE

Numerical rating	Verbal judgment preferences
1	Equally important
3	Moderately more important
5	Strongly more important
7	Very strongly more important
9	Extremely more important
2, 4, 6, 8	Intermediate values between adjacent scale values
Reciprocal	*Aij* indicates the importance of *ith* factor over *jth*, then *aij* can be calculated as the reciprocal of *aij*

- Step 3 : Projects Selection

This step consists of evaluating the projects in accordance with the predetermined criteria to assign them a weighting and determine their importance and priority.

A scoring technique is proposed to determine the value of projects and prioritize the most important ones. The weight of each project is measured by the percentage of its contribution to achieving the goal.

- Step 4 : Projects Prioritization

The results from the previous step are compiled and the projects are compared. The one with the highest score is considered the one that generates the most value and must be achieved first. For that aim, TOPSIS method has been proposed.

- Step 5 : Portfolio Adjustment

Once projects are selected and prioritized, a portfolio adjustment is made. This step necessitates the reorganization of the portfolio as a result of the analysis done in the previous steps and aims to provide a better aligned portfolio that supports the university's strategy. The criteria for portfolio balancing depend on the weight of each project and its contribution to achieving the objectives.

IV. IMPLEMENTATION

The proposed approach was implemented in the field of higher education and specifically in a Moroccan public university. In response to accountability requirement, university decision makers have to manage the risks associated with the allocation of limited resources.

Indeed, the objective of this study is to contribute to the optimization of these resources and to achieve the right balance of investments by means of a portfolio aligned with the strategic direction and contributing to the performance of the university. Thus, it presents the procedure followed to build the project portfolio.

- Step 1 : Projects Identification

The selection of alternatives is one of the most important decision of portfolio construction because it influences the success of all the portfolio. Five alternatives will be analyzed to build the IT project portfolio. To determine this list, we consulted key people with different responsibilities and functions in the process of portfolio management of IT projects in a Moroccan university. A list of alternatives or candidate projects has been established (Table IV).

TABLE IV. LIST OF CANDIDATE PROJECTS

Code	Alternative
A1	Upgrading LANs in institutions
A2	Institutional messaging
A3	Inter-site interconnection
A4	Strengthening the security platform
A5	Upgrading the student and teaching management platform

- Step 2 : Criteria Evaluation

Information system project experts have evaluated these criteria using a pairwise comparison. A weight was then calculated for each criterion. The pairwise comparison results obtained are shown in Table V.

TABLE V. PAIRWISE COMPARISON

	C1	**C2**	**C3**	**C4**	**C5**	**C6**
C1	1	1/3	5	5	5	5
C2	3	1	5	5	5	5
C3	1/5	1/5	1	3	1/3	5
C4	1/5	1/5	1/3	1	1/5	3
C5	1/5	1/5	3	5	1	5
C6	1/5	1/5	1/5	1/3	1/5	1

AHP method was used to determine criteria weight. Based on Saaty scale decision making matrix was prepared. Table VI presents weights calculated according to AHP approach.

TABLE VI. WEIGHT CALCULATION WITH AHP METHOD

Criterion	**Criterion weight**	**Priority**
C1	28%	2
C2	41%	1
C3	8%	4
C4	5%	5
C5	14%	3
C6	4%	6

- Step 3 : Projects Selection

We describe trough Table VII the dataset of the selected projects and the scoring of each alternative on different criteria.

TABLE VII. DATA SET DESCRIPTION

	C1	**C2**	**C3**	**C4**	**C5**	**C6**
A1	80%	60%	80%	50%	50%	80%
A2	80%	80%	80%	50%	50%	80%
A3	80%	70%	80%	50%	60%	80%
A4	80%	50%	50%	60%	60%	80%
A5	80%	50%	50%	60%	50%	80%

- Step 4 : Projects Prioritization

The dataset is used as decision matrix, and then normalized decision matrix is calculated (Table VIII).

TABLE VIII. NORMALIZED DECISION MATRIX

r_{ij}	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
A1	0.45	0.43	0.51	0.41	0.41	0.45
A2	0.45	0.57	0.51	0.41	0.41	0.45
A3	0.45	0.50	0.51	0.41	0.49	0.45
A4	0.45	0.35	0.32	0.49	0.49	0.45
A5	0.45	0.35	0.32	0.49	0.41	0.45

TABLE IX. WEIGHTED DECISION MATRIX

V_{ij}	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
w	**28**	**41**	**8**	**5**	**14**	**4**
A1	12.522	17.438	4.1141	2.062	5.7735	1.7889
A2	12.522	23.251	4.1141	2.062	5.7735	1.7889
A3	12.522	20.345	4.1141	2.062	6.9282	1.7889
A4	12.522	14.532	2.5713	2.4744	6.9282	1.7889
A5	12.522	14.532	2.5713	2.4744	5.7735	1.7889

TOPSIS weighted Decision Matrix is calculated using priorities derived by AHP Method in Step 2 (Table IX).

Positive ideal A^+ and Negative ideal A^- solutions are defined according to the weighted decision matrix.

$$A^+ = \{12.522, 23.251, 4.1141, \quad 2.4744, \quad 6.9282, \quad 1.7889\}$$

$$A^- = \{12.522, \quad 14.532, 2.5713, \quad 2.062, \quad 5.7735, \quad 1.7889\}$$

Then for each competitive alternative the separation distance is calculated (Table X).

TABLE X. SEPARATION DISTANCE OF ALTERNATIVES

	S^+	S^-
P1	5.94	3.29
P2	1.22	8.85
P3	2.93	6.12
P4	8.85	1.22
P5	8.92	0.41

Finally, the relative closeness of each location to TOPSIS ideal solution is measured and projects are ranked in a descending order (Table XI).

- Step 5 : Portfolio Adjustment

After evaluating projects and approving investment programs, projects must undergo regular evaluations to adjust the portfolio and continually align with strategic factors that change over time. Hence, the portfolio is reorganized as a result of the performed analysis in the previous steps. Projects with the greatest weight will be implemented as a priority.

TABLE XI. PRIORITIZED PORTFOLIO

Projects		**Ranking**
Institutional messaging	0.87	1
Inter-site interconnection	0.67	2
Upgrading LANs in institutions	0.35	3
Strengthening the security platform	0.12	4
Upgrading the student and teaching management platform	0.04	5

V. DISCUSSION OF RESULTS

IT project portfolio governance in universities is essential to ensure that programs and projects deliver expected benefits and make an optimal contribution to the performance of the university. Effective IT governance has been shown to have a positive impact on financial performance [36]. However, for non-profit organizations such as public universities, other dimensions beyond the material value of projects need to be considered [37]. Their mission requires a balance between material and immaterial dimensions to achieve their educational, research and management goals.

Programs and projects are part of the university's ecosystem. They must be initiated by taking into account the needs of stakeholders. The objective of this paper is to identify how portfolio management can benefit from using Cobit 5 as an IT governance framework and how to leverage its processes, principles, and enablers in designing our approach.

Cobit 5 enablers can be applied in this practical situation and can be used to implement effective and efficient IT governance. They were used as a determining factor in the preparation of the proposed approach which is relevant and fits perfectly within the framework of project portfolio management.

Cobit 5 is an integrated framework that not only covers all of the organization's processes, but also separates them into governance and management processes, which makes it possible to distinguish between portfolio management, which is more a function of governance, and the management of program and project, which is more operational.

In fact, the process APO 05 "*Manage Portfolio"*, member of COBIT domain "*Align, Plan and Organize (APO) domain*", consists of aligning investments with the organization's strategic goals, manage programs according to constraints and available resources. In addition, it aims to prioritize projects, balance the portfolio and optimize its performance by proposing any adjustments.

Although Cobit 5 covers all the areas to be piloted and proposes effective practices to detect processes to be improved, it does not provide a practical approach for the implementation of the proposed practices. Therefore, it comes back to the organization managers to analyze, according to their context, the technological and organizational choices and implement the change. The papers' contribution responds to this need by offering a practical tool to manage the portfolio of IT projects. It is based on the process APO 05 and specifically addresses the needs of sub-process APO05.03 related to project selection and prioritization.

Based on the literature on Cobit 5 and the opinion of IT project management experts in universities, a set of criteria has been identified: alignment with strategy, response to stakeholder needs, value creation for stakeholders, resource optimization, risk optimization and respect for the values, culture and ethics of the university.

The multi-criteria aspect is important when making portfolio management decisions. The proposed approach combined both AHP and TOPSIS assessment techniques to facilitate decision making. This combination made it possible to select and prioritize IT projects by the experts who participated in this study by evaluating the criteria and prioritizing each alternative. This scoring technique has helped to determine the value of projects and to focus efforts and resources on urgent and important projects.

The findings show that the proposed approach allows to measure and evaluate the benefits and risks, to select and prioritize the projects successfully. Indeed, it can eliminate projects with low values and concentrate available resources exclusively on those meeting current and immediate needs. It provides an innovative way for universities to make the best selection of projects to be executed.

It seems that the combination of Cobit practices, AHP and TOSPIS approaches can offer a better solution to align the portfolio with the strategic objectives of universities.

VI. Conclusion

The selection and prioritization of projects is a strategic decision for universities. This paper aims to contribute to IT governance in universities by developing a project portfolio management approach adapted to public institutions. Several alternatives were evaluated using different criteria for projects selection. The evaluation of alternatives was conducted in the case of five projects and it was based on new multi-criteria analysis using the AHP-TOPSIS method and based on COBIT 5 framework practices.

This method, based on the identified criteria has determined the order of alternatives and identified the best ranked project among these alternatives. Based on the obtained ranks, decision-makers can conclude which of the alternatives must be prioritized. This paper provides an overview of the aspects that must be taken into consideration during the process of selecting and prioritizing projects.

The implementation of such an approach will be beneficial for project managers. It will enable them to lead information technology with effective practices and a standardized management framework.

It is found that the combined use of MCDM methodologies AHP and TOPSIS proves to be suitable for the implementation of the sub-process APO05.03 "*Evaluate and selects programs to fund*", it can correctly guides decision makers for evaluating projects and visualizing the importance of each criterion on alternatives before reaching a final decision.

Future research will focus on the implementation of the sub-process APO05.04 "*Monitor, optimize and report on investment portfolio*".

Acknowledgment

We express our deep sense of gratitude to Mr. J. EL KAISSI and Ms. N. ABOULFADL IT managers for their assistance with the projects evaluation and their contribution to carry out this study. They provided helpful comments and suggestions.

References

[1] M. D. Srivannaboon S, "A two-way influence between business strategy and project management," International Journal of Project Management vol. 24, pp. 493-505, 2006.

[2] P. Patanakul, "Key attributes of effectiveness in managing project portfolios", International Journal of Project Management Elsevier Publishing., vol. 33, p. 1084 – 1097, 2015.

[3] B. K. A. G. H. J. D. Unger, "Enforcing strategic fit of project portfolios by project termination: an empirical study on senior management

involvement", International Journal of Project Management , vol. 30, n° %16, p. 675–685, 2012.

[4] D. Jonas, "Empowering project portfolio managers: How management involvement impacts project portfolio management performance," International Journal of Project Management, vol. 28, pp. 818-831, 2010.

[5] C. e. a. Austin, " Application of Project Management in Higher Education.," Journal of Economic Development, Management, IT, Finance and Marketing, vol. 5, n° 12, pp. 75-99, 2013.

[6] S. D. H. a. W. V. Grembergen, "IT Governance and Its Mechanisms," Information Systems Control Journal, vol. 1, 2004.

[7] W. &. D. H. S. Van Grembergen, "Enterprise Governance of Information Technology: Achieving Strategic Alignment and Value," Springer, New York, 2009.

[8] ITGI, "Board Briefing on IT Governance," IT Governance Institute, 2001. [En ligne]. Available: www.itgi.org. [Accès le 03 2018].

[9] I. 38500, "International Organization for Standardization (ISO), ISO/IEC 38500 Corporate governance of information technology," Switzerland, 2008.

[10] S. H. T. J. A. &. V. G. W. De Haes, "Adoption and Impact of IT Governance and Management Practices: A COBIT 5. Perspective," International Journal of IT/Business Alignment and Governance (IJITBAG), vol. 7, n° 11, pp. 50-72, 2016.

[11] P. Z. K. &. K. R. L. Zhang, "Impact of IT Governance and IT Capability on Firm Performance," Information Systems Management, vol. 33, n° 14, pp. 357-373, 2016.

[12] P. &. R. J. Weill, "IT Governance: How Top Performers Manage IT Decision Rights for Superior Results," Harvard Business School Press, Boston, 2004.

[13] J. &. C. V. Bhattacharjya, "Adoption and Implementation of IT Governance: Cases from Australian Higher Education," chez 17th Australasian Conference on Information Systems, Association for Information Systems, , Adelaide, 2006.

[14] K. P. P. a. P. S. Jairak, "Information Technology Governance Practices Based on Sufficiency Economy Philosophy in the Thai University Sector," Information Technology & People, vol. 28 , n° 11, pp. 195-223, 2015.

[15] K. P. P. a. P. S. Jairak, "Information Technology Governance Practices Based on Sufficiency Economy Philosophy in the Thai University Sector," " Information Technology & People , vol. 28, n° 11, pp. 195-223, 2015.

[16] C. &. H. H.-H. Svensson, " Establishing a Business Process Reference Model for Universities," Procedia Technology, vol. 5, n° %10, pp. 635-642, 2012.

[17] E. C. f. A. Research, "Process and politics: IT governance in higher education.," chez ECAR (Educause Center for Applied Research), Colorado, 2008.

[18] ISACA, "COBIT Five: A Business Framework for the Governance and Management of Enterprise IT," ISACA, 2012. [En ligne].

[19] I. G. Institute, "COBIT 5: Enabling Process," IT Governance Institute, 2012. [En ligne]. Available: http:\\www.ITGI.org. [Accès le 2018].

[20] S. d. S. A. Ramlaoui, "Comparative study of COBIT with other IT Governance Frameworks," International Journal of Computer Science Issues, vol. 11, n° %16, pp. 95-101, 2014.

[21] I. 38500, "Corporate Governance of Information Technology. ISO/IEC. 2008," ISO/IEC 38500, 2008. [En ligne]. Available: www.iso.org . [Accès le 02 2018].

[22] M. J. P. Simonsson, "Model ¬Based IT Governance Maturity Assessments with COBIT," Department of Industrial Information and Control Systems, KTH, Royal Institute of Technology, 2008.

[23] R. M. Gomes J, "Improving project success: A case study using benefits and project management," Procedia Computer Science , vol. 100, pp. 489-497, 2016.

[24] V. C. G. b. R. Q. C. c. J. R. B. Pinyarat Sirisomboonsuk a, "Relationships between project governance and information technology governance and their impact on project performance," International Journal of Project Management JPMA, p. 14, 2017.

[25] PMI, " A guide to the project management body of knowledge," 2004.

[26] Project Management Institute, The standard for portfolio management, Pennsylvania, USA, 2006.

[27] B. S. B. a. P. Eskerod, "Project portfolio management - there's more to it than what management enact," International Journal of Project Management, vol. 26, p. 357–365, 2008.

[28] A. EL YAMAMI, S. AHRZ, K. MANSOURI, M. QBADOU et E. ILLOUSSAMEN, "Representing IT Projects Risk Management Best Practices as a Metamodel," Engineering, Technology & Applied Science Research, vol. 7, n° 15, pp. 2062-2067, 2017.

[29] T. Saaty, The analytical hierarchy process, New York: Mc Graw Hill, 1980.

[30] T. L. Saaty, "Decision making with the analytic hierarchy process," Services Sciences, p. 84, 2008.

[31] C. L. H. a. K. Yoon, "Multiple attribute decision making methods and applications," Springer-verlag, 1981.

[32] A. EL YAMAMI, S. AHRIZ, K. MANSOURI, M. QBADOU et E. ILLOUSSAMEN, "Developing an Assessment Tool of ITIL Implementation in Small Scale Environments," International Journal of Advanced Computer Science and Applications(IJACSA), vol. 8, n° 19, pp. 183-190, 2017.

[33] T. B. Robert J. Benson, From Business Strategy to IT Action: Right Decisions for a Better Bottom Line, wiley, 2004.

[34] P. G. Eng K. Chew, Information Technology Strategy and Management: Best Practices, IGI Global, 2009, p. 496.

[35] G. S. Lavoisier, Gestion de portefeuille de projets informatiques, 2007.

[36] G. L. B. J. L. M. A. C. G. &. P. C. Lunardi, " The impact of adopting IT governance on financial performance: An empirical analysis among Brazilian firms," International Journal of Accounting Information Systems, 15(1), 66-81, 2014.

[37] R. Peterson, "Crafting Information Technology Governance," Information Systems Management, 21(4), 7-22., 2004.

An Explorative Study for Laundry Mobile Application

Laundry Process Change

Doaa M. Bamasoud*, Asma M. Alqahtani, Eman A. Aljdea, Reem A. Alshomrani, Maha S. Alshahrani, Zohoor A. Alghamdi, Ameerah M. Alghamdi, Shahd F. Almaawi, Asrar D. Alshahrani

Computer Science Department
Faculty of Computers & Information Technology
University of Bisha, Asir, Saudi Arabia

***Abstract*—With the current rapid development of technology, many services need redesigning in order to keep up with customer demands. Therefore, organizations nowadays resort to redesigning services and business processes in order to maintain their competitiveness and success. With the recent advances in smartphone capabilities, and their growing penetration rate among individuals, organizations intend to take advantage of these devices by designing mobile applications to help evolve their business. The laundry business is one sector which has great potential for further development. Turning the ordinary routine of laundry into a service obtainable through a mobile application will contribute to reducing the burden of laundry tasks on individuals. This paper reviews the relevant literature and has design an instrument which investigates an individual's need for such mobile applications.**

Keywords—Business process change; smartphones; mobile application; laundry

I. Introduction

Through the fast-paced evolution of information technology, and the modern style of living, people face dilemmas when finding sufficient time to do home duties, such as laundry, cleaning and cooking. With the advent of smartphones over the past years, and with their enhanced capabilities and growing penetration rates among individuals, organizations intend to take advantage of these devices for developing mobile applications which evolve and market their business services and goods [1]-[3]. While laundry is a daily need for everyone, it is at the same time a burdening and time-consuming home duty. Consequently, developing a mobile application for bringing laundry services to a customer's doorstep will be effective and beneficial for both customers and service providers [4]. Through browsing literature, it can be determined that research seldom investigates the laundry sectors, and the redesigning of its business processes, through providing services online through mobile applications. Therefore, this research has been conducted in order to reinforce this area of knowledge. This paper has been structured into four main parts. The first section is an introduction, which introduces the topic, the research problem, and the research's significance. The second section presents the research's theoretical background, through a review of related literature. After that the research and data collection methodologies explain research methods, data collection procedures are undertaken, and the research's results are reported. Finally, a discussion and conclusion section presents the study's results.

A. Research Problem

People nowadays do not have sufficient time to complete home duties after work [4]. The development of information technology and mobile technologies has led to the development of electronic business (E-business), and has extended such business's competitive advantages [2]-[4]. Organizations intend to redesign their ordinary processes, and to change their business processes, in order to remain competitive, to obtain success, and to keep up with technology innovations [5], [6]. Therefore many business sectors have developed mobile applications designed to market their business services and goods [1], [3], [7].

B. Research Significance

One business sector which has not yet gained interest in research is the laundry business sector. As is commonly known, laundry is both a daily requirement for everyone, and a very burdensome task and time-consuming home duty. Therefore, developing a mobile application for bringing a laundry service to a customer's doorstep will be effective and beneficial for both the customer and the service provider [4], [8]. Research has seldom investigated the laundry sector, and its business process redesign through providing services online through mobile applications. Therefore, this research has been conducted in order to reinforce this area of knowledge.

II. Theoretical Background

A. Business Process Change

The currently rapid change and growth of technology has revolutionized interactions between customers and business providers. Consequently, organizational changes have been motivated by exterior innovations, rather than by the internal factors of organizations. In [9], [10], the authors emphasized that organizations which decline to adopt innovative changes and advancements can be subsequently left behind. Therefore, organizations which have attempted business process changes are likely to achieve spectacular performance improvements and attain competitive advantages [9], [11]. For the time being, organizations must face recent technology developments, and must determine how to adapt to these changes [3].

In [6], the authors defined business process change as being the "methodological process that uses information technology to radically overhaul business processes, and thereby attain major business goals". Therefore, the goal of process change serves to maintain competitiveness and maximize organization benefits [9]. The employment of IT in any business plays a key role in changing organizational business processes and achieving success [5]. In [3] the author stated that mobile applications are valued by mobile users, and therefore mobile applications should be paid full attention. Additionally, it is notable that laundry is substantially important for everyone. However, literature has seldom covered the relationship between laundry, and the services that can be provided to customers via mobile applications.

B. Impact of Technology

People have started to use technology to serve themselves. Technology has invaded all aspects of our lives and has become substantial for everyone [12]. Therefore technology has greatly influenced people's lifestyles. Presently, great progress has been witnessed in smartphone technology, and smartphones are no longer just a means of voice communication. Rather, they provide various capabilities which help transform traditional services into electronic ones. With the rapid development of the internet and mobile technologies, these technologies have led to the development of mobile business, which is becoming increasingly innovative and diverse [2], [4].

C. Laundry Services

The evolution of a technology-focused life has helped solve most problems faced by individuals and has harnessed technology to help serve them and accomplish their goals [3], [4], [5]. In [5], the researcher emphasized the importance of service innovations to customers, by including an addition to the current service mix, or by changing the existing services offered. Laundry is a daily and time-consuming home duty. However, research has seldom investigated the laundry sector, and the means through which its business processes have been redesigned via digital applications. Developing a mobile application for bringing laundry services to a customer's doorstep will be effective and beneficial for both customers and services provider [4].

III. Methodology

A. Instrument Development and Data Collection

The survey instrument has been developed based through prior literature, refer to Fig. 1. In this particular case, the research has been collected through web-based surveys. The targeted respondents have been from different age ranges, and both genders. The instrument has been divided into two sections, as presented in Table I. The first section includes the demographic details of respondents, including their age, gender, marital status, employment, place of living, monthly income, whether or not the respondents possess smartphones, and what operating systems their smartphones use. The second section is divided into six constructs. The first of these constructs relates to perceived usefulness. This construct is designed to measure the degree to which a person believes that using a proposed mobile application will enhance the process of doing their laundry. The second construct relates to the perceived ease of use, which measures the degree to which a person believes that using a mobile application for laundry would minimize their expended efforts. The third construct relates to behavioral intention towards using a laundry mobile application, measuring the user's intention to use a laundry mobile application. The fourth construct involves redesigning the original laundry routine, which concerns measuring the degree to which a person believes using a proposed mobile app would change their ordinary laundry routine. The service's fifth construct concerns the user's personal opinion about services included within the proposed mobile application. The last construct is that of user satisfaction, which measures the user's satisfaction with services provided through the laundry mobile application.

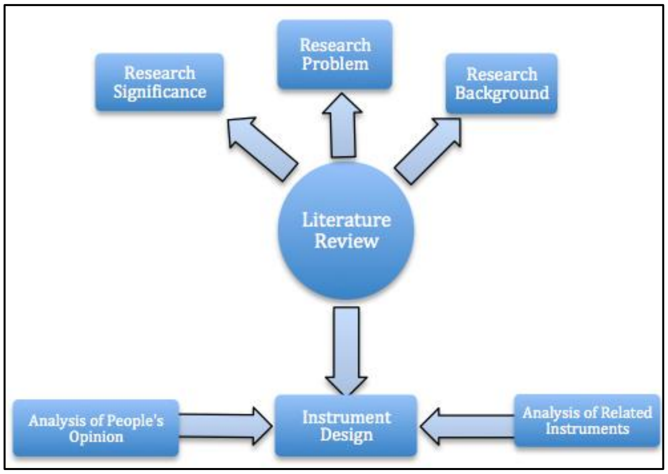


Fig. 1. Research methodology.

To ensure the developed instrument's clarity and content validity, experts must validate it [13]. In this study, three experts have validated the instrument. A 5-point Likert-type scale has been used, with a scale ranging from 1 (strongly disagree) to 5 (strongly agree). Table I describes the developed instrument.

TABLE I. Survey Instrument

Classification	**Survey Items**
Gender	Male Female
Age	Less than 18 years 18 years to 25 years 25 years to 35 years More than 35 years
Marital Status	Single Married
Living	Alone With Family With Friends
Employment	Student Employed Unemployed
Monthly Income	Less than 3000 S.R 3000 S.R to 5000 S.R 5000 S.R to 10000 S.R More than 10000 S.R
Possesses	Yes

Smartphone	No
Smartphone Operating System	iOS Android
Perceived Usefulness Construct	The degree to which the Laundry mobile application will enhance the effectiveness of clients' laundry. The laundry mobile application will facilitate the laundry work of clients. The laundry mobile application will increase client productivity. The laundry mobile application will be useful for clients.
Perceived Ease of Use Construct	I can easily learn how to use the Laundry mobile application. I can easily become a proficient in using the Laundry mobile application. I can use the Laundry mobile application very well if I use it for enough time. The laundry mobile application will make my laundry process easier. The use of the Laundry mobile application will not require excessive effort.
Usage Behavior Intention Construct	I plan to use the Laundry mobile application very often. I do not intend to take full advantage of the Laundry mobile application. I need to have a mobile app for laundry services. I will strongly recommend others to use a Laundry mobile application.
Redesigning Ordinary Laundry Routine Construct	Using a Laundry mobile application is likely to contribute to the success of changing my ordinary laundry routine. The Laundry mobile application will be useful in identifying the process of redesigning the ordinary laundry process. The laundry mobile application will facilitate identifying new opportunities for improvement.
Service Desired Construct	Do you support a partnership between the application and the laundry shops spread throughout the region? Do you prefer having a chat box with the person who does the service? Do you encourage the addition of an electronic payment service?
User's Satisfaction Construct	Do you support having an urgent laundry service? Would you like to have your clothes returned in more private bags? Do you support notifying clients when requests are completed? Do you support evaluating the service after completion? Do you support having separated laundry for clients, if desired?

B. Data Analysis

IBM SPSS 14 statistical software has been used to carry out the statistical analysis. The descriptive statistics of the collected data and the reliability test were assessed through SPSS. A total of 70 successful responses were used in the analysis. The collected data has been examined for missing data and for respondents' test bias. A description of the sample has been provided in Table II.

TABLE II. SAMPLE DESCRIPTION

	Label	Frequency (N=70)	Percentage (100%)
Gender	Male Female	22 48	31.43% 68.57%
Age	Less 18 years 18 -25 years 25 - 35 years More 35 years	8 30 13 19	11.43% 42.86% 18.57% 27.14%
Marital Status	Single Married	38 32	54.29% 45.71%
Living	Alone With Family With Friends	4 66 0	5.71% 94.29% 0%
Employment	Student Employed Unemployed	33 30 7	47.14% 42.86% 10%
Monthly Income	Less 3000 S. R 3000-5000 S. R 5000-10000 S.R More 10000 S. R	38 4 8 20	54.29% 5.71% 11.43% 28.57%
Possesses Smartphone	Yes No	70 0	100% 0%
Smartphone Operating System	IOS Android	44 26	62.86% 37.14%

TABLE III. SURVEY RESULTS

Construct	Measure	Cronbach's Alpha	Cronbach's Alpha if Item is Deleted
Perceived Usefulness	PU1 PU2 PU3 PU4	.866	.850 .828 .797 .844
Perceived Ease of Use	EoU1 EoU2 EoU3 EoU4 EoU5	.884	.858 .861 .843 .845 .885
Usage Behavior Intention	BI1 BI2 BI3 BI4	.722	.562 .866 .595 .550
Redesigned Original Laundry Routine	RedPro1 RedPro2 RedPro3	.778	.705 .669 .728
Services Desired	Dsrdsrvc1 Dsrdsrvc2 Dsrdsrvc3	.685	.606 .732 .419
User's Satisfaction	Usrsat1 Usrsat2 Usrsat3 Usrsat4 Usrsat5	.744	.655 .681 .700 .701 .751

Out of 70 respondents, 48 or 68.57% were females, and 22 or 31.43% were males. 11.43% of the respondents were younger than 18 years, while 42.86% were aged between 18 and 25 years, 18.57% were between the ages of 25 and 35 years, and 27.14% were over 35 years. 54.29% of the respondents were single, and 45.71% are married. 5.71% of the respondents lived alone, while 94.29% lived with their families, and none lived with their friends. 47.14% of the respondents were students, 42.86% were employed and 10% were unemployed. 54.29% of respondents had a monthly income of less than 3,000 SR, while 5.71% of them had a monthly income of between 3,000 and 5,000 SR, and 11.43% had a monthly income of 5,000 to 10,000 SR. 28.57% of respondents had a monthly income which exceeded 10,000 SR. All respondents possessed smartphones, with 62.86% owning iOS smartphones, and 37.14% owning android smartphones.

The authors used SPSS14 to validate the developed instrument, and deployed a scale reliability test for the proposed instrument measures. An internal reliability analysis has helped assess goodness of fit [14]. The most popular assessment of inter-item consistency reliability is the Cronbach's alpha, and accordingly it has been employed for the pilot data. In [15] the researchers stated that a Cronbach's alpha value of at least 0.5 is considered good for measures.

Table III has shown the statistical results of reliability. The overall Cronbach's alpha value for the perceived usefulness construct is 0.866, which is considered a high value. Within the perceived ease of use construct, the overall Cronbach's alpha value is 0.884, which is also considered to be a high value. Furthermore, the overall Cronbach's alpha of the usage behavior intention construct is 0.722, which is considered a good value. The redesign original laundry routine construct has an overall Cronbach's alpha value of 0.778, which is considered to be a high value. The services' desired construct has an overall Cronbach's alpha value of 0.685, which is considered a good value. Finally, the user's satisfaction construct has an overall Cronbach's alpha value of 0.744, which is considered a high value. By referring to the results in Table III, the Cronbach's alpha values for all six constructs are considered good. Therefore, there is no need to exclude and delete items of the constructs to raise the Cronbach's alpha of those constructs. Consequently, the researchers retain all items and measures of constructs.

IV. Discussion and Conclusions

This paper has highlighted change of business process, and its importance in meeting the requirements of individuals regarding technological advancements and innovations in the laundry sector. Furthermore, the paper has highlighted the current importance of smartphones, and the electronic services which can be offered through mobile applications and their contributions to redesigning and changing ordinary processes of business sectors. The authors emphasized meeting the needs of individuals, by exploiting and utilizing the smartphones, and by providing electronic services using mobile applications. Therefore, this paper aims to draw attention to the importance of how mobile applications can facilitate the changing of business processes within business sectors.

V. Future Work

Future work will focus on conducting a main study. Research has been planned in order to further explore the theoretical and practical aspects of how smartphones' mobile applications for the laundry sector can help ease the burden laundry places on people and can deliver services wherever they are. Moreover, the research plans to explore the influence of such mobile applications particularly in Saudi Arabia.

Acknowledgment

The authors Asma M. Alqahtani, Eman A. Aljdea, Reem A. Alshomrani, Maha S. Alshahrani, Zohoor A. Alghamdi, Ameerah M. Alghamdi, Shahd F. Almaawi, and Asrar D. Alshahrani would like to express their deep gratitude to: 1) University of Bisha; and 2) Ministry of Education Kingdom of Saudi Arabia for their support in pursuing their undergraduate study.

References

[1] Liébana-Cabanillas, F. Muñoz-Leiva and F.- Sánchez-Fernández, J., "A global approach to the analysis of user behavior in mobile payment systems in the new electronic environment," Service Business, vol. 12, 2018, no. 1, pp. 25-64.

[2] Ramakrishna, V. and Dey, K., "Mobile application and user analytics. In mobile application development, usability, and security," 2017, pp. 231-259.

[3] Rupnik, R. "Mobile applications development methodology," handbook of research in mobile business: technical, methodological and social perspectives, 2009, pp. 160–172.

[4] Priambodo, B. and Ani, N., "M-Laundry advertisement system utilizing location based service," international journal of advanced research in computer science and software engineering, vol. 6,2016, no. 6, pp.59–63.

[5] Ye, H. and Kankanhalli, A., "User service innovation on mobile phone platforms : investigating impacts of lead userness, Toolkit Support, and Design Autonomy," MIS quarterly, vol. 42, 2018, no. 1, pp. 1–47.

[6] Grover, V. Teng, J. T. C. and Fiedler, K. D., "Information technology enabled business process redesign: an integrated planning framework,"Omega,vol. 21, 1993, no. 4, pp.433–447.

[7] Baloh, P. and Trkman, P.,"Influence of internet and information technology on work and human resource management," Informing Science, vol. 6, 2003, pp. 498-505.

[8] Jamgade, A., "Online laundry service," international journal of scientific research in computer science. Engineering and Information Technology, vol. 2, 2017, no. 1, pp. 195–197.

[9] Kasemsap, K., "The role of business process reengineering in the modern business world," In I. management association (Ed.). Project management: concepts, methodologies, tools, and applications, 2016, pp. 1802-1829.

[10] Skelsey, D., "Why do people in business resist change?," Project Laneway. Available on: http://archive.is/jzSul (2003). Accessed 9 January 2018

[11] Grover, V. Teng, J., Segars, A. and H.- Fiedler, K., "The influence of information technology diffusion and business process change on perceived productivity: The IS executive's perspective," Information and management, 1998, pp. 141–159.

[12] Zulkefly S. N. and Baharudin R., "Mobile phone use amongst students in a university in Malaysia: Its correlates and relationship to psychological health," European Journal of scientific research, vol. 27, 2009, no. 2, pp. 206-218.

[13] Boudreau, M. C. Gefen, D. and Straub, D. W., "Validation in information systems research: A state-of-the-art assessment," MIS quarterly, vol. 25, 2001, no. 1, pp. 1-16.

[14] Cavana R. Delahaye B. and Sekaran U., "Applied business research: qualitative and quantitative methods," Queensland: John Wiley and Sons, Australia, 2001.

[15] Schmidt F. and L.-Hunter J., "General mental ability in the world of work: Occupational attainment and job performance," Journal of personality and social psychology, vol.86, 2004, no. 1, pp. 162-173.

Extraction of ERP Selection Criteria using Critical Decisions Analysis

Motaki Noureddine, Kamach Oualid
National School of Applied Sciences of Tangier
Tangier, Morocco

***Abstract*—Companies use ERP systems to automate business processes in order to increase productivity, reduce costs, and meet customer requirements. ERP selection for an enterprise is a decision-making project that is both risky and expensive, a wrong selection of this system or project partners can lead to the failure of the ERP implementation project. In this paper, we combined the theoretical findings of ERP selection issue with expert's practical recommendations to determine the critical decisions that need to be made in the pre-implementation phase. Then we presented a methodology to determine ERP selection criteria based on the critical decisions analysis. A part of this work was performed within a company that has just launched an ERP implementation project.**

Keywords—ERP selection; criteria; critical decisions; implementation; information system

I. Introduction

Companies need to increase productivity, decrease total costs, reduce stock, meet the customers' requirements, maximize return on investment (ROI) and reduce lead times. To overcome these challenges and improve efficiency to be competitive, companies often use ERP systems.

An enterprise resource planning system (ERP) is a software package composed of business applications (modules) that automate core corporate activities and the set of business tasks with the aim of better support enterprise's top managers' decision making.

ERP systems are regarded as a way of becoming and maintaining competitiveness, with the use of ERP, different components of information system coming from different parts of the organization can be unified and stored in a centralized database. Indeed, main corporate activities such as inventory control, purchase, production planning, sales, manufacturing, supply chain management, human resources and finance can be integrated and automated through a several modules included in ERP system and meet most of company's requirements.

Many improvements can be achieved by successfully implementing an ERP: this system offers many benefits for enterprise such as improved supply chain management, instant access to reliable information, elimination of multiple data entries and redundant operations, time saving and costs reduction. In addition, productivity is achieved through ERP systems by automating, integrating and standardizing business processes.

ERP system's adoption considered as one of the most critical investment projects due to the high cost, complexity and adaptation risks. Companies are spending huge budgets and hiring work teams for a considerable period of time using both of its internal resources and external consulting to implement an ERP system.

Given the diversity of companies' business processes, strategies, goals, services and business sectors, ERP systems in market cannot fully meet all special business requirements of companies and satisfy completely their needs. Many companies implement their ERP software without given enough importance to the selection phase and without understanding requirements that must be included in chosen ERP package. A wrong selection may lead to unsuccessful ERP implementation project that can affect the performance of the company.

From this aspect, to make ERP implementation project successful, managers must initiate an ERP selection project whose objective is to choose suitable ERP software that most closely suits its requirements. Moreover, Selecting an ERP system is an extremely risky, high cost and difficult decision-making problem for managers. It's a whole process which requires more than interviewing a few vendors: it is considered as the most critical success factor for ERP implementation.

For ERP selection problem, researchers proposed various methods that differ from each other based on their complexity and tools used. A part of these methods are based on mathematical optimization, scoring and ranking techniques while others use multi-criteria decision analysis approaches. All ERP selection methods use selection criteria that can evaluate ERP alternatives.

During the ERP selection process, the criteria list must be determined; however, there is no unique way to classify these criteria. There are little works that has proposed a comprehensive methodology to extract and determine the criteria for selecting an ERP in an industrial context [15], [16], [24].

In this paper, we will present a methodology for extract the selection criteria based on the critical decisions that must be made by steering team. We will start by determining the list of critical decisions according to the relevant literature on ERP selection in one hand, and the practical recommendations of the ERP experts we interviewed in other hand. Then we will describe our methodology which generates selection criteria from each critical decision.

II. Literature Review

In the recent literature there are various studies proposing different techniques and methods to provide a solution to ERP selection problem, several evaluation methods have been proposed, such as scoring, ranking, mathematical optimization and multi-criteria decision analysis [8].

Numerous studies proposed multi-criteria decision-making methods (MCDM) to prioritize alternatives and calculate the relative efficiencies of ERP solutions [1], [2], [3], [7], [9]. A part of these studies are successfully combined more than one multi-criteria decision-making methods (hybrid methodology) [2], [5], [6], [7], [21], while other studies have focused on identifying success factors in ERP selection process [10], [11], [20].

The review of the state of the art on ERP selection Methods revealed the following methods:

AHP (Analytic Hierarchy Process) [1], [13], [16], [17], [23], ANP (Analytic Network Process) [5], [14], [15], [23], PROMETHEE [2], [5], SHERPA (Systematic Help ERP Acquisition), FL (Fuzzy Logic) [3], [12], [15], [17], [18], PM (Priority Matrix), TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) [24], [25], SMART [4].

In previous years, researchers conducted theoretical and empirical studies to identify and analyze the selection criteria for a suitable ERP system [22].

Based on data from 19 Canadian organizations that have adopted the ERP system or are evaluating the adoption of an ERP system, Kumar, Kumar, and Maheshwari presented the most used ERP selection criteria. According to this study, some of these criteria include Functionality of the system (79%), Systems reliability (64%), Fit with organization systems (64%), Available business best practices in the system (50%), Cross module integration (50%), System using latest technology (43%), Vendor reputation (43%) [24].

Little research has addressed the issue of criteria identification in ERP selection context: (Chun-Chin Wei et al.) proposed a method to systematically identify the appropriate selection criteria, through the construction of ERP selection objectives.

The method proposed is based on enterprise strategic objective and presented in several steps [16]:

1) Define enterprise strategic objectives.

2) Define ERP system scope by project team.

3) Drive the ultimate goals from the strategic objectives

4) Structure the objectives by distinguishing fundamental-objectives from means-objectives in the objective development process.

5) Establish the hierarchy of ERP system fundamental-objectives, using top-down decomposition method or bottom-up synthesis method.

6) Create a means-objective from the fundamental-objectives and determine linkages among them.

7) Extract the attributes (criteria) used for ERP evaluation: quantitative and qualitative attributes that satisfy the strategies and goals of the company should be involved.

In this study, the authors explained step by step how to construct a specific objective structure relating to the company's strategies and how to extract the proper criteria for evaluating the fulfillment of company's requirements.

Baki and Çakar, presented results from a study on ERP package selection criteria in 55 Turkish manufacturing companies from variety of industries, they proposed a criteria list that include fit with parent/allied organization systems, better fit with organizational structure, functionality, system reliability, technical criteria, compatibility with other systems, cost, vision, ease of customization, service and support, market position of the vendor, domain knowledge of vendor, references of the vendor, methodology of the software and consultancy, cross module integration, implementation time [26].

To determine a set of ERP selection criteria Ayağ and Özdemi analyzed a set of companies that have already implemented an ERP system. They observed how companies defined the selection criteria for the adoption of their ERP, According to the authors, the ERP selection criteria can be classified into three determinants that have relationships with each other: competitive advantage, productivity, profitability. Under the three determinants, seven dimensions are listed: system cost, vendor support, flexibility, functionality, reliability, ease of use and technology advance. These dimensions play an important role for each determinant and affect each other (Fig. 1). Ultimately, 22 criteria are determined. All the criteria of a dimension are linked and influence each other positively or negatively [15].

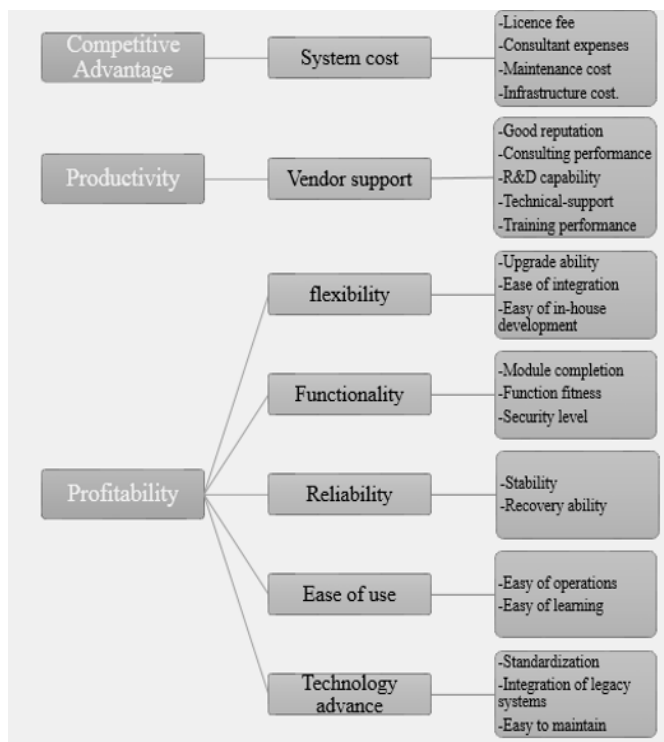


Fig. 1. Determinants, dimensions and criteria according to Ayağ and Özdemi.

Based on the research related to ERP selection topic, it is clear that the diversity of selection criteria complicates their classification into standard groups.

Frequently, ERP selection criteria are defined according to researchers specialty or based on the results found in literature, these criteria are divided in general into several categories which differs from a research work to another, and as we mentioned above, there are little works that has proposed a comprehensive methodology to extract and determine the criteria for selecting an ERP in an industrial context.

In a large number of the works carried out to find a solution to ERP selection issue, researchers used a list of the criteria to evaluate ERP systems through different methods (AHP, ANP, TOPSIS, and fuzzy Logic) but without explaining how they have obtained these criteria.

III. Research Methodology

The objective of this paper is to present critical decisions for selecting ERP system and to propose a simple and practical method that uses these critical decisions to determine the selection criteria list. Through this method, steering team can extract and evaluate principal criteria for ERP selection by making decisions that will take into account company constraints and requirements.

Our proposal can be summarized by the following research questions:

- Are there techniques for defining and evaluating selection criteria?
- What are the decisions to make before starting the ERP selection process?
- What is the impact of the critical decisions on ERP selection criteria?
- How can we extract the selection criteria from the decisions made by the steering team?

To perform this work, the relevant literature on ERP selection was analyzed, summarized and complemented by information available in cases studies and surveys, we raised up the most popular ERP selection methods and criteria. Furthermore, we reviewed several ERP selection processes to determine their strengths and limitations.

Moreover, theoretical findings related to the ERP selection were combined with ERP practical recommendations in order to derive insights.

Indeed, we conducted a series of meeting and interviews with ERP Consultants, projects managers, various department heads, who already participated in ERP implementation projects.

Accordingly, we were able, through these interviews, to extract the best practices followed, and identify different problems encountered in the ERP selection phase to improve the proposed method.

In fact, a part of this work was performed within a company that has just launched an ERP implementation project. This enterprise is located in the north of Morocco, and is considered as a member of a holding company that already uses a SAP ERP.

IV. Selection Process

The purpose of the ERP selection process is to choose the appropriate ERP system that can meet the enterprise requirements. This process involves several stages starting from the constitution of the selection team until signing the contract with the vendor [19].

Different models available to represent the selection process, we present below the most common stages in these selection process models:

Step 1: Constitution of the selection team: The first step in the selection process is to form the selection team that include IT experts and key members from each department with suitable competencies and knowledge. This team must collect the needs of each department and identify the characteristics of the ERP system, as they have to establish a short list of vendors…

Step 2: Determination of the ERP required functionalities: The objective of this step is to determine functional specifications related to the new information system (based on ERP), including functionalities, constraints, and management rules associated to the company functions. These technical and functional specifications will be used to evaluate each alternative ERP solution.

Step 3: Definition of the evaluation criteria used to select the suitable ERP provider: ERP selection criteria will be determined by the selection team in collaboration with the steering team based on the company constraints, the consultants' advices and the critical decisions (purpose of our contribution) made in the pre-implementation phase.

Step 4: Definition of the ERP shortlist: In order to create an ERP short list, selection team conducts market analysis and identifies potential suppliers. Indeed, this team takes care of the collection of all available information about ERP systems, and identifies those who are specific to the industrial activity of the company, and to find out the ERP systems employed by similar companies (benchmarking).

Step 5: Evaluation of the ERP according to the selection criteria: In this step, weights associated to the selection criteria will be determined after a consultation session performed by the selection team. Thereafter, alternative ERP systems will be evaluated using multi-criteria decision methods such as AHP, ANP, and TOPSIS. These methods can be applied in order to deal with the ambiguities involved during the evaluation of ERP alternatives.

Step 6: Selection of the suitable ERP: After evaluating each ERP product on the short list, the company may start by inviting the chosen ERP vendors in order to do a Demonstration of the Product and then later, they can start the Negotiations before signing the contracts.

V. Criteria List

We propose in this section an original vision of criteria, based on the stakeholders of ERP implementation project: vendor, integrator (consultants), client and partners (Fig. 2):

- Vendor: The enterprise that has developed the ERP system, it provides technical support for the ERP. For some criteria, we will confuse vendor and ERP product.
- Integrator: Represent functional and technical consultants and experts involved in various business processes, it ensures the accompaniment of the client in the ERP implementation project
- Client: Enterprise that will use the ERP (project owner); it generally initiates the ERP selection process.
- Partners: Customers, suppliers and group's members in case the client is part of a holding company. In the selection process, partners' information systems should be taking into account.

Fig. 2. Stakeholders of ERP implementation project.

Table I summarizes the selection criteria of the ERPs according to the stakeholders.

This structuring allows organizing selection criteria according to the chronological order of ERP selection process. Indeed, criteria related to the Client (the company that will use the ERP) and its partners, are the most critical and must be evaluated first. These criteria represent the results of a deep analysis of the company's business needs and must reflect its objectives and strategy.

The ERP product and vendor criteria focus on the evaluation of functional and technical aspect of the ERP and the quality of services offered by vendor, as well as its market position. The fourth element of this structure is the integrator criteria that assess the quality of experienced integrators for a given ERP, its related implementation methods and the overall implementation costs.

For our proposed method, this structuring will be used as a reference to determine selection criteria list that better matches the company's requirements. Thus, we can see that the criteria list generated according to our method will be a subset of the reference list.

VI. CRITICAL DECISIONS

Several critical decisions (Table II) must be taken by selection team before starting the selection process; these decisions should allow the determination of the ERP selection criteria adopted for the assessment of each alternative ERP (Fig. 3). Indeed, for each decision element, selection team can choose from several options. Subsequently, for each chosen option, one or more ERP selection criteria will be generated (Fig. 5).

TABLE I. ERP SELECTION CRITERIA ACCORDING TO THE STAKEHOLDERS

Stakeholder	Criteria
Vendor/Product	• Vendor market position. (C1) • References. (C2) • Financial position. (C3) • Reputation in the field (C4) • Technical support (C5) • Training support (C6) • Service & support cost (C7) • Product License Cost (C8) • Functionality. (C9) • Implementation of Desired Business Processes. (C10) • Short Implementation Time. (C11) • Ergonomic Software. (C12) • The provision of best practices. (C13) • Latest trends in the IT industry. (C14) • Modules independency. (C15) • The ability to integrate different platforms and data. (C16) • System stability. (C17) • Flexibility in adjusting demands according to business requirements. (C18) • Ease of use / implementation. (C19)
Client	• Enterprise size. (C20) • Activity area. (C21) • Desired Business Processes. (C22) • Enterprise budget. (C23) • Technical Infrastructure. (C24)
Integrator	• The provision of experienced integrators in the alternative ERP implementation. (C25) • The implementation methodology adopted by the integrator. (C26) • Integrator's ERP implementation experience in a similar industry. (C27) • Implementation cost. (C28) • Training cost. (C29) • Development cost (C30) • Average duration of alternative ERP implementation. (C31) • Integrator's support after going live. (C32)
Partners	• ERP systems used by customers and/or suppliers. (C33) • Consultant's suggestions. (C34) • The level of use of the ERP by competing enterprises or enterprises whose business sector is the same. (C35) • Customer and Supplier Needs. (C36)

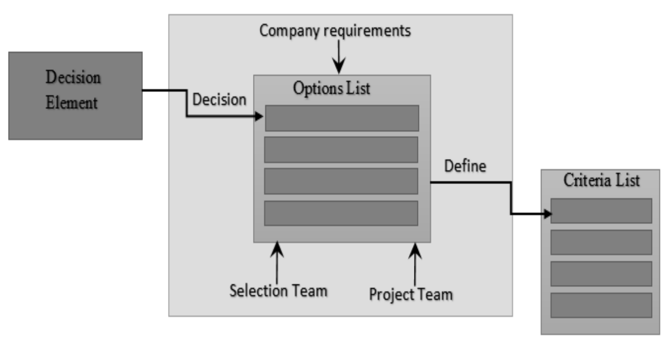


Fig. 3. General model: extraction of ERP selection criteria.

TABLE II. ERP SELECTION CRITICAL DECISIONS

Critical Decision	Code
ERP Type	CD1
Deployment Type	CD2
Target Processes	CD3
ERP Adaptation Level	CD4
Existing SI situation	CD5
Existing Provider situation	CD6
Partners recommendations	CD7
Technical Implementation type	CD8
Implementation team involved	CD9
ERP Project Budget	CD10
ERP Project Duration	CD11

A. Critical Decision 1: ERP Type

There is a wide variety of ERP systems designed to help SMEs and large companies to manage their information systems, these ERPs differ from each other according to various aspects such as the business sector, management best practices, target company size, integrated process complexity, and functional coverage of each proposed module.

The first critical decision is to determine the type of ERP that will be implemented: The steering committee must choose between an ERP for SMEs and an ERP for large companies. This decision makes it possible to define two major criteria: the type of ERP that must fit with the company size, and the ERP functional coverage according to the company activity area.

B. Critical Decision 2: Deployment Type

Firstly, the project team must define the strategic vision of the enterprise towards the ERP implementation project, and determine the parts of the company (sites, departments) concerned by this project. They will answer the basic questions regarding the aspect of the deployment: Will the ERP deployment be performed by module, by process, by site? Or it is a full deployment that concerns all functional parts of the enterprise.

This decision allows to determine specific ERP modules that will be used, and to specify if these modules must be independent or not, and finally, to determine the ERP implementation method that corresponds to the deployment type adopted.

C. Critical Decision 3: Target Processes

What are the organizational processes that will be managed by the ERP? What functionalities would be needed to manage each process? Through these questions, the project team can describe in details a list of processes that the enterprise wants to manage, and then determine ERP modules that will be used to achieve this objective.

The project team should determine all the business processes that will be subject to automation through the ERP: a preliminary study of existing user environment and various departments' needs must take place to be able to define the company's requirements, extract the requested functionalities, and then formulate the ERP functional criteria.

This Decision element will be used to establish multiple selection criteria used during the selection process. Moreover, it makes it possible to fill in associated criteria such as ERP compatibility with the enterprise processes, and inclusion of management best practices.

D. Critical Decision 4: ERP Adaptation Level

The project team will determine the functional implementation strategy by specifying whether the ERP must adapt to the company's processes through specific developments, or it is the company that has to adapt to the ERP by investing in business process reengineering (BPR).

The determination of functional implementation strategy of the ERP will have a direct impact on the adopted criteria: if we choose a strategy based on the adaptation of the ERP then the criteria adopted will be the development cost and the flexibility of the ERP, while in case we chose to adapt the company's business processes, two criteria will be adopted: functionality and inclusion of management best practices (Fig. 4).

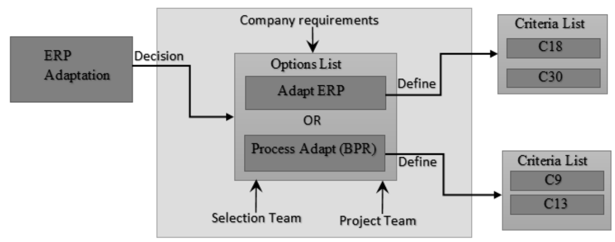


Fig. 4. Criteria extraction example according to CD4.

E. Critical Decision 5: Existing SI situation

Generally, a company already uses several application packages to manage data from different departments. Each application provides several functions, and uses its own database, but does not necessarily communicate with other applications.

Based on the needs specified in a previous step, the steering team must decide about which specific applications will be replaced by the ERP modules, and which ones need to be interfaced with the ERP.

If the steering committee decides to replace all the application packages with the ERP then this decision makes it possible to choose the appropriate criteria from functional coverage and integrity (C9), and implementation of desired business processes (C10), whereas in the case where the ERP must communicate with other applications, the adopted criterion will be the ability to integrate different platforms and data (C16).

F. Critical Decision 6: Existing Provider situation

Some companies already have previous experience with an existing ERP vendor, it consists in providing and implementing ERP product to manage one or more business processes in the company sites and subsidiaries. The success or failure of this previous experience can have an impact on the enterprise decision regarding the acceptance of this vendor for a new ERP implementation project.

Indeed, if this vendor receives highly favorable reviews, then its ERP product will be privileged compared to other ERP alternatives, otherwise this ERP will be eliminated from the alternatives shortlist automatically.

G. Critical Decision 7: Partners recommendations

Each company has different type of partners (suppliers, customers, subsidiaries, group members in the case of a holding company), the steering team can discuss if the ERP adopted by the business partners of the enterprise will be taken into consideration in the selection process.

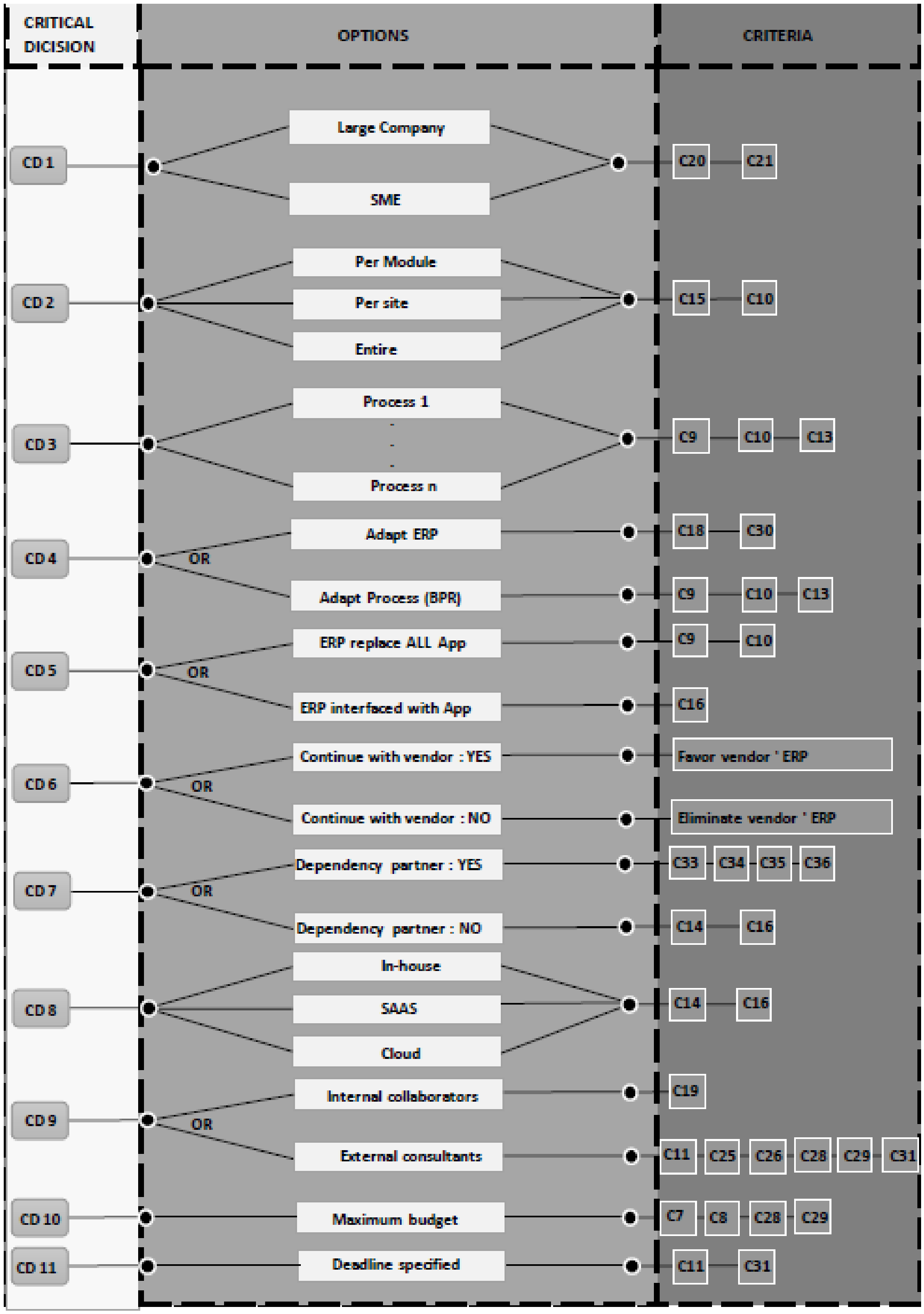


Fig. 5. Criteria generated according to the critical decisions.

This decision has a critical effect on the ERP selection Process, especially if partners require that a particular ERP must be used, in this case the ERP adopted by the partner will be favored compared to other alternatives. In the opposite case, the selected ERP must be able to communicate with the Information System of the partners and support the exchange with different systems and data, hence the need to adopt the C14 and C16 criteria.

H. *Critical Decision 8: Technical Implementation Type*

Before launching the selection process, the technical implementation type of the ERP must be defined; several types of installations can be used: in-house installation, outsourced installation based on datacenters, software as a service (SAAS), and ERP installation based on Cloud.

In the case of an internal installation: does the company keep the same IT strategy: operating system, DBMS, network architecture? Or will it be aligned with the requirements of the chosen ERP?

The suitable ERP must comply with all the technical requirements of the company, and particularly the installation method (Fig. 6).

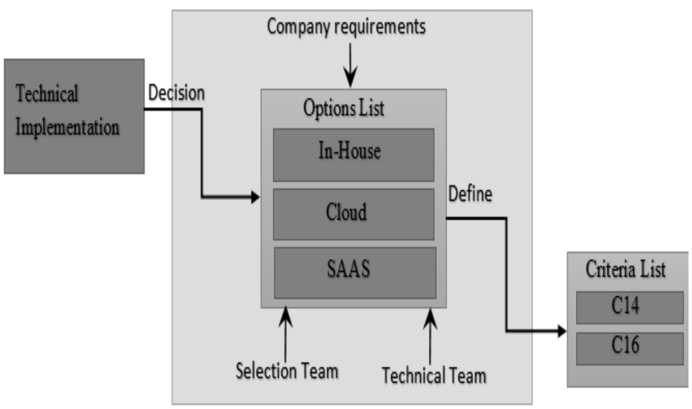


Fig. 6. Criteria extraction example according to CD8.

I. *Critical Decision 9: Implementation team involved*

During an ERP project, the integrator can support the company in the different phases of ERP implementation: starting from the requirements specification to the Go Live phase through the collaboration with a set of technical and functional consultants that have a good experience in this type of projects.

Some companies prefer that the ERP implementation should be performed entirely by their internal teams. In this case, ERP ease of use and implementation method simplicity are two essential criteria in the ERP evaluation Process. In the opposite case, the company can delegate the implementation of the ERP to an experienced integrator, therefore six ERP selection criteria will be defined (Fig. 5). These criteria evaluate implementation methodology of each alternative ERP and the quality of associated integrators.

J. *Critical Decision 10/11: ERP Project Budget and Duration*

The selection team must be aware of the maximum budget set by the company for ERP project as well as the planned date to start the Go Live phase. These two elements (budget and deadline) make it possible to define criteria evaluating methodology and duration of the ERP implementation on the one hand, and all the explicit and hidden costs of the ERP implementation project on the other hand.

The following table (Table III) summarizes all the questions that must be answered in the decision-making phase (critical decisions).

TABLE III. CRITICAL DECISIONS QUESTIONS

Decision element	Questions
CD1	What is the type of ERP to adopt? An ERP for SMEs or an ERP dedicated to large companies?
CD2	What kind of deployment will be used?
CD3	What are the organizational processes that will be managed by the ERP? What functionalities would be needed to manage each process?
CD4	What level of adaptation will be considered? What is the implementation strategy that should be adopted?
CD5	What will be the new situation of the business applications of the company? Will all these applications be replaced by ERP? Are there any applications that need to be interfaced with ERP?
CD6	In case of a previous experience with an existing ERP vendor, will this ERP provider be taken into consideration in the selection process?
CD7	The fact of "ERP adopted by enterprise partners?" will be taken into consideration in the selection process? The fact of "ERP recommended by consultants or partners?" will be taken into consideration in the selection process?
CD8	What kind of technical implementation will be adopted by the company?
CD9	Will the ERP implementation project be performed by an internal team or by an external integrator?
CD10	What is the budget allocated by the company for this ERP project?
CD11	What is the desired duration for the implementation of the ERP solution?

VII. DISCUSSION AND CONCLUSION

Usually the ERP selection has a profound impact on the success or failure of the implementation project, and should be planned very carefully by the selection team.

The strategic vision of the enterprise towards the ERP implementation project should be clear and unique for all the people involved in this project. Moreover, the requested functionalities and the business processes that will be managed by the ERP must be clearly identified before starting the selection process.

ERP system should be flexible and customizable and generally encompasses best practices in a given industry. the steering committee must determine adaptation level of the ERP that will be adopted, If changes to the company business process are required, the means involved and the change's cost must be clearly identified and communicated to financial decision maker, Else if changes to the ERP itself will be needed, then in this case, the cost of customizing the ERP through specific developments must be determined. The ERP selection team must understand that the decision to customize the ERP or change the business process will have an impact on the criteria that will be used to evaluate the ERP.

In addition, the selection team must identify the type of technical implementation that will be used to install the ERP then determine the new situation of legacy system: identify the applications that will be replaced by ERP, and those that must be interfaced with this new system.

Several charges and expenses can be seen during an ERP project: license cost, integration cost, maintenance cost, training cost, and data migration cost. The budget allocated to the implementation project is a key factor for the ERP selection, the estimated cost of implementation of each alternative ERP must be calculated carefully and compared against the budget allocated by the company for the ERP project.

The complexity of the implementation methodology of a given ERP has a direct impact on the project duration, the selection team must take into account this estimated duration to evaluate each alternative ERP.

Selection team and steering committee must understand that the selection criteria adopted for the ERP evaluation depends on the critical decisions made before starting the selection process.

In this work we have determined a list of critical decisions that must be made by the selection team and the steering committee, these decisions concern all the functional, technical and organizational aspects encountered in an ERP implementation project and which has a direct impact on the selection process.

The proposed methodology was partially developed in a company that just started the ERP selection project, we were able to take advantage of the opinions of the ERP consultants and the business process managers on the problems encountered and the critical decisions that must be made to succeed the project Selection.

On the theoretical side, a part of our methodology has been based on the analysis of the literature on ERP selection and the results obtained in case studies and surveys.

We have raised the most popular ERP selection criteria and then structured them according to the stakeholders of an ERP project in order to respect the chronological order of the ERP implementation project evolution.

In addition, after presenting the critical decisions for the selection of the ERP, we proposed various possible options for each decision, and thus determine the criteria generated according to the chosen option.

This methodology has the advantage of being practical and simple to use, moreover it takes into account the technical, functional and organizational constraints of the company and its specific needs. All the criteria generated reflects the company's requirements and will be determined from an in-depth discussion by the selection and steering teams. This will ensure that the ERP system selected through these criteria will be most suited to the company needs.

REFERENCES

[1] Maria Manuela Cruz-Cunha, "ERP Selection using an AHP-based Decision Support System", IRMJ journal, vol. 29, Issue 4, 2016.

[2] Sandarbh Shukla, Mishra, Rajeev Jain, Yadav, "An integrated decision making approach for ERP system selection using SWARA and PROMETHEE method", International Journal of Intelligent Enterprise, vol. 3, Issue 2, 2016.

[3] Burak Efe, "An integrated fuzzy multi criteria group decision making approach for ERP system selection", Applied Soft Computing, vol. 38, pp. 106–117, January 2016.

[4] Moutaz Haddara, "ERP Selection: The Smart Way" Procedia Technology, Vol. 16, pp. 394-403, 2014.

[5] Huseyin Selcuk Kilic , Selim Zaim and Dursun Delen, "Selecting "The Best" ERP system for SMEs using a combination of ANP and PROMETHEE methods", , Expert Systems with Applications, Vol. 42, Issue 5, pp. 2343–2352, 1 April 2015.

[6] Kilic, Zaim, and Delen, "Development of a hybrid methodology for ERP system selection: The case of Turkish Airlines", Decision Support Systems Journal, vol. 66, pp. 82–92, October 2014.

[7] Gürbüz, Alptekin and Isıklar Alptekin, "A hybrid MCDM methodology for ERP selection problem with interacting criteria", Decision Support Systems Journal, vol. 54, Issue 1, pp. 206–214, December 2012.

[8] Lin, Chen, Ting, "An ERP model for supplier selection in electronics industry", Expert Systems with Applications Journal, vol. 38, Issue 3, pp. 1760–1765, March 2011.

[9] Karsak, Özogul, "An integrated decision making approach for ERP system selection", Expert Systems with Applications, vol. 36, Issue 1, pp. 660–667, January 2009.

[10] Malhotra and Temponi, "Critical decisions for ERP integration: Small business issues", International Journal of Information Management, vol. 30, Issue 1, pp. 28–37, February 2010.

[11] Hakim and Hakim, "A practical model on controlling the ERP implementation risks", Information Systems Journal, vol. 35, Issue 2, pp. 204–214, April 2010.

[12] Sen, C. G., Baraçlı, H., Sen, S., & Baslıgil, H, "An integrated decision support system dealing with qualitative and quantitative objectives for enterprise software selection". Expert Systems with Applications, vol. 36, pp. 5272–5283, 2009.

[13] Ünal and Güner, "Selection of ERP suppliers using AHP tools in the clothing industry", International Journal of Clothing Science and Technology, vol. 21 Issue: 4, 2009.

[14] Perçin, "Using the ANP approach in selecting and benchmarking ERP systems", Benchmarking: An International Journal, vol. 15 Issue: 5, pp. 630-649, 2008.

[15] Ayag˘ and Özdemir, "An intelligent approach to ERP software selection through fuzzy ANP", International Journal of Production Research, vol. 45, pp. 2169-2194, Issue 10, 2007.

[16] Wei, Chien and Wang, "An AHP-based approach to ERP system selection", International Journal of Production Economics, vol. 96, Issue 1, 18 April 2005, pp. 47–62, 2005.

[17] Özdağoğlu and al, "Comparison of AHP and Fuzzy AHP for the Multi-Criteria Decision Making Processes with Linguistic Evaluations", 2007.

[18] Yu-Rong Zeng, Lin Wang & Xian-Hao Xu, "An integrated model to select an ERP system for Chinese small- and medium-sized enterprise under uncertainty", Technological and Economic Development of Economy, pp. 38–58, 2 November 2015.

[19] Igor Rivera and María del Rosario Pérez Salazar ,"Guía de Selección de Erp En Las Pequeñas Y Medianas Empresas Mexicanas" , 2013.

[20] Wen-Hsien Tsai, Pei-Ling Lee, Yu-Shan Shen and Hsiu-Ling Lin , "A comprehensive study of the relationship between entrerprise resource planning selection criteria and enterprise resource planning success" , 2013.

[21] Huseyin Selcuk Kilic, Selim Zaim and Dursun Delen ,"Selecting "The Best " ERP system for SMEs using a combination of ANP and PRPMETHEE methods", vol. 42, Issue 5, pp. 2343-2352, 1 April 2015.

[22] Moutaz Haddara and Ondrej Zach, "ERP Systems in SMEs: A Literature Review", 2011.

[23] Cristina Lópeza, Alessio Ishizakab , "GAHPSort: A new group multi-criteria decision method for sorting a large number of the cloud based ERP solutions", Computers in Industry , pp. 12–24, 2017.

[24] V. Kumar , B Maheshwari & U. Kumar , "Enterprise resource planning systems adoption process: A survey of Canadian organizations", International Journal of Production Research, vol. 40, Issue 3, pp. 509-523, 2010.

[25] Rekha Gupta and S. Kazim Naqvi, "A Framework for Applying CSFs to ERP Software Selection: An Extension of Fuzzy TOPSIS Approach",International Journal of Intelligent Information Technologies (IJIIT), vol. 13, Issue 2 ,2017.

[26] Baki, B., & Çakar, K., "Determining the ERP package-selecting criteria: The case of Turkish manufacturing companies", Business Process Management Journal, vol 11, Issue 1, pp. 75–86, 2005.

Performance Analysis of Artificial Neural Networks Training Algorithms and Transfer Functions for Medium-Term Water Consumption Forecasting

Lemuel Clark P. Velasco, Angelie Rose B. Granados, Jilly Mae A. Ortega, Kyla Veronica D. Pagtalunan
Mindanao State University-Iligan Institute of Technology
Iligan City, The Philippines

***Abstract*—Artificial Neural Network (ANN) is a widely used machine learning pattern recognition technique in predicting water resources based on historical data. ANN has the ability to forecast close to accurate prediction given the appropriate training algorithm and transfer function along with the model's learning rate and momentum. In this study, using the Neuroph Studio platform, six models having different combination of training algorithms, namely, Backpropagation, Backpropagation with Momentum and Resilient Propagation and transfer functions, namely, Sigmoid and Gaussian were compared. After determining the ANN model's input, hidden and output neurons from its respective layers, this study compared data normalization techniques and showed that Min-Max normalization yielded better results in terms of Mean Square Error (MSE) compared to Max normalization. Out of the six models tested, Model 1 which was composed of Backpropagation training algorithm and Sigmoid transfer function yielded the lowest MSE. Moreover, learning rate and momentum value for the models of 0.2 and 0.9 respectively resulted to very minimal error in terms of MSE. The results obtained in this research clearly suggest that ANN can be a viable forecasting technique for medium-term water consumption forecasting.**

Keywords—Artificial neural network; backpropagation; water consumption forecasting

I. Introduction

Artificial Neural Networks (ANN) is a mathematical model inspired from how brain neurons learn and perform pattern recognition. ANN have been used as a technique for predicting and forecasting in various areas including finance, power generation, medicine, water resources and environmental science [1]-[3]. ANNs are composed of one or more processing units called artificial neurons or perceptrons. Basically, ANNs consist of three layers, namely, the input layer, the hidden layer and the output layer. The input layer represents the model inputs and the output layer represents the model outputs. The hidden layer consists of nodes that during optimization attempt to functionally map the model inputs to the model outputs. The basic idea of an ANN is that the network learns from the input data and the associated output data with the help of training algorithms and transfer functions [3]-[6]. Back propagation training algorithm is a supervised learning method based on the gradient descent of the quadratic error function and is considered as the universal function approximator [4], [5]. During the learning process, the gradient descent method is used to minimize the total error or mean error of the output computed by the network [1], [6]. The activations of the input nodes are multiplied by the weighted connections and are passed through a transfer function at each node in the first hidden layer. The activations from the first hidden layer are then passed to the neurons in the next layer, and this process is repeated until the output activations are obtained from the output layer. The output activation values and the target pattern are compared and the error signal is calculated based on the difference between target and calculated pattern. This error signal is then propagated backwards to adjust network weights so that network will generate correct output for the presented input pattern. The training patterns are presented repeatedly until the error reaches an acceptable value or other convergence criteria are satisfied [5]. As this technique involves performing computation backwards it is named as backpropagation.

ANN modelling approaches have been embraced enthusiastically by practitioners in water resources, as they are perceived to overcome some of the difficulties associated with traditional statistical approaches [1], [4], [7]. With the changing landscape and climate brought about by weather phenomena and unprecedented human activities, water as a very important environmental resource should be managed scientifically with the use of tools and techniques that will optimize usage management and conservation. Decision-makers can utilize machine learning platforms and models in analyzing huge volumes of data related to water management that can in turn be used to develop applications that will generate valuable inputs for short, medium and long term planning. These applications involved in water consumption forecasting use historical data to predict medium-term consumption helpful for decision makers in making critical decisions involving supply planning, reservoir or urban infrastructure changes, staging treatment and distribution system improvements [3], [6,]. Predictive applications involving forecasting water consumption are as important as descriptive applications since these applications give foresight on trends and patterns using machine learning models [8]. With the use of ANN an accurate and reliable prediction of future water consumption can help decision-makers to take necessary measures according to the possible crises and limitations.

Neuroph, a lightweight Java neural network framework for developing common neural network architectures implements multilayer perceptron having various backpropagation training algorithms and transfer functions. A major challenge in the implementation of ANN in water consumption forecasting is the choice of the appropriate ANN model design involving training algorithms and transfer functions that can yield the smallest error from the actual water consumption. Consequential to this challenge is water consumption data preparation through data normalization to ensure the avoidance of slow neural network training [7]-[9]. Data normalization is a process of final data preparation for network training so that the normalized data are shaped to meet the network input layer requirements. With the proper analysis of the water consumption data and the formulation of the appropriate ANN model, an ANN technique can be a viable solution to generate close to perfect prediction values for the prediction of water consumption. The aim of this research then is to determine an appropriate data normalization and ANN model design for medium term water consumption forecasting. This study aims to contribute to the recent technology researches in machine learning by evaluating performance of ANN models that could help water utility companies in their decision-making, proper planning and management of water resources.

II. Methodology

A. Water Consumption Data Preparation

Eighteen-year water consumption data from a city's Waterworks System in the Philippines based from the monthly bills of January 1998 to December 2015 was used in this study. As shown in Table I, a total of 1,080 rows of data with eleven (11) corresponding columns containing the month, the billed accounts and the respective water consumed in cubic meter from five water consumption categories namely domestic, commercial, industrial, bulk and whole water consumed.

TABLE I. Water Consumption Data

Month	Domestic Billed Accounts	Domestic Water Onsumed	Commercial Billed Accounts	Commercial Water Onsumed	Industrial Billed Ccounts	Industrial Water Consumed	Bulk Billed Account	Bulk Water Onsumed	Whole Billed Ccounts	Whole Water Onsumed
1	14657	830613	1138	152931	20	137422	2	4173	15817	1125139
2	14487	833589	1173	155840	19	124218	2	3249	15681	1116896
3	14153	761834	1107	131138	19	120886	2	3652	15281	1017510
4	13337	816552	1111	141278	19	121263	2	3479	14469	1082572
5	14430	789987	1137	125028	19	114681	2	2803	15588	1032499

Data normalization is a means of fitting the data within unity so that all data values will take on a value range of 0 to 1 [9]-[11]. It is one of the most significant pre-processing strategy which has a significant impact on the accuracy and performance of any model such that the sole purpose of data normalization is to guarantee the quality of the data before it is fed to a model [11]. Furthermore, the normalization process for the raw inputs has great effect on preparing the data to be suitable for the training. Due to the different units of the data, it is important to normalize the input and output data in the model development. It is required to normalize all the datasets between 0 and 1 to fit the data within unity [6], [12]. Two normalization methods namely Min-Max normalization and Max normalization were used and tested in this study. For the Min-Max normalization, a function was used to normalize the water consumption values using equation (1):

$$\bar{x} = \frac{x - x_{MIN}}{x_{MAX} - x_{MIN}} \qquad (1)$$

where $\underline{x}$ is the normalized data point, x is the original data point, x_{MIN} and x_{MAX} are the minimum and maximum of the dataset, respectively. On the other hand, Max normalization was used to normalized the water consumption values using equation (2):

$$L_s = \frac{L}{L_{max}} \qquad (2)$$

where L is the actual load, L_{max} is the maximum load during the day and L_S is the normalized load. Each of these normalization methods were applied into different models. Each of these normalization methods was applied into different formulated models having different training algorithms and transfer functions with comparison conducted using Mean Square Error (MSE) for every neural network model testing.

After normalizing the dataset, the data was then partitioned into training and testing sets. Approximately 95% of the dataset was assigned as the training set containing 1026 records of the water consumption data from January 1998 to December 2014 while 5% of the dataset was assigned as the testing set containing 54 records of the water consumption data from January 2015 to December 2015.

B. ANN Model Design Evaluation

The type of neural network used in this study was multilayer perceptron neural network with three layers: an input layer, one hidden layer and an output layer. The number of variables used as input parameters were then determined. There is no general rule for selecting the number of neurons in a hidden layer. It only depends on the complexity of the system being modeled [13]. The most popular approach in finding the optimal number of neurons in hidden layer is by trial and error [4], [6], [13]. Moreover, according to research, researchers conducted a study evaluating the number of neurons in the hidden layer but still none was accurate [14]. Thus, trial and error approach was used in this study to determine the optimum neurons in hidden layer of the network. In order to determine the optimum number of hidden neurons, several formulae on how to ascertain the optimum neurons in the second layer was also considered in this study. Hidden neuron formulae were gathered and determined from the academic journals as shown in Table II. Also, there are five neurons in the output layer representing the next month water consumption in each category.

There are several types of transfer functions, however this study only used Gaussian and Sigmoid function since these transfer functions can produce positive values between 0 and 1 which corresponds to the training and testing data sets that were normalized in a scale of 0 and 1. Training then commenced with an important consideration that the size of the steps taken in weight space during training is a function of a number of internal network parameters which includes the learning rate and momentum [1], [12]. Factors such as learning rate and momentum affect the performance in the learning process of the network. Learning rate is a parameter that determines the size of the weights adjustment each time the weights are changed during training while the momentum is a factor used to speed up convergence and maintain generalization performance of the network [15], [16]. The choice of appropriate parameters has a major impact on the performance of the backpropagation algorithm [1], [15], [16]. A good selection of these parameters could speed up and improve in great measure the learning process to reach the goal, although a universal answer does not exist for such configuration [16]. Furthermore, authors believe that choosing the learning rate can be done by trial and error [1], [3]. In this case, the learning rate and momentum value used in this study was done by trial and error. Both learning rate and momentum parameter were usually in the range between 0.1 and 0.9 [12]. Training attempts were conducted using all combinations of learning rate and momentum. This was done to select the learning rate parameter to be used in training the models. After each training run, the training results was then evaluated and compared with the results achieved in the previous runs to select the best run.

TABLE II. FORMULAE FOR HIDDEN NEURONS

Authors	Hidden Neuron Formula	No. of Hidden Neurons
Lipae, J.L. and Deligero, E.P. (2012)	$\sqrt{\boldsymbol{N_i N_o}}$ Where $\boldsymbol{N_i}$ is the input neuron and $\boldsymbol{N_o}$ is the output neuron	7
Sheela, K.J. and Deepa, S.N. (2013)	N-1 where N is the input-target relation	10
	$N_h = n + n_o - 1/2$ where n is the number of inputs and n_o is the number of outputs	16
	$N_h = (N_{in} + \sqrt{N_p}) / L$ where L is the number of hidden layer, N_{in} is the number of input neuron and N_p is the number of input sample	26
Param, Sowjanya (2015)	2/3 the size of the input layer plus the size of the output layer	12

This research evaluated the performance of different ANN models based on the type of training algorithms and transfer functions of the neural network. Training algorithms such as Backpropagation, Backpropagation with Momentum and Resilient Propagation were used in this study. Backpropagation is one of the most widely used training algorithms for training feedforward neural networks. This type of network configuration is the most commonly in use due to its ease of training [10]. The Backpropagation algorithm modifies network weights to minimize the mean squared error between the desired and the actual outputs of the network. Furthermore, Backpropagation uses supervised learning in which the network is trained using data for which the input as well as the desired outputs is known, one of the most well-known variants is the Backpropagation with Momentum [15],[17]. Momentum was added that for faster training. With this change, the weight change continues in the direction it was heading. This weight change, in the absence of error, would be a constant multiple of the previous weight change. The momentum term is an attempt to try to keep the weight change process moving, and thereby not gets stuck in its local minima which make the convergence faster and the training more stable in some cases [17]. On the other hand, Resilient Propagation (Rprop) was one of the best performing first order learning algorithms for multilayer neural networks[10],[18],[19]. The basic principle of Rprop is to eliminate the harmful influence of the size of the partial derivative on the weight step in which only the sign of the derivative is considered to indicate the direction of the weight update.

These training algorithms were paired with Sigmoid and Gaussian transfer functions. There are several types of transfer functions, this study however only used Sigmoid and Gaussian function since these transfer functions can produce positive values between 0 and 1 which corresponds to the training and testing data sets that were normalized in a scale of 0 and 1. Each combination of training algorithm and transfer function represent one model. Shown in Table III are the formulated models.

TABLE III. FORMULATED ANN MODELS

Model	Training Algorithm	Transfer Function
Model 1	Backpropagation	Sigmoid
Model 2	Backpropagation	Gaussian
Model 3	Backpropagation with Momentum	Sigmoid
Model 4	Backpropagation with Momentum	Gaussian
Model 5	Resilient Propagation	Sigmoid
Model 6	Resilient Propagation	Gaussian

During evaluation, test runs were conducted in each model by feeding the training dataset into the network and trained using the Backpropagation algorithm, Backpropagation with Momentum and Resilient Propagation. Backpropagation algorithm was used to modify the network weights in order to decrease the value of the error function on subsequent tests of the inputs. The process of adjusting weights was continued until the error is less than some desired limit after which the network is considered trained. The training process of the ANN models will stop when the network output error has reached its minimal value [5], [12]. Error measure was then computed to assess the neural network's accuracy since accuracy is the most important criteria in evaluating forecasting models and in choosing between competing models. In each test run, error measure was calculated to determine and be compared with the predictive capability of the models. In order to evaluate the ANN models, Mean Square Error (MSE) was calculated as the error measure. MSE was used as it penalizes extreme errors obtaining partial derivative with respect to the weights and that it lies close to the heart of the normal distribution [1], [18]. Among the designed models, the model that produces the smallest MSE was chosen as the neural network model.

III. Results and Discussion

A. Water Consumption Data Preparation Results

A comparison of two normalization techniques namely the Min-Max normalization and Max normalization was performed for the purpose of determining which of the two techniques yields a more accurate model. Each of the normalization techniques was tested to see whether the normalization technique has a significant effect on the neural network accuracy based on MSE values. The comparison was made by training the neural network using different combination of transfer functions namely Sigmoid and Gaussian and training algorithms such as Backpropagation, Backpropagation with Momentum and Resilient Propagation with 7 hidden neurons. The training dataset was used and then fed into the network. The learning parameters like learning rate and momentum of 0.2 and 0.7 respectively, was used during training. The MSE for each network that used the Min-Max normalization technique was calculated and presented in Table IV, respectively showing the MSE results using different training algorithms and transfer functions.

TABLE IV. Min-Max Normalization MSE Results

Training Algorithm	Mean Square Error	
	Sigmoid	**Gaussian**
Backpropagation	0.003874424	0.003885264
Backpropagation with momentum	0.003914189	0.003913525
Resilient propagation	0.005360615	0.006531334

In using the Min-Max normalization in normalizing the water consumption data, the lowest mean square error was achieved using the combination of Backpropagation training algorithm and Sigmoid transfer function while the highest mean square error was achieved using the Resilient Propagation as the training algorithm and Gaussian transfer function. The researchers observed that the MSE values of Sigmoid and Gaussian activation are below 0.0039 when Backpropagation and Backpropagation with Momentum were used as the training algorithm. As shown in Fig. 1, even though Sigmoid and Gaussian activation have close MSE results, Sigmoid with Backpropagation as the training algorithm has lower MSE values than Gaussian.

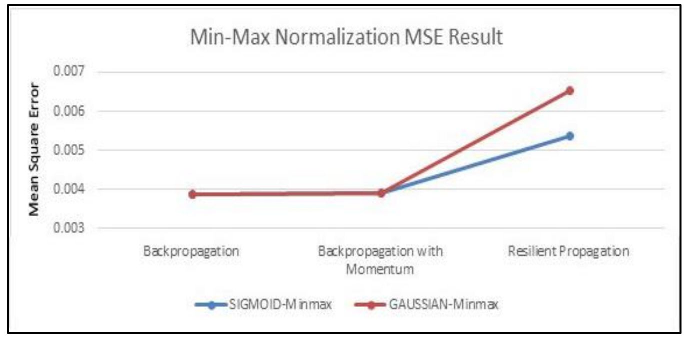


Fig. 1. Min-max normalization MSE result graph.

The MSE results using different training algorithms and transfer functions for each network that used the Max normalization technique was calculated and presented in Table V.

TABLE V. Max Normalization MSE Results

Training Algorithm	Mean Square Error	
	Sigmoid	**Gaussian**
Backpropagation	0.004959902	0.005191838
Backpropagation with momentum	0.005250859	0.005312595
Resilient propagation	0.003896138	0.003882556

In using the Max normalization in normalizing the water consumption data, the lowest mean square error was achieved using the Resilient Propagation training algorithm and Gaussian transfer function while the highest mean square error was achieved using the Backpropagation with Momentum algorithm and Gaussian transfer function. As shown in Fig. 2, the researchers observed that the MSE of Sigmoid and Gaussian activation showed lowest values when Resilient Propagation training algorithm was used. Even though the Sigmoid and Gaussian activation have close MSE results, Gaussian activation has lower MSE values than Sigmoid.

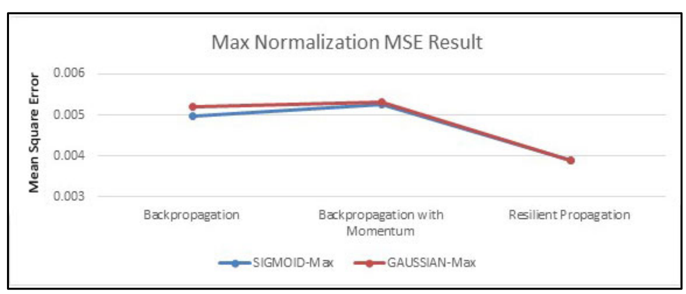


Fig. 2. Max normalization MSE result graph.

In comparing the two normalization techniques, the Min-Max normalization technique yielded better MSE values than the Max normalization technique. As shown in Fig. 3, among the four normalization tests, the data that used Min-Max normalization and Backpropagation as the training algorithm with Sigmoid transfer function produced the lowest MSE while the data that used the Max normalization and Backpropagation with Momentum as the training algorithm with Gaussian transfer function produced the highest error.

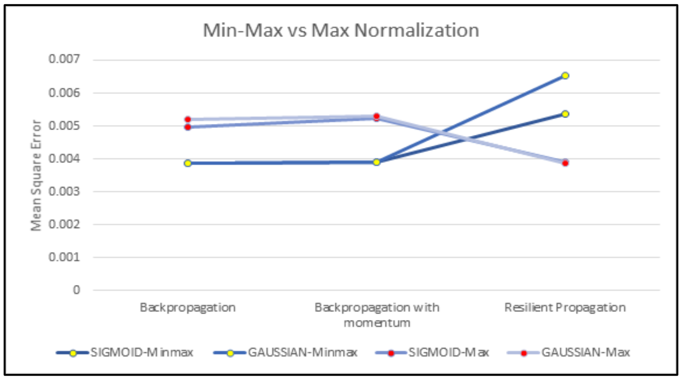


Fig. 3. Min-max normalization MSE result graph.

The researchers observed that error curves are not in almost linear pattern. In Resilient Propagation, Sigmoid and Gaussian using Min-Max and Max normalization technique have opposite results. The values of Sigmoid and Gaussian using Min-Max normalization are higher than the values of Sigmoid and Gaussian using Max normalization. Even though the MSE values are within 0 to the 0.007 range, the values have a significant difference between one another. Thus, among Min-Max and Max normalization techniques, the type of normalization method used has a huge and significant effect on the performance of the network in terms of accuracy. Furthermore, since Min-Max normalization has the lowest MSE, it was used as the normalization technique for the training dataset. Studies on different types of normalization techniques in data mining as a preprocessing engine conducted also concluded that Min-Max normalization is the best design for training data set because it has a higher percentage of accuracy compared to Max normalization, Z-score normalization and decimal point normalization [10], [20]. Moreover, the Min-Max normalization technique was also used in water demand prediction using artificial neural networks and support vector regression to avoid having more weight being assigned to features with larger values [21]. Studies conducted on predictive analytics also support the results of this study that Min-Max normalization is better than Max normalization [9]-[11], [20], [21].

B. ANN Model Design Evaluation Results

Designing the architecture of an ANN model includes the identification of the number of neurons for input, hidden and output layers, as well as the performance analysis of the training algorithms and the transfer functions. As shown in Fig. 4, there were 11 input neurons which represent the month, billed accounts in each category and the water consumed in each category while the number of neurons in the output layer was 5 representing the next month water consumption in each category. A neural network with one hidden layer has the tendency to perform very well [2], [5], [7]. Thus, the researchers used only 1 hidden layer.

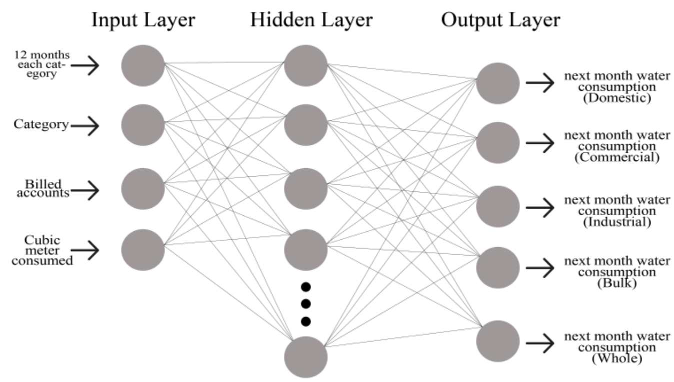


Fig. 4. The proposed ANN architecture.

Determining the number of hidden neurons does not have a standardized or theoretical approach to calculate the number of neuron in the hidden layer [14]. In order to select the appropriate number of hidden neurons to be used in this study, the researchers conducted series of tests with results shown in Table VI.

TABLE VI. HIDDEN NEURONS TEST RESULTS

Hidden Neuron Equation	Hidden Neurons	Total MSE (M^3)
$\sqrt{N_i N_o}$ where N_i is the input neuron and N_o is he output neuron	7	0.00394277592437603
N-1 where N is the input-target relation	10	0.00396177000639483 25
$N_h = n + n_o - ½$ where n is the number of inputs and n_o s the number of outputs	16	0.00401964300187572 7
$N_h = (N_{in} + \sqrt{N_p}) / L$ where L is the number of hidden layer, N_{in} is the number of input neuron and N_p is the number of input sample	26	0.00424731991227425 9
2/3 the size of the input layer plus the ize of the output layer	12	0.00449276241990014 6

The hidden neuron equations presented in the first column of the table was calculated to determine the hidden neuron. Each number of hidden neuron was tested and yielded results in MSE. Among the hidden neuron choices, the first equation with 7 hidden neurons yielded the lowest MSE while the last equation with 12 hidden neurons produced the highest MSE. As a result, 7 hidden neurons were selected because it yielded the least MSE that is 0.00394.

Neuroph, a Java open source framework designed to develop artificial neural network was used in training the ANN models. It contains an implementation for most of the mainstream ANNs and learning algorithms, such as multilayer perceptron network and backpropagation learning algorithm. Training dataset was fed into each model during training. The models used the same set of learning parameters to maintain the credibility of the evaluation results. Learning parameters such as maximum error, learning rate and momentum were set. Maximum error was set as the stopping criteria during training. If the error on the training or selection test drops below the given target values, the network is considered to have trained sufficiently well, and training is terminated. The error never drops to zero or below, so the default value of zero is equivalent to not having a target error [22]. The learning rate parameter determines the size of the weights adjustment each time the weights was updated during training. A learning rate of 0.0 does not learn [16]. The momentum parameter was a factor used to speed network training for escaping the local minima to avoid error fluctuation [13]. Both learning rate and momentum parameter were usually in the range between 0.1 and 0.9 and choosing the learning rate can be done by trial and error [3], [12]. Fig. 5 shows the graphical representation of the overall results for the test of the learning rates. Different combination of learning rate and momentum yielded varying results in terms of MSE. The researchers observed that the higher the learning rate, the bigger the MSE was produced. Generally, a small learning rate can ensure the reduction of the error function but may slow the convergent process, while a large learning rate can speed up the learning process but may cause divergence [6], [16]. According to an author, if the selected learning rate is too large, then the local minimum may be overstepped constantly, resulting in oscillations and slow convergence to the lower error state [23]. As a result, learning rate of 0.9 yielded the highest MSE among the others. Likewise, if the learning rate is too low, the number of iterations required may be too large, resulting in slow performance.

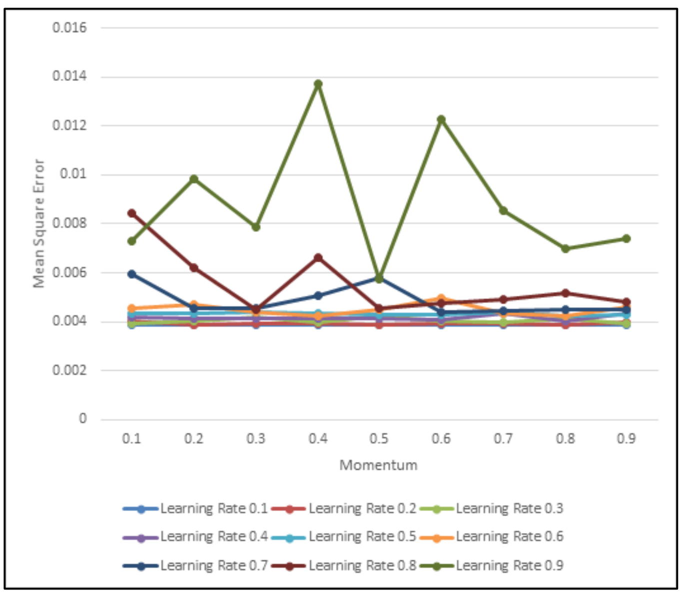


Fig. 5. Learning parameters test results.

Contrary to results shown in learning rate 0.1 and 0.2 combined with momentum 0.1 to 0.9, the number of iteration was not too large and yielded the least MSE. Thus, this study used the combination of momentum 0.1 to 0.9 and learning rate 0.1 and 0.2 in training the models because it performed well in terms of MSE. Each combination of learning rate and momentum corresponds to 1 training attempt. Generally, there is a total of 108 training attempts where each model has 18 training attempts. Table VII shows the training attempts with the smallest MSE obtained in each model.

TABLE VII. ANN MODEL TRAINING RESULTS

Model	Maximum Error	Learning Rate	Momentum	MSE
Model 1	0.01	0.2	0.9	0.003870079156963867
Model 2	0.01	0.1	0.3	0.0045561730198 7756
Model 3	0.01	0.1	0.4	0.003870627416201587
Model 4	0.01	0.1	0.2	0.004689457453932001
Model 5	0.01	0.1	0.9	0.00398609747754698
Model 6	0.01	0.2	0.2	0.003990659899236666

As observed, Model 1 has better precision on prediction which was trained using Backpropagation algorithm and Sigmoid transfer function with 0.2 and 0.9 values of learning rate and momentum respectively. The researchers found out that the values of 0.2 and 0.9 for the learning rate and momentum yielded the fastest learning convergence to the minimal number of errors during training attempts. Although the selection of the learning rate and momentum is an essential task, other factors like training algorithm and activation function are more vital as the results shows that different combination of this yielded differential result. The researchers observed that Sigmoid transfer function paired to Backpropagation and Backpropagation with Momentum algorithms yielded the least MSE compared to Gaussian activation function. In other words, Sigmoid transfer reached a very good overall approximation [10], [24]. Moreover, when Sigmoid and Gaussian were trained using Resilient Propagation algorithm, it yielded results that were close to each other as shown in Fig. 6. Among the training algorithms, Backpropagation outperformed the Backpropagation with Momentum and Resilient Propagation. Backpropagation training algorithm yielded the least MSE value while Resilient Propagation yielded the highest MSE value. The result shows that the MSE values obtained by the Backpropagation and Backpropagation with Momentum were very close to each other but Backpropagation training algorithm has smaller values than the Backpropagation with Momentum algorithm. This implies that Backpropagation performed better than Backpropagation with Momentum. This was supported by studies who also used Backpropagation algorithm as the best model for water demand prediction [8], [25].

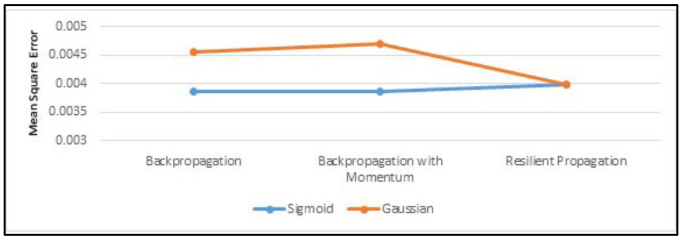


Fig. 6. Comparison of training algorithms with transfer functions.

IV. Conclusion and Recommendations

This study attempted to conduct a performance analysis of different ANN training algorithms, transfer functions, learning rates and momentum to discover a suitable ANN model for month-ahead water consumption prediction. Two normalization techniques namely Min-Max normalization and Max normalization were compared in the water consumption data preparation phase with results showing that Min-Max scaling yielded better results in terms of MSE values. Six models were then formulated with different combination of training algorithms and transfer functions. The type of neural network used in this study was a multilayer perceptron having three layers in each model. Neuroph Studio, a Java based Neural Network IDE was then used to simulate the designed models. Learning parameters such as maximum error, learning rate and momentum were set, showing that a learning rate and momentum value of 0.2 and 0.9 respectively, performed as the better combination of learning parameter with a minimal error in terms of MSE. After the models were compared and tested, among the 6 models, Model 1 which was composed of Sigmoid transfer function and Backpropagation training algorithm has the best precision in forecasting and yielding the smallest value of MSE equal to 0.003870079156963867. The results of this study show that Sigmoid activation function being paired with different training algorithm yielded better results compared to Gaussian. Moreover, Backpropagation algorithm performs better when compared to other training algorithms.

It is recommended that the use of additional input factors could yield better results in training the network model. The study could potentially be improved if other variables that affect water consumption are to be examined. Other ANN frameworks could also be used to expand performance analysis conducted in this study. One or more ANN frameworks can be compared with the performance results of this study and contrast if other frameworks have better or the same performance with that of Neuroph. Other normalization techniques, combination of learning parameters, training algorithms and activation functions could also be explored as enhancements for future work. With the results of the performance analysis from the different training algorithms and activation functions being shown in this research, future directions can be geared towards the use of the best performing model for a chosen validation set and an evaluation of how close is its prediction to the corresponding actual water consumption.

References

[1] H. R. Maier, and G. C. Dandy, "Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications," Environmental Modelling & Software 15, Elsevier Science Ltd., pp. 101-124, 2000.

[2] L. C. P. Velasco, N. R. Estoperez, R. J. R. Jayson, C. J. T. Sabijon, and V. C. Sayles, "Day-ahead Base, Intermediate, and Peak Load Forecasting using K-Means and Artificial Neural Networks," International Journal of Advanced Computer Science and Applications, Volume 9 Issue 2, The Science and Information Organization Limited, DOI : 10.14569/IJACSA.2018.090210, pp. 62-67, 2018.

[3] J. R. Mohammed and H. M. Ibrahim, "Multilayer Perceptron - Multiactivation Function Artificial Neural Network Model for Municipal Water Demand Forecasting," ARPN Journal of Engineering and Applied Sciences, Vol. 8, No. 12, pp. 1019-1038, 2013.

[4] M. A. Al-Zahrani, and A. Abo-Monasar, "Urban residential water demand prediction based on artificial neural networks and time series models," Water Resources Management, 29(10), pp. 3651-3662, 2015.

[5] M. Djeddou and B. Achour, "The Use of a Neural Network Technique for the Prediction of Sludge Volume Index in Municipal Wastewater Treatment Plant," Larhyss Journal, pp. 351-370, 2015.

[6] J. Liu, H. Savenije, and J. Xu, "Forecast of water demand in Weinan City in China using WDF-ANN model," Physics and Chemistry of the Earth, 28, pp. 2019-224, 2003.

[7] M. Firat, M. E. Turan, and M. A. Yurdusev, "Comparative analysis of neural network techniques for predicting water consumption time series," Journal of Hydrology, 384, pp. 46-51, 2010.

[8] L. C. P. Velasco, A. R. B. Granados, J. M. A. Ortega and K. V. D. Pagtalunan, "Medium-Term Water Consumption Forecasting using Artificial Neural Networks for a Water Utility," Proceedings of the 17th Conference of the Science Council of Asia, National Research Council of the Philippines, 2017

[9] T. Jayalakshmi and A. Santhakumaran, A. "Statistical Normalization and Back Propagation for Classification," International Journal of Computer Theory and Engineering, pp. 89-93, 2011.

[10] L. C. P. Velasco, J. A. A. Dagaang , C. R. Villezas, P. N. C. Palahang, "Performance Analysis of Different Combination of Training Algorithms and Activation Functions in Predicting the Next Day Electric Load,", Proceedings of the 16th Philippine Computing Science Congress, 2016

[11] S. C. Nayak, B. B. Misra, and H. S. Behera, "Impact of Data Normalization on Stock Index Forecasting," International Journal of Computer Information Systems and Industrial Management Applications, Volume 6, pp. 257 – 269, 2014.

[12] A. Jain, A. K. Varshney, and U.C. Joshi, "Short-Term Water Demand Forecast Modelling at IIT Kanpur Using Artificial Neural Networks," Water Resources Management, 15, pp. 299-321, 2011.

[13] J. Cheng, Q. S. Li, R, Xiao, "A new artificial neural network-based response surface method for structural reliability analysis,", Probabilistic Engineering Mechanics, Volume 23 Issue 1, Elsevier, pp. 51-63, 2008.

[14] S. Karsoliya, "Approximating Number of Hidden layer neurons in Multiple Hidden Layer BPNN Architecture," International Journal of Engineering Trends and Technology,3(6), pp. 714-717, 2012.

[15] E. Istook, T. Martinez, "Improved backpropagation learning in neural networks with windowed momentum," International Journal of Neural System, 12(3–4): pp. 303–318, 2002.

[16] N. O. Attoh-Okine, "Analysis of Learning Rate and Momentum Term in Back Propagation Neural Network Algorithm Trained to Predict Pavement Performance," Journal of Advances in Engineering Software, 30(4), pp.293–302, 1999.

[17] S. K. Lahiri, K. C. Ghanta,"Artificial neural network model with parameter tuning assisted by genetic algorithm technique: study of critical velocity of slurry flow in pipeline," Asia-Pacific Journal of Chemical Engineering, Volume 5 Issue 5, 2009.

[18] L. C. P. Velasco, P. B. Bokingkito and J. T. Vistal, "Week-Ahead Load Forecasting using Multilayer Perceptron Neural Network for a Power Utility", Proceedings of the 17th Conference of the Science Council of Asia, National Research Council of the Philippines, 2017

[19] A. D. Anastasiadis, G. D. Magoulas and M. N. Vrahatis, "An Efficient Improvement of the RPROP Algorithm,", In Proceedings of the First International Workshop on Artificial Neural Networks in Pattern Recognition, 2003.

[20] L. Al Shalabi, Z. Shaaban and B. Kasasbeh, "Data Mining: A Preprocessing Engine," Journal of Computer Science, 2 (9), pp. 735-739, 2006.

[21] I. S. Msiza, F. V. Nelwamondo and T. Marwala "Water Demand Prediction using Artificial Neural Networks and Support Vector Regression," Journal of Computers, Volume 3 Number 11, 2018.

[22] P. P. Balestrassi, E. Popova, A.P. Paiva, and J. W. Marangon Lima, J.W., “Design of Experiments on Neural Network’s Training for Nonlinear Time Series Forecasting” pp. 1160-1178, 2009.

[23] A. Abrahan, "Artificial Neural Networks", John Wiley & Sons, 2005.

[24] A. K. Palit, and D. Popovic, “Computational Intelligence in Time Series Forecasting Theory and Engineering Applications” Springer, XXII, 2005.

[25] S. Behboudian, M. Tabesh, and M. Falahnezhad, “A Long-Term Prediction of Domestic Water Demand Using Preprocessing in Artificial Neural Network,” Journal of Water Supply: Research and Technology-Aqua, pp31-42, 2014.

A Linear Array for Short Range Radio Location and Application Systems

Saad Hassan Kiani
Electrical Engineering Department
Iqra National University
Peshawar, Pakistan

Khalid Mahmood
Electrical Engineering Department
University of Technology
Nowshehra, Pakistan

Ahsan Altaf
School of Electrical and Electronics Engineering, Istanbul Medipol University, Istanbul, Turkey

***Abstract*—Patch array antennas have primarily been good candidates for higher performance results in communication systems. This paper comprises of linear 1x4 patch antenna array study constructed on 1.575mm thick Roggers 5880 substrate with high gain of 12.8dB and focused directivity of 12.9dBi. The array network is fed using T Junction method showing well matched input impedance results. With higher performance parameters and reflection coefficient, voltage standing wave ratio, the proposed antenna array is suited for short range radiolocation and radio services application.**

Keywords*—Linear array; gain; Roggers 5880; voltage standing wave ratio; directivity; radiolocation; short range radio applications

I. Introduction

With recent advancement in communication systems, antenna design has seen a paid growth with higher usage of mm wave, high altitude application systems and higher frequency band technologies [1]. Patch antenna dominance is because they offer unique features. They are extremely light weighted, offering less space and good parameters performance and easy array assembling fabrications. On one single substrate, entire patch array can be constructed to enhance performance of patches, however very closed assembling give arises to unwanted surface currents also be known as mutual coupling which should be reduced as minimum as possible [2].

Different techniques and designs have been made like multiband antennas [3], [4], antennas with wearable characteristics [5] and miniaturized antennas as smaller antennas [6]. However, a single element patch antenna usually doesn’t deliver the desired results [7] and lower gain and other poor performance parameters like narrow impedance bandwidth, Lower efficiency keep them on disadvantage.

This is due to a reason of mismatch losses. A well designed antenna can lead to poor performance parameters with high mismatch losses. Microstrip feed line and co axial cable are used usually while designing array network and single feed elements as these two methods are direct contacting feeding schemes [8].

Antenna to be used as arrays can lead to desired results but designing an array of antenna is a challenging task as their mutual coupling effect especially in multiple input multiple output (MIMO) antennas can degrade antenna performance but with help of isolation enhancement [9]-[11] MIMO patch antenna s are widely used in up to date communication technology. Radiolocation Services (RCS) and Radio Location Services (RLS) have been assigned a range of 5GHz to 6 by IEEE 802.11a standard protocols.

This study presents a single input multiple output planar 1x4 linear antenna arrays for short range Radio services and short services Radio location applications systems. A single design of square patch antenna is constructed and its result are measured and compared with proposed array. The proposed linear single input multiple output arrays showed enhance performance parameters results as compared to single element. The paper comprises of four stages, namely, Introduction, Antenna Design Results and Conclusion. Future works remarks are mentioned at the end.

II. Antenna Design

A. Single Element Design

Roggers RT Duroid 5880 is taken as a substrate for our antenna design due to its cost and atmospheric properties.

The single element square shape patch antenna is shown in Fig. 1.

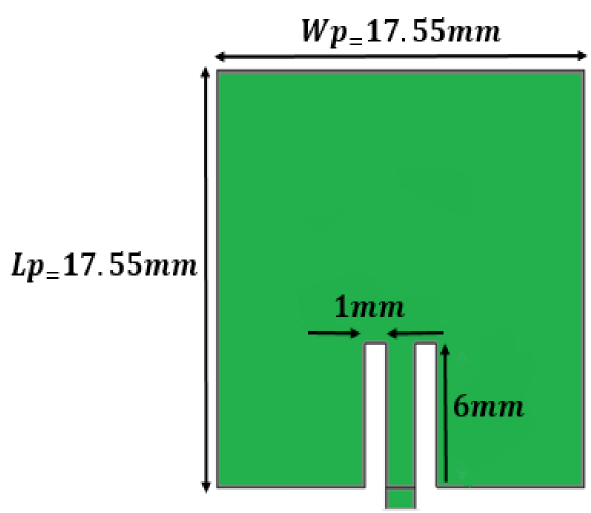


Fig. 1. Square patch antenna.

Before designing a patch antenna, there are known parameters with the help of which antenna is designed. These known parameters are shown in Table I.

TABLE I. KNOWN ANTENNA PARAMETERS

S.No	Parameter	Data
1	Resonant Frequency	5.5GHz
2	Dielectric Constant	2.3
3	Substrate Thickness	1.575

With the known parameters, antenna design parameters are calculated. As described in [12], the length and width of the patches can be described by the following equations. However for our design purpose, we have design a square patch array rather than rectangular since impedance matching is much easy in it. The calculated parameters are shown in Table II.

$$Width, W = \frac{c}{2f0\sqrt{\frac{\varepsilon r+1}{2}}} \tag{1}$$

$$Length. L = L(eff) - 2\Delta L \tag{2}$$

Where

$$L(eff) = \frac{c}{2f0\sqrt{\varepsilon_{(reff)}}} \tag{3}$$

And

$$\varepsilon_{(reff)} = \frac{\varepsilon r+1}{2} + \frac{\varepsilon r-1}{4}\left(1+\frac{12h}{W}\right)^{-1/2} \tag{4}$$

TABLE II. ANTENNA CALCULATED PARAMETERS

S. No	Parameter	Data
1	Patch Length, Lp	17.5mm
2	Patch Width, Wp	17.5mm
3	Ground Plane Length, GP	35.0mm
4	Ground Plane Width, Gw	35.0mm
5	Feed Length, Lf	5.00mm
6	Feed Width, Fw	1.412mm
7	Inset Feed Length, IFl	6mm

Usually while designing a single element, input impedance of 50 Ω is desirable but in our design but consider that each one of the elements of an array is fed by an output of some feed network, rather than direct connection to a 50Ω SMA connector, or something similar. Consequently, our single element was fed with an input impedance of 100Ω.

B. Array Design

The array is composed of four identical square patches separated by half wavelength distance of 28mm as shown in Fig. 2. The *Sl* and *Sw* length and widths of Roggers 5880 are taken 50mm and 70mm respectively. The overall dimension of array is 3500mm^2.

For power splitting network in an array system, power divider as corporate feed system plays a key role. It simply splits power the between "n" numbers of outputs ports with a certain distribution.

For impedance matching purposes, maximum power transfer theorem is used. Using quarter transformer, the 100 Ω transmission lines are matched to 70.7 Ω transmission line which are again splitted to 100 Ω and finally to inset feed of 50 Ω. Dimensions of power divider are shown in Table III.

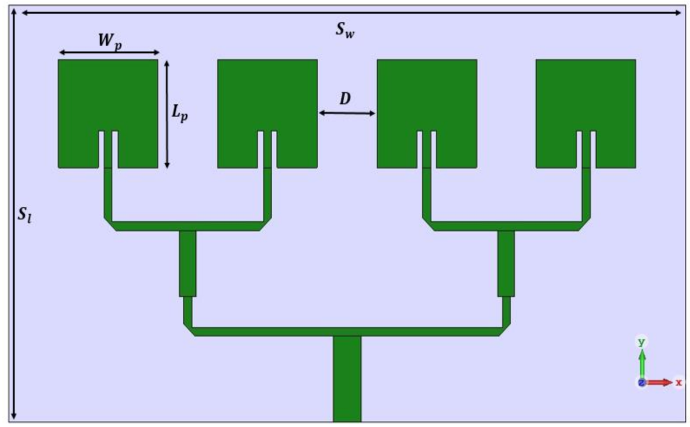


Fig. 2. 1x4 Linear patch array.

TABLE III. DIMENSIONS OF POWER DIVIDER OF ARRAY

S. No	Parameters	Data
1	Length of 50 Ω Transmission Line	12.50mm
2	Width of 50 Ω Transmission Line	4.852mm
3	Length of 70 Ω Transmission Line	10.721mm
4	Width of 70 Ω Transmission Line	2.932mm
5	Length of 100 Ω Transmission Line	5.00mm
6	Width of 100 Ω Transmission Line	1.412mm

III. RESULTS AND DISCUSSION

The array was designed in Computer simulation Technology 2014. The array showed good performance parameters and an excellent impedance matching. The results are summarized in Table IV.

TABLE IV. PERFORMANCE PARAMETERS OF ANTENNA ARRAY

Parameters	Conventional	Proposed
Return Loss	-12.25dB	-20.0dB
Gain	7.18dB	12.82dB
Directivity	7.15dBi	12.96dBi
VSWR	1.05	1.034
Bandwidth	200MHz	245MHz

The return loss plot is shown in Fig. 3. The antenna showed good reflection co efficient of -20dB with satisfactory bandwidth of 245MHz. As compared to single element proposed array is well matched with power divider corporate feed.

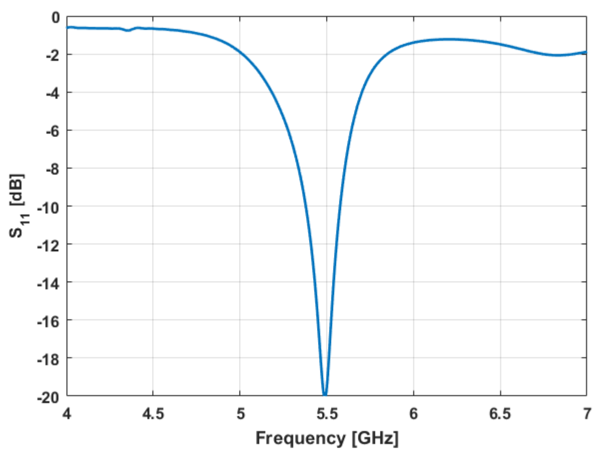


Fig. 3. S parameters of proposed antenna array.

The E and H plane polar pattern fields of our proposed 1x4 linear array are shown in Fig. 4 and directivity graph in both E plane and H plane is shown in Fig. 5.

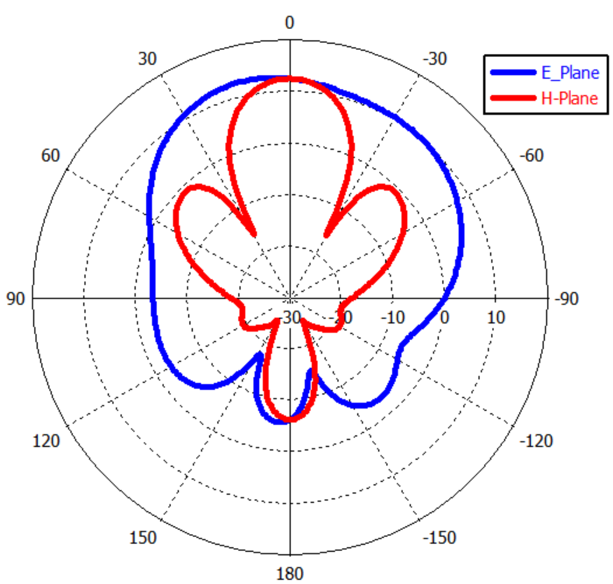


Fig. 4. E field and H field polar pattern of linear array.

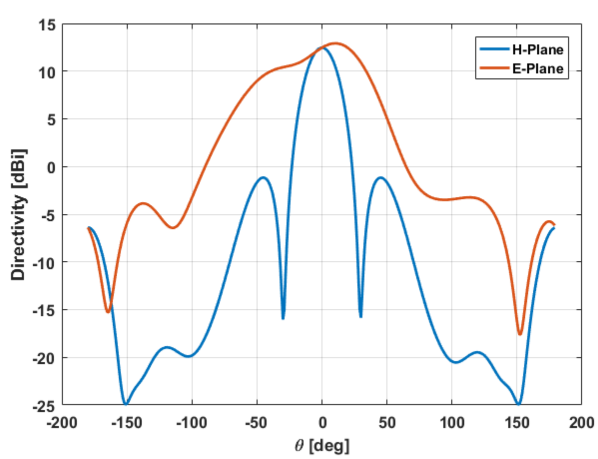


Fig. 5. E and H plane directivity plot.

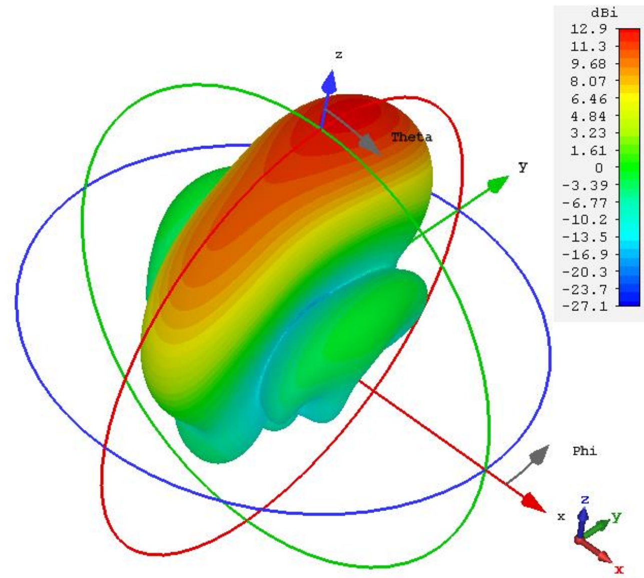


Fig. 6. 3 Dimensional plot of directivity.

In H field polar pattern, the main lobe direction is 0.0 deg with angular width of 24.3deg and side lobe level of-12.9dB while in E plane, the main lobe direction is 11.0 deg with angular width of 77.6deg and side lobe level of -15.8deg. The antenna showed broad sided direction of 12.9dB of directivity with high gain of 12.82dB. The percentage bandwidth of proposed linear array is 4.82%. The directivity 3D graph is shown in Fig. 6.

IV. Conclusion

In this paper, a square patch linear 1x4 array of 5.5GHz is proposed. A single element was designed and its results were compared to that of linear array results. The antenna array is composed by four antenna elements fed with the parallel method that allows the exciting signal to reach equally each element. The structure is implemented over Roggers 5880 substrate with 2.3 as relative permittivity, 1.575 mm for thickness and 0.0009 for loss tangent. It was found that with addition in patch elements, proposed array showed better performance results. With higher bandwidth and gain, the proposed array can be used for Short Range Radio Location and Short Range Radio Services applications.

V. Future Work

In future, this linear array can be tested with increased size of patch elements. Furthermore, the proposed patch array structure can de designed with 2x2 array configuration and same technique can be implemented through aperture coupled feed network.

References

[1] Wang, H., X. B. Huang, and D. G. Fang. "A single layer wideband U-slot microstrip patch antenna array." IEEE antennas and wireless propagation letters 7 (2008): 9-12.

[2] Qian, Y., Coccioli, R., Sievenpiper, D., Radisic, V., Yablonovitch, E., & Itoh, T. (1999). A microstrip patch antenna using novel photonic band-gap structures. *Microwave Journal*, *42*(1), 66-72.

[3] Kiani, S. H., Mahmood, K., Khattak, U. F., Burhan-Ud-Din, & Munir, M. (2016). U Patch Antenna using Variable Substrates for Wireless

Communication Systems. International Journal of Advanced Computer Science and Applications, 7(12), 286-291.

[4] Kiani, S. H., Mahmood, K., Munir, M., & Cole, A. J. (2017). A Novel Design of Patch Antenna using U-Slot and Defected Ground Structure. *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, *8*(3), 17-20.

[5] Salonen, P., Kim, J., & Rahmat-Samii, Y. (2005, July). Dual-band E-shaped patch wearable textile antenna. In Antennas and propagation society international symposium, 2005 IEEE (Vol. 1, pp. 466-469). IEEE.

[6] Zhang, Huiying, et al. "Miniaturized implantable antenna integrated with split resonate rings for wireless power transfer and data telemetry." Microwave and Optical Technology Letters 59.3 (2017): 710-714.

[7] George, Jemima Nissiyah, and M. Ganesh Madhan. "Analysis of single band and dual band graphene based patch antenna for terahertz region." Physica E: Low-dimensional Systems and Nanostructures 94 (2017): 126-131.

[8] Singh, Gurdeep, and Jaget Singh. "Comparative analysis of microstrip patch antenna with different feeding techniques." International Conference on Recent Advances and Future Trends in Information Technology, iRAFIT. 2012.

[9] Veeramani. A, Afsane Saee Arezomand, Vijayakrishnan. J, Ferdows B. Zarrabi, "Compact S-shaped EBG Structures for Reduction of Mutual Coupling," 2015 Fifth International Conference on Advanced Computing & Communication Technologies, 2015.

[10] Altaf, Ahsan, M. A. Alsunaidi, and Ercument Arvas. "A novel EBG structure to improve isolation in MIMO antenna." USNC-URSI Radio Science Meeting (Joint with AP-S Symposium), 2017. IEEE, 2017.

[11] Tu, D. T. T., Van Hoc, N., Quan, H., & Van Yem, V. (2016, July). Compact MIMO antenna with low mutual coupling using defected ground structure. In Communications and Electronics (ICCE), 2016 IEEE Sixth International Conference on (pp. 242-247). IEEE.

[12] Balanis, Constantine A. Antenna theory: analysis and design. John Wiley & Sons, 2016.

Factors Influencing Cloud Computing Adoption in Saudi Arabia's Private and Public Organizations: A Qualitative Evaluation

Mohammed Ateeq Alanezi
College of Computing and Information Technology
Shaqra University, Kingdom of Saudi Arabia

***Abstract*—Cloud Computing is becoming an important tool for improving productivity, efficiency and cost reduction. Hence, the advantages and potential benefits of cloud computing are no longer possible to be ignored by organizations. However, organizations must evaluate factors that influence their decisions before deciding to adopt cloud computing technologies. Many studies have investigated cloud computing adoption in developed countries compared with few studies that have concentrated on examining the factors that influence cloud computing adoption in developing countries. It is not clear to see whether these factors that have been identified by these studies, can be applied in developing countries. The motive of this study is to contribute to the adoption of cloud computing, and to elevate the consciousness of cloud computing technology amongst authorities, researchers, administrators, business enterprise managers and service carriers, particularly within the Saudi Arabian context. This study explores factors that encourage the implementation of cloud or have the capacity to detract from adopting cloud computing in private and public organizations in Saudi Arabia. A qualitative approach through IT professional representatives' interviews was adopted in this study, which explored two categories, namely, a) the negative impact category which includes: security and privacy, government policy, lack of knowledge, and Loss of control; and b) the positive impact category which includes three factors: reduce expenses, improve IT performance, and promote scalability and flexibility.**

Keywords—Cloud computing; private and public organization adoption; qualitative evaluation; Saudi Arabia

I. Introduction

Organizations are, nowadays, changing their strategy in IT towards the adoption of cloud computing. The strategy is mainly connected with technological innovation, its flexibility in implementation and growth in economy. Hence it is not possible anymore for organizations to ignore the advantages of cloud computing [1]. Various areas such as businesses, institutions and government are attracted to cloud computing [2]. Cloud computing offers a shift from the current installed software in the personal computers to the cloud-based servers which can be accessed from anywhere at any time. In addition, cloud computing provides an on-demand provisioning of resources and scalability [3]. Various advantages such as business process transformation, real world applications delivery, lower expenditures in infrastructure, and mobilization are more likely to achieve higher efficiency in business strategy [4].

Cloud computing is attached to the scope of e-service as a potential means in order to serve the public and private sectors with more services [5]. It is a method of green technology that has a potential to improve utilization of the current resources in the data centre and lower their consumption. Energy savings in computing and storage service can be offered by the cloud computing through the use of large shared servers and storage units [6]-[8]. In order to realize these benefits, more information, development and promotion is needed to encourage the adoption [9]. The economic slump that hit recent years in IT has increased the interest of cloud services adoption, which can be used to reduce the costs [10]. There is a very limited number of theoretical studies that identified and evaluated the factors that impact the implementation of cloud computing. Most of the published work so far is white papers, newsletters and policy statements lacking academic methodology and analysis [3]. Most of these studies, which are empirical in nature, have focused only on the advantages and the associated risks of cloud computing [9]. A few studies have been found in the context of developing countries including Saudi Arabia [11]-[13]. Most of these studies have taken a well-established theory (such as Technology-Organization Environment framework (TOE), Diffusion of Innovation (DOl), and Technology Acceptance Model (TAM)) and collected data from a developing country in order to test if the theory or the model was applicable in that context. They failed to focus on the contextual details and the specific factors that are affecting the implementation of cloud computing in the local context. Thus, the main goal of this study is to identify the factors which influence the implementation of cloud computing in private and public organizations in Saudi Arabia.

II. Literature Review

Whenever the development takes place on a structural societal change in a place where various diverse social, and economic changes play a major role, it can be predicted that information and communication technology development is a very important factor [14]. Accessing the computing resources or developing robust IT infrastructure is a very difficult approach in developing countries. However, the new computer paradigms, such as cloud computing, had provided access to various resources that are available remotely or have been in any other case inaccessible. The evolution of cloud computing will change facing of this, which includes small and large-scale businesses, entrepreneurs, researchers and government systems [15]. However, it is not yet clear how the implication of cloud computing in public and private sectors will be handled. Still it can be predicted that information and communication technology have the potential to access the information that was not possible few years back. All the public and the private enterprises are trying to use the cloud-based services which is becoming increasingly important. As stated in [16], the concept of cloud computing is becoming the main strategy for various organizations. This short survey of latest literature and research findings on cloud computing is offered in three parts. The first part involves the conceptualization of cloud computing, the second part outlines its benefits, and the third part explores organizations' adoption of cloud computing.

A. Overview of Cloud Computing

Cloud computing can be defined as a technology paradigm that enables ubiquitous, appropriate, on-demand network access to share computing resources and services. Various computing resources like storage, servers, networks, applications and services which required to be rapidly provisioned and managed with minimal amount of efforts are offered in this paradigm [17]. Through the use of cloud computing technology, there is no need for users bother about managing resources and they get an access to all the available resources over the internet [1]. Cloud computing is defined as *"a parallel and distributed computing system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on Service-Level-Agreements (SLA) established through negotiation between the service provider and consumers"*[18]. The service providers are trying to make software and technology available on demand. The computing and storage over the internet can be used by the end-user who does not require having the knowledge about the actual location and the configuration of the system, where his/her data is coming from and being saved.

Could computing provides three types of services [19], [20]. They include infrastructure as a service, Platform as a service and Software as a service.

- Infrastructure-as-a-Service (IaaS): In this service, the cloud offers some physical devices such as computers which are connected virtually, various servers, and devices used for storage as services. All these services are located in one physical place (data center) but they could be accessed only through the internet. The clients of cloud computing only pay for what resources and services they use. For example, organizations can use cloud computing technique in email management, website hosting, and data management.

- Platform-as-a-Service (PaaS): This service typically stands for computing platform. Through PaaS, developers can use all services over the cloud computing instead of purchasing it. In PaaS, licenses for various platforms such as all the operating systems, databases management, and other programming software, become easily available on the internet [11].

- Software-as-a-Service (SaaS): In SaaS, cloud providers maintain infrastructure and platform to provide applications in business such as word, emails, CRM, and ERP. All these apps are carried over the internet which can be available for the end-user. The end user can easily get an access to the apps anywhere at any time. This is the most utilized Cloud Model [11].

Free email services like Google and Microsoft are an example of a cloud computing approach that is already in action. However, there are various other services that can be used by private and public organizations which are launched recently that make use of cloud computing services such as data management, data analysis, mobile telecommunication services and various computing-based services that makes use of grid computing and parallel computing systems [21].

B. Benefits of Cloud Computing

In latest literature, it could be observed that challenges faced by organizations could be resolved by cloud computing which has been proved to be reliable [8]. The main advantages of cloud computing include offering lower expenses [22], better services standardization [23], and greater business adaptability [24]. Other advantages such as increased accessibility to cutting-edge technology and scalability of services, makes cloud computing technologies more appealing to be widely implemented in various sectors [6], [25]. Some of the other main advantages of cloud computing system are particularly in the application of e-government services which includes infrastructure minimization, improvement of network security, service scalability and implementation speed [26]. As suggested in [27], low total cost of ownership, low initial investment costs, faster time to benefits realization, continuous enhancement of the service offerings are usual benefits of implementing cloud computing. A study by Rath et al. [28] tried to find the advantages/benefits of the adoption of cloud computing by Indian companies. They concluded that the advantages for SMEs includes reduced investment in hardware, effective use of computing systems in current data centers, services and cost saving related to technology infrastructure and faster software upgrades with low expense.

C. Cloud Computing Adoption in Organizations

Many studies have been conducted on cloud computing adoption by private and public organizations. Neves et al. [29], have tried to identify various issues that are associated with social, economic, political and technological factors affecting the adaption of cloud computing and had given various ideologies to cope up with these issues. Makena [30] has identified the factors that has an effect on the adoption of cloud computing by Kenyan SMEs. Those factors are: technology, management support, relative advantage, size, preparedness, competitiveness, complexity and compatibility. Azarnik et al. [31] concluded that a risk-management program that is well-defined by which it focuses on cloud computing is essential for the success of organizations' cloud computing transformation. Services such as data mining, optimization, risk modeling and simulation which are computationally intensive are the most important benefits for cloud computing adoption in Indian organizations [28]. The support of software functionalities, the ubiquitous access to data, responsive solutions to customer support queries, maintaining huge data and reducing cost are key success factors for organization's customers to adopt cloud computing [32]. Stieninger and

Nedbal [33] depicted that the factors influencing an implementation of cloud computing are perceived by safety and security, cost, trust management, and efficiency of energy. Hussein et al. [34] used qualitative and quantitative methods to investigate the awareness of cloud computing within SMEs present in Malaysia. As a result, the majority of the users only know about its benefits in terms of data backup and storage. A survey among seventeen IT managers was conducted in Spain concluded that they were also not aware of the cloud computing applications [35] .

From the earlier publications, it can be noticed that many articles which are published are using different theories and models to study the effective implementation of cloud computing. Technology-Organization-Environment framework (TOE) is used in [36] to investigate the effective implementation of cloud computing within the SMEs, especially in UK. Some researchers proposed the implementation of the cloud computing technology based on the already existing TOE framework as shown in [36], [38]. Furthermore, the theory of Diffusion Of Innovation (DOI) has also been used [33]. Some researchers have proposed the process of studying the implementation of cloud computing based on the already defined two models together in a combined form [33], [37]. For example, Saedi and Iahad [37] found that the Technology-Organization-Environment Framework (TOE) and Actor Network Theory (ANT) are suitable for examining cloud computing implementation in various SMEs. Furthermore, Technology Acceptance Model (TAM) applied among the various models and methodologies to study cloud computing adoption [39].

III. METHODOLOGY

This study has utilized a qualitative approach to identify factors that can impact cloud computing adoption. In this approach, IT professional representatives who were invited from various private and public organizations in Saudi Arabia were interviewed. In order to make sure that the sample reflects the geographical spread and size classifications of private and public organizations across Saudi Arabia, the selection of organizations across different industries in Saudi Arabia has been carried out. A summary of the profile of these organizations is presented in Table I.

TABLE I. ORGANIZATIONS' PROFILE

Organization Domain	Organization type	No. of participants
ICT	Private	3
	Public	4
Health	Private	3
Construction	Private	1
Manufacturing	Public	5
	Private	3
Education	Public	4
Finance	Private	3
Telecommunication	Private	2
Traveling	Private	1
Training centers	Private	1
Research centers	Public	2

A. Interview Guide and Process

To improve the reliability of this research, only voluntary participants who were responsible for the planning and the cloud computing adoption in their organizations such as IT directors, IT practitioners, IT executive, IT managers, and system analysts were interviewed (Table II).

TABLE II. PARTICIPANTS' POSITION

Participants' position	Quantity
IT directors	2
System analysts	4
IT practitioners	13
IT executive	7
IT managers	6

An interview guide was developed and used to guide the interview method. In the guide, information about the study aim, rights of participant, responsibilities of the researcher and a confidentiality guarantee on the interview were provided. Interviewees were clearly notified that they may decline to answer any question and provided answers would not be classified as right or wrong (i.e. there is no right or wrong answers in this interview).

To explore the factors that impact cloud computing adoption, the semi-structured interviews incorporated a range of sub-questions through various themes (see appendix A). Pilot interviews with five academic staff in Shaqra University were conducted to examine the research questions' relevancy as well as to evaluate the guide that was given to interviewees. The pilot interviews led to modifications and improvements on the interview questions as well as the interview guide.

The interviews kicked off with a general question about cloud computing; this was followed by questions related to the factors which led to encourage or discourage the organization to adopt cloud computing. Moreover, interviewees were asked to list the criteria which used by them to evaluate the readiness of their organization to adopt cloud computing. Questions about their experiences (positive and negative) with cloud computing adoption were asked. However, the questions as described were posed to the participants who responded in such a manner that the semi-structured nature of the interviews proved beneficial for the researcher with the flexibility of asking further questions to avoid bias by eliciting an obvious response [40]. Few interviewees were re-contacted to confirm or elucidate information in order to validate data.

The 32 personal in-depth interviews were conducted in various locations in Saudi Arabia where it was convenient to the participants between July and September 2017. Since most of the participants speak English fluently, English language used during the interviews, which were between 35 and 50 minutes. Only one participant preferred to be interviewed in Arabic. The interviews were tape-recorded with the permission of all interviewees. The interviews were carried out until the point of data saturation was reached [41].

B. Data Analysis

The analysis of the interview data identified the factors that impact cloud computing and the manner of that impact. Content analysis was utilized in the study as it is suitable for analyzing different communication materials including "*narrative responses, open-ended survey questions, interviews, focus groups, observations, printed media such as articles, books, or manuals*" [42]. Using the steps suggested by [43] for a content analysis process, to draw out the relevant items, themes and categories, the data collected was analyzed (Fig. 1).

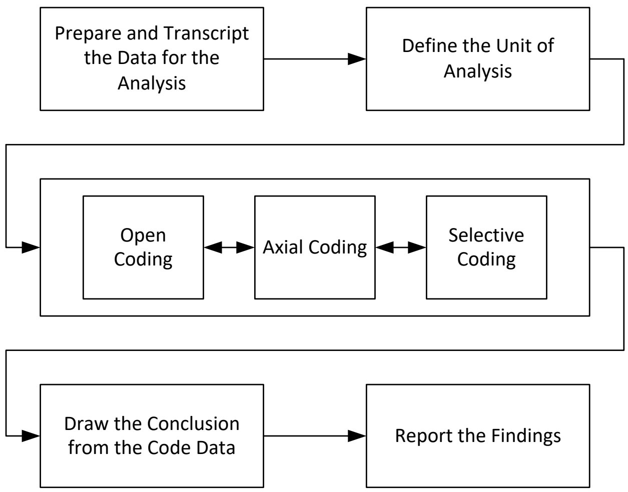


Fig. 1. Content analysis processes.

According to [43], usually, individual themes are utilized in qualitative content analysis as the unit for analysis; these themes might be articulated in a word, a phrase, a sentence, or a paragraph. Each interview was transcribed and prepared for data analysis. In this study, categories and a coding scheme were generated inductively from the data. In order to code qualitative data, the researcher has applied the three coding techniques proposed by [44] which are: open coding, axial coding and selective coding.

1) Open Coding

Open coding technique is used to identify discrete concepts and categories [44]. The open coding was accomplished by reading each transcript to note any interesting or relevant information by making a brief note in the page border about the nature of this information. From these border notes, the researcher made a list of codes from different types of information. After reading through the lists of open code that excerpted from transcripts, the researcher sorted and categorized these initial codes in the way that describe what they are about. Based on these processes, the researcher managed to generate 79 notes based on how different concepts are related and links.

2) Axial Coding

Axial coding is used to connect categories (79 concepts in the first stage) for the relationships by generating subcategories [44]. In this process all the initial concepts in open coding process were reduced by considering possible relationships with other concepts. Based on these processes, the initial 79 concepts were further refined and incorporated into 32 concepts. These 32 subcategories (concepts) were further classified into 7 categories (factors). These factors are security and privacy, lack of knowledge, government policy, loss of control reduce expenses, improve IT performance, and promote scalability and flexibility. The purpose of this process is to group together all subcategories with similar occurrences and incidents into categories. The concepts under each of the factors were then validated by scrutinizing possible relationships with other factors. This process was repeated for each factor to ensure the relationship is held. After repeating this process, all related factors from the transcribed interviews were saturated and no further relationships emerged. Finally, the original interview transcripts were reviewed to identify any possible outliers that could be incorporated into the factors. Further, direct quotes were extracted from the transcripts to support the qualitative data analysis findings [42].

3) Selective Coding

Selective coding is the third coding process which is used to discover the core categories. This stage was accomplished by integrating the related categories (7 factors in the axial coding stage) to discover the core-categories and refining them. The researcher linked these factors under two major categories such as positive impact and negative impact. The coding procedure begins with exploring concepts, followed by connecting the concepts for relationships under main factors, and finally connecting these factors to discover the core-categories. Fig. 2 was translated from the analysis processes to clarify the coding process.

The researcher compared the emerging theme from the data and the current concepts in cloud computing literature. Whenever consistency is founded, the researcher modified the concepts to match with the literature. For example, government policy was initially conceptualized in this study as regulation and law; however, after considering the same meaning concepts in the literature, this concept was modified to government policy. The researcher followed this way in order to be consistent with the literature and to add theoretical soundness to the conceptualization.

IV. Result and Discussion

Cloud computing is a new technology which is implemented in the field of ICT. The effective adoption of cloud computing gradually changed the way of services presented by ICT. Many organizations have already started to realize the importance of implementing cloud computing more effectively [28]. Some of the organizations have also started to implement the technology [36]. However, various issues should be taken into consideration when implementing cloud computing technology. One of the interviewees (participant 16) answered a question about the meaning of cloud computing by replying "*cloud computing provides an on-demand provisioning of resources and scalability. It is a form of IT services delivery over the internet. These services include servers, emails, storage, databases, networking, software, analytics and more* ".

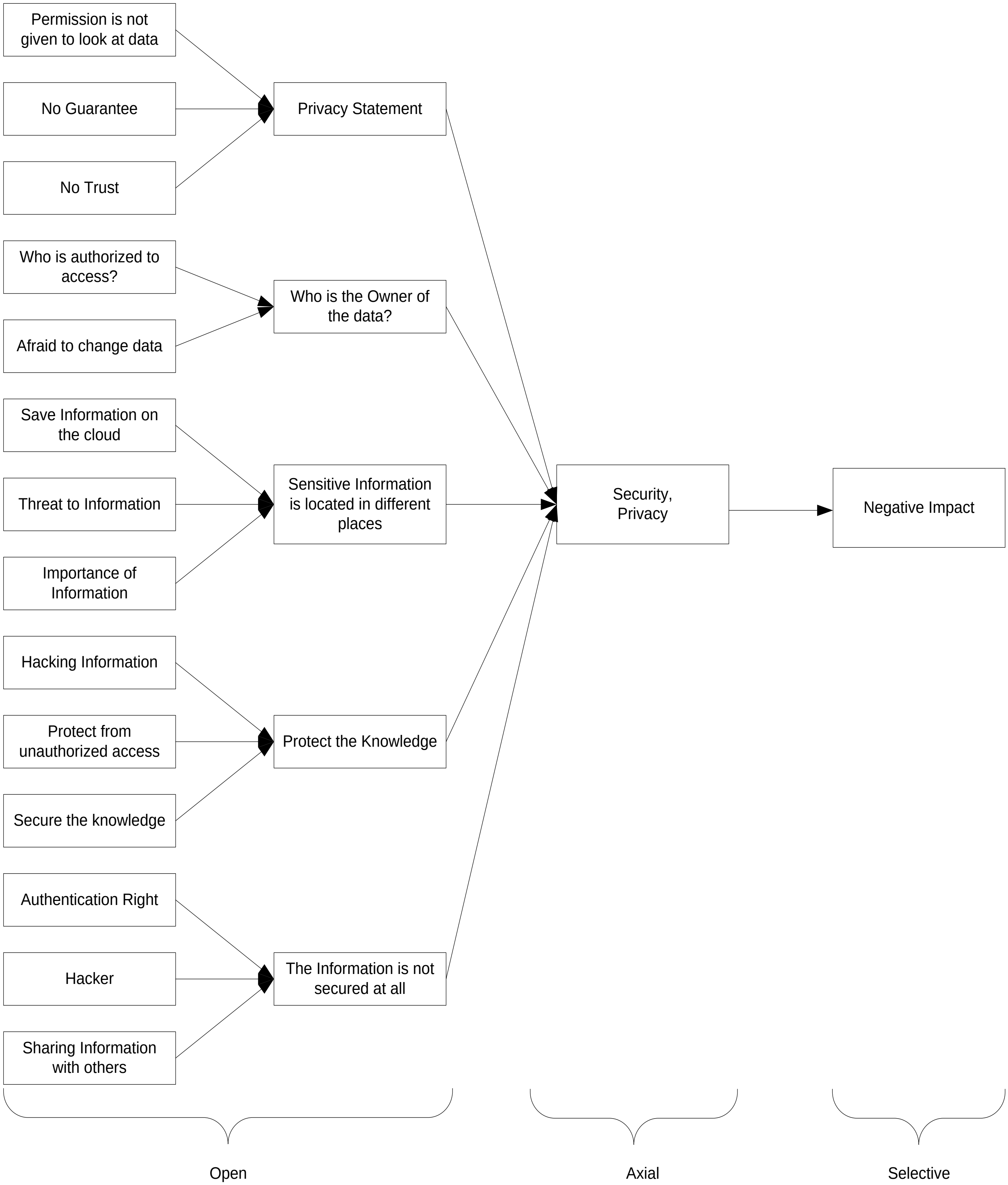


Fig. 2. Illustration of coding (privacy and security).

In this study, the interviewees used different criteria to assess the factors that affect the adoption of cloud of computing. These criteria are sorted into two major categories and 7 factors as follows:

A. *Negative Impacts*

This category has the factors that hinder the adoption of cloud computing by organizations. The results revealed after analyzing the data show that cloud computing can be negatively affected by four factors as shown in Fig. 3.

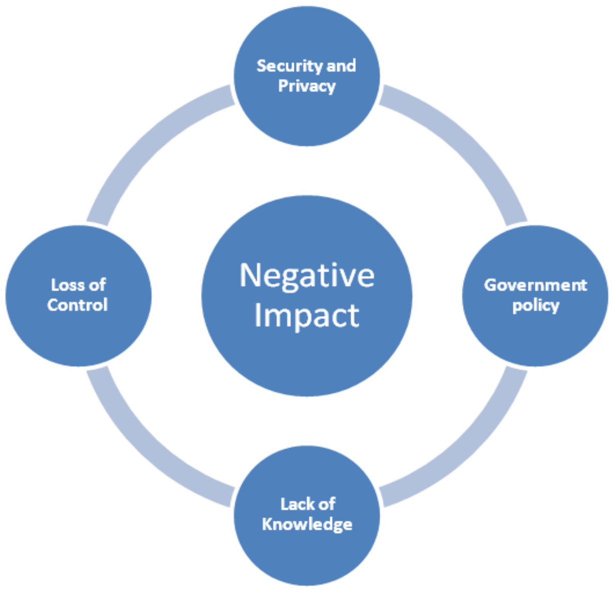


Fig. 3. Negative impact factors.

1) *Security and privacy*

In this study, security and privacy have been identified as one of the main factors impeding the adoption of cloud computing. Unsurprisingly, the same has been found in many previous research [31], [36]. As per [45], security and privacy is one of the key challenges that keep organization away from the adopting cloud computing. Most of the participants in this study saw the security and privacy as major obstacle in adopting cloud. One of the interviewees (participant 1) mentioned that "*before we agree to shift to cloud technologies we need to ensure that our data is secured in the cloud and what the ability does the cloud provider have to deal with denial of service attacks*". As cloud computing includes various technologies including databases, virtualization, resource scheduling, networks, operating systems, memory management, load balancing, concurrency control and transaction management. Consequently, security issues for many of these systems and technologies are applicable to cloud computing. This was confirmed by participant 9 who said "*we need to understand the security requirements of the exit process. For instance, cloud service provider must allow secured network traffic and block malicious network traffic. Also, provider should give customers the essential tools to protect their system*". Similar to this result, Seifu et al. [46] confirmed that the vulnerabilities and threats from infrastructure and network, up to the service platform was major concern in cloud computing adoption.

One of the major concerns is private leakage that has raised a great restriction for many organizations to execute cloud applications [47]. Likewise, many interviewees expressed their concerns on privacy as shown below:

"*The secure processing of personal data and organizational data in the cloud computing represents a very big challenge*" (participant 13).

"*In my opinion, it is difficult to provide your data and business application to a third party because the security and privacy issues become a critical concern these days*" (participant 21) .

2) *Government policy*

Government policy issue has been seen from participants perspective as a threatening factor for the successful implementation of cloud computing. This is consistent with [22], [46]. Government policies regarding how cloud computing applications are imported and implemented can affect the decision of adopting cloud computing within an organization negatively [46]. The interviewees (participant 4 and 9) agree that the government policy related issues such as geographic location of systems and servers cannot be controlled. Furthermore, some governments-imposed restrictions on cross border data flows for privacy protection and security. One interviewee (participant 9) said that "*Cloud computing has grown to be a new phenomenon used by every organization in Saudi Arabia. In this case the government definitely need to control and organize its use*". Whereas as (participant 6) said that "*there are some restricted policies by Saudi government to store the governmental data out of country*".

3) *Lack of knowledge*

The rate of adoption of cloud computing is directly linked to levels of skills and expertise about cloud computing according to [30]. In supporting to this claim one of the interviewees (participant1) said that "*lack of awareness is a key obstacle to cloud usage*". Whereas, (participant 3) said "*many businesses are not able to reap the full benefits of the cloud service because they do not fully understand what it provides*". This is consistent with a study conducted by [1], [48] and concluded that lack of knowledge had a substantial effect on the attitude towards trust and the usage of cloud services. Additionally, the lack of standardization makes it more difficult to know if someone has the required knowledge. One of the interviewees (participant 30) confirmed that *"different cloud services providers use different terminologies".* Therefore, researchers and cloud computing agencies should play an important role to raise the level of cloud technology skills. Moreover, organizations also have to take time to train their employees if they want properly identified, utilize and expertise the service cloud computing.

4) *Loss of control*

Cloud service means that your data and applications are being hosted by someone else's server, not exclusively on your own server. One of the interviewees (participant 11) said *"no one believes that cloud platforms as they exist today are perfect in controlling their data".* To be fair, those who complain about lack of control are often thinking about direct hardware access. Another interviewee (participant 28) said *"Loss of control over data: service user does not know where exactly its data is stored and processed in the cloud*". However, many managers believe that access control is more problematic with cloud computing. In addition, the biggest issue is that the data travel across the internet and we have no control on neither physical infrastructure nor entry/exit outlet points [49].

B. Positive Impact

This category has the factors that encourage the adoption of cloud computing in organizations. The results revealed after analyzing the data show that cloud computing can be positively affected by three factors as shown in Fig. 4.

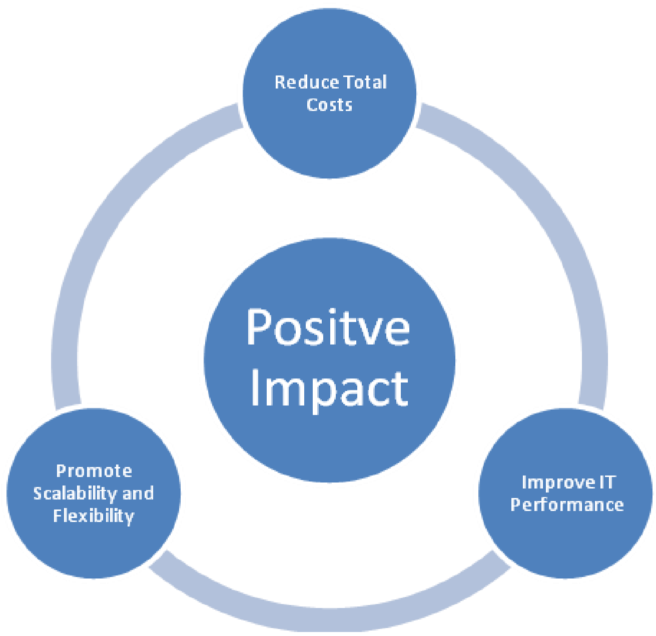


Fig. 4. Positive impact factors.

1) Total cost reduction

Reduction in total cost has been identified as one of the major factors that encourage organizations to adopt cloud computing. For most organizations, traditional process often involves buying technology equipment and tools, getting consulting services, paying for installation and license, and signing up to long maintenance contract. One interviewee (participant 21) mentioned that *"with cloud services, there are no software licensing to purchase, no extra cost for installation services and no annual maintenance contract to renew"*. This is consistent with [50] who claimed that better effective use of cloud computing technology leads to cost savings and productivity gains. The key advantage of cloud computing is to increase the utilization compared to the process of traditional data centers, in term of the cost of electricity, bandwidth utilization, related infrastructure for operations, software, and hardware required. One of the interviewees (participant 13) mentioned *"moving to cloud helps to have better resources and efficient bandwidth availability with less cost"*. In cloud computing, an organization pay based on the hour for computing resources, which leads to cost savings since the rate to rent a machine from a cloud provider is much lower than the cost to own one [22]. Thus, cloud computing offers services with minimum costs and maximize the profit in contrast to a data center [50].

2) Improvement of IT performance

There are several reasons that make cloud computing a beneficial tool for any organization. One of those reasons identified by this study is that cloud computing provides lesser cost computers to users. In other words, the user does not have to have a high-end computer to run cloud computing web-based applications since these applications run in the cloud instead of running them on a user computer. Therefore, cloud computing technologies improve the IT performance by sharing resources [10]. The interviewees agree that cloud computing provides low cost for IT infrastructure with the use of cloud computing recourses. Consequently, the IT department of an organization does not have to invest in a very large number of powerful servers, software and hardware resources. This helps the IT department to enhance or change the internal computing resources. This beneficial tool of cloud computing improves IT performance by increasing the throughput, avoiding load balancing issues, minimizing the delay and response time for transactions, since it is considered as a type of parallel, virtual, distributed, configurable, and flexible systems [50]. This is confirmed by an interviewee (participant 10) who mentioned that *"when the organization needs to improve their performance, they should to think about adopting the cloud computing"*.

3) Promote scalability and flexibility

Scalability is one of the cloud computing performance measurements which is directly perceived by users as technical service reliability and availability, as well as by scalability of the applications [4]. Scalability is one of the important keys to the success of many organizations that are doing business in web applications. Furthermore, maintaining sufficient resources to meet the requirements might be costly [51]. This study discovered that most of the interviewees' participants agreed that promoting scalability and flexibility in any organization is the key factor of providing efficient services, which has a direct impact for capacity, reliability, and availability in cloud computing environment. Cloud computing enables organization employees to be more flexible by sharing documents and other online services, which can also help support internal and external collaboration. Furthermore, cloud computing allows the organization to upgrade or reduce IT requirements as needed. For example, most cloud service providers will allow organizations to increase their existing resources to accommodate increased business needs or changes. This will allow organizations to support their economy growth without making costly changes to their existing IT systems. One of the interviewees (participant 32) added that *"one of the keys successful in adoption cloud computing is IT resource scalability which any organization are in need to provide high processing and large storage capabilities at the lowest possible cost"*.

V. Conclusion and Future Work

There are many factors which need to be considered before adopting cloud computing by private and public organizations in Saudi Arabia. The research findings show some important issues that are influencing the adoption of cloud computing by these organizations. Security and privacy is one of the top leading issues in determining the decision for adopting cloud computing in these organizations. The vulnerabilities and threats from infrastructure and network was a major concern in adopting such technology. This research also shows that an increase in the awareness of cloud computing among private and public organizations serves as the perfect recipe for increased adoption and acceptance of this technology. The Saudi government shall play important role in determining which areas of improvement must be carried out to encourage the spread of ICT and cloud-based solutions. The results depicted that there are some restricted policies by Saudi government to store the governmental data out of country. Therefore, it is recommended that the Saudi government should adopt a new regulatory strategy that will promote the new technology of cloud computing and its impact on security, privacy, and personal data protection. Furthermore, the potential finding of this research illustrates that by adopting cloud computing technologies, IT performance in organizations can be improved, since it provides low cost for IT infrastructure with the use of cloud computing recourses.

The findings of this research recommend that strategies aimed at upgrading the regulatory approach should be introduced. Furthermore, launching this should be strongly advocated in the interests of Saudi Arabia to face the challenges of a successful migration to cloud computing while maintaining international standards. This study finds that mangers and cloud services providers should invest considerably greater resources in adopting cloud computing technologies. Aspects for consideration include improving security and ensuring the privacy are mandatory to guarantee a safe environment for organizations to adopt cloud computing technologies.

In light of the research outcomes, the factors identified are now being tested quantitatively and will be reported in due course. Using this data in various environments and contexts to provide cross-cultural comparisons may enrich the literature and result in understanding the adoption of cloud computing. Also future research can investigate whether the organizations' characteristics affect the adoption of cloud computing.

ACKNOWLEDGMENT

The author would like to thank Shaqra University, Saudi Arabia for providing a great research environment with excellent infrastructure. The author would also like to thank all participants who have attended the interviews.

REFERENCES

[1] Armbrust, M., et al., A view of cloud computing. Communications of the ACM, 2010. 53(4): p. 50-58.

[2] Buyya, R., et al., Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation computer systems, 2009. 25(6): p. 599-616.

[3] Ali, O., et al. Collaborative cloud computing adoption in Australian regional municipal government: An exploratory study. in Computer Supported Cooperative Work in Design (CSCWD), 2015 IEEE 19th International Conference on. 2015: IEEE.

[4] Low, C., Y. Chen, and M. Wu, Understanding the determinants of cloud computing adoption. Industrial management & data systems, 2011. 111(7): p. 1006-1023.

[5] Sabi, H.M., et al., Conceptualizing a model for adoption of cloud computing in education. International Journal of Information Management, 2016. 36(2): p. 183-191.

[6] Lin, A. and N.-C. Chen, Cloud computing as an innovation: Percepetion, attitude, and adoption. International Journal of Information Management, 2012. 32(6): p. 533-540.

[7] Wang, C., et al. Privacy-preserving public auditing for data storage security in cloud computing. in Infocom, 2010 proceedings ieee. 2010: Ieee.

[8] Avram, M.-G., Advantages and challenges of adopting cloud computing from an enterprise perspective. Procedia Technology, 2014. 12: p. 529-534.

[9] Rittinghouse, J.W. and J.F. Ransome, Cloud computing: implementation, management, and security. 2016: CRC press.

[10] Nikkhah, H.R. and R. Sabherwal, Mobile Cloud-Computing Applications: A Privacy Cost-Benefit Model. 2017.

[11] Alkawsi, G.A., A.K. Mahmood, and Y.M. Baashar. Factors influencing the adoption of cloud computing in SME: A systematic review. in Mathematical Sciences and Computing Research (iSMSC), International Symposium on. 2015: IEEE.

[12] Alhammadi, A., C. Stanier, and A. Eardley, The Determinants of Cloud Computing Adoption in Saudi Arabia. Computer Science & Information Technology, 2015. 55.

[13] Al-Ruithe, M., E. Benkhelifa, and K. Hameed, Current State of Cloud Computing Adoption–An Empirical Study in Major Public Sector Organizations of Saudi Arabia (KSA). Procedia Computer Science, 2017. 110: p. 378-385.

[14] Sumner, A., M.A. Tribe, and M.A. Tribe, International development studies: Theories and methods in research and practice. 2008: Sage Los Angeles, CA.

[15] Greengard, S., Cloud computing and developing nations. Commun. ACM, 2010. 53(5): p. 18-20.

[16] Dahiru, A.A., J.M. Bass, and I.K. Allison, Cloud computing: adoption issues for sub-Saharan African SMEs. 2014.

[17] Mell, P. and T. Grance, The NIST definition of cloud computing. 2011.

[18] Voorsluys, W., J. Broberg, and R. Buyya, Introduction to cloud computing. Cloud computing: Principles and paradigms, 2011: p. 1-41.

[19] Gupta, P., A. Seetharaman, and J.R. Raj, The usage and adoption of cloud computing by small and medium businesses. International Journal of Information Management, 2013. 33(5): p. 861-874.

[20] Malligai, V. and V.V. Kumar, Cloud Based Mobile Data Storage Application System. International Journal of Advanced Research in Computer Science and Technology, 2014. 2: p. 126-128.

[21] McLeod, C., Cloud Computing Service for SMEs Launched in Uganda. 2013.

[22] Saeed, I., G. Juell-Skielse, and E. Uppström, Cloud enterprise resource planning adoption: Motives & barriers. Advances in Enterprise Information Systems II, 2012. 429.

[23] Aljabre, A., Cloud computing for increased business value. International Journal of Business and social science, 2012. 3(1).

[24] Sether, A., Cloud Computing Benefits. Browser Download This Paper, 2016.

[25] Miller, M., Cloud computing: Web-based applications that change the way you work and collaborate online. 2008: Que publishing.

[26] Greer Jr, M.B., Software as a service inflection point: Using cloud computing to achieve business agility. 2009: iUniverse.

[27] Seethamraju, R. Determinants of SaaS ERP Systems Adoption. in PACIS. 2013.

[28] Rath, A., et al. Decision points for adoption cloud computing in small, medium enterprises (SMEs). in Internet Technology And Secured Transactions, 2012 International Conference for. 2012: IEEE.

[29] Neves, F.T., et al., The adoption of cloud computing by SMEs: identifying and coping with external factors. 2011.

[30] Makena, J.N., Factors that affect cloud computing adoption by small and medium enterprises in Kenya. International Journal of Computer Applications Technology and Research, 2013. 2(5): p. 517>< meta name=.

[31] Azarnik, A., et al., Associated risks of cloud computing for SMEs. Open International Journal of Informatics, 2012. 1(1): p. 37-45.

[32] Agostino, A., K.S. Søilen, and B. Gerritsen, Cloud solution in Business Intelligence for SMEs–vendor and customer perspectives. Journal of Intelligence Studies in Business, 2013. 3(3).

[33] Stieninger, M. and D. Nedbal. Diffusion and acceptance of cloud computing in SMEs: towards a valence model of relevant factors. in System Sciences (HICSS), 2014 47th Hawaii International Conference on. 2014: IEEE.

[34] Hussein, W.N., R. Sulaiman, and A.K. Hamzah. E-business and cloud computing awareness for Malaysian SMEs: A recommendation from academic and industry perspectives. in Research and Innovation in Information Systems (ICRIIS), 2013 International Conference on. 2013: IEEE.

[35] Trigueros-Preciado, S., D. Pérez-González, and P. Solana-González, Cloud computing in industrial SMEs: identification of the barriers to its adoption and effects of its application. Electronic Markets, 2013. 23(2): p. 105-114.

[36] Alshamaila, Y., S. Papagiannidis, and F. Li, Cloud computing adoption by SMEs in the north east of England: A multi-perspective framework. Journal of Enterprise Information Management, 2013. 26(3): p. 250-275.

[37] Saedi, A. and N.A. Iahad. Developing an instrument for Cloud Computing adoption by Small and Medium-sized Enterprises. in Research and Innovation in Information Systems (ICRIIS), 2013 International Conference on. 2013: IEEE.

[38] Kihara, T. and D. Gichoya. Adoption and use of cloud computing in small and medium enterprises in Kenya. in IST-Africa Conference and Exhibition (IST-Africa), 2013. 2013: IEEE.

[39] Tjikongo, R. and W. Uys. The viability of cloud computing adoption in SMME's in Namibia. in IST-Africa Conference and Exhibition (IST-Africa), 2013. 2013: IEEE.

[40] Easterby-Smith, M., R. Thorpe, and P.R. Jackson, Management research. 2012: Sage.

[41] Guest, G., A. Bunce, and L. Johnson, How many interviews are enough? An experiment with data saturation and variability. Field methods, 2006. 18(1): p. 59-82.

[42] Hsieh, H.-F. and S.E. Shannon, Three approaches to qualitative content analysis. Qualitative health research, 2005. 15(9): p. 1277-1288.

[43] Zhang, Y. and B. Wildemuth, Qualitative analysis of content. 2009. USA: Libraries Unlimited Inc.

[44] Strauss, A.L. and J. Corbin, Pesquisa qualitativa: técnicas e procedimentos para o desenvolvimento de teoria fundamentada. 2008: Artmed.

[45] Schneiderman, R., For Cloud Computing, the Sky Is the Limit [Special Reports]. IEEE Signal Processing Magazine, 2011. 28(1): p. 15-144.

[46] Seifu, S.D., et al., Cloud-Computing: Adoption Issues for Ethiopian Public and Private Enterprises. The Electronic Journal of Information Systems in Developing Countries, 2017. 78(1): p. 1-14.

[47] Qiu, M., et al., Privacy-preserving smart data storage for financial industry in cloud computing. Concurrency and Computation: Practice and Experience, 2018. 30(5).

[48] Despotović-Zrakić, M., V. Milutinović, and A. Belić, Handbook of research on high performance and cloud computing in scientific research and education. 2014: IGI Global.

[49] Ahmed, A.A. and C. Xue Li, Analyzing data remnant remains on user devices to determine probative artifacts in cloud environment. Journal of forensic sciences, 2018. 63(1): p. 112-121.

[50] Chieu, T.C., et al. Dynamic scaling of web applications in a virtualized cloud computing environment. in E-Business Engineering, 2009. ICEBE'09. IEEE International Conference on. 2009: IEEE.

[51] Subashini, S. and V. Kavitha, A survey on security issues in service delivery models of cloud computing. Journal of network and computer applications, 2011. 34(1): p. 1-11.

Appendix A

Interview Questions:

1) What does the cloud computing technologies mean for you?

2) Do you use any of these technologies before?

3) If yes, can you describe of these technologies and how it was deployed and how it was delivered to you via the internet?

4) If no, why did you not use these technologies ?

5) How secure do you think ~~of~~ the use of cloud computing technologies?

6) What specific criteria did you use to evaluate the readiness of your organization to adopt cloud computing ?

7) What aspects of adoption of cloud technologies led you to positive experiences?

8) What aspects of adoption of cloud technologies led you to negative experiences?

9) From your experience, what factors that will encourage organization to adopt cloud computing?

10) From your experience, what factors that will discourage organization to adopt cloud computing?

11) Can you describe the challenges, risks, privacy and trust issues associated with cloud computing adoption?

Multi-Level Encryption Framework

Ahmad Habboush
Faculty of Computer and Information Technology
Jerash University
Jerash, Jordan

***Abstract*—Multi-level encryption approaches are becoming more popular as they combine the strength of multiple basic/traditional approaches into a complex one. Many multi-level encryption approaches have been introduced for different systems, like Internet of Things, sensor networks, big data, and the web. The main obstacles in building such approaches are to have a secure as well as a computationally efficient multi-level encryption approach. In this paper, we propose a computationally efficient multi-level encryption framework that combines the strength of symmetric, the encryption algorithm AES (Advance Encryption Standard), Feistel network, Genetic Algorithm's Crossover and Mutation techniques, and HMAC. The framework was evaluated and compared to a set of benchmark symmetric encryption algorithms, such as RC5, DES, and 3-DES. The evaluation was carried out on an identical platform and the algorithms were compared using the throughput and running time performance metrics and Avalanche effect security metric. The results show that the proposed framework can achieve the highest throughput and the lowest running time compared to the considered benchmarked symmetric encryption algorithms and passes the avalanche effect criterion.**

Keywords*—*Multi-level encryption; Advance Encryption Standard (AES); Feistel encryption; symmetric encryption algorithm

I. Introduction

Nowadays, different types of communication are available at the same time such as people to people, people to objects, and objects to objects. Communication is used worldwide and it is spread in almost all areas; such as industry, academia, health…etc. One of the main challenges in communication is security.

As the Internet or any network channels are considered to be unsafe, the aim of having security measurements is to ensure the data are transmitted and received without being stolen, manipulated or deleted. The most commonly used security measurement for network communication is encryption [1].

Encryption is the process of manipulating a plain text message into a ciphered one, usually this process is done with the use of a key. Based on the type of the key used in the encryption/decryption process, the encryption algorithms can be categorized in to symmetric key and asymmetric key. In symmetric key algorithms, the same key can be used to encrypt/decrypt a plain text. Consequently, these algorithms are secure and computationally efficient. Examples of such algorithms: AES, RC6, MARS, Bluefish, DES, and 3-DES [2]-[4]. On the other hand, asymmetric encryption algorithms require two keys for encryption and decryption, thus, making the generation of these keys computationally high and not efficient for encrypting large data. Example of asymmetric encryption algorithm is RSA [5].

With the advancement of technology and computation power, hackers have developed different models to attack such basic encryption algorithms. One of the solutions to deal with this problem is to build a complex model that consists of multi-level encryption algorithms [6]-[9]. The idea is to combine the strength of multiple basic encryption algorithms together to build a complex, sophisticated encryption approach. Two challenges are facing such approaches; computation efficiency and security. The earlier is needed when sending large data and the later ensures that the combination between these levels (algorithms) will lead into a complete system that is secure. Many multi-level encryption approaches were proposed in literature, but they were suffering from computational or security problems or being specified for a certain type of networks.

In this paper, we propose a multi-level encryption framework that is secure and computationally efficient. The framework consists of using a symmetric encryption algorithm AES, Feistel network, Crossover and Mutation techniques from genetic algorithm, and HMAC for encrypting and decrypting data.

The proposed framework was evaluated against benchmarked symmetric encryption algorithms like: RC6, DES, and 3-Des for performance and security metrics. In terms of performance, the proposed framework is compared against them using throughput and running time metrics. The results show that the proposed framework has the highest throughput and the lowest running time. For security purposes, the Avalanche effect is used. The proposed framework has the highest avalanche effect and passes that criterion.

The rest of the paper is as follows: Section 2 review some related work. Section 3 introduces the system framework. Section 4 presents the experimental evaluation. Finally, Section 5 concludes the paper.

II. Related Work

A. *Selecting a Template*

Many algorithms and approaches have been proposed to deal with the encryption/decryption problem. Besides the classical symmetric and asymmetric encryption algorithms, new encryption algorithms categories are being introduced in recent years. Some of these categories include: Homomorphic encryption [10]-[14], attributes-based encryption [15]-[20], and multi-level encryption algorithms. In this section, some recent approaches that uses multi-level encryption are introduced.

S. Aljawarneh, et al. [6] proposed a resource efficient encryption algorithm for multimedia big data. The algorithm is composed of a framework with multi-level encryption that includes: Fiestel encryption scheme, AES with Sbox, and genetic algorithm. There is no key generation since the key is generated from the plain text. The authors propose a multi-threaded version of the scheme to increase performance [7]. The main drawback of this scheme is that the scheme does not preserve confidentiality. If two senders are using this scheme, they can encrypt/decrypt the messages from the other sender.

S. Masadeh, et al. [8] introduced a multi-level security approach that consists of PGP, WIFI, and HMAC systems for authentication and communication encryption. The approach uses the PGP to generate messages and the message is encrypted using the public key of the receiver.

S. Aljawarneh, et al. [9] proposes a multi-level encryption system for WiFi called secure WiFi. The system is composed of Feistel encryption and HMAC. Their proposed system uses 64bit blocks to be encrypted which are not suitable for encrypting large data.

III. System Model

The system model is similar to the one proposed by [6] with an additional security level to solve their confidentiality problem. Fig. 1 shows our proposed system model. The original file to be encrypted is split into chunks, where each one has the size of 256bit. The 256bit plaintextchunksis divided into two equal 128bit blocks. The first block is encrypted using a 3-level non-key Feistel network. This encryption is composed of bit and bytes shifting and rotation processes. The second 128bit block is encrypted using Advance Encryption Standard (AES) where the key for this encryption is the output of the Feistel network encryption. The third step is to use the crossover and mutation techniques from genetic algorithm to scrapple the two blocks together. Lastly, we use HMAC hash function to encrypt the result using a hash key. The resulted encrypted file is sent over network channel to the receiver for encryption. The decryption process is opposite of the encryption process. Next, we explain these levels in more details.

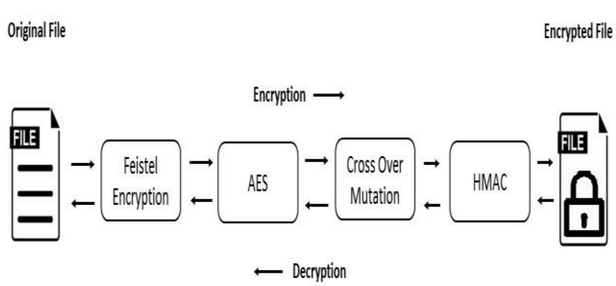


Fig. 1. Proposed framework.

A. Feistel Encryption

As a start, the file is divided into 256bit chunks and each chunk is encrypted using the proposed framework. For a chunk, the first step is to split it into two 128bit blocks (right and left). The left block is encrypted using 3-level Feistel network as shown in Fig. 2. The first level splits the block into individual 16 bytes and defines the highest and lowest bits within a byte. The second level shifts right the highest bits within each byte, so they become the lowest bits. At the same time, the lowest bits of each byte are rotated to become the highest bits of another byte as shown in the figure. Lastly, the bytes are shifted and rotated and combined to form a ciphered key that will be used in the AES encryption.

B. Advance Encryption Standard (AES)

The Advance Encryption Standard (AES) is a symmetric encryption algorithm that needs only one key for encryption and decryption. AES is considered to be fast and secure encryption algorithm. It uses substitution and permutation network where it iterates the encryption process for 10 rounds to generate the cipher text. In AES, the size of the key is very crucial since it will determine the number of encryption rounds. In general, AES uses 10 rounds for 128bit key and 12 rounds for 192bit key.

In our proposed framework, AES will use the ciphered key generated from the Feistel encryption as a key to encrypt the right 128bit block. The result will be an encrypted version of the right block.

AES consists of three main stages: Add Round Key, Rounds, and Final Round. The add key round or the initial round uses a bitwise XOR to combine the block of the round key with each byte of the state.

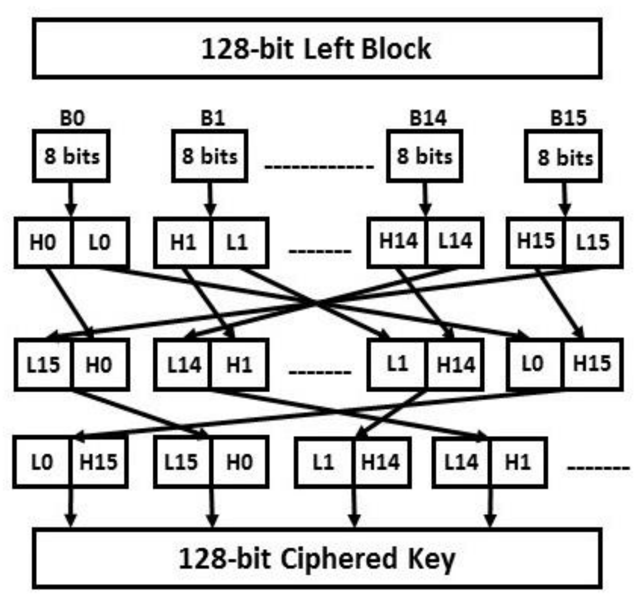


Fig. 2. Feistel encryption.

The rounds stage is mainly consist of four main substages: the sub-byte process, shift rows process, mix columns process, and the add round key process. In our framework, these substages are repeated for 8 times. Lastly, the final round which is similar to one round of the rounds stage with a small difference that the mix columns process substagewill not be performed. By the end of the AES encryption level, both left and right blocks are encrypted.

C. Cross Over and Mutation

The aim of this stage is to scatter the encrypted block generated from the Feistel encryption and the one generated by the AES. The cross over and mutation techniques are parts of genetic algorithm. The cross over is an operation where two blocks are changed together in a random way. Fortwo blocks, cross over starts by randomly choosing the cross over points. These points are recorded and stored for the decryption process. Next, the blocks from the cross over points to the least significant bit are interchanged. For example: for two blocks 010100101011 and 100101010101. If the cross over point is 8, the resulting blocks are: 010100100101 and 100101011011.

The mutation operation is used in genetic algorithm to maintain the generic diversity of results from one generation to another in a random way. For two blocks, the mutation operation starts by selecting one mutation point in a random way. At each block, the bit corresponding to the mutation point is flipped and the mutation point is recorded for the decryption process. An example of the mutation operation: for two blocks 010100101011 and 100101010101. If the mutation point is 8 then the resulting two blocks are 010100100011 and 100101011101.

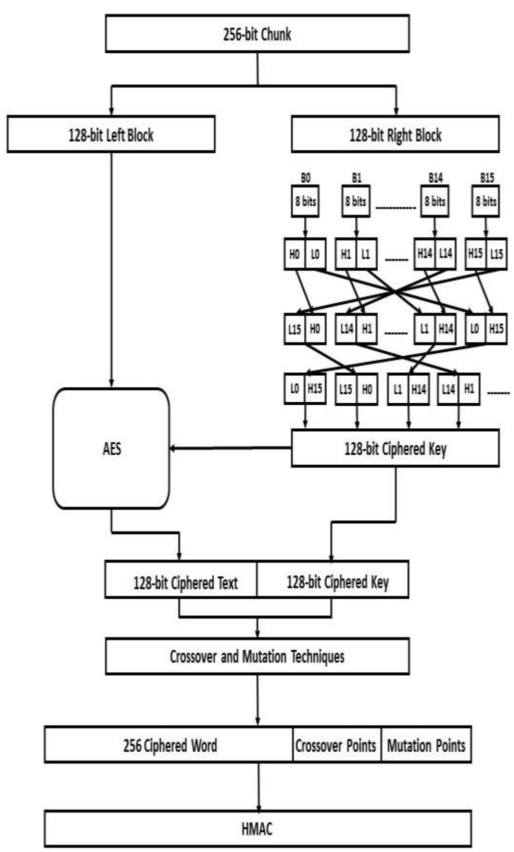


Fig. 3. Architecture of the proposed encryption framework.

The crossover and mutation operations are repeated for many times. In our experiments we repeat these operations for 5 times. By the end of this stage, the encrypted data that need to be sent to the receiver includes: the encrypted 256bit chunk, the cross over points, and the mutation points.

D. HMAC

HMAC is message authentication code that involves the use hash function and key. The strength of HMAC is based on the strength of the hash function used. In our proposed framework we use SHA-1 hash function for authenticate the encrypted message. The aim of this stage is to maintain confidentiality of the encrypted messages. By the end of this stage, the encrypted message is ready to be sent to the receiver. Fig. 3 shows the architecture of the encryption part of our proposed framework.

Note that the decryption process is the opposite of the encryption process. The receiver just need to have the key of for HMAC and perform the decryption in a reverse order of the encryption process. Fig. 4 presents the architecture of the decryption process of our proposed framework.

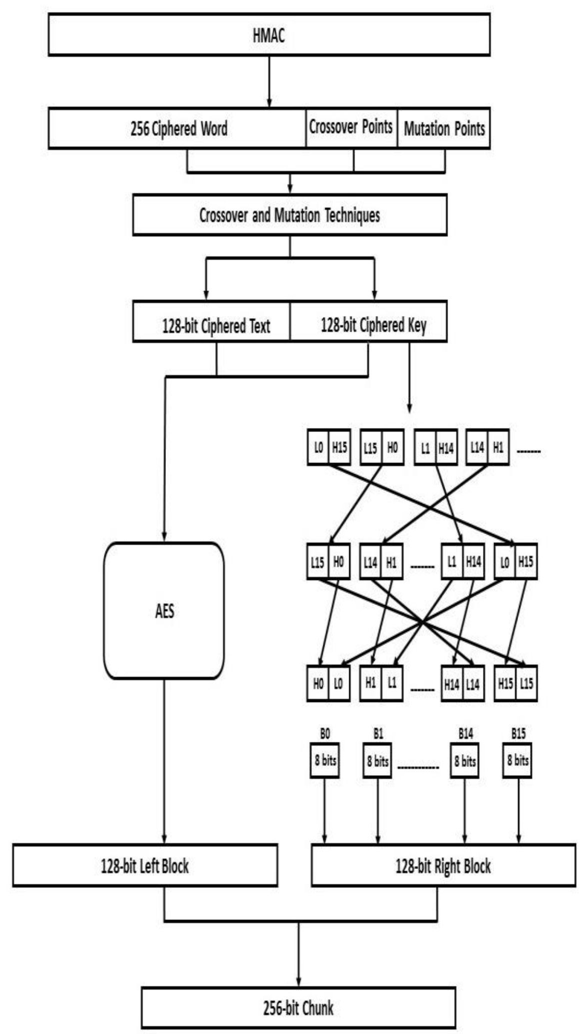


Fig. 4. Architecture of the proposed decryption framework.

IV. Experimental Evaluation

The evaluation is conducted to show the efficiency and performance of the proposed framework. The proposed framework was implemented using JAVA. All experiments were conducted on an identical platform; a Windows based machine that is equipped with 32 Gb of memory and an Intel i7 4770 3.4 GHz CPU.

The proposed framework results were compared to a number of symmetric encryption algorithms such as CR6, DES, 3-DES using file sizes ranges from 1MB to 1GB. The encryption running time and throughput performance metrices are used in this evaluation. For security evaluation, we compare the Avalanche effect security metric.

Fig. 5 (a)-(e) illustrates the encryption running time for our proposed framework compared to RC6, DES, and 3-DES encryption algorithm for files range from 1 MB to 1GB. The results clearly show that the proposed framework outperforms all other encryption algorithms for file sizes 1 MB to 1 Gb. AES is very fast and secure as it has very strong resistance against attacks. On the other hand, RC6 is less secure, needs more rounds, and it uses extra multiplication operation that increases the encryption running time. DES uses a 56bit key, has a S-Box structure where the encryption operation needs to use a lookup mechanism. This lookup mechanism will lead to a slow software implementation. The 3-DES is the slowest encryption algorithm, it uses 168bit key size and it runs DES three times. Fig. 6 shows the average encryption running time for all ranges.

The encryption process throughput for files range from 1MB to 1GB is shown in Fig. 7 (a)-(e). Similar to the encryption running time, our proposed framework has the highest throughput in comparison with the other encryption algorithms. Two main factors affect the throughput: the file size and running time. As our proposed framework has the lowest running time, it is expected that it will provide the highest throughput. It should be noted that the results only show the encryption process since the decryption process is just the reverse of the encryption process and it provides almost the same results. Fig. 8 shows the average encryption throughput for all ranges.

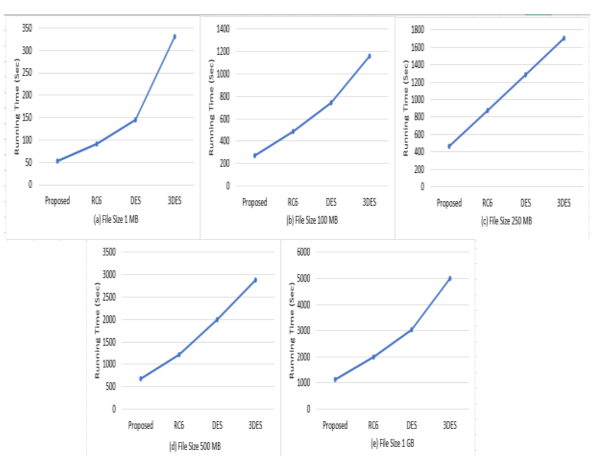


Fig. 5. Average Encryption part running time for file ranges: (a) 1MB (B) 100MB (c) 250MB (d) 500MB (e) 1GB.

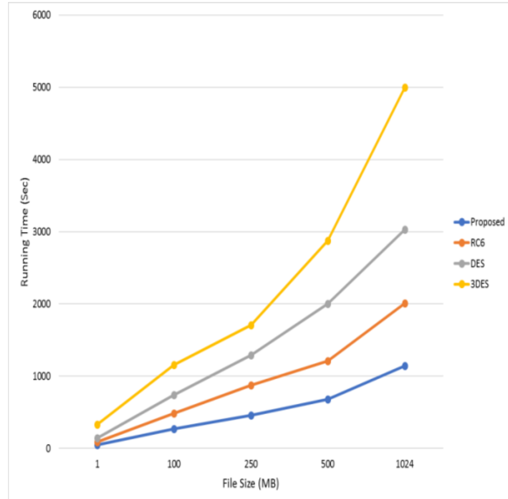


Fig. 6. Average encryption running time for all ranges.

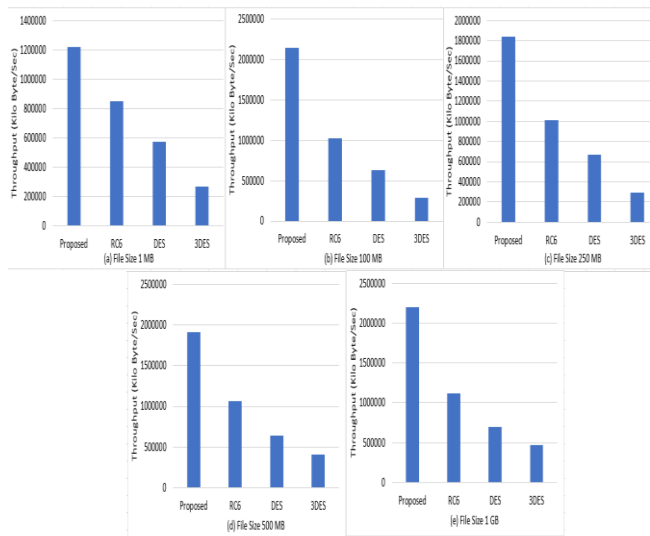


Fig. 7. Average throughput for file ranges: (a) 1MB (B) 100MB (c) 250MB (d) 500MB (e) 1GB.

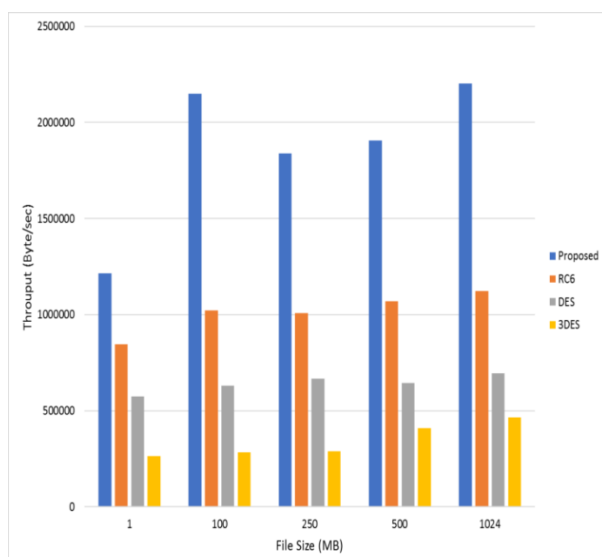


Fig. 8. Average encryption throughput for all ranges.

To evaluate security, the Avalanche effect is calculated for our proposed framework and other symmetric encryption algorithms. The idea of the Avalanche effect is to show the effect of changing one bit in the plaintext to the ciphertext. This will show how solid the encryption algorithm is against cracking and hacking threats and real time attacks. For an algorithm to pass the avalanche effect criteria, changing one bit in the plaintext is expected to affect half of the bits in the cipher text [21].

Fig. 9 shows the avalanche effect of the proposed framework compared to benchmarked algorithms; RC6, DES, 3-DES. The avalanche effect of our proposed framework was 58% compared to 44% for RC6, 27% for DES, and 35% for 3-DES. From these results, it can be seen that the proposed framework is the only encryption that satisfies the avalanche effect criterion.

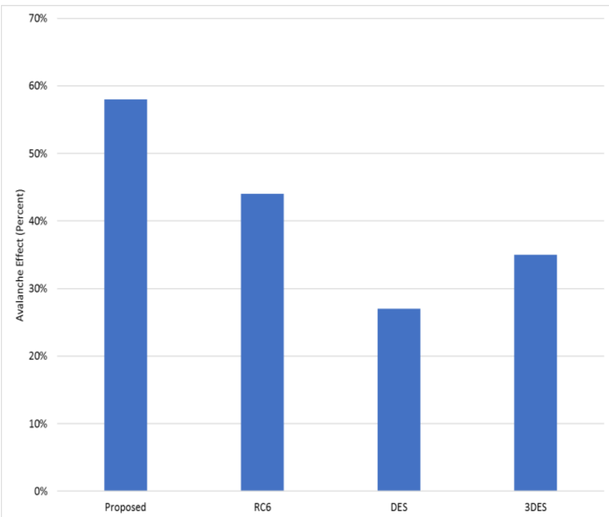


Fig. 9. Avalanche effect.

V. Conclusion

Multi-level encryption algorithms are popular since they combine the strength of many encryption techniques at the same time. In this paper, we proposed a multi-level encryption framework that combines the strength of Feistel encryption, AES, Crossover and mutation, and HMAC. The framework was evaluated against symmetric encryption algorithm CR6, DES, 3DES for performance and security metrics. The results show that the proposed framework has the lowest running time, highest throughput, and passes the Avalanche effect criterion.

References

[1] William Stallings “Network Security Essentials (Applications and Standards)”, Pearson Education, 2004.

[2] Mohan, H.S. and Reddy, A.R., 2011. Performance analysis of AES and MARS encryption algorithms. IJCSI International Journal of Computer Science Issues, 8(4), pp.1694-0814.

[3] Ebrahim, M., Khan, S. and Khalid, U.B., 2014. Symmetric algorithm survey: a comparative analysis. International Journal of Computer Applications (0975 – 8887) Volume 61– No.20, January 2013.

[4] Daemen, J., and Rijmen, V. "Rijndael: The Advanced Encryption Standard." Dr. Dobb's Journal, March 2001.

[5] Vincent, P.D.R., 2016. RSA Encryption Algorithm-A survey on its various forms and its security level. *International Journal of Pharmacy and Technology*, *8*(2), pp.12230-12240.

[6] Aljawarneh, S. and Yassein, M.B., 2017. A resource-efficient encryption algorithm for multimedia big data. Multimedia Tools and Applications, 76(21), pp.22703-22724.

[7] Aljawarneh, S. and Yassein, M.B., 2017. A multithreaded programming approach for multimedia big data: encryption system. Multimedia Tools and Applications, pp.1-20.

[8] Masadeh SR, Azzazi A, Alqaralleh BA, Al Sbou AM. A novel Paradigm In Authentication System Using Swifi Encryption/ Decryption Approach. International Journal of Network Security & Its Applications. 2014 Jan 1;6(1):17.

[9] Aljawarneh S, Masadeh S, Alkhateeb F. A secure wifi system for wireless networks: an experimental evaluation. Network Security. 2010 Jun 30;2010(6):6-12.

[10] Brakerski, Z. and Vaikuntanathan, V., 2014. Efficient fully homomorphic encryption from (standard) LWE. *SIAM Journal on Computing*, *43*(2), pp.831-871.

[11] Brakerski, Z., Gentry, C. and Vaikuntanathan, V., 2014. (Leveled) fully homomorphic encryption without bootstrapping. ACM Transactions on Computation Theory (TOCT), 6(3), p.13.

[12] Ducas, L. and Micciancio, D., 2015, April. FHEW: bootstrapping homomorphic encryption in less than a second. In Annual International Conference on the Theory and Applications of Cryptographic Techniques (pp. 617-640). Springer, Berlin, Heidelberg.

[13] Dowlin, N., Gilad-Bachrach, R., Laine, K., Lauter, K., Naehrig, M. and Wernsing, J., 2017. Manual for using homomorphic encryption for bioinformatics. Proceedings of the IEEE, 105(3), pp.552-567.

[14] Lepoint, T. and Naehrig, M., 2014, May. A comparison of the homomorphic encryption schemes FV and YASHE. In International Conference on Cryptology in Africa (pp. 318-335). Springer, Cham.

[15] J. S. Su, D. Cao, X. F. Wang, Y. P. Su, and Q. L. Hu, “Attribute-based encryption schemes,” Journal of Software, vol. 6, pp. 1299–1315, 2012.

[16] Q. Tang and D. Ji, “Verifiable attribute-based encryption,” International Journal of Network Security,

vol. 10, no. 2, pp. 114–120, 2010.

[17] G. Wang, Q. Liu, and J.Wu, “Hierarchical attribute-based encryption for fine-grained access control in cloud storage services,” in Proceedings of the 17th ACM conference on Computer and communications security, pp. 735–737, 2010.

[18] G.Wang, Q. Liu, J.Wu, and M. Guo, “Hierarchical attribute-based encryption and scalable user revocation for sharing data in cloud servers,” Computer & Security, vol. 30, pp. 320–331, 2011.

[19] B. Waters, “Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization,” Public Key Cryptography V PKC, vol. 6571 of LNCS, pp. 53–70, 2011.

[20] Yao X, Chen Z, Tian Y. A lightweight attribute-based encryption scheme for the Internet of Things. Future Generation Computer Systems. 2015 Aug 1;49:104-12.

[21] Heys HM, Tavares SE. Avalanche characteristics of substitution-permutation encryption networks. IEEE Transactions on Computers. 1995 Sep;44(9):1131-9.

A Developed Collaborative Filtering Similarity Method to Improve the Accuracy of Recommendations under Data Sparsity

Hael Al-bashiri,
Mansoor Abdullateef Abdulgabber, Awanis Romli
Faculty of Computer Systems & Software Engineering,
Univirsity Malaysia Pahang, Gambang, Kuantan, Malaysia

Norazuwa Binti Salehudin
Kolej Komuniti Bentong,
Km6 Karak Setia,
28600 Karak, Pahang

***Abstract*—This paper presented a new similarity method to improve the accuracy of traditional Collaborative Filtering (CF) method under sparse data issue. CF provides the user with items, that what they need, based on analyses the preferences of users who have a strong correlation to him/her preference. However, the accuracy is influencing by the method that use to find neighbors. Pearson correlation coefficient and Cosine measures, as the most widely used methods, depending on the rating of only co-rated items to find the correlations between users. Consequently, these methods have lack of ability in addressing the sparsity. This paper presented a new proposed similarity method based on the global user preference to address the sparsity issue and improve the accuracy of recommendation. Thus, the novelty of this method is the ability to solve the similarity issue with a capability of finding the relationship among non-correlated users. Furthermore, to determine the right neighbors during the process of computing the similarity between a pair of users, the developed method considered two main factors (fairness and proportion of co-rated). The MovieLens 100K benchmark dataset is used to evaluate the developed method accuracy. The experiments' result showed that the accuracy of the developed method is improved compared to the traditional CF similarity methods using a specific common CF evaluation metrics.**

Keywords—Recommendation system; collaborative filtering; similarity method; data sparsity

I. Introduction

Nowadays, information overload is a big challenge suffered by people in their daily life. Therefore, to make their decisions to filter information, they tend to ask friends, scan newspapers, follow advertising, and so forth. This may help them to alleviate this issue [1]. However, the information available on the Internet is still growing daily in a tremendous amount which is progressively becoming a big challenge in people's daily life. This motivates researchers to automate the recommendation way to assist users to address these challenges. The Recommender System (RS) can suggest a set of items that may be related to their favorites, among the tremendous amount of data available [2]-[16]. RS is worked by creating the target user's profile preference and matching it with other users' profiles preferences in the database to locate his/her similar users.

Moreover, the RS can be classified into three approaches, based on the state-of-the-art in recommender systems, which are Content-based (CB), Collaborative Filtering (CF) and Hybrid approaches [6], [8], [17]-[24]. CF is one of the most successful methods used to provide service of recommendation. It proposes the items based on the analysis the feedback provided by the users [3], [25]-[28]. Moreover, CF can be grouped into two main models: model-based and memory-based models [1], [29]. Where model-based need to build a model that will be used later to predict what the users will be preferred. Whereas, the pre-built model in memory-based not required. The correlations between the users/items are calculated directly based on the feedback provided by the users. This correlation can be computed in space of users (user-based) or items (item-based) [1], [8], [30].

In general, the main idea behind CF is that the users who have similar preferences in the past they will share similar preferences in the future [31]. Consequently, finding the nearest neighbors is the critical phase in the CF approach. Therefore, the similarity method in the CF is fundamental to its performance. In the existing methods, there are several proposed methods, such as Pearson's Correlation (PCC), Cosine, their derivatives methods and others [32]. However, providing high-quality recommendations to users with whom the system does not have enough information about their preferences is a key challenge faced by the CF system. Since most users do not rate enough number of items in the database, then the user-item rating matrix will be usually sparse [2], [19], [33]-[44]. As a result, finding the correlation among users who have a small number of ratings will be a problem that might lead to locating unsuccessful neighbors and in turn lead to weak recommendations. Moreover, there is still a room for the development of recommendation accuracy, as well as the similarity methods in locating the accurate neighbors.

Therefore, the primary goal of this paper is to developed a new similarity CF method to enhance the recommendation in term of accuracy. This developed method will be reliant on the global preferences to address the issue of data sparsity. Disparate the traditional similarity methods that depend on the ratings. The developed method builds users' profile preferences by adopting the item types more than rating data. In addition, the new similarity measure considered two main factors. First one is the proportion of the number of items rated by the target user to the number of items taken by both users. It is taken into account to ensure the fairness when calculating the correlations between a pair of users. The correlation weight between the pair of users should be increased as the number of ratings for each of them is close and vice versa. The second one, to devalue the correlation value when the number of co-rated items is small the percentage of common items is considered. Several experiments will be contacted on MovieLens benchmark datasets to evaluate the developed method.

The structure of the remainder of this work is as follows: Literature review is discussed in Section 2. The developed similarity method and its phases are presented in Section 3. In Section 4, the evaluation process and experimental results are discussed. This paper is concluded in Section 5.

II. Literature Review

Due to the similarity measures have a significant effect on the quality of recommendations several developments on the similarity methods have been done. For example, in [45], the contextual information is used to improve the traditional similarity methods via analyzing the singularity of user ratings. The researchers classified the ratings into positive and non-positive. Next, calculate the singularity of each user and item in the database. Finally, they combined the singularity values with actual ratings to calculate the similarity weight between users. Their philosophy says that "if 95% of users voted positively for the item, the similarity derived (for this item) between two users who belong to the 5% (very singular) must be greater than the similarity derived between two users who belong to the 95% (not very singular)". Moreover, in work [46], the authors developed new similarity method depends on the three types of significances: the significance of an item, the significance of each user in providing recommendations to other users and the significance of an item for a user. Then, according to these significances, the PCC and Cosine similarities methods are used. Choi and Suh [47] introduced a combination of traditional methods to give a new similarity method. It considered the correlation between the target item and each co-rated item in the process of computation similarity between users. Get a different set of neighbors with each different target item. They combined PCC, Cosine and Distance methods to compute the item and user correlation, respectively. Another improved similarity measure was presented by Mao and Cui in [48]. To solve the issue of data sparsity, the authors added impact factor to the traditional similarity measure. This impact factor, ε, represents the ratio of co-rated items by the pair of users. However, if they do not have common items, then the correlation value between them is zero. Moreover, Huang and Dai [49] proposed Weight Distance Model (WDM) to calculate the correlation between users. In this proposed method, the ratio of co-ratings and the similarity between a target item and each item in the co-rated set are considered. Additionally, a new weighting method takes into account the compromise factor was introduced in [50]. The compromise is the fraction of the number of common items on non-co-rated items. Nevertheless, there is still a drawback, may lead to low recommendations when the number of common not more enough.

Other works improved similarity measures are introduced. Shunpan, Lin [51] introduced a singularity-based similarity measure. A pair of users should be having strong correlation if they rate items which are rated by only a few users compared to when they rate items which are rated by many users. Next, PCC is improved and used based on these singularity values. Moreover, the ratio of co-rated items is considered using adjusted Jaccard measure. Whereas, the authors in [52] proposed a new method for choosing neighbors depend on intersection and union neighborhood. Firstly, the neighbors are the ones who share same items of the target user. Secondly, the neighbors are the ones who share at least one item of the target user. However, all these measures still depend on common items when finding neighbors of a target user. Therefore, if there are no common items between a pair of users the similarity between them cannot be calculated and will be zero.

Mahara [53] introduced a new combined similarity method to enhance the recommendation accuracy under data sparsity issue. This new method utilized Mean Measure of Divergence that takes into account the user' rating behavior (low or high). The PCC, Jaccard and Measure of Divergence are combined to find the correlation between users. [54] is another example of a linear combination that combined PCC and Jaccard measures for web service recommendation. In addition, NWSM is a new weight similarity model proposed by Zang, Liu [55]. It considers the percentage of common rating (Jaccard), user rating preference (PCC), and the different contributions of other users to the target (mean and variance of the rating). Cao, Deng [56] used Bhattacharyya Coefficient (BC) to improve the similarity method to solve the issue of sparse data. Two main steps are adopted in this method. First, finding the nearest neighbors of items by calculating the CB similarity between two items and take the top N items to identify the neighborhood of the target item. Second, locating the nearest neighbors of users using the similarity method in [36].

Recently, some other methods in the improvement of similarity measures have been proposed. Koohi and Kiani [33] presented a new method based on the subspace clustering technique to address the problem of data sparsity and high dimensionality. The item space divided into three subspace Interested, Neither Interested nor Uninterested, and Uninterested. Next, based on these subspaces the correlations between users are calculated. Moreover, Bilge and Yargıç [57] to improve the multi-criteria CF accuracy, authors applied z-score and decoupling normalization to overwhelm the negative effects of varying rating habits of users. Zhang and Yuan [43] improved similarity method by analyzing the shortcomings of traditional memory-based CF. In the improved method, the correlation between co-rated items and all items rated by the target user is considered. In the study [58], a new linear combination similarity method is proposed to overcome the issue of data sparsity. The global preferences, local context of the user behavior and proportion of common ratings between two users are considered based on PSS, Bhattacharya Coefficient, and Jaccard, respectively.

From this quick discussion, almost every paper that has been discussed includes a different similarity method attempting to improve the recommendation accuracy. Overall, these studies highlight the importance of similarity measure in improving the recommendation accuracy. However, the process of similarity calculating among users in the most of those measures depends on the co-rating. Therefore, the process of finding the relationship between a pair of users who do not have co-rating will be complicated. Moreover, most of the users do not have enough ratings. Therefore, the chance of users to have enough common items will be decreased. Consequently, the computation correlation will be more difficult and might lead to a fake relationship.

III. Proposed Similarity Method

Some existing works related to improving similarity measure have been discussed in the previous section. Therefore, this section will start with a brief introduction about the motivation and assumptions of the proposed method. Next subsection will present the global preference representation. Finally, the proposed similarity method and prediction method will be presented.

A. Motivation

From related work section, we can note that the correlation between a pair of users in the most improved similarity methods depends on the PCC or Cosine measures which can be computed as shown in (1) and (2), respectively.

$$S(x,y)^{PCC} = \frac{\sum_{i\in I_{x,y}}(r_{x,i}-\bar{r_x})(r_{y,i}-\bar{r_y})}{\sqrt{\sum_{i\in I_{x,y}}(r_{x,i}-\bar{r_x})^2}\sqrt{\sum_{i\in I_{x,y}}(r_{y,i}-r_y)^2}} \quad (1)$$

$$S(x,y)^{\text{Cosine}} = \frac{\sum_{i\in I}(r_{x,i})(r_{y,i})}{\sqrt{\sum_{i\in I_{x,y}}(r_{x,i})^2\sum_{i\in I_{x,y}}(r_{y,i})^2}} \quad (2)$$

Where $s(x,y)$ is the similarity between user x and user y, and $I_{x,y}$ represents a set of items which rated by both users *x* and *y*. The symbols $\bar{r_x}$ and $\bar{r_y}$ symbolize the average rating of user *x* and *y*, respectively. $r_{x,i}$ denotes to the rating value of the item *i* by the user *x*.

Although those similarity methods have been proposed based on PCC and Cosine to improve the accuracy of recommendation, there are some shortcomings. First, the similarity calculation in most of those measures depends on the co-rating. Therefore, the similarity calculation will suffer when the user has few ratings. This reduces the chance of users to have common items, and then the computation correlation will be more difficult. Second, some users rate items randomly so it is not right to say that the rating value reflects their interest. This may lead to locating unsuccessful neighbors. Third, ignoring the proportion of common ratings and not considering absolute value also will lead to low accuracy. Next, ignoring the fairness factor, users who have the same number of rating items should have more strong correlation than others, through similarity calculation process also may come out with the fake relationship. Finally, the memory-based CF mechanisms still have an open room for enhancement in the accuracy of recommendations. Therefore, the researcher going to develop new similarity method depends on global preferences. The next subsection explains how the rating matrix utilized to present global preference of users.

B. Global Preference Presentation

Typically, in common CF, the rating value usually represents the degree of preference of a user on an item. Whereas, in this paper, the proposed method utilized the type of items to express this preference. Three main steps to build the global preferences profile will be expressed. In this case, the MovieLens dataset is used to explain these steps. The MovieLens dataset has 18 types of movie such as action, crime, comedy, documentary, etc. Each movie can be belonging at least to one or more types. All ratings of users on movies are utilized to build their global preferences. This process passes through three sub-processes as shown in Fig. 1, which are presented as follow:

User-item matrix: we defined U to represent a set of *n* users in the dataset, and I is a set of *m* items that are rated within the interval [*Min*, …, *Max*]. The rows indicate the vector ratings of users. Likewise, the columns indicate the item' ratings. Therefore, the cells intersection will be filled by $r_{i,j}$ values that symbolizes the rating of user *i* on item *j* where the absence of ratings will be symbolled by the symbol *.

Frequent rating matrix: let's assume the following hypothesis. In e-commerce, the users purchase their commodity based on the type of color, style, brand, etc. Therefore, we can say that their preferences may be represented depend on this behavior (type of their purchases). Similarly, MovieLens domain is classified into 18 types. Thus, the users who like to watch documentary movies will prefer to watch this type more than the others. To explain this step, we assume that $\vec{I^j}$ is a vector represents the types' information of an movie *j*, where $\vec{I^j} = (c_1^j, \ c_2^j, \ \ldots, c_g^j, \ \ldots, \ , c_{k-1}^j, c_k^j)$. The $\vec{c}$ vector indicates the types of movies in the dataset, where $\vec{c} = (c_1, c_2, \ldots, c_g, \ldots, c_{k-1}, c_k)$, and k is the total number of types of movies in that dataset. Where , the value of c_g^j will be equal to 1 if the item *j* belongs to the g^{th} category and 0 if it is otherwise. The frequent matrix values will be represented by $t_{i,g}$ that represents the number of movies rated by user *i* and belongs to type g^{th}. The $t_{i,g}$ value can be counted using $t_{i,g} = \sum_{j\in I_i} c_g^j$. Where, I_i is the set of movies chosen by user *i* and c_g^j denotes the information value of movie *j* that belongs to g^{th} category *i*.

Normalization matrix: normalizing the frequent matrix to transform the rating count value into ratio value between zero and one. The normalization will be done to preserve the standardization in the process of compare. The normalized values will be utilized to represent the global preference of users which are used as inputs in the calculation of similarity process. For example, If the $\vec{T_i}$ vector represents the category information of user *i* where $\vec{T_i} = (t_{i,1}, t_{i,2}, \ldots, t_{i,g}, \ldots, t_{i,k-1}, t_{i,k})$, Therefore, the preferences of users on each category presented in the vector space model by a user-type normalization matrix,. Where the normalized value, $w_{i,g}$ is the percentage preference of user *i* on type g which can be calculated using $w_{i,g} = \frac{t_{i,g}}{\sum_{g=1}^{k} t_{i,g}}$, Where k is the number of item types in the database. Next, the normalized matrix will be used as a main input to defined the developed similarity measure.

	i_1	$\cdots$	i_j	$\cdots$	i_m
u_1	*	$\cdots$	$r_{1,j}$	$\cdots$	$r_{1,m}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
u_i	$r_{i,1}$	$\cdots$	$r_{i,j}$		$r_{i,m}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
u_n	$r_{n,1}$	$\cdots$	$r_{n,j}$	$\cdots$	*

	c_1	$\cdots$	c_g	$\cdots$	c_k
u_1	$t_{1,1}$	$\cdots$	$t_{1,g}$	$\cdots$	$t_{1,k}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
u_i	$t_{i,1}$	$\cdots$	$t_{i,g}$	$\cdots$	$t_{i,k}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
u_n	$t_{n,1}$	$\cdots$	$t_{n,g}$	$\cdots$	$t_{n,k}$

	c_1	$\cdots$	c_g	$\cdots$	c_k
u_1	$w_{1,1}$	$\cdots$	$w_{1,g}$	$\cdots$	$w_{1,k}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
u_i	$w_{i,1}$	$\cdots$	$w_{i,g}$	$\cdots$	$w_{i,k}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
u_n	$w_{n,1}$	$\cdots$	$w_{n,g}$	$\cdots$	$w_{n,k}$

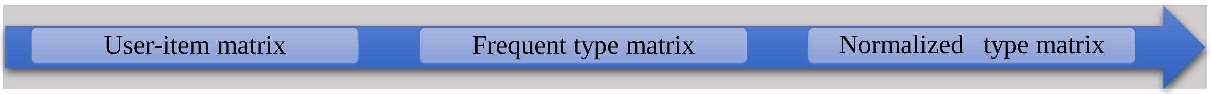


Fig. 1. Global preference presentation.

C. User Similarity Measurement based on the Global Preference

The similarity measure is required to calculate the correlation between a pair of users. In this work, the PCC and Cos will be adjusted and adopted based on the normalized matrix data and fairness and co-rated proportion factors, respectively. First, adopting fairness factor to the proposed similarity measure makes it more accurate. The correlation between a pair of users who has a close number of ratings should be stronger than the others. In this study, the fairness factor can be defined as the proportion of the number of items rated by the target user to the number of items taken by both users. For example, let u is the target user and $|I_u|$ is the set of items rated by u. Moreover, v is the compared user, where $|I_v|$ is the set of items rated by him/her. Therefore, the fairness factor (Ff) of each user can be defined as per (3) below:

$$Ff(u,v) = \frac{|I_u|}{|I_{u+}I_v|} \text{ , and } Ff(v,u) = \frac{|I_v|}{|I_{u+}I_v|} \tag{3}$$

Where, $Ff(u,v)$ is the weight of fairness of user u compare to user v, and $Ff(v,u)$ is the weight of fairness of user v compare to user u.

Second, the proportion of co-rated items is also considered in the proposed measure. Further, if the similarity between user $sim(u,v) = sim(u,l)$, for example, and users u and v have more co-rated than users u and l. It is obvious that the $sim(u,v)$ weight should be stronger than $sim(u,l)$. The sigmoid function will be used to devalue the weight of similarity when the co-rating small as shown in (4). Where the denominator ε will be utilized to limit the minimum size of co-rated. If the size of the set of common items equal or bigger than the ε threshold then the sigmoid weight would be bigger than 0.9 and vice versa. For example, if the ε equalled to 1 and size of co-rated of a pair of users equaled to 0, then the sigmoid value would be 0.5. But, if the size of co-rated is more than 3, the sigmoid value would be bigger than 0.95. The sigmoid function (Sf) can be computed as shown in (4).

$$Sf(u,v) = \frac{1}{1 + Exp^{(-\frac{|I_{u,v}|}{\varepsilon})}} \tag{4}$$

where $Sf(u,v) = Sf(v,u)$ and $|I_{u,v}|$ represents the set of co-rated items of users u and v.

As in above mentioned, based on the global preference presented in the normalized matrix and to adopt the aforementioned factors, the similarity between the pair of users u and v can be defined as DPcc and DCos methods in (5) and (6), respectively. Where, DPcc and DCos represent the developed similarity measures based on PCC and Cosine, respectively.

$$S(u,v)^{DPcc} = \frac{\sum_{g\in k}(w_{u,g} * Ff(u,v) - \overline{w_u})\,(w_{v,g} * Ff(v,u) - \overline{w_v})}{\sqrt{\sum_{g\in k}(w_{u,g} * Ff(u,v) - \overline{w_u})^2}\sqrt{\sum_{g\in k}(w_{v,g} * Ff(v,u) - \overline{w_v})^2}} * Sf(u,v) \tag{5}$$

$$S(u,v)^{DCos} = \frac{\sum_{g\in k}(w_{u,g} * Ff(u,v))(w_{v,g} * Ff(v,u))}{\sqrt{\sum_{g\in k}(w_{u,g} * Ff(u,v))^2 \sum_{i\in I_{x,y}}(w_{v,g} * Ff(v,u))^2}} * Sf(u,v) \tag{6}$$

After the similarity measurement is formulated, the correlation between users in the database will be computed to determine the most similar users. The users who have highest weight similarity with the target user will be located as neighbors. The adjusted weighted method is used to compute the predictions score for the user u on each neighbors' item. Equation (7) has been used to compute predictions.

$$P_{u,i} = \overline{r_u} + \frac{\sum_{v\in k}(w_{v,g} * Ff(v,u) - \overline{w_v})sim(u,v)}{\sum_{v\in N}|sim(u,v)|} \tag{7}$$

Where, $P_{u,i}$ is the prediction value for u about a specific item i, and N is the nearest neighbor of user u.

In the final phase of this method, M-top items will be provided to the target user as a set of recommendation.

IV. Experiments And Evaluation

A. Dataset and Metrics

Several experiments were contracted to evaluate the performance of proposed similarity method (DPcc and DCos). The MovieLens 100K dataset, as a public dataset available and widely used in the processes of CF system [33], was used in the process of evaluation. It includes 100,000 ratings provided by 943 users on 1,682 movies. Released in April 1998. The ratings were provided by users on a scale ranging from 1 to 5 stars. The sparsity level of 100k is 93.7%. Moreover, a specific widely used evaluation metrics are used to test the proposed method. Which are the: Mean Absolute Error (MAE); Recall; Precision; and F-measure measurements. MAE is used to calculate the difference between the actual rating and the predicted rating by users in the test, see (8). While, the precision is the fraction of retrieved items in a recommendation list that the user would rate useful, see (9). Moreover, the recall is the fraction of relevant items that are retrieved to the relevant items see (10). F-measure metric is a combined metric of precision and recall, it gives different information, the weighted mean of precision and recall, compared to precision and recall, see (11).

$$MAE = \frac{\sum_{i=1}^{N} |p_i - r_i|}{N} \quad (8)$$

Where N represents the number of items that have been selected for the work test and rated by the target user, p_i and r_i are the predicted ratings and actual ratings for the item i, respectively.

Table I illustrates the recommendation confusion matrix and how precision and recall metrics are defined relate to this matrix.

TABLE I. Recommendation Confusion Matrix

		Results in the test set	
		Rated	**Unrated**
Recommendation Results	**Recommended**	True Positive (TP)	False Positive (FP)
	Not recommended	False Negative (FN)	True Negative (TN)

$$P = \frac{TP}{TP + FP} \quad (9)$$

$$R = \frac{TP}{TP + FN} \quad (10)$$

$$\text{F} - \text{measure} = \frac{2 * (P * R)}{(P + R)} \quad (11)$$

A. Experimental and Results

In this section, the experimental results conducted on MovieLens datasets 100K will be reported. This dataset is divided randomly into two training set 80%, and testing set 20% to show the proposed methods accuracy improvement. Regarding inputs parameters, the ε, sigmoid function parameter, is tested with various values to identify the best value which was 9. Another main inputs they were the size of recommended items and neighbors. The number of recommended items were 10, 20, 30, 40, and 50 and the final results are represented by averaging variation size of neighbors (10, 20, 30, 40 and 50) in terms of precision, recall, and F-measure. The Pearson's correlation (PCC), Cosine and Constrained Pearson Correlation Coefficient (CPCC) are used for comparison as most common traditional CF [43], [53], [58]. The "DPcc" and "DCos" denote to the results from our proposed methods based on Pearson's correlation and Cosine methods, respectively.

Fig. 2 illustrates the MAE rate of the proposed methods compare to PCC, Cosine, and CPCC. The size of neighbors was presented on a horizontal axis with variation in sizes, which are 30, 50, 70, 100, and 150. There is an improvement in the MAE rate when the number of neighbors increases. The MAE rate of proposed methods has notable enhancement. They have the lowest proportion of MAE; it is evident that the DPcc and DCos methods have the best accuracy in term of MAE.

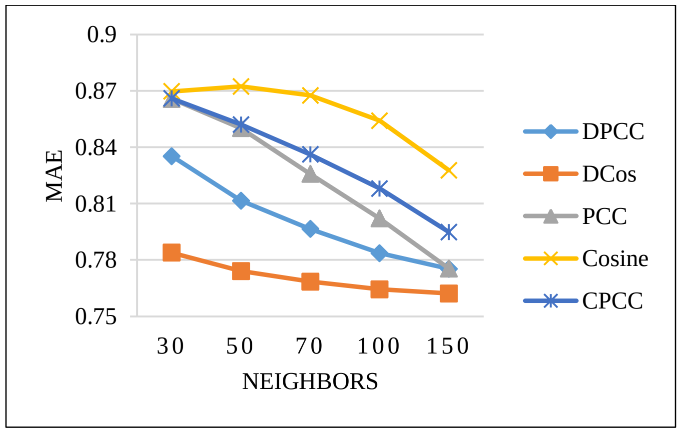


Fig. 2. Comparison of MAE with respect to the size of neighbors.

Fig. 3 present the comparison of recall between PCC, Cosine, CPCC, and the proposed methods. In general, for all methods, the rate rose gradually to reach to the highest rate when the number of recommendations was 50. It can be observed that the recall rates of DPcc and DCos were the highest respect to all size of recommendation. To sum up, the recall rate improves as the number of recommended items increases.

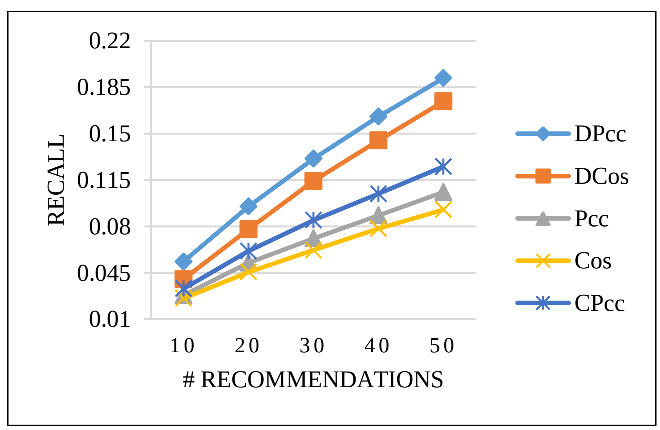


Fig. 3. Comparison of recall with respect to the size of recommendations.

The graph in Fig. 4 gives comparative information about the precision rate for CF based on PCC, Cosine, CPCC, and the proposed methods. At first glance, it is clear that the precision rate, for all methods, declines from the initial point, when the number of recommendations was 10, to reach to lowest value when the number of recommendations was 50. With regards to the proposed methods, it can be observed from the graph that the rate of precision was the highest compared to others.

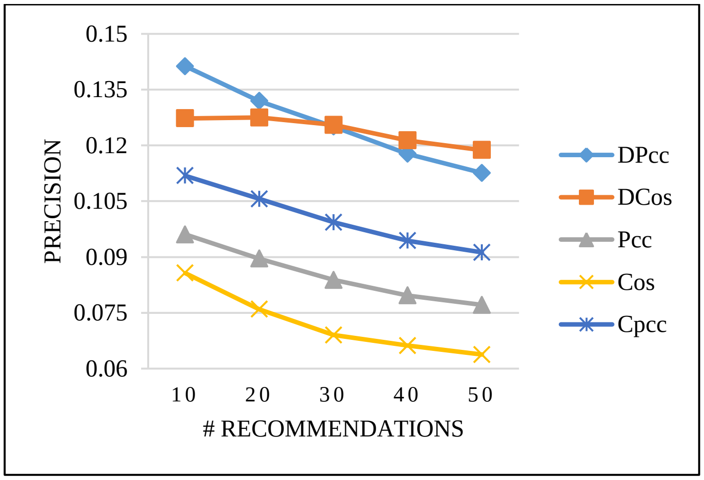


Fig. 4. Comparison of precision with respect to the size of recommendations.

Fig. 5 shows the percentage of F-measure for CF using PCC, CPCC, Cosine, and the proposed similarity methods. It can be observed from the graph that, for all methods, there is a significant rise in the percentage of F-measure for all methods from the initial point when the sizes of recommended items were 10 to 30. However, after that, it rose slightly within the next two sizes of recommendations. As a conclusion, the F-measure rate of CF-BSF is the best when compared to other methods PCC, Cosine and CPCC. Nevertheless, the F-measure rate of the proposed method was the overall highest as it can be seen in the figure.

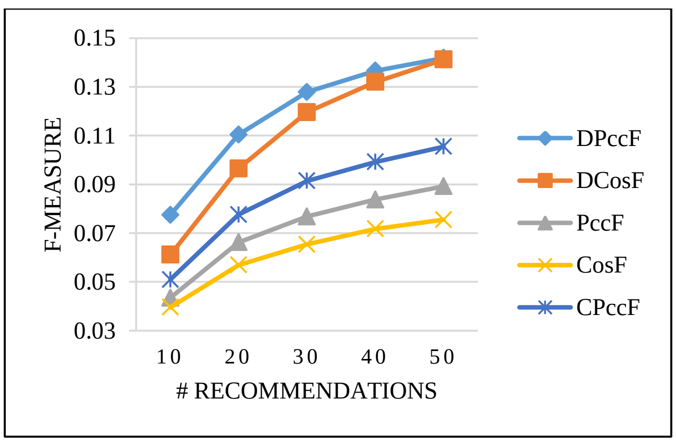


Fig. 5. Comparison of MAE with respect to the size of recommendations.

V. Conclusion

In CF method, locating the successful neighbors is an essential step in the improvement of recommendation accuracy. Thus, the critical step is how to develop an appropriate similarity measure. Several similarity methods in the state-of-the-art have been improved. However, most of them are still suffering from the negative impacts of sparsity data issue. Therefore, this work introduced a proposed similarity method utilized the global preferences of users to solve this issue. This global preference is inferred based on ratings of users to represent their preferences. Next, these preferences are used as input for the proposed similarity measure. Consequently, the correlation between a pair of users is calculated even they do not have common items. In addition, adopting two factors, fairness and proportion of co-rated items, in the proposed similarity to improve the accuracy of the recommendations has a positive effected as shown in the result. To conclude that, we can say, the problem of data sparsity is solved, and the accuracy is improved as shown in the experiments result. This result showed that the proposed method improved the accuracy when compared to the common traditional CF similarity methods using specific common evaluation metrics (MAE, Recall, Precision, and F-Measure).

However, the only main limitation of this work is related to the type of the dataset that can be worked with. The dataset that includes unclassified items will require pre-processing action before implementing it. The items classifications are needed to be revealed the dataset to meet the aspects of the proposed method. Therefore, the proposed method will be generalized to work on a different type and a larger size of datasets such as Netflix and 10M MovieLens through considering an accurate classification step. Moreover, the further research can be improved the accuracy of this proposed method by considering the factor of different degree of ratings of both users. Moreover, the singularity and significance also will be considered.

Acknowledgment

The research reported in this study is conducted by the researchers at University Malaysia Pahang (UMP), it is funded by RDU170381 and PGRS170384 grants. The researchers would like to thank UMP for supporting this research.

References

[1] X. Su and T.M. Khoshgoftaar, "A survey of collaborative filtering techniques," Advances in artificial intelligence. 2009.

[2] L. Chen, G.L. Chen, and F. Wang, "Recommender systems based on user reviews: the state of the art," User Modeling and User-Adapted Interaction. vol. 25, pp. 99-154, 2015.

[3] P. Resnick, et al. "GroupLens: an open architecture for collaborative filtering of netnews". in Proceedings of the 1994 ACM conference on Computer supported cooperative work, Chapel Hill, North Carolina, USA: ACM, 1994, pp. 175-186.

[4] W. Hill, et al. "Recommending and evaluating choices in a virtual community of use". in Proceedings of the SIGCHI conference on Human factors in computing systems: ACM Press/Addison-Wesley Publishing Co., 1995, pp. 194-201.

[5] U. Shardanand and P. Maes. "Social information filtering: algorithms for automating “word of mouth”". in Proceedings of the SIGCHI conference on Human factors in computing systems, Denver, Colorado, USA ACM Press/Addison-Wesley Publishing Co., 1995, pp. 210-217.

[6] M. Balabanović and Y. Shoham, "Fab: content-based, collaborative recommendation," Communications of the ACM. vol. 40, pp. 66-72, 1997.

[7] J.A. Konstan, et al., "GroupLens: applying collaborative filtering to Usenet news," Communications of the ACM. vol. 40, pp. 77-87, 1997.

[8] P. Resnick and H.R. Varian, "Recommender systems," Communications of the ACM. vol. 40, pp. 56-58, 1997.

[9] Y.R. Murti and Z. Baizal, "Compound Critiquing for Conversational Recommender System Based on Functional Requirement," Advanced Science Letters. vol. 22, pp. 1892-1896, 2016.

[10] S. Sivapalan, et al. "Recommender systems in e-commerce". in World Automation Congress (WAC), 2014, Waikoloa, HI, USA: IEEE, 2014, pp. 179-184.

[11] X.W. Yang, et al., "A survey of collaborative filtering based social recommender systems," Computer Communications. vol. 41, pp. 1-10, 2014.

[12] R. Zhang, et al. "Collaborative filtering for recommender systems". in Advanced Cloud and Big Data (CBD), 2014 Second International Conference on, Huangshan, China: IEEE, 2014, pp. 301-308.

[13] T. Arekar, M.R. Sonar, and N. Uke, "A Survey on Recommendation System," International Journal of Innovative Research in Advanced Engineering (IJIRAE). vol. 2, pp. 1-5, 2015.

[14] P. Nagarnaik and A. Thomas. "Survey on recommendation system methods". in Electronics and Communication Systems (ICECS), 2015 2nd International Conference on, Coimbatore, India: IEEE, 2015, pp. 1496-1501.

[15] O.S. Revankar and Y. Haribhakta, "Survey On Collaborative Filtering Technique In Recommendation System," IJAIEM. vol. 4, pp. 85-91, 2015.

[16] M.K. Najafabadi, et al., "Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data," Computers in Human Behavior. vol. 67, pp. 113-128, 2017.

[17] L. Chen, G. Chen, and F. Wang, "Recommender systems based on user reviews: the state of the art," User Modeling and User-Adapted Interaction. vol. 25, pp. 99-154, 2015.

[18] F. Ricci, L. Rokach, and B. Shapira, Introduction to recommender systems handbook. 2011: Springer.

[19] L. Sharma and A. Gera, "A survey of recommendation system: Research challenges," International Journal of Engineering Trends and Technology (IJETT). vol. 4, pp. 1989-1992, 2013.

[20] R. Burke, Hybrid web recommender systems, in The adaptive web. 2007, Springer. p. 377-408.

[21] J.A. Konstan, et al. "Recommender systems: A grouplens perspective". in Recommender Systems: Papers from the 1998 Workshop (AAAI Technical Report WS-98-08), 1998, pp. 60-64.

[22] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," Ieee Transactions on Knowledge and Data Engineering. vol. 17, pp. 734-749, 2005.

[23] J. Bobadilla, et al., "Recommender systems survey," Knowledge-Based Systems. vol. 46, pp. 109-132, 2013.

[24] L. Lü, et al., "Recommender systems," Physics Reports. vol. 519, pp. 1-49, 2012.

[25] J.B. Schafer, et al., Collaborative filtering recommender systems, in The adaptive web. 2007, Springer. pp. 291-324.

[26] M.D. Ekstrand, J.T. Riedl, and J.A. Konstan, "Collaborative filtering recommender systems," Foundations and Trends in Human-Computer Interaction. vol. 4, pp. 81-173, 2011.

[27] B. Sarwar, et al. "Item-based collaborative filtering recommendation algorithms". in Proceedings of the 10th international conference on World Wide Web, Hong Kong, Hong Kong: ACM, 2001, pp. 285-295.

[28] J.L. Herlocker, et al., "Evaluating collaborative filtering recommender systems," Acm Transactions on Information Systems. vol. 22, pp. 5-53, 2004.

[29] R. Zhang, Q.-d. Liu, and J.-X. Wei. "Collaborative Filtering for Recommender Systems". in Advanced Cloud and Big Data (CBD), 2014 Second International Conference on: IEEE, 2014, pp. 301-308.

[30] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," Knowledge and Data Engineering, IEEE Transactions on. vol. 17, pp. 734-749, 2005.

[31] D. Jannach, et al., Recommender systems: an introduction. 2010: Cambridge University Press.

[32] Q. Cheng, et al. "The new similarity measure based on user preference models for collaborative filtering". in Information and Automation, 2015 IEEE International Conference on: IEEE, 2015, pp. 577-582.

[33] H. Koohi and K. Kiani, "A new method to find neighbor users that improves the performance of Collaborative Filtering," Expert Systems with Applications. vol. 83, pp. 30-39, 2017.

[34] L. Gu, P. Yang, and Y. Dong. "An dynamic-weighted collaborative filtering approach to address sparsity and adaptivity issues". in 2014 IEEE Congress on Evolutionary Computation (CEC): IEEE, 2014, pp. 3044-3050.

[35] H. Sobhanam and A. Mariappan. "Addressing cold start problem in recommender systems using association rules and clustering technique". in Computer Communication and Informatics (ICCCI), 2013 International Conference on: IEEE, 2013, pp. 1-5.

[36] B.K. Patra, et al., "A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data," Knowledge-Based Systems. vol. 82, pp. 163-177, 2015.

[37] H. Luo, et al., "A collaborative filtering framework based on both local user similarity and global user similarity," Machine Learning. vol. 72, pp. 231-245, 2008.

[38] H.J. Ahn, "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem," Information Sciences. vol. 178, pp. 37-51, 2008.

[39] H.F. Liu, et al., "A new user similarity model to improve the accuracy of collaborative filtering," Knowledge-Based Systems. vol. 56, pp. 156-166, 2014.

[40] T. Arekar, M.R. Sonar, and N. Uke, "A Survey on Recommendation System." . 2014.

[41] O.S. Revankar and Y. Haribhakta, "Survey On Collaborative Filtering Technique In Recommendation System," IJAIEM, ISSN 2319-4847. vol. 4,. 2015.

[42] N. Polatidis and C.K. Georgiadis, "A multi-level collaborative filtering method that improves recommendations," Expert Systems with Applications. vol. 48, pp. 100-110, 2016.

[43] B. Zhang and B. Yuan. "Improved collaborative filtering recommendation algorithm of similarity measure". in AIP Conference Proceedings: AIP Publishing, 2017, pp. 020167-1–020167-6.

[44] S. Ghazarian and M.A. Nematbakhsh, "Enhancing memory-based collaborative filtering for group recommender systems," Expert systems with applications. vol. 42, pp. 3801-3812, 2015.

[45] J. Bobadilla, et al., "A collaborative filtering approach to mitigate the new user cold start problem," Knowledge-Based Systems. vol. 26, pp. 225-238, 2012.

[46] J. Bobadilla, et al., "Collaborative filtering based on significances," Information Sciences. vol. 185, pp. 1-17, 2012.

[47] K. Choi and Y. Suh, "A new similarity function for selecting neighbors for each target item in collaborative filtering," Knowledge-Based Systems. vol. 37, pp. 146-153, 2013.

[48] J. Mao, et al. "An improved similarity measure method in Collaborative Filtering Recommendation Algorithm". in Cloud Computing and Big Data (CloudCom-Asia), 2013 International Conference on, Fuzhou, China: IEEE, 2013, pp. 297-303.

[49] B.-H. Huang and B.-R. Dai. "A Weighted Distance Similarity Model to Improve the Accuracy of Collaborative Recommender System". in Mobile Data Management (MDM), 2015 16th IEEE International Conference on, Pittsburgh, PA, USA: IEEE, 2015, pp. 104-109.

[50] P. Pirasteh, D. Hwang, and J.E. Jung, "Weighted Similarity Schemes for High Scalability in User-Based Collaborative Filtering," Mobile Networks and Applications. vol. 20, pp. 497-507, 2015.

[51] L. Shunpan, M. Lin, and Fuyongyuan, "A singularity-based user similarity measure for recommender systems," IJICIC. vol. 11, pp. 1629-1638, 2015.

[52] Y.E.M. El Alami, E.H. Nfaoui, and O. El Beqqali. "Improving Neighborhood-Based Collaborative Filtering by a Heuristic Approach and an Adjusted Similarity Measure". in BDCA, 2015, pp. 16-22.

[53] T. Mahara, "A New Similarity Measure Based on Mean Measure of Divergence for Collaborative Filtering in Sparse Environment," Procedia Computer Science. vol. 89, pp. 450-456, 2016.

[54] K. Saranya, G.S. Sadasivam, and M. Chandralekha, "Performance comparison of different similarity measures for collaborative filtering technique," Indian Journal of Science and Technology. vol. 9, pp. 2016.

[55] X. Zang, et al. "A New Weighted Similarity Method Based on Neighborhood User Contributions for Collaborative Filtering". in Data Science in Cyberspace (DSC), IEEE International Conference on, Changsha, China: IEEE, 2016, pp. 376-381.

[56] H. Cao, et al. "An improved recommendation algorithm based on Bhattacharyya Coefficient". in Knowledge Engineering and Applications (ICKEA), IEEE International Conference on, Singapore, Singapore: IEEE, 2016, pp. 241-244.

[57] A. Bilge and A. Yargıç, "Improving accuracy of multi-criteria collaborative filtering by normalizing user ratings," nadolu Üniversitesi Bilim Ve Teknoloji Dergisi A-Uygulamalı Bilimler ve Mühendislik. vol. 18, pp. 225-237, 2017.

[58] S. Kg and G.S. Sadasivam, "Modified Heuristic Similarity Measure for Personalization using Collaborative Filtering Technique," Appl. Math. vol. 11, pp. 317-325, 2017.

An Optimization of Audio Classification and Segmentation using GASOM Algorithm

Dabbabi Karim, Cherif Adnen
Research Unity of Processing and Analysis of Electrical and Energetic Systems
Faculty of Sciences of Tunis, University Tunis El-Manar
2092 Tunis El-Manar, Tunis, Tunisia

Hajji Salah
School of Engineers of Tunis, 3000
University Tunis El-Manar, Tunis, Tunisia

***Abstract*—Now-a-days, multimedia content analysis occupies an important place in widely used applications. It may depend on audio segmentation which is one of the many other tools used in this area. In this paper, we present an optimized audio classification and segmentation algorithms that are used to segment a superimposed audio stream according to its content into 10 main audio types: speech, non-speech, silence, male speech, female speech, music, environmental sounds, and music genres, such as classic music, jazz, and electronic music. We have tested the KNN, SVM, and GASOM algorithms on two audio classification systems. In the first audio classification system, the audio stream is discriminated into speech no-speech, pure-speech/silence, male speech/female speech, and music/ environmental sounds. However, in the second audio classification system, the audio stream is segmented into music/speech, pure-speech/silence, male speech/female speech. For pure-speech/silence discrimination, it is performed in the two systems according to a rule-based classifier. Concerning the music segments in both systems, they are discriminated into different music genres using the decision tree as a classifier. Also, the first audio classification system has succeeded to achieve higher performances compared to the second one. Indeed, in the first system using the GASOM algorithm with leave-one-out validation technique, the average accuracy has reached 99.17% for the music/environmental sounds discrimination. Moreover, in both systems, the GASOM algorithm has always reached the best results of performances compared to KNN and SVM algorithms. Therefore, in the first system, the GASOM algorithm has been contributed to obtain an optimized consumption time compared to that one obtained using the two HMM and MLP methods.**

Keywords—Segmentation and classification audio; features extraction; features discrimination; GASOM algorithm

I. Introduction

In order to facilitate and help the users to be more accurate and efficient in their research for multimedia contents on search engines, content-based indexing and retrieval technologies is a good way to help them to access directly to the required multimedia contents. Recently, the research in the multimedia content relies on the content-based audio retrieval and other relevant techniques such as the audio segmentation, the audio indexing, the audio browsing, and the audio annotation. Generally, there are many techniques to categorize the audio content into speech, music or other sounds, and there are different methods to process each type of them. Concerning the retrieve of speech and spoken documents, they are transformed into texts by automatic speech recognition systems. For the retrieve of music, an approximate string matching algorithm has been proposed in [1] to solve a string matching problem and to match strings of features, such as the rhythm, melody, and chord strings of musical objects in a music database. Also, besides speech and music, we can find general sounds that represent the major audio type. In some research, such sounds has been dedicated to the classification and in others, it has been used in more specific areas, such as the classification of piano [2] and ringing [3] sounds. Furthermore, in order to face the growing size of audio databases with a huge amount of audio data, an efficient organization and manipulation of data is required. For example, a discrimination of speech and non-speech segments with a high accuracy is required for such applications, such as the automatic transcription instance of broadcast news (BN), automatic speech and speaker recognition, recovery audio requests, and so forth. As the audio data contains alternating sections of different audio types, an automatic classification of its content into appropriate audio classes is a fundamental step in the processing of audio streams. Thus, this kind of separation is called audio content classification. Regarding the audio stream segmentation, it is often substantial with the classification process in the recovery system and they are together useful for many classification tasks. Moreover, the feature extraction process is a conditioning element for the overall classification performance as it includes three types of features which can be extracted from temporal, frequency, and coefficient domains. Concerning the time domain features, they include the Zero-Crossing Ratio (ZCR), the Silence Ratio (SR), the Root Mean Square (RMS), and so on. As for the frequency domain features, they contain the pitch, the bandwidth, the Spectral Centroid (SC), and so on. Also, the linear prediction coefficients (LPC) and the Mel-Frequency Cepstral Coefficients (MFCC) are widely exploited in automatic speech recognition and automatic classification of general sounds. Recently, the wavelet coefficients have attracted much attention of researchers thanks to its multi-resolution property and its better time-frequency resolution [4], [5]. Furthermore, a major change in the online service has been created by the excessive increase of multimedia data on the internet. Therefore, the audio information becomes an important part of most multimedia applications, especially music, which is the most common and popular example of online information. Thus, the segmentation and classification of audio streams according to their content is a useful means for analyzing audio, video, and understanding content. However, performing this task requires an efficient and accurate technique. Such a technique is called audio segmentation which splits an audio stream into homogenous regions. Also, an emerging increase in digital data is caused by the advent of multimedia and network technology, which in turn begets a growing interest in multimedia content-based information retrieval. Indeed, the discrimination of audio signal according to its content is the fundamental step for its analysis and understanding. For audio segmentation and classification, it is considered as a pattern recognition problem and it includes two main stages: feature extraction and extracted-features-based classification [6]. Also, the categorization of audio content analysis applications can be performed in two parts: the first part is the discrimination of an audio stream into homogenous regions and the second part is the discrimination of a speech stream into segments of different speakers. In [7], [8], the discrimination of an audio stream into different audio types has been performed using Support Vector Machine (SVM) algorithm and K-Nearest Neighbor (KNN) algorithm. Moreover, the characterization of various audio content levels of a sound track has been carried out by frequency tracking in an audio indexing system proposed in [9]. This system has the specificity that it does not need any prior information. In [10], the authors have proposed a fuzzy approach that uses a hierarchical segmentation and classification according to automatic audio analysis. In [11], an extracted-features-based music and speech discrimination has been performed using a multi-dimensional Gaussian Maximum A posteriori (MAP) estimator, a Gaussian Mixture Model (GMM), a k-d tree-based spatial partitioning scheme, and a KNN classifier. Also, the change point detection is a process which splits the audio stream into homogenous and continuous temporal regions by searching for temporal boundaries. On the other hand, it has a problem which arises in the definition of homogeneity criteria. For this purpose, stream segmentation can be performed by calculating the Generalized Likelihood Ratio (GLR) statistics without prior knowledge of classes [12]. However, computing statistics using MFCC coefficients requires a large amount of data for training [12].

For a transcript of meetings and automatic camera tasks, the segmentation of the meeting of a group of persons according to their voices is required. Indeed, the segmentation of feature vectors has been carried out using Bayesian Information Criterion (BIC), which has required a large amount of data for training [13], [14]. Also, the Structures Support Vector Machine (SSVM) has been used by structured discriminator models for large-vocabulary speech recognition tasks and the determination of features has been performed by Hidden Markov Models (HMMs) [15], [16] and a Viterbi decoding [17]. The human auditory systems rely principally on perception, while audio retrieval systems are traditionally text-based, which is not sufficient to achieve perceptual similarity between two audio clips because it only elaborates the high-level audio content. Thus, a query technique has been used to solve this problem and it was a very different approach to audio classification. In [18], modeling of continuous probability distribution of audio characteristics has been performed by a Gaussian mixture model (GMM). Also, a MMI-supervised tree-based vector quantizer and a feed-forward neural network have been proposed in [15], [19], [20], [21] for the task of detecting speech and environmental sounds on a sound stream. Indeed, a Kernel Fisher discriminator-based regularized kernel has been used for an unsupervised change detection task [22], [23].

Speech is not only limited to be used as a mode of transmission words of messages, but it can be also used as a means of transmitting emotions, personality, etc. Indeed, in many speech applications, mainly in speech segmentation and speaker verification, words containing vowel regions have a vital importance. For this, dividing an audio stream into segments is possible by a vowel regions-based audio segmentation. In fact, the audio segmentation algorithms can be divided into three general categories: the first category includes the features extraction stage in which the time and frequency domain features are extracted, and then their classification is performed by a classifier in order to discriminate the different audio signals according to their content. For the second audio segmentation category, it includes the feature extraction statistics which are used for discrimination by a classifier. Thus, these types of features are called posterior probability-based features. In this category, the classifier requires a large amount of data for training in order to reach accurate results. Concerning the third category of audio segmentation algorithms, it requires the use of efficient discriminators, such as BIC, Gaussian Likelihood Ratio (GLR), and Hidden Markov Model (HMM). In fact, good results are given by these classifiers if a large amount of data for training is provided. Also, many applications have been performed using audio segmentation and classification. Among these applications we can find the content-based audio classification and retrieval which are most used in the entertainment industry, managing audio archives, use of commercial music, supervising, and so forth. Nowadays, millions of databases on the World Wide Web are presented for audio search and indexing, and for audio segmentation and classification. In the monitoring of broadcasts news programs, the audio classification has contributed to reach efficient and accurate navigation through the archives of broadcasts news. The analysis of superimposed speech is a complex problem, and consequently improved-performance systems are required. Also, the audio stream segmentation is a preprocessing step in many audio processing applications in which it has a significant impact on the speech recognition performance. For this, the proposed audio segmentation and classification algorithm must be optimized, efficient, and fast in order to be used in real-time multimedia applications. Indeed, the hybridization of Self-Organization-Map (SOM) algorithm with Genetic Algorithm (GA) (called GASOM algorithm) is such algorithm which meets these requirements. To deal with complex data characteristics, the GASOM algorithm allows avoiding weakness such as slow convergence time being always trapped in the local minima. Moreover, this algorithm requires less training data, and consequently a high accuracy and a reduced-consumption time can be achieved. Indeed, the weights of the SOM algorithm have been optimized using GA algorithm, which allows obtaining a better mapping quality of classification and labeling data. In this work, the input data in the first audio segmentation and classification system is segmented, and then classified into nine basic audio types: speech, silence, music, environmental sounds, speech male, speech female, electronic music, classic music, and jazz music. Concerning the second audio segmentation and classification system, the input data is segmented, and then classified into eight basic audio types: speech, music, silence, speech male, speech female, electronic music, classic music, and jazz music. In this paper, we also exhibit possible solutions for classifying the audio stream using the two KNN and SVM classifiers. Furthermore, different descriptors have been proposed to face the audio variety and discriminate very well between the different audio types.

The remaining sections of this paper are organized as follows: in Section I, audio segmentation and classification steps, feature extraction process, classification approaches (KNN, SVM, and GASOM) are presented, and then discussed. In the next section, an exhibition of different evaluations used to assess the experimental tests. In last section, the experimental results are discussed.

II. RESEARCH METHOD

A. Pre-classification

At first, the audio signal has been segmented into 1-s frames by applying the growing-window technique with a sample rate of 16 KHz. Consequently, the DCT coefficients at each frame have been calculated by Fast Fourier transform (FFT).Indeed, these last steps form together the short-term Fourier transform (STFT) which is a category of short-term processing techniques. Thus, we have obtained a matrix of the STFT coefficients from which their magnitudes are calculated to form a resulting matrix that can be treated as an image. This image is called spectrogram of signal.

B. Audio Classification and Segmentation Step

A separated analysis of each widowed frame in the audio clip has been performed as a pre-classification step before the classification. After that, the normalized feature vectors have been extracted, and then the classification step has been performed by selecting one of the algorithms SVM, KNN, and GASOM. Concerning the classification of audio clip/frames into speech and non-speech segments, it has been performed using a SVM, KNN, or GASOM classifier. For the speech segments, they have been discriminated into silence and pure-speech segments according to a rule-based classifier as the speech signal contains mostly silence frames. After that, the pure-speech segments have been used by the SVM, KNN, or GASOM classifier in order to discriminate between male speech and female speech. Also, the SVM, KNN, or GASOM classifiers have been then used to classify the non-speech segments into musical and environmental sounds. At the end, music genre discrimination has been carried out by a decision tree using music segments. Fig. 1 illustrates the block diagram of the first proposed audio classification system. Indeed, the audio stream has been each time down sampled to 16000 KHz and the features {Zero-Crossing rate, short-time energy, spectrum flux, Mel-frequency cepstral coefficients, vector chroma, spectral centroid, harmonic ratio, energy of entropy, spectral energy, and periodicity analysis}have been extracted, and then classified. These features {Mel-frequency cepstral coefficients, spectral flux, zero-crossing rate, and short time energy} have been used by the selected classifier (KNN, SVM, or GASOM algorithm) to classify the audio stream into speech and non-speech segments. For the discrimination between silence and pure-speech segments, it has been performed by a rule-based classifier, and then the pure-speech segments have been discriminated into male speech or female speech using the KNN, SVM, or GASOM algorithm as a classifier and {harmonic ratio and frequency estimator} as features. Also, the discrimination of non-speech segments into music and environmental sounds has been performed by the KNN, SVM, or GASOM algorithm as a classifier and {spectrum flux and Mel-frequency cepstral coefficients} as features. Moreover, the features {the minimum of the sequence entropy values and the mean value of the spectral flux sequence} have been used by the decision tree as a classifier in order to discriminate between different musical genres.

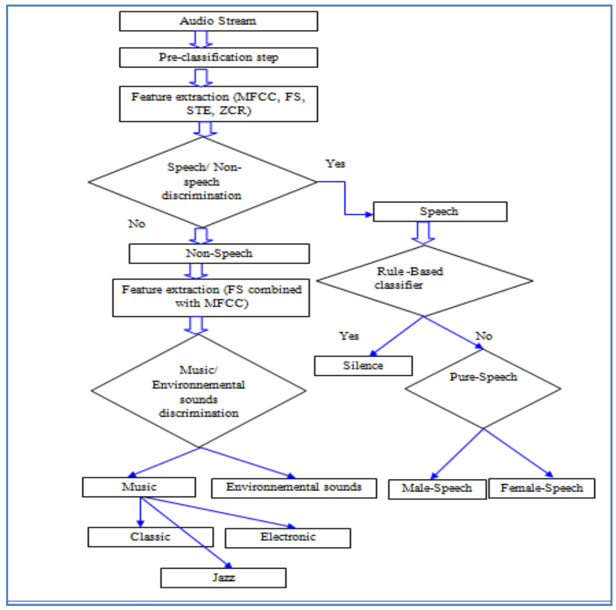


Fig. 1. Block scheme of the first audio classification and segmentation system.

C. Feature Extraction Step

At first, the audio signal has been divided into mid-term windows, and then the short-term processing technique has been applied for each segment. After that, the feature statistics have been calculated using feature sequences from each mid-term segment. Therefore, we obtain a set of statistics which represents each mid-term segment. In this work, the audio input has been divided into short-term windows and 23 audio features have been calculated per window. Thus, two mid-term statistics have been drawn per feature and a 46-dimensional vector has been obtained as output of the mid-term function. Also, the sizes of windows were 2 seconds and 0.05 seconds for mid-term and short-term processing, respectively. Moreover, the mid-term and short-term window steps were respectively set to 1 second and 0.025 seconds.

1) The Energy: The calculation of the short-term energy is given by the following expression:

$$E(i) = \sum_{n=1}^{W_L} |x_i(n)|^2, \ n = 1 \dots W_L \qquad (1)$$

Where x_i and W_L are respectively the sequence of audio samples of the ith frames and the length of the frame. The normalization of the energy is usually performed in order to eliminate the dependence on the frame length. Thus, the expression of (1) becomes as follows:

$$E(i) = \frac{1}{W_L}\sum_{n=1}^{W_L}|x_i(n)|^2 \quad (2)$$

For the short-term energy variation, it is faster for speech frames than those of music because the speech signals contain weak phonemes and short periods of silence between words.

2) Zero Crossing-Rate (ZCR)

This feature is defined as a measure of the occurrences of signal changes from positive to negative or vice versa. Also, another more general definition is the amount of zero-crossing in the frame. Moreover, the ZCR feature is a good discriminator for a speech and music separation and it is higher for speech than to music as it contains more silent regions [24], [25]. Indeed, the ZCR feature is expressed as follows:

$$ZCR = \frac{1}{2(M-1)}\sum_{n=1}^{M-1}|sgn[x(n+1)] - sgn[x(n)]| \quad (3)$$

Where $x(n)$ and $sgn[.]$ represent respectively the discrete signal that is in the range of $n = 1, \dots, M$ and the sign function.

3) The Entropy of Energy

The interpretation of the measure of abrupt changes in the level-energy of an audio signal represents the short-term entropy of energy. Indeed, the calculation of this feature is carried out at first by dividing each short-term frame into k sub-frames of fixed duration. After that, the energy of each sub-frame j is calculated and divided by the total energy of the short-term frame ($E_{shortFrame}$)as in equation (1). Thus, the resulting sequence of sub-frame energy values e_j, j=1,…,K, is treated by a division operation (a standard procedure)as a sequence of probabilities such as in (4):

$$e_i = \frac{E_{subframe_j}}{E_{shortFrame_i}} \quad (4)$$

Where $$E_{shortFrame_i} = \sum_{k=1}^{K} E_{subframe_k} \quad (5)$$

At the end, the calculation of the entropy $H\,(i)$ of a sequence $e_i$is carried out according to the following equation:

$$H(i) = -\sum_{j=1}^{k} e_j\, log_2(e_i) \quad (6)$$

4) The Spectral Centroid and Spread:

The two simple measures of the spectral position and shape are carried out by the spectral centroid and the spectral spread. For the spectral centroid, it is defined as the center of ‘gravity’ of the spectrum. Indeed, the value of the spectral centroid C_i of the i^{th} audio frameis givenby the following expression:

$$C_i = \frac{\sum_{k=1}^{W_{f_L}} kX_i(k)}{\sum_{k=1}^{W_{f_L}} X_i(k)} \quad (7)$$

Concerning the second central moment of the spectrum, which is the spectral spread, it can be calculated by taking the derivation of the spectrum from the spectral centroid according to the following equation:

$$S_i = \sqrt{\frac{\sum_{k=1}^{W_{fL}}(k-C_i)^2 X_i(k)}{\sum_{k=1}^{W_{fL}} X_i(k)}} \quad (8)$$

5) The Spectral Entropy (SE)

The calculation of the spectral entropy is similar to that one of the entropy of energy with a difference that this latter is performed in the frequency domain [26]. Indeed, the spectrum of the short-term frame is at first divided into L sub-bands (bins), and then the energy E_f of the fth sub-band, $f = 0, \dots, , L-1$, is normalized by the total spectral energy, which is

$$n_f = \frac{E_f}{\sum_{f=0}^{L-1} E_f}, f = 0, \dots, L-1.$$

At the end, the entropy of the normalized spectral energy $n_f$is carried out according to the following equation:

$$H = -\sum_{f=0}^{L-1} n_f \,. log_2(n_f) \quad (9)$$

In [27], [28], an efficient discrimination between speech and music has been performed by the variant of the spectral entropy called chromatic entropy.

6) The Spectral Flux (SF)

The measure of the spectral change between two successive frames is performed by spectral flux which is calculated as the squared difference between the normalized magnitudes of the spectra of two successive short-term windows such as:

$$Fl_{(i,i-1)} = \sum_{k=1}^{w_{f_L}}(EN_i\,(k) - EN_{i-1}(k))^2 \quad (10)$$

Where, $$EN_i(k) = \frac{X_i(k)}{\sum_{l=1}^{w_{f_L}} X_i(l)} \quad (11)$$

$EN_i(k)$ is defined as the kth normalized DTF coefficient at the ith frame.

7) The Spectral Rolloff

The frequency below which a certain percentage (usually around 90%) of the magnitude distribution of the spectrum is concentrated, is defined as a spectral rolloff. Each time that the $mth\ DFT$coefficient corresponds to the spectral rolloff of the ith frame, the expression satisfying this condition is given by the following equation:

$$\sum_{k=1}^{m} X_i(k) = C\sum_{k=1}^{w_{f_L}} X_i(k) \quad (12)$$

Where C is the adopted percentage. Also, the normalization of the spectral rolloff frequency is usually performed by dividing it withw_{f_L}so that it takes values between 0 and 1.

8) MFCC Coefficients

This feature represents the cepstral representation of the signal where the distribution of frequency bands is carried out according to the Mel-scale instead of the linearly spaced approach. Let $\widetilde{O_k}$ the power at the output of the kth frame filter, the resulting MFCC coefficients are expressed by the following equation:

$$c_m = \sum_{k=1}^{L}(log\,\widetilde{O_k})\cos\left[m\left(k-\frac{1}{2}\right)\frac{\pi}{L}\right], m = 1, \dots, L. \quad (13)$$

Furthermore, the MFCC coefficients are defined according to (13) as the coefficients of the discrete cosine transform of Mel-scaled log-power spectrum. Also, the MFCC coefficients have been used in many audio analysis applications, such as speaker clustering [29], music genre classification [30], and speech recognition [31].

9) The Chroma Vector

The chroma vector is defined as the 12-element representation of the spectral energy [32]. Moreover, this descriptor has been widely applied in music-related applications [33]-[36]. Indeed, the computation of the chroma vector is performed by grouping the DFT coefficients of a short-term window into 12 bins: one of the 12 equal-tempered pitch classes of Western-type music is represented by each bin. Therefore, the mean of the log-magnitudes of respective DFT coefficients is produced by each bin such as:

$$v_k = \sum_{n \in S_k}\frac{X_i(n)}{N_k}, k \in 0, \dots, 11. \quad (14)$$

Where S_k and N_k represent respectively a subset of frequencies that correspond to the DFT coefficients and the cardinality of S_k.

10) Periodicity Estimation and Harmonic Ratio

In general, we can categorize the audio signals into a periodic (noise-like) and quasi-periodic. Despite the fact that some signals have a periodic behavior, it is so hard to find the same periods for two signals. Concerning the voiced signals and the majority of music signals, they are included in the category of quasi-periodic signals. For the estimation of the fundamental frequency, it is carried out according to the autocorrelation function, which calculates the correlation between the shifted signal and the original one [37]. After that the fundamental period which exhibits the maximum autocorrelation is chosen to be the lag. Indeed, the correlation $R_i(m)$ can be defined as the correlation of the i^{th}frame with itself at time-lag m such as:

$$R_i(m) = \sum_{n=1}^{W_L} x_i\,(n)x_i(n-m) \quad (15)$$

Therefore, the calculation of the normalized autocorrelation function for thei^{th}frame is given by the following equation:

$$\Gamma_i(m) = \frac{R_i(m)}{\sqrt{\sum_{n=1}^{W_L} x_{i(n)^2}\sum_{n=1}^{W_L} x_{i(n-m)^2}}} \quad (16)$$

Where W_L is the number of samples per frame and m is the time-lag.

Also, the harmonic ratio is defined as the maximum value of Γ_i *and it* is determinate by the following equation:

$$\mathrm{HR_i} = \max_{\mathrm{T_{min}} \le \mathrm{m} \le \mathrm{T_{max}}}\{\Gamma_\mathrm{i}(\mathrm{m})\} \quad (17)$$

Where T_{min} and T_{max} are the allowable values of the fundamental period.

Therefore, the position of the occurrence of the maximum value of Γ_i is used to determinate the selected fundamental frequency as follows:

$$\Gamma_0^\mathrm{i} = \arg\max_{\mathrm{T_{min}} \le \mathrm{m} \le \mathrm{T_{max}}}\{\Gamma_\mathrm{i}(\mathrm{m})\} \quad (18)$$

III. Classification Approaches

We have designed two audio classification systems: in the first one, the SVM/KNN/GASOM classifiers are at first applied to classify segments into speech/non-speech segments, and then the non-speech segments are used for music/environmental sounds discrimination using the SVM, KNN or GASOM algorithm as a classifier. After that, the music segments are used by the decision tree classifier to discriminate between the different music genres. For the features of speech segments, they are discriminated by a rule-based classifier into pure-speech and silence, and then the SVM, KNN or GASOM algorithm, is also used to discriminate between the pure-speech segments into male speech and female speech. Concerning the second audio classification system, a speech and music discrimination is at first performed using the KNN, SVM or GASOM algorithm as a classifier, and then the music segments are classified into different music genres using the decision tree classifier. For the speech segments, they are used by a rule-based classifier to discriminate between the silence and pure-speech segments. After that, the pure-speech segments are used to discriminate between male speech and female speech using KNN, SVM or GASOM algorithm as a classifier.

A. Super Vector Machine (SVM) Algorithm

The learning of an optimized separation hyper plan for given positive and negative examples is performed by the Super Vector Machine (SVM) [38], [39]. Indeed, this classifier minimizes the probability of misclassifying unseen patterns for a fixed data that has an unknown probability distribution. Thus, the SVM allows obtaining an optimized performance on training data, and consequently the structural risks are minimized. In fact, this characteristic makes the difference between SVM and other traditional pattern recognition techniques in term of optimization. Also, we distingue two types of SVM: linear and kernel-based non-linear. The complication of the distribution of features in the audio data causes areas of overlap between the different classes and there is no possibility to separate them linearly. Such a situation can be manipulated by a kernel support vector machine. Moreover, the kernel has been used by SVM in order to create an optimal separation hyper plane [40], [41]. Indeed, the kernel function implicitly maps the input vectors to a high-dimensionality feature space in which they are linearly separable. Among the most well-known and used functions of kernel, we can mention: polynomial, function-based Gaussian radial, and a multilayer perception. In fact, the kernel-based Gaussian radial has empirically shown its high performance compared to other types of kernel. For this, we have used it in our proposed models. Furthermore, the expression of the kernel-based Gaussian radial is given as follows:

$$k(x, y) = \exp\left(-\frac{x-y^2}{2\sigma^2}\right) \quad (19)$$

Where, σ is the width of the Gaussian function.

B. K-Nearest Neighbor (KNN) Algorithm

The KNN classifier is a non-parametric classifier which works as follows: for each input vector to be classified, a search is started in order to find the location of the k nearest training examples, and then the class which has the largest members in this location is assigned to the input. Indeed, the measure of the neighborhood is performed using the Euclidian distance. Also, the domination of certain features due to their range of values during the calculation of the Euclidian distance, requires the use of the linear method (20) as a remedy of this issue by normalizing the jth feature, $j = 1, \ldots, L$, to zero mean and the standard deviation to 1:

$$\hat{\gamma_i} = \frac{\gamma_i(j) - \mu(j)}{\sigma(j)}, i = 1, \ldots, M, j = 1, \ldots, L. \quad (20)$$

Where, $\mu(j)$is the mean value of the jth feature, $\sigma(j)$ is the respective standard deviation, L is the dimensionality of the feature space, and M is the number of training samples.

C. Self-Organized Mapping (SOM) Algorithm

The neural network map SOM was inspired from biology by Teuvo Kohonen. It is assimilated as many elementary processors represented by the neurons which are connected to each other in order to exchange information. In fact, the parallel and massive work of many formal neurons offers them the capacity for learning and deciding in the recognition task [42], [43]. In general the activation function is non-linear and it differs from an application to another. Moreover, the neural weights in the vicinity of the activated neuron (winner neuron) are updated by the learning rules, which make them close to the input vector:

$$\Delta w_i = \gamma h_{iv}(x(t) - w_i) \quad (21)$$

Where γ is the learning ratio and h_{iv} is the neighborhood function which relies on the distance between the units i and v on the map.

Furthermore, the map SOM network can be a universal tool of representation and recognition by virtue of its non-linear activation function. Thus, this algorithm can be applied in an unsupervised manner and it can be used for the recognition of voluminous input data.

D. GASOM Algorithm

To avoid the degradation of the diversity of genetic population in early generations, the SOM algorithm in order is used to maintain it thanks to its observed approximation property. Also, in order to increase the space research towards an optimal solution and avoid premature convergence, the Genetic Algorithm (GA) was hybridized to the SOM algorithm. This suggested algorithm allows the introduction of feature vectors into the SOM map in order to perform learning and testing operations. Indeed, there is an activation of a single neuron of the SOM map at each iteration, and consequently an appointing of the best matching unit (BMU). Among other neurons of the map, the best representative of the data inputs at this iteration is called the winning neuron. Also, every time we obtain a BMU neuron via the training iterations, which is special to each input and we will get an individual (a chromosome) assigned to this input for the reconstruction of population to be treated by the Genetic Algorithm (GA). Indeed, the representation of each chromosome is performed by a matrix of criteria which corresponds to the matrix of criteria for each neuron of a SOM map type during the iterations of learning or test [44]. After that the equation of changes and the update of the vectors of weights determinate the new chromosomes forming the new population for the next generation. Moreover, the modification of the update equation for the training of SOM map is performed by adding new coefficients according to the fitness values of the chromosomes of the current population. Furthermore, the ability of an input data is completely simulated by the weight of neuron as it is the largest organelle in the unit. Therefore, the diversification of population in the SOM topology has a huge effect on the evolution of the results of data recognition of the weights of units in the evolutionary process. Indeed, the explanatory diagram of the GASOM hybridization is shown in Fig. 2.

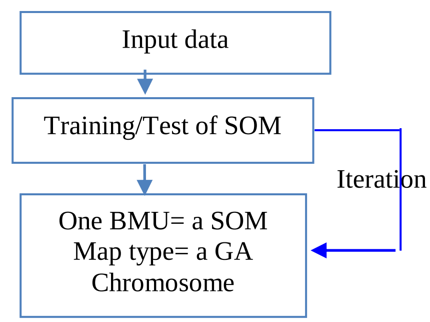


Fig. 2. Explanatory diagram of the GASOM hybridization.

E. Discrimination Steps for First Audio Classification System

1) Speech and Non-Speech Discrimination

This discrimination has been performed by the KNN, SVM or GASOM classifiers which have been applied with MFCC coefficients, SF, ZCR, and STE. Concerning the training databases, they were used to generate speech and non-speech code books.

2) Speech and Silence Discrimination

The detection of silence was performed according to features STE and ZCR by using 1-s window. For the classification, it has been performed by a rule-based classifier: each time when STE and ZCR exceed the predefined threshold value, then they were classified as pure-speech frame, otherwise they were classified as silence frame.

a) Male and Female Speech Discrimination

We describe in this sub-section a voice-based gender identification approach which can be used for the annotation of multimedia content-based indexing. Typically, the range of values of the fundamental frequency for a male speaker is quite narrow (between 80-200 Hz) and large for a female speaker (150-350Hz). The gender identification system proposed in this work is based on n general audio classifier and it consists of three main steps: In the first step, the features {harmonic ratio and the periodicity estimation} are extracted and normalized (statistics). After that the different segments are clustered using GASOM, KNN or SVM algorithm as a classifier. In this work, we have used the correlation-based pitch estimation feature since it relies considerably on the speech quality.

After the segmentation of the signal, each window obtained of duration T is modeled by a vector composed of two fundamental frequencies in ascending order (low and high frequency) representing the Harmonic Ratio (HR) in that frame. To avoid the incorrect peak selection caused by the existence of sub-harmonics in the spectrum and to look for a single peak representing exactly the sum of the harmonics and sub-harmonics, the sufficiently strong sub-harmonics are examined to see if they can be considered as a pitch candidate or not. Indeed, if the estimated HR in each frame exceeds the HR_thresold value (0.4), then the sub_HR is considered as an f0 candidate, otherwise the harmonic is favored. Therefore, we obtain two matrixes containing the f0 and HR candidates for each frame. After that, the values of the averages and variances of HR are calculated in each frame, and then normalized by their respective maximum so that the classifier captures the relation between the peak in the spectrum and other frequency bands. For the test stage, we have used 50 pairs of voice samples. While, 25 pairs of voice samples has been used to train the gender speech classifier in the training stage. Moreover, each sample is regarded containing a single speaker and the T window used in this stage is a training of basic units, and it is similar to that used in the test stage.

3) Discrimination of Music and Environmental Sounds

This discrimination was performed using non-speech segments. Also, the FS feature was combined with MFCC coefficients and they are used as descriptors for this discrimination. Moreover, one of the algorithms KNN, SVM and GASOM was used as a classifier in this stage. Experiences have proved that the SF feature for music is lower than that for environmental sounds.

a) Discrimination of Music Genres

We have used the long-term feature for each segment of music such as the minimum entropy values and the average SF values of the sequences to discriminate the different musical genres. Also, the decision tree was used as a classifier since it is self-exploratory and easy to interpret. It has to mention here that the long-term feature for classic music has higher values compared to those for electronic music and this can be explained by the smoother energy changes (high-entropy) in the classic music and, these long-term feature values cannot be reached by the Jazz music. Also, we have tried the spectral Rolloff descriptor besides the entropy and the spectral flux, and we have found out that these latter were the best for this kind of discrimination.

F. Discrimination Steps for Second Audio Classification System

1) Music and Speech Discrimination

The statistic values (mean) of the sequences of spectral flux segments were used to discriminate between music and speech. Furthermore, the values obtained for the spectral flux were higher for speech than for music due to the fast alternation of local spectral changes between the speech phonemes. Moreover, we have tried the flux centroid and the chroma vectors as descriptors for this kind of discrimination, and the best discrimination result has been also reached by the spectral flux. Also, one of the algorithms SVM, KNN, and GASOM was used each time as a classifier in this discrimination.

2) Speech and Silence Discrimination, Male and Female

3) Speech Discrimination, and Discrimination of Music Genres

These discriminations have been performed in the same way as those of the first audio classification system.

The two audio classification systems are given in Fig. 3 and 4.

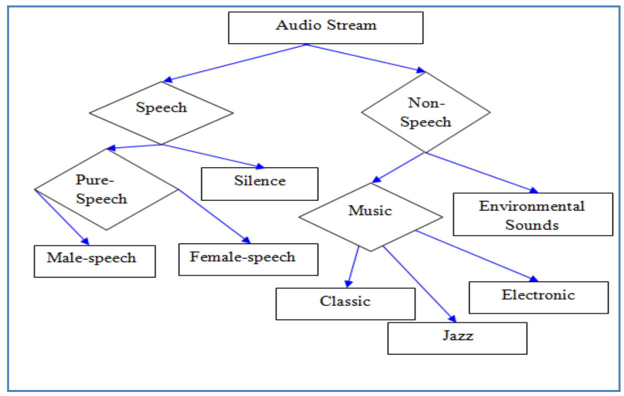


Fig. 3. First audio classification system.

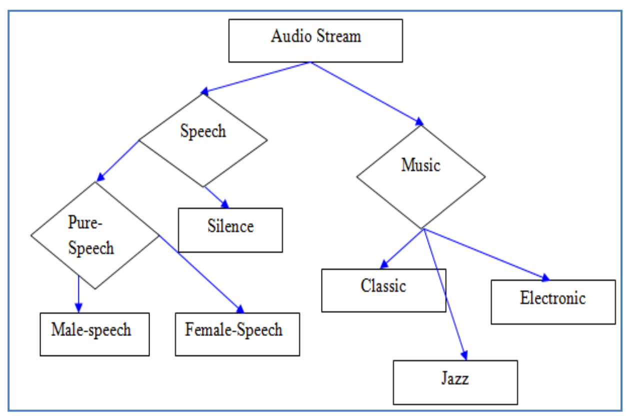


Fig. 4. Second audio classification system.

IV. Evaluations

A. Measures of Performance

To know the type of errors during the training and testing phases, we have used the CM confusion matrix, which is a $N_c \times N_c$ matrix whose rows and columns refer to the true and predicted class labels, respectively, of the dataset. Indeed, the confusion matrix is expressed as follows:

$$CM_n(i,j) = \frac{CM(i,j)}{\sum_{n=1}^{N_c} CM(i,n)} \tag{22}$$

Where, $CM(i,j)$ is the number of samples of class i, which are assigned to class j by the adopted classification method. Also, we have used the *overall accuracy* (*Acc)* which is defined as the ratio of the samples of dataset that have been correctly classified. Indeed, this evaluation criterion has the following expression:

$$Acc = \frac{\sum_{m=1}^{N_c} CM(m,m)}{\sum_{m=1}^{N_c}\sum_{n=1}^{N_c} CM(m,n)} \qquad (23)$$

Moreover, in order to describe how well the classification algorithm performs on each class, we define two class-specific measures: the first measure is the *class Recall*,Re(i), which is expressed as the proportion of data with true class label i that are correctly assigned to class i*:*

$$Re(i) = \frac{CM(i,i)}{\sum_{m=1}^{N_c} CM(i,m)} \qquad (24)$$

Where $\sum_{m=1}^{N_c} CM(i,m)$ is the total number of samples that are recognized to belong to class i. Concerning the second measure, it is the *class precision, pr(i),* which is defined as the ratio of samples that are correctly classified to class i with taking into account the total number of samples that are classified to that class.

$$Pr(i) = \frac{CM(i,i)}{\sum_{m=1}^{N_c} CM(m,i)} \qquad (25)$$

Where, $\sum_{m=1}^{N_c} CM(m,i)$ is the total number of samples that are classified to class i.

For the F_1*-measure,* it is defined as the harmonic mean values of precision and recall, such as:

$$F_1(i) = \frac{2Re(i)\Pr(i)}{Re(i)+\Pr(i)} \qquad (26)$$

B. Validation Methods

To generalize the performance of classifiers outside the training dataset, we have applied in this work two validation approaches:

1) Leave-One-Out Approach

It can be defined as a variation of k-fold cross-validation which splits randomly the dataset into non-overlapping k subset of equal size. Also, this technique is an exhaustive validation technique which is known by producing very reliable validation results.

2) Repeated-Hold-Out Approach

This approach allows refining and repeating k-times the Hold-out approach which splits the dataset into non-overlapping subsets: one for the test and the other for the training. Thus, the division of the dataset into two subsets is performed randomly at each iteration.

V. Results and Analysis

The first audio database used for the evaluation of our algorithms contains many audio types such as speech, music, environmental sounds, others1, others2, others3, which are extracted from different audio events. For the others1 type, it includes low-energy environmental sounds, such as wind, rain, silence, background sound, etc. Concerning the others2 type, it includes environmental sounds with abrupt changes in signal energy such as the sound of thunder, a door closing, an object breaking, etc. While, the others3 type contains high- energy sounds, non-abrupt environmental sounds, such as machine sounds. Also, the audio data in this data set are provided as 4-second chunks at two sampling rates (48 kHz and 16 kHz) with 48 kHz and 16 kHz for respectively the data in stereo and mono. Indeed, the 16 kHz recordings were obtained by down sampling the right-hand channel of the 48 kHz recordings. Thus, each audio file corresponds to a single chunk [45]. Moreover, we have used another data set containing sounds of different music genres, which are extracted from film soundtracks and music effects. Indeed, this dataset consists of 1000 audio tracks each 30 seconds long and it contains 10 genres whose each one is represented by 100 tracks. Furthermore, the tracks are all 22050Hz Mono 16-bit audio files in .wav format [46]. More details about this dataset can be found in [46]. In fact, we have used 2/3 of the dataset for training and 1/3 for testing different classifiers. In this work, we have used KNN, SVM, and GASOM algorithms as classifiers to test our models. We can note from Ttable I that for speech/non-speech discrimination, all algorithms have reached good classification results. Also, for speech/silence discrimination, all algorithms have reached the best classification result which is 100%. Moreover, for male/female speech discrimination, there is a little confusion between the two genres and the best classification value (98.8%) has been reached by GASOM algorithm with the leave-one-out validation technique. Good classification results have been also reached by the GASOM algorithm for music/environmental sounds discrimination in which it has reached the best value (99.4%). In the discrimination of music genres, the best results were 96.4% for classic music, 100% for jazz music, and 94.6% for electronic music, which were all obtained using a decision tree and a GASOM algorithm as classifiers in all previous levels of the audio discrimination process. Also, we can mention from Table I that all algorithms give good classification results in the speech/non-speech, speech/silence, and male/female speech discriminations. Moreover, the SVM algorithm has exceeded the KNN algorithm and it was competitive to GASOM algorithm in all audio discrimination types. Furthermore, the best discrimination results for all discrimination types have been achieved with all algorithms using leave-one-out as a validation technique. For the repeated-hold-out technique, the discrimination results have been always under those obtained with the leave-one-out validation technique.

From Table II, we can show a slight difference between GASOM algorithm and other algorithms in the classification results for the speech/music discrimination. Indeed, the percentage of speech which was recognized as speech is 97.85% for GASOM algorithm with the leave-one-out validation technique against 92.7% and 97.7%, respectively for the KNN and SVM algorithms. In speech/music discrimination, we have also tested the centroid flux and chroma vector, but the best result has been obtained by the spectral flux as it is recorded in Table II. For the silence/speech discrimination, the best results (100%) have been obtained by all algorithms like in the first proposed system. Concerning the male/female speech discrimination, the best result (95.7%) has been obtained using the GASOM algorithm as a classifier and leave-one-out as a validation technique. Also, this algorithm has proved its dominance by contributing to reach the best classification result using the decision tree as a classifier for the discrimination of music genres in which this classifier has reached the best value (94.2%) for the classic music. For the jazz music, 93.5% was the best classification result achieved by the decision tree as a classifier in the phase of discrimination of musical genres and the KNN algorithm as a classifier in all previous levels of the audio discrimination process. Furthermore, the best classification result for the electronic music (93.3%) has been reached by the decision tree as a classifier in the discrimination of different music genres and the KNN and SVM algorithms as classifiers in all previous levels of the audio discrimination process. Like in the first proposed system, the leave-one-out validation technique in this second audio classification system has mostly reached the best discrimination results compared to the repeated-hold-out validation technique.

Now, we can summarize the efficiency of the two proposed systems by comparing the performance results. From Tables III and IV, we can note that the first audio classification system has proved its success as it has reached the best performance results using different classification algorithms in all levels of the audio discrimination process by comparison to the second audio classification system. Also, the GASOM algorithm has reached the best F1-measure average for the music/environmental sounds discrimination with the leave-one out validation technique. For the male/female speech discrimination in the second audio classification system, the F1-mesure average has reached the best value (94.99%) using GASOM algorithm as a classifier and repeated hold-out as a validation technique. However, it has reached 98.04% in the first audio classification system using the same algorithm and leave-one out as a validation technique. Furthermore, for the discrimination of musical genres, the F1-measure average in the first audio classification system has reached the best value (97.04%) using the decision tree as a classifier and the GASOM algorithm as a classifier (with the leave-one-out validation technique) in all previous levels of the audio discrimination process. However, it has only reached 93.22% in the second audio classification system using the same algorithm and the same validation technique. We can note also that the performance results (for the discrimination of male/female speech and musical genres) were better for the first audio classification system as it contains more stages of audio discrimination. Thus, these discrimination stages have contributed to pure the audio segments from one level of audio discrimination to another until the discrimination of musical genres. For this, the results for discrimination of musical genres in the first audio classification system were better than in the second one.

TABLE I. CONFUSION MATRIX FOR DIFFERENT AUDIO CLASSIFICATION STEPS USING DIFFERENT ALGORITHMS IN THE FIRST AUDIO CLASSIFICATION SYSTEM

Confusion Matrix for Different Audio Classification Steps Using KNN Algorithm								
Leave-One-Out (best K=11)			*leave-one-out (best K=3)*			*Leave-One-Out (best K=3)*		
Speech	97.10	2.90	**Speech**	100	0.00	**Female-Speech**	98.10	1.90
Non-Speech	7.10	92.90	**Silence**	0.00	100	**Male-Speech**	5.80	94.20
Repeated-Hold-Out (best K=7)			*Repeated-Hold-Out (best K=7)*			*Repeated-Hold-Out (best K=3)*		
Speech	97.10	2.90	**Speech**	100	0.00	**Female-Speech**	97.60	2.40
Non-Speech	7.10	92.90	**Silence**	0.00	100	**Male-Speech**	6.00	94.00
Leave-One-Out (best K=3)				*Leave-One-Out (best K=3)*				
Music	95.80	4.20			**Classic**	92.50	7.50	0.00
Environmental Sounds	6.50	93.50			**Jazz**	0.00	100	0.00
Repeated-Hold-Out (best K=3)					**Electronic**	3.40	2.60	94.00
Music	94.60	5.40		*Repeated-Hold-Out (best K=3)*				
Environmental Sounds	6.10	93.90			**Classic**	89.50	9.70	0.80
					Jazz	0.40	98.30	1.30
					Electronic	3.40	2.60	94.00
Confusion Matrix for Different Audio Classification Steps Using SVM Algorithm								
Leave-One-Out			*Leave-One-Out*			*Leave-One-Out*		
Speech	98.1	1.9	**Speech**	100	0.00	**Female-Speech**	98.7	1.3
Non-Speech	5.4	94.6	**Silence**	0.00	100	**Male-Speech**	3.8	96.2
Repeated-Hold-Out			*Repeated-Hold-Out*			*Repeated-Hold-Out*		
Speech	97.1	2.9	**Speech**	100	0.00	**Female-Speech**	97.6	2.4
Non-Speech	7.5	96.9	**Silence**	0.00	100	**Female-Speech**	6.00	94.00
Leave-One-Out					*Leave-One-Out*			

Music	97.9	2.1			**Classic**	93.1	6.9	0.00
Environmental Sounds	2.6	97.4			**Jazz**	0.00	100	0.00
Repeated-Hold-Out					**Electronic**	2.40	1.00	96.6
Music	97.1	2.9				repeated-hold-out		
Environmental Sounds	2.8	97.2			**Classic**	89.5	9.7	0.8
					Jazz	0.4	98.3	1.3
					Electronic	3.4	2.6	94.0
Confusion Matrix for Different Audio Classification Steps Using GASOM Algorithm								
Leave-One-Out			*Leave-One-Out*			*Leave-One-Out*		
Speech	98.3	1.70	**Speech**	100	0.00	**Female-Speech**	98.80	1.20
Non-Speech	4.4	95.60	**Silence**	0.00	100	**Male-Speech**	2.80	97.20
Repeated-Hold-Out			*Repeated-Hold-Out)*			*Repeated-Hold-Out*		
Speech	97.10	2.90	**Speech**	100	0.00	**Female-Speech**	97.60	2.40
Non-Speech	7.50	92.50	**Silence**	0.00	100	**Female-Speech**	2.90	97.10
Leave-One-Out					*Leave-One-Out*			
Music	**99.40**	**0.60**			**Classic**	96.4	3.6	0.00
Environmental Sounds	1.05	98.95			**Jazz**	00.00	100	0.00
Repeated-Hold-Out					**Electronic**	4.40	1.00	94.60
Music	97.80	2.20			Repeated-Hold-Out			
Environmental Sounds	2.40	97.60			**Classic**	91.50	7.70	0.80
					Jazz	0.40	98.8	0.80
					Electronic	3.40	2.60	94.00

TABLE II. CONFUSION MATRIX FOR DIFFERENT AUDIO CLASSIFICATION STEPS USING DIFFERENT ALGORITHMS IN THE SECOND AUDIO CLASSIFICATION SYSTEM

Confusion Matrix for Different Audio Classification Steps Using KNN Algorithm								
Leave-One-Out (best K=13)			*leave-one-out (best K=13)*			*Leave-One-Out (best K=13)*		
Speech	92.70	7.30	**Speech**	100	0.00	**Female-Speech**	93.10	6.90
Music	9.25	90.75	**Silence**	0.00	100	**Male-Speech**	5.80	94.20
Repeated-Hold-Out (best K=15)			*Repeated-Hold-Out (best K=15)*			*Repeated-Hold-Out (best K=15)*		
Speech	97.10	2.90	**Speech**	100	0.00	**Female-Speech**	92.60	7.40
Music	7.50	92.50	**Silence**	0.00	100	**Male-Speech**	6.20	93.80
Leave-One-Out (best K=13)				*Repeated-Hold-Out (best K=15)*				
Classic	92.50	7.50	0.80		**Classic**	91.50	7.70	0.80
Jazz	6.50	93.50	0.00		**Jazz**	4.40	90.70	4.90
Electronic	4.20	2.50	93.30		**Electronic**	4.40	2.60	93.00
Confusion Matrix for Different Audio Classification Steps Using SVM Algorithm								
Leave-One-Out			*Leave-One-Out*			*Leave-One-Out*		
Speech	97.70	2.30	**Speech**	100	0.00	**Female-Speech**	94.80	5.20
Music	3.60	96.40	**Silence**	0.00	100	**Male-Speech**	5.80	94.20
Repeated-Hold-Out			*Repeated-Hold-Out*			*Repeated-Hold-Out*		
Speech	97.10	2.90	**Speech**	100	0.00	**Female-Speech**	94.60	5.40
Music	4.10	95.90	**Silence**	0.00	100	**Female-Speech**	6.05	93.95
Leave-One-Out					*Repeated-Hold-Out*			
Classic	93.20	6.80	0.00		**Classic**	92.10	7.10	0.80

Jazz	7.50	97.4	0.00		**Jazz**	4.10	90.70	5.20
Electronic	4.20	2.50	93.30		**Electronic**	4.80	2.60	92.60
Confusion Matrix for Different Audio Classification Steps Using GASOM Algorithm								
Leave-One-Out			*Leave-One-Out*			*Leave-One-Out*		
Speech	97.85	2.15	**Speech**	100	0.00	**Female-Speech**	94.80	5.20
Music	3.40	96.60	**Silence**	0.00	100	**Male-Speech**	5.80	94.20
Repeated-Hold-Out			*Repeated-Hold-Out)*			*Repeated-Hold-Out*		
Speech	97.30	2.70	**Speech**	100	0.00	**Female-Speech**	95.70	4.30
Music	4.00	96.00	**Silence**	0.00	100	**Female-Speech**	5.80	94.20
Leave-One-Out					*Repeated-Hold-Out*			
Classic	94.20	5.80	0.00		**Classic**	92.60	7.00	0.40
Jazz	6.90	98.95	0.00		**Jazz**	3.20	91.40	5.40
Electronic	5.25	2.50	92.25		**Jazz**	4.80	2.60	92.60

TABLE III. DIFFERENT PERFORMANCE RESULTS OBTAINED USING DIFFERENT ALGORITHMS FOR THE FIRST AUDIO CLASSIFICATION SYSTEM

The Performance Results Using KNN Algorithm					
Classification Type	**Validation Method**	**Overall accuracy**	**Average Precision**	**Average Recall**	**Average F1 measure**
Speech-Non-Speech	**Repeated-Hold-Out**	95.00	95.10	95.00	95.04
	Leave-One-Out	95.00	95.10	95.00	95.04
Speech -Silence	**Repeated-Hold-Out**	100	100	100	100
	Leave-One-Out	100	100	100	100
Female and Male Speech	**Repeated-Hold-Out**	95.80	95.90	95.80	95.84
	Leave-One-Out	96.15	96.25	96.15	96.19
Music and Environmental Sounds	**Repeated-Hold-Out**	97.15	97.25	97.15	97.19
	Leave-One-Out	94.65	94.75	94.65	94.69
Classic, Jazz and Electronic Music	**Repeated-Hold-Out**	90.10	91.2	94.1	92.62
	Leave-One-Out	95.26	95.36	95.26	95.30
The Performance Results Using SVM Algorithm					
Classification Type	**Validation Method**	**Overall accuracy**	**Average Precision**	**Average Recall**	**Average F1 measure**
Speech-Non-Speech	**repeated-hold-out**	97.00	97.10	97.00	97.04
	leave-one-out	96.35	96.45	96.35	96.39
Speech -Silence	**repeated-hold-out**	100	100	100	100
	leave-one-out	100	100	100	100
Female and Male Speech	**repeated-hold-out**	95.80	95.90	95.80	95.84
	leave-one-out	97.45	97.55	97.45	97.49
Music and Environmental Sounds	**repeated-hold-out**	97.65	97.75	97.65	97.69
	leave-one-out	96.56	96.66	96.56	97.11
Classic, Jazz and Electronic Music	**repeated-hold-out**	93.93	94.00	93.93	93.96
	leave-one-out	96.56	96.70	96.56	96.63
The Performance Results Using GASOM Algorithm					
Classification Type	**Validation Method**	**Overall accuracy**	**Average Precision**	**Average Recall**	**Average F1 measure**
Speech-Music	**Repeated-Hold-Out**	96.35	96.45	96.35	96.39

	Leave-One-Out	96.95	97.00	96.95	96.97
Speech –Silence	Repeated-Hold-Out	100	100	100	100
	Leave-One-Out	100	100	100	100
Female and Male Speech	Repeated-Hold-Out	97.35	97.45	97.35	97.39
	Leave-One-Out	98.00	98.10	98.00	98.04
Music and Environmental Sounds	Repeated-Hold-Out	97.70	97.80	97.79	97.74
	Leave-One-Out	99.17	99.27	99.17	99.21
Classic, Jazz and Electronic Music	Repeated-Hold-Out	94.76	94.86	94.76	94.80
	Leave-One-Out	97.00	97.10	97.00	97.04

TABLE IV. THE OBTAINED PERFORMANCES USING DIFFERENT ALGORITHMS FOR THE SECOND AUDIO CLASSIFICATION SYSTEM

The Obtained Performance Using KNN Algorithm					
Classification type	**Validation Method**	**Overall accuracy**	**Average Precision**	**Average Recall**	**Average F1 measure**
Speech-Music	**repeated-hold-out**	91.22	91.32	91.22	91.26
	leave-one-out	97.05	97.15	97.05	97.09
Speech -Silence	**repeated-hold-out**	100	100	100	100
	leave-one-out	100	100	100	100
Female and Male Speech	**repeated-hold-out**	93.20	93.30	93.20	93.24
	leave-one-out	93.95	94.05	93.95	93.99
Classic, Jazz and Electronic Music	**repeated-hold-out**	90.1	91.2	94.1	92.62
	leave-one-out	93.10	93.20	93.10	93.14
The Obtained Performance Using SVM Algorithm					
Classification type	**Validation Method**	**Overall accuracy**	**Average Precision**	**Average Recall**	**Average F1 measure**
Speech-Music	**Repeated-Hold-Out**	96.50	96.60	96.50	96.54
	Leave-One-Out	97.05	97.15	97.05	97.09
Speech -Silence	**Repeated-Hold-Out**	100	100	100	100
	Leave-One-Out	100	100	100	100
Speech Female and Speech Male	**Repeated-Hold-Out**	94.27	94.37	94.27	94.31
	Leave-One-Out	94.50	94.60	94.50	94.54
Classic, Jazz and Electronic Music	**Repeated-Hold-Out**	91.86	91.96	91.86	91.90
	Leave-One-Out	93.00	93.10	93.00	93.04
The Obtained Performance Using GASOM Algorithm					

Classification type	Validation Method	Overall accuracy	Average Precision	Average Recall	Average F1 measure
Speech-Music	**Repeated-Hold-Out**	96.65	96.75	96.65	96.69
	Leave-One-Out	97.22	97.32	97.22	97.26
Speech -Silence	**Repeated-Hold-Out**	100	100	100	100
	Leave-One-Out	100	100	100	100
Speech Female and Speech Male	**Repeated-Hold-Out**	94.95	95.05	94.95	94.99
	Leave-One-Out	94.50	94.60	94.50	94.54
Classic, Jazz and Electronic Music	**Repeated-Hold-Out**	92.20	92.30	92.20	92.24
	Leave-One-Out	93.18	93.28	93.18	93.22

To make a comparison, the first audio classification system has been developed using Hidden Markov Model (HMM) and Multilayer Perceptron *(*MLP) classifier. For the MLP classifier, it is a Multilayer Perceptron Feed Forward Fully Connected Neural Network (MLPFFFCNN) with a sigmoid activation function. Indeed, it is a neural network with 3 hidden layers with 4 neurons for each one and a number of output units equals to the number of classes. Concerning the training of this classifier, it has been carried out using back propagation algorithm and the stopping criterion has been set according to the Mean Square Error (MSE) when it reaches the zero value. For the second classifier, it is a background HMM with 4 states in order to represent the sequences of observation vectors. Moreover, a refinement stage has been added using a Viterbi decoding as a resegmentation stage in order to refine the segmentation results. The Esperance-Maximization (EM) algorithm has been also used in order to learn the parameters of HMM. As it is shown in Table V, the HMM classifier has succeeded to achieve good results in terms of measured performances by comparison to MLP algorithm. Indeed, it has reached the best F1-measure averages in all levels of the first audio classification system. However, these results remain competitive to those obtained with GASOM algorithm which has succeeded to reach the best results as it was mentioned above.

Furthermore, the GASOM algorithm has outperformed the HMM and MLP algorithms in terms of time consumption for which it has reached the best results in all audio classification levels as it shown in Table VI. Thus, this speed of processing makes this algorithm so desired in real-time applications.

TABLE V. DIFFERENT PERFORMANCE RESULTS OBTAINED USING THE MLP AND HMM ALGORITHMS FOR THE FIRST AUDIO CLASSIFICATION SYSTEM

Classification type	Classifier	Overall accuracy	Average Precision	Average Recall	Average F1 measure
Speech-Non-Speech	*MLP*	95.50	95.60	95.50	95.54
	HMM	96.00	96.10	96.00	**96.04**
Speech – Silence	*MLP*	100	100	100	100
	HMM	100	100	100	**100**
Female speech/Male Speech	*MLP*	95.8	95.9	95.8	95.84
	HMM	97.00	97.10	97.00	**97.04**
Music /Environmen-tal Sound	*MLP*	96.75	96.85	96.75	96.79
	HMM	97.30	97.40	97.30	**97.34**
Classic, Jazz and Electronic Music	*MLP*	93.80	93.90	93.80	93.84
	HMM	95.50	95.60	95.50	**95.54**

TABLE VI. THE OBTAINED TIME CONSUMPTION IN ALL AUDIO CLASSIFICATION LEVELS USING MLP, HMM AND GASOM ALGORITHMS

Audio classification step Algorithm	Speech-Non Speech	Speech - Silence	Female speech/ Male Speech	Music /Environ--mental Sounds	Classic, Jazz and Electronic Music
MLP	1.6	1.0	1.9	1.1	1.8
HMM	1.2	0.9	1.1	0.8	1.4
GASOM	0.5	0.6	0.7	0.5	1.0

VI. Conclusion and Future Work

Two audio classification systems have been proposed in this work in which an audio stream is discriminated into homogenous regions and classified into basic audio types such as speech, non-speech, silence, music, environmental sounds and so on. The principle goal is to exploit audio segmentation algorithms which can be integrated in multimedia content analysis applications and audio recognition systems. Indeed, three algorithms have been used for the two audio classification and segmentation systems. For the first system, the audio stream has been discriminated into speech/non-speech, pure-speech/silence, male/female speech, environmental sounds/music, music genres: classic, jazz, and electronic music. Concerning the second system, the audio stream has been segmented into speech/music, pure-speech/silence, male/female speech, music genres: classic, jazz, and electronic music. For the discrimination of musical genres and pure-speech/silence, the decision tree and a rule-based classifier are respectively used as classifiers. While, in the retaining levels of two audio classification systems, one of the algorithms KNN, SVM, and GASOM has been used as a classifier. Experimental results have shown that the GASOM algorithm is so efficient for most audio discrimination types in terms of accuracy and time consumption. Thus, this advantage plus the no-requirement of much training data makes this algorithm very useful for real-time multimedia applications. In future work, the two proposed systems can be exploited to perform many applications, such as the automatic speech recognition, the human-computer interaction systems, the speaker tracking, and so on. Also, the GASOM algorithm can be combined with k-means algorithm in order to access more data and achieve better performances.

References

[1] Chih-Chin Liu, Jia-Lien Hsu, Chen ALP (1999) An approximate string matching algorithm for content-based music data retrieval, IEEE IntConf Multimedia Comp Syst 1: 451–456.

[2] Kosugi N, Nishihara Y, Kon'ya S, Yamamuro M, Kushima K (1999) Music retrieval by humming-using similarity retrieval over high dimensional feature vector space, IEEE Pacific Rim ConfCommun, Comp Signal Processing 404–407.

[3] Kataoka M, Kinouchi M, Hagiwara M (1998) Music information retrieval system using complex-valued recurrent neural networks, IEEE IntConfSyst, Man, and Cybernetics, 5: 4290–4295.

[4] Guohui Li, Khokhar AA (2000) Content-based indexing and retrieval of audio data using wavelets, ICME 2000 2: 885–888.

[5] Subramanya SR, Youssef A (1998) Wavelet-based indexing of audio data in audio/multimedia databases, ProcInt Workshop on Multi-Media Database Management Sys 46–53.

[6] I.McLoughlin,*AppliedSpeechandAudioProcessing:WithMATLAB Examples*, Nanyang Technological University, CambridgeUniversity Press, 2009.

[7] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on Speechand Audio Processing*, vol. 10, no. 7, pp. 504–516, 2002.

[8] L. Lu, H.-J. Zhang, and S. Z. Li, "Content-based audio classification and segmentation by using support vector machines," *Multimedia Systems*, vol. 8, no. 6, pp. 482–492, 2003.

[9] M. L. Coz, J. Pinquier, R. Andre-Obrecht, and J. Mauclair, "Audio indexing including frequency tracking of simultaneous multiple sources in speech andmusic," in *Proceedings of the 11thInternationalWorkshop on Content-Based Multimedia Indexing(CBMI '13)*, pp. 23–28, IEEE, Veszprem, Hungary, June 2013.

[10] S. Kiranyaz, A. F. Qureshi, and M. Gabbouj, "A generic audio classification and segmentation approach for multimedia indexing and retrieval," *IEEE Transactions on Audio, Speech andLanguage Processing*, vol. 14, no. 3, pp. 1062–1081, 2006.

[11] Scheirer E, Slaney M (1997) Construction and evaluation of a robust multifeature speech/music discriminator, ICASSP-97 2: 1331–1334.

[12] A. Dessein and A. Cont, "An information-geometric approach to real-time audio segmentation," *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 331–334, 2013.

[13] H. Vajaria, T. Islam, S. Sarkar, R. Sankar, and R. Kasturi, "Audio segmentation and speaker localization in meeting videos," in *Proceedings of the 18th International Conference on PatternRecognition (ICPR '06)*, vol. 2, pp. 1150–1153, IEEE, Hong Kong, August 2006.

[14] J. Huang, Y. Dong, J. Liu, C. Dong, and H.Wang, "Sports audio segmentation and classification," in *Proceedings of the IEEEInternational Conference on Network Infrastructure and DigitalContent (IC-NIDC '09)*, pp. 379–383, IEEE, Beijing, China, November 2009.

[15] J. Hennebert, M. Hasler, and H. Dedieu, "Neural networks in speech recognition," in *Proceedings of the 6th MicrocomputerSchool of Neural Networks, Theory and Applications (Micro-Computer '94)*, pp. 23–40, Prague, Czech Republic, 1994.

[16] P. Nguyen, G. Heigold, and G. Zweig, "Speech recognition with flat direct models," *IEEE Journal on Selected Topics in SignalProcessing*, vol. 4, no. 6, pp. 994–1006, 2010.

[17] S.-X. Zhang and M. J. F. Gales, "Structured SVMs for automatic speech recognition," *IEEE Transactions on Audio, Speech andLanguage Processing*, vol. 21, no. 3, pp. 544–555, 2013.

[18] P. Hu, W. Liu, W. Jiang, and Z. Yang, "Latent topic model for audio retrieval," *Pattern Recognition*, vol. 47,no. 3, pp. 1138–1143, 2014.

[19] Birkenes, T. Matsui, K. Tanabe, S. M. Siniscalchi, T. A. Myrvoll, andM. H. Johnsen, "Penalized logistic regression with HMM log-likelihood regressors for speech recognition," *IEEETransactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1440–1454, 2010.

[20] Z. Ping, T. Li-Zhen, and X. Dong-Feng, "Speech recognition algorithm of parallel subband HMM based on wavelet analysis and neural network," *Information Technology Journal*, vol. 8, no. 5, pp. 796–800, 2009.

[21] V. R. V. Krishnan and P. BabuAnto, "Features of wavelet packet decomposition and discrete wavelet transform for Malayalam speech recognition," *International Journal of Recent Trends inEngineering*, vol. 1, no. 2, pp. 93–96, 2009.

[22] Z. Harchaoui, F. Vallet, A. Lung-Yut-Fong, and O. Capp´e, "A regularized kernel-based approach to unsupervised audio segmentation," in *Proceedings of the IEEE International Conferenceon Acoustics, Speech, and Signal Processing (ICASSP '09)*, pp. 1665–1668, IEEE, Taipei, Taiwan, April 2009.

[23] T. Giannakopoulos and S. Petridis, "Detection and clustering of musical audio parts using Fisher linear semi-discriminant analysis," in *Proceedings of the IEEE 20th European SignalProcessing Conference (EUSIPCO '12)*, pp. 1289–1293, IEEE, Bucharest, Romania, August 2012.

[24] L. Lu, H.-J. Zhang, and S. Z. Li, "Content-based audio classification and segmentation by using support vector machines," *Multimedia Systems*, vol. 8, no. 6, pp. 482–492, 2003.

[25] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-basedmusic information retrieval: current directions and future challenges," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

[26] Hemant Misra, ShajithIkbal, HervéBourlard, HynekHermansky, Spectral entropy based feature for robust ASR, in: Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'04, vol. 1, IEEE, 2004, pp. I–193.

[27] A. Pikrakis, T. Giannakopoulos, S. Theodoridis, A computationally efficient speech/music discriminator for radio recordings, in: International Conference on Music Information Retrieval and Related Activities, ISMIR06, 2006.

[28] A. Pikrakis,T. Giannakopoulos, S.Theodoridis,A speech/music discriminator of radio recordings based on dynamic programming and bayesian networks, IEEETransactions on Multimedia 10 (5) (2008) 846–857.

[29] TheodorosGiannakopoulos,SergiosPetridis,Unsupervised speaker clustering in a linear discriminant subspace,in:Proceedings of the 2010 Ninth International Conference on Machine Learning and Applications, ICMLA '10, 2010, pp. 1005–1009.

[30] G. Tzanetakis,P. Cook,Musical genre classification of audio signals, IEEETransactions on Speech and Audio Processing 10 (5) (2002) 293–302.

[31] David Pearce, Hans günterHirsch,The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, in: ISCA ITRWASR2000, 2000, pp. 29–32., Ericsson Eurolab Deutschland Gmbh.

[32] GregoryH.Wakefield,Mathematical representation of joint time-chroma distributions, in: SPIE's International Symposium on Optical Science,Engineering, and Instrumentation, International Society for Optics and Photonics, 1999, pp. 637–645.

[33] Mark A. Bartsch, Gregory H.Wakefield,Audiothumbnailing of popular music using chroma-based representations, IEEETransactions on Multimedia 7 (1) (2005) 96–104.

[34] Mark A. Bartsch, Gregory H.Wakefield,To catch a chorus: using chroma-based representations for audio thumbnailing, in: 2001 IEEEWorkshop on the Applications of Signal Processing to Audio and Acoustics, IEEE, 2001, pp. 15–18.

[35] Meinard Müller, Frank Kurth, Michael Clausen, Audio matching via chroma-based statistical features, in: Proceedings of ISMIR, London, GB, 2005, pp. 288–295.

[36] A. Pikrakis,T. Giannakopoulos, S.Theodoridis,A speech/music discriminator of radio recordings based on dynamic programming and bayesian networks, IEEETransactions on Multimedia 10 (5) (2008) 846–857.

[37] Lawrence R. Rabiner, RonaldW. Schafer, Introduction to digital speech processing, Foundations andTrends in Signal Processing,Now Publishers Inc, 2007.

[38] J.Weston and C. Watkins, "Support vector machines for multiclass pattern recognition," in *Proceedings of the 7th EuropeanSymposium on Artificial Neural Networks (ESANN '99)*, vol. 99, 1999.

[39] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on NeuralNetworks*, vol. 13, no. 2, pp. 415–425, 2002.

[40] K.-B.Duan and S. S. Keerthi, "Which is the bestmulticlass SVM method? An empirical study," in *Multiple Classifier Systems*, vol. 3541 of *Lecture Notes in Computer Science*, pp. 278–285, Springer, Berlin, Germany, 2005.

[41] P. Clarkson and P. J. Moreno, "On the use of support vector machines for phonetic classification," in *Proceedings of theIEEE International Conference on Acoustics, Speech, and SignalProcessing (ICASSP '99)*, vol. 2, pp. 585–588, IEEE, Phoenix, Ariz, USA, March 1999.

[42] Marie Cottrell et al : 'Cartes auto-organisées pour l'analyse exploratoire de données et la visualisation', Université Paris 1, 2003.

[43] Gaussier E, et al: 'A hierarchical model for clustering and categorising documents', Advances in Information Retrieval,Proceedings of the 24th BCS-IRSG European Colloquium on IR Research ECIR-02, Glasgow. Lecture Notes in Computer Science 2291, p.229-247, Springer, 2002.

[44] Mohamed Salah Salhi, NajetArous, NoureddineEllouze: 'A suitable model of evolutionary SOM for phonemes recognition', journal IRECOS Napoly-Italy, (Indexé COMPENDEX – Elsevier– Copernicus). Laboratoire LSTS- ENIT_Tunis, septembre 2009.

[45] http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-real-life-audio.

[46] George Tzanetakis, Gtzan genre collection. http://marsyas.info/download/data_sets.

The Effectiveness of Cloud-Based E-Learning towards Quality of Academic Services: An Omanis' Expert View

Qasim A. Alajmi[1]
Faculty of Computer Science & MIS,
Oman College of Management & Technology,
Sultanate of Oman

Adzhar Kamaludin[2], Ruzaini Abdullah Arshah[3], Mohammed A. Al-Sharafi[4]
Faculty of Computer Systems & Software Engineering,
University Malaysia Pahang,
LebuhrayaTunRazak, 26300, Kuantan, Pahang, Malaysia

***Abstract*—The purpose of this paper is to understand the importance, relevance and the need of implementing Cloud-Based E-Learning (CBEL) in Higher education institutions (HEIs) in Oman. The paper maintains its emphasis on addressing the effectiveness of the cloud based e-learning system and also takes into account the activities of comparing and contrasting the before and after effects of the implementation of the CBEL on higher educational sector in Oman. Method: The methodological approach of this paper follows qualitative approach of data collection and the data analysis techniques used in this paper are interpretivist approach and thematic analysis. Results and Findings: The data analysis techniques used in this research paper helped in understanding and gathering meaningful insights relating to the need and significance of implementing CBEL in educational sectors.**

Keywords*—Cloud computing; e-learning; cloud-based e-learning; higher education institutions; quality of academic services

I. Introduction

A wide array of traditional educational platforms today is becoming obsolete and least suitable and worthy for fitting the requirements of consistent educational advancements [1], [2]. These conventional educational platforms are also becoming unable to go in-line with the variations of the learning demand in time and for this matter, the developments in the computer networking systems have brought revolutionary opportunities for it. E-learning in this case serves to be as a promising pattern for education [3], [4]. Cloud computing on the other hand, is associated with being as an enhanced base that enables the provision of and utilization of an adequate pool of computing resources [5]. Cloud computing promotes this practice with its effective scalability and its enhanced capability of utilizing resources that are virtualized, and these resources are used as a service through the Internet [6], [7]. The applications of cloud computing relate with the provision of flexibility to educational platforms as a whole. Cloud computing applications help in the prevention of challenges that are generally faced by conventional educational platforms [4]. For this matter, this research paper maintains its focus of attention on understanding the implementation of Cloud-Based E-Learning (CBEL) in HEIs in Oman and the relative effectiveness of the work operations achieved with its implementation. The discussion carried out reflects on the information collected from authorities in the educational arena that provides insights on the combination of e-learning and cloud computing in an educational arena.

Following are the research questions that are proposed in the study: 1) How effective was the implementation of cloud computing based e learning? 2) What were the specific areas where cloud computing based e-learning enhanced work operations? The proposed study aims to understand the 'application and implementation of CBEL', specifically its adoption in the HEIs in Oman. Following are the objectives of this study: To understand the effectiveness of the CBEL that is implemented in educational sectors in Oman, to highlight specific key areas of improvement with HEIs in Oman and to compare and contrast the before and after effects of the implementation of CBEL in Oman.

II. Literature Review

A. Cloud Computing and E-learning

The cloud computing technology makes full potential use of technology and its relevant developments, which further makes use of existing remote servers that are centralized and internet in order to maintain the applications and information data. The cloud computing application enables and permits businesses and consumers to utilize applications, irrespective of the installation requirements. This practice also provides access to consumers and business so that they are enabled to access their personal informational data files on any computer, provided with an Internet connection. The cloud computing technology assures the efficiency of the computing processes by enabling the centralization of bandwidth, processing and data storage. Thus for this reason, cloud computing is associated with being as an element that makes use of hardware and software computing resources, that are transported through a network connection or Internet in the form of a service. Cloud computing derives its name from the shape of a cloud which serves to be as an image of the complex and compound infrastructure and framework. It contains in its system diagrams Cloud computing with user's informational data, computation and software applications, highly relies on the remote services.

This research has been supported by the Universiti Malaysia Pahang Research Grant RDU180310.

E-learning on the other hand makes use of technologies that assists and helps in online assessment, education, learning and knowledge extraction. In this way, e-learning makes potential use of the existing innovative educational technology and enormous amount of resources; this in turn enables newer and better ways for teaching [8]. E-learning thus can be associated with the form of education and learning that is network based, assuring the use of technological advancements. In addition to this, the cloud computing system in educational institutions and campuses is delighted with the provision of an effective infrastructure and framework, moreover it also relates with the provision of a delivery model that reflects on their dynamic demands. There exist wide arrays of benefits and advantages that are associated with the application of cloud computing. As far as the use of cloud computing in educational and learning arena is concerned, it provides increased level of support for educational platforms, so as to help them in devising certain prevention strategies against some of the commonly held challenges. Furthermore, the challenges faced by educational institutions range from rapid and effective communication, reduction in cost incurring, provision of flexibility, provision of accessibility, privacy and security concerns.

B. Higher Education in Oman

The expansion associated with the (HEI's) is regarded with being as rapid and swift in the Arab economies. The underlying reason behind this rapid expansion is the increased education's social demand and the emerging governmental preferences in order to support and develop the effectiveness and efficiency of the human resources as per the requirements of their relative economies. It is the administrative and financial capacity and capability of these economies to successfully respond to this emerging need and this has been well discussed and has been a part of literature. A wide variety of approaches are adopted and used in order to upgrade the quality of higher education. Majority of the Arab countries make use of wide arrays of approaches in order to improve the quality of higher education. One of the approaches reflects on the shift and transmission of education that has been made to private sector, this helps in preventing the issue arising due to emerging demand. In addition to this, there still exists a universal disapproval associated with the quality of higher education in Gulf countries and in Oman in particular, despite of the consistent efforts and attempts made for upgrading the quality of higher education [9].

All of the member states of the Gulf Cooperation Council (GCC) are experiencing an increased and swift transformations and expansions in their higher educational sectors. Moreover, there lies multiple questions that are related with what factors to focus on when teaching and how to carry out the activities associated with teaching process. With the development and formation of the educational institutions, including setting up their infrastructure, there also exist certain emerging concerns that arise due to maintenance and sustainability of the quality control and its accountability. A wide number of Higher Education Sectors now seek to assure the containment of wide range, diverse and varying levels of Higher Education Institutions with varying curricula [10]. The central questions that emerge could relate to explaining the enforcement of the teaching standards and the measurement of the research output. The concept and idea of forming international partnerships are of high significance to the Gulf Cooperation countries' universities and colleges. Some of the Gulf member countries have resulted in successfully attracting the western/foreign universities, on the other hand, other member states are still stuck with unaccredited educational platforms from around the world [11]. In Oman there are 29 private universities/colleges affiliated with western universities.

It is because of the apparent failure of the public educational institutions in meeting the requirements for highly qualified graduates of universities that are required for the development in social and economic terms. On the other hand, a majority of the international and private educational institutions including colleges and universities have efficiently succeeded in attracting a wide range of students and have also increased in number, because of their ability to naturally respond to the escalating ideas and concepts of globalization, privatization, agreeableness of adoption towards the foreign culture. As far as the higher education system is Oman is concerned, it is undergoing significant refinement phase. Oman is making relevant efforts to accommodate and catering the increasing young national population by enlarging the base for primary and secondary schools. As per the data census of the past three decades, Oman is seen as flourishing and observed to be followed by swift developments in its higher education system [11]. The country has formed an overall independent system which includes more than 60 Higher Educational Institutes (Public and Private) they successfully provide varying degree programs. For a majority number of authorities belonging from different parts of the world, the improvements and enhancement of quality of the overall education system has served to be as a prior concern. In order to fortify the national and social identity among a wide range of individual learners, a wide majority of countries are taking necessary steps in order to strengthen and refine the education systems and quality. As a matter of fact, this practice further leads to enabling individual learners to efficiently respond to consistently changing political, economic and technological environments and overall swiftly changing demands of the future [4].

C. Analysis of Higher Education in Oman

As per the analysis of the above mentioned Higher Education system in Oman, a complete SWOT analysis can be drawn out to understand the overall environment of higher education in Oman and also identify the need and growth level of implemented CBEL in Oman. As discussed in the section above, the growing population of the young nationals in Oman serves to be as a dynamic opportunity for the country. This opportunity can thus be used as a strengthening element to cash on making further developments in the higher education system and make use of the implementation and application of the CBEL. In addition to this another strengthening element for Oman is its ability and efforts in developing independent educational system that comprises of wide variety of (HEI). The increased range of developed higher educational institutes in Oman can catch the benefits of implementing and successfully utilizing the CBEL.

In addition to this, one of the weaknesses of the higher education system in Oman is the widely accepted level of criticism on the quality of education among majority of the

Gulf countries. This element also serves to be as an emerging threat for Oman, because of the perceived quality of education system and the resulting impact of it. Furthermore, another emerging potential threat can be the possible inability of Oman's education system to understand, identify and develop relevant prevention strategies for problems and risks that are associated with the implementation of CBEL. As a matter of fact, Oman's overall educational system is flourishing and is under development, thus there lie higher chances of less familiarity with the issues and risks that are related with the CBEL.

D. Current E-Learning Environment

As briefly discussed in the sections above, e-learning is associated with being as a learning system that is electronically supported-learning undertakes the necessary intersections with varying elements of practice, such as, learning, teaching, assessment, technology and knowledge [12]. The transmission of relevant network based skills, knowledge base, learning abilities is enabled by the computer based e-learning processes. These network based processes reflect the containment of applications which include computer based learning, web and network based learning, virtually enabled education opportunities and thus promoting digital based collaboration. Synchronous and asynchronous are the two broad categories that are used to understand the concepts and usage of e-learning. Furthermore, synchronous e-learning related with the learning process that enables and individual learner to indulge in a course available online and with the help of streaming video and audio online and communicating with the faculty member online at a specific period of time. The synchronous e-learning appears to be as a learning process that provides numerous advantages for the learner, but on the other hand it does have a major disadvantage associated with it too. The most crucial disadvantage associated with synchronous e-learning is the inability of this e-learning category to provide time flexibility [13], as the entire learning process activities are carried out at a specific time period. On the other hand, asynchronous e-learning process provides a student with the advantage of time flexibility, thus an individual learner is enabled to participate in the online learning process by assuring the learning processes are carried out at an appropriate and suitable time for a learner. Moreover, asynchronous e-learning is also exposed to a major disadvantage, which relates with the lack of immediacy and timely respond rate of a faculty member.

E. Cloud Computing Educational Environment

Cloud computing has served to be as a dynamically developing subject that has successfully attracted a wide range of individuals from different parts of the world and belonging from different disciplines. The provision, transmission and management of varying computer resources such as, processing, servers, storage, applications and networking, are all transformed with the cloud computing applications. Moreover, educational cloud is associated with being as one of the most attractive and prominent applications of cloud computing [14]. The underlying reason of educational cloud being as the most interesting application of cloud computing is its power of undertaking and possessing the emphasis of thousands of computers on one problem. Thus, this practice followed by educational cloud, enables it to permit the searches of different researchers, seeking relevant models, discovering at a rapid rate and hence helping in the development of a smarter planet. A wide variety of educational platforms can now unlock doors for their technological infrastructures for public and private sectors in order to undertake the research developments.

The consistently changing, increasing and demanding resources requirements and ever growing energy costs can now efficiently be met by universities with the efficiencies of cloud computing application. It is because of this matter that now students and individual learners are now enabled to make use of their mobile devices and stay connected with the campus's educational services [15]. Furthermore, the faculty members are now consistently demanding and requiring to be provided with efficient level of flexibility and access while undertaking technology integration into the classes. The services of the educational cloud computing reflect upon the increasing variety of services that exist on the Internet, along with the most swiftly growing innovative part of education and technology. Moreover, educational cloud also assures the provision of varying levels of services, which are associated with being as advantageous and beneficial for individual learners, staff and faculty members. The overall benefits and the role that cloud computing plays should be prominently highlighted because; it enables the provision of direct accessibility to wide arrays of multiple academic resources, educational tools and research applications.

The educational and learning arena is undertaken by the educational cloud computing on a next level, because of the advancements and developments in the applications, platforms, services and overall academic based cloud computing services. Today, a wide variety of individual learners/students, faculty members and researchers are making complete potential use of cloud computing based services and applications. In addition to this, a heavy investment in cloud computing is followed by these applications because of the increased dependability of the future of academic learning research on cloud computing [16]. On the other hand, the application of cloud computing is regarded with being as a new model that is threatening the role performance of educational institutions and faculty members to be very specific. This practice of cloud computing poses varying levels of challenges to the computing service personnel who in one way might fear the outcomes of their job roles being outsourced. Furthermore, some of the developers might also feel increased discomfort level the transference of operationally critical informational data and services beyond the institution's boundaries. Hence, there exist certain challenges that are associated more with the perceptions that are developed rather than the reality.

III. METHODOLOGY

A. Research Design

The research design of this study is designed on the basis of qualitative data gathering through Semi-structured interviews and qualitative analysis of the collected data. The data collection instrument used is the interview questions. In order to contribute to the meaningfulness of the subject study and motive of the study, the use of qualitative research was followed. The use of qualitative research data collection also contributes to the significance of this study, because it enables to cater and include the fresh information and the new findings that will emerge with time, thus the study overall will assure rapid revision. The semi-structured interview questions designed to take into consideration the change and development in the effectiveness of higher education system in Oman and on what specific areas the refinement of the education system is highlighted. In addition to this, the interview questions also developed in a way that they will gather information on the areas for further improvements that are needed to be taken into account in order to assure efficient implementation of the CBEL.

B. Sample

As a matter of fact, the sampling process in carrying out a research study on data collection is regarded with being as both, crucial, complex and of high relevance. Thus in order to meet the data collection requirements of the subject study, the process of sampling needs to be adequate, it needs to highly reflect and generate relevant and meaningful information. For this reason, to maintain quality of the respondent's views and to get a deeper and meaningful insight into each interview question, a sample of 10 experts, which serve as regulatory authority in the higher education system in Oman. The study conducted maintained its focus on primary data collection and dependent on making use of qualitative data collection techniques and qualitative analysis, so as in order to generate meaningful qualitative views and responses from respondents. Table I contains further details on the kinds of participants in this study. The identity information of the experts was detached and a distinctive ID for differentiating the quotes of each participant was included for the purpose of confidentiality.

TABLE I. INTERVIEWEES' PROFILE

# ID	Description	Experience	Reason for Inclusion
1	Director General for Public and Private universities and colleges in MoHE in Oman.	22 yrs	Due to her direct involvement in HEIs issues.
2	Director General for Applied colleges in MoHE in Oman.	30 yrs	Due to his various designations had in IT developments for higher education in Oman.
3	Director General for strategic planning in MoHE in Oman.	28 yrs	Due to his membership in lots of committees of education nature.
4	Director for IT directorate in MoHE in Oman.	26 yrs	Due to his achievement in ELearning in some of the Public colleges.
5	Ex-Under secretary in MoHE in Oman.	35 yrs	Due to his actively involved in Educational council.
6	Ex- Director for IT directorate in MoHE in Oman.	35 yrs	He is one of the figures in Oman IT programs initiatives.
7	Consultant in MoHE in Oman.	30 yrs	Full experienced in Cloud services.
8	Expert in MoHE in Oman.	32 yrs	Dues to his involvement with top decision makers in the ministry.
9	Expert in MoHE in Oman	38 yrs	Chair of design curriculum committee.
10	Academic advisor in MoHE in Oman.	40 yrs	Supervisor of Academic issues for HEIs in Oman.

C. Reliability and Validity

Content validity established by seeking help from experts to review the research instruments so that any obscure and unclear questions are revised. Internal validity established through triangulation as different sources are used to collect data that are interviews and survey in case of proposed study. Also, researcher's biasness controlled by practicing impartiality when collecting, analyzing and interpreting data. External validity ensured by applying the findings in other settings and assessing the similarities between current research context and another context; besides validity, proposed study also ensured reliability; external reliability established by describing the participants fully about the study and by explicitly explaining the methods of data collection. Internal reliability established though recording of data mechanically so that it is possible to re-analyze it if required. Also, peer examination followed by applying and utilizing findings from other researchers.

D. Data Analysis

The discussion carried out in the sections above suggest that the data collection was qualitative in nature and for this matter the data collected from a sample of 10 expert's members with regulatory authority belonging from the higher education system in Oman in the form of interview questions. The analysis of a qualitative data is crucial to researchers but on the other is equally of high value and significance too. To ensure the qualitative collection of information and extraction of meaningful data, the data analysis done based on interpretivist research methodology and thematic analysis. Thus in order to make the data analysis of the qualitative information/data extracted, a theme based structure will be designed for the 8 [see Appendix A] interview questions and based on the theme of those questions the answers generated is analyzed and evaluated. Thematic analysis maintains focus of attention on highlighting, identifying, scrutinizing and in recording themes or patterns within the data [17]. Themes are developed for each interview question and they are of significance because they help in explaining a specific phenomenon that is related with the research question. And thus in this way the themes turn into separate broad segments for analysis.

Whereas on the other hand, interpretivist approach is also used in this research paper to qualitatively analyze the data. As a matter of fact, interviews among other primary data collection methods are used in the interpretivist approach. The interpretivist approach to data analysis of the research study is subjective based study that relates with deriving meaning and understanding from the research study data. With this approach that is being used as a qualitative data analysis technique, the primary goal of the analysis is to develop and understanding instead of making predictions. Using an interpretivist approach, the information generated from the sample data is not considered to be permanent but s rather taken as an information piece that is relative to time, context, culture and the overall situation in which the research study was carried out.

IV. RESULTS AND DISCUSSION

This section deals with highlighting the key results generated from the interview questions and analyzing the answers of respondents, keeping in view the research study's subject. As mentioned in the discussion earlier that the respondent's answers and views analyzed by using interpretivist approach and the overall analysis based on a thematic analysis. For a number of obvious reasons, the subject of the research is associated with the implementation of the CBEL systems in higher educational system in Oman specifically, thus the interview questions are also designed on the similar pattern. Multiple themes are developed and used to extract meaningful answers and insights from the responses of respondents. These themes roughly revolve around the importance of CBEL for educational sector as a whole, how effective the system is in promoting flexibility and innovation in operations and how efficient the CBEL is promoting cost reduction. Table II summarizes the respondent's responses to the interview questions:

TABLE II. RESPONDENTS' RESPONSES

Qs	Respondents' responses
How effective is the implementation of (CBEL) in (HEIs) in Oman?	70 Present, Highly supported views on the fact that implementation of CBEL system has increased the effectiveness and efficiency.
What is the level of significance of ELBCC for successful undertaking of operations in HEIs?	50 Present, Shows high level of agreeableness in supporting the significance of ELBCC for successful implementation of in educational sectors in Oman.
What are the future expectations for upgrading the implementation process and operations of CBEL?	50 Present, expressed that no need to worry al all about upgrading and implementation, as this became part of SLA.
What is the significance of heavy investments made by educational sector to implement e-learning system?	50 Present, Agreeableness and supporting statements to making increased investments in promoting the implementation of CBEL application in educational platforms.
How significant is CBEL in promoting flexibility and innovation in routine operations?	50 Present, Agreeableness and supporting statements to making increased in promoting the flexibility and innovation of CBEL application in educational platforms in routine operations
How efficient is the adoption of e-learning over cloud in providing increased and convenient access to resources for students and faculty members?	50 Present, Supported on the fact that the implementation and operations of CBEL provide convenience and increased access to relevant informational data to individual learners and faculty members altogether.
How efficient is CBEL in cost reduction for educational institutions?	70 Present, Highly supported views on the cost reduction due to adopt ELBCC in HEIs in Oman.
What is the level of impact of the risks and threats associated with CBEL in Oman?	50 Present, Emphasized on the impact and significance of understanding the risks and threats that are associated with the implementation and functioning of the CBEL systems.

The first question emphasizes upon generating deeper insights from the respondents on their opinion about the effectiveness of implementation of CBEL. The responses generated from the respondents showed highly supporting views on the fact that implementation of cloud based e-learning system has increased the effectiveness and efficiency of the operations that are undertaken by the educational sectors. This directs the attention on the past researchers that have been carried out. As discussed in the earlier sections, an CBEL proves to be effective in carrying out the processes in the educational institutions [18]. The implementation of e-learning highly supports and increased the effectiveness of the operations of assessment, learning and educating the individual learners. The responses of individuals on this question also support the past researches by proving the increased effectiveness of learning and educating processes to the individual learners that are geographically dispersed.

Another key interview question is themed on understanding the rate of effectiveness of CBEL in promoting the elements of flexibility and innovation in operations that are undertaken in overall higher educational system. The respondents showed high level of agreeableness and expressed positive responses on the ability of e-learning and its implementation effects in higher education system of Oman. The positive responses from respondents direct the focus of attention on the past researches and prove that the implementation of the cloud based e-learning system in educational platforms assures the use of technological advancements and innovative measures to carry out the routine operations [19]. As mentioned in the earlier section, the use CBEL because of its increased dependency and reliance on utilization of technology has resulted in successfully attracting significant momentum. The CBEL has also proved to provide time flexibility to individual learners. The responses of respondents on this questions supports the previous researches which claims that e-learning provides the advantage of time flexibility to students that are geographically dispersed. As per the discussion carried out in the paper earlier, an asynchronous e-learning process enables a student to get involved in a learning practice as per time that suite the student. Thus e-learning implementation in educational arena helps in promoting the elements of innovation and providing flexibility of operations, especially in terms of time.

In addition to the above, the respondents also showed high level of agreeableness and gave positive answers in supporting the significance of making increased investments in implementation of CBEL systems in educational sectors in Oman [20]. As a matter of fact, the market today is focused on directing its efforts on understanding the technological changes that are taking place and also on how to make full potential use of these technological advances in order to reap the benefits of technological developments. The educational sector in Oman is flourishing and so is the population growth rate of young nationals, thus it is of high significance for the regulatory bodies (respondents) to identify and understand the rapid developments and changes in the technological demands of the educational sector and also assure the timely implementation of the technological systems. The respondents supported the fact that it is of high importance for governing bodies and regulatory authorities in the educational sector to make necessary input and promote increased investments in order to support the implementation plan related with the CBEL.

Besides this, the respondents' agreeableness and supporting statements to making increased investments in promoting the implementation of CBEL application in educational platforms also reflected on highlighting the further importance of the e-learning system in relation to distance learning. With the emerging need of distance education, there arises a need of assessment and learning too. It suggests that the concept and applications of learning and educating with the help of technological advances can now be applied to varying levels of instructional settings, without the geographical constraints [21]. In addition to this, there exist certain key elements of e-learning that are of utmost relevance and significance in distance education. The e-learning strategy to be followed is intent on by the media sources used for its delivery, available resources and the availability of time and its adequate management.

Furthermore, the respondents also emphasized on the impact and significance of understanding the risks and threats that are associated with the implementation and functioning of the CBEL systems. The responses generated from the interview questions, reflected on the fact that is highly important to not only identify the risks and threats associated with the e-learning system but on also devising prevention strategies to cater them and minimize their adverse impact [22]. The responses generated under this question direct the focus of attention on varying levels of security threats and risks that adversely affect the functionality of CBEL system in educational institutions. There exist wide arrays of security risks, such as user authentication or authorization, manipulation on part of students, confidentiality of the information data that is kept private and the threat of integrity in terms of editing or making alterations in the informational data that is kept private.

These are some of the most commonly held prominent risks and threats associated with the implementation of e-learning system. In addition to this, the previous researches also suggest the need for the adoption of certain prevention strategies to overcome or minimize the effect of these security and authenticity threats [22]. The prevention strategies include the information security mechanism using SMS, this helps in the provision of generating special passkeys to registered cellular phone numbers and helps in providing authorized access to informational data. Another prevention strategy relates with undertaking information security mechanism using biometrics, which helps in providing authorized access to informational data depending on its ability to recognize an individual based on physical and behavioral features. And lastly, the use of mechanism that is related with access control list also provides prevention measures for avoiding the issues arising due to security threats of the CBEL system.

Moreover, the respondents also supported on the fact that the implementation and operations of CBEL provide convenience and increased access to relevant informational data to individual learners and faculty members altogether. This directs the attention on the overall benefits that cloud based e-learning brings for students and faculty members despite of the geographical boundaries that exist. The CBEL application helps support students in providing them with resources that include the online access to video and audio lectures and services of an online assessing examinee to assess and mark the performance of an individual learner [23]. The access of resources for students and faculty members relates with the idea of promoting the benefits that are associated solely with the use of e-assessment, e-marking and e-submission through the implementation of CBEL systems in educational institutions.

The respondents also generated positive responses that were in support of the question related to the significance, importance and need for implementing the CBEL system. The responses reflect upon the wide range of benefits are associated with the usage of electronic marking and electronic submission. First relates with the provision of convenience, cost management and flexibility for students, thus they do not need to carry out the printing processes of their work or submitting assignments and seeking access to feedback. In addition to this, [24] states that the processes of e-submission and e-marking provides benefits to both, instructors and individual learners in keeping adequate track record of the relevant submissions and progress made. Also, authenticity of the content can also be assessed by using plagiarism software to identify any text matches. This practice helps in maintaining the academic integrity levels. Another advantage of using e-submission and e-marking is the retained access that is provided to staff specifically in order to provide further feedback, marking and accessing the files for future references. Another benefit relates with the portability element for instructors and examiners. Online submission and marking of the papers provides examiners with ease of carrying out the assessment processes with convenience. Thus, as per the data collection and gathering relevant opinions and responses from the respondents, the entire research study and the data collection process reflects on showing increased support for the need and importance for educational sectors to undertake the implementation and implication of the CBEL system.

V. Recommendations

In addition to extracting the meaningful and deeper insights of the opinions of the regulatory authorities in the higher education system of Oman and also making additional arguments in support of the literature data that is of relevance to the research study's subject, the paper also highlighted on making necessary recommendations that could positively contribute to the study subject. The recommendations revolve around maintaining focus on identifying and understanding the threats that are associated with the implementation process, functionality and overall operations of the CBEL system. The educational platforms need to follow and support the application of e-learning in educational sectors but there also lies an immense need to keep into consideration the threats and risks of security that come along with e-learning implementation and operations. Thus, it is highly recommended to educational platforms to devise necessary steps and prevention strategies in order to eliminate or minimize the impact of risk and threats.

VI. Conclusion

This study maintained its focus on understanding the implementation of the CBEL system, with an emphasis on its adoption in HEIs in Oman. As mentioned in the sections above that the data on the research study's subject was collection through carrying out qualitative interviews from the regulatory authorities in the HEIs in Oman. For this matter, the results generated from the respondents showed high level of agreeableness with the level of significance and importance that is associated with CBEL implementation in educational arena. The responses of the respondents reflected on positivity of the importance of making use e-learning systems in the educational institutions and the benefits that are generated with its application. In addition to this the data collection on the research paper also highlights the benefits that the implementation of cloud based CBEL brings for educational sector as a whole. The understanding and efficient utilization of the e-learning system has helped Oman's educational arena to go parallel with the consistent changes that have been taking place in technological, social, political and economic environments. Also, the implications of the CBEL also reflect on promoting and contributing towards the development and growth of Oman's education system and have also contributed in promoting the Oman's overall educational sector refinement. Moreover, the research paper not only contributes in the extraction of the respondent's views and limits the responses on the interview questions, but the paper it helps in generating deeper meaningful insights from the open ended interview questions and helps provide supporting arguments in support of the past relevant researches that have been carried out previously. Hence, the research paper helps in proving the importance, significance and need for educational sectors (specifically for the higher education system in Oman); to encourage decision makers for adoption of CBEL.

Acknowledgment

The authors would like to express their cordial thanks to Oman College of Management & Technology, Sultanate of Oman. We also would like to thank Universiti Malaysia Pahang (UMP) for supporting this work through Research Grant Scheme RDU180310.

References

[1] Q. Alajmi, Sadiq, S., & Alsharafi, M., "Cloud Computing delivery and Delivery Models: Opportunity and Challenges," in International Conference on Science, Engineering, Management and Social Sciences, Malaysia: university Teknologi Malaysia., 2016, pp. 15-16.

[2] Q. Alajmi and A. Sadiq, "What should be done to achieve greater use of cloud computing by higher education institutions," in Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016 IEEE 7th Annual, 2016, pp. 1-5: IEEE.

[3] A. P. Rovai, M. K. Ponton, M. J. Wighting, and J. D. Baker, "A comparative analysis of student motivation in traditional classroom and e-learning courses," International Journal on ELearning, vol. 6, no. 3, p. 413, 2007.

[4] Q. AlAjmi, R. A. Arshah, A. Kamaludin, A. S. Sadiq, and M. A. Al-Sharafi, "A Conceptual Model of E-Learning based on Cloud Computing Adoption in Higher Education Institutions."

[5] M. A. Al-Sharafi, R. A. Arshah, and E. A. Abu-Shanab, "Factors Influencing the Continuous Use of Cloud Computing Services in Organization Level," presented at the Proceedings of the International Conference on Advances in Image Processing - ICAIP 2017, Bangkok, Thailand, 25 - 27 Augst, 2017.

[6] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: state-of-the-art and research challenges," Journal of internet services and applications, vol. 1, no. 1, pp. 7-18, 2010.

[7] M. A. Al-Sharafi, R. A. Arshah, and E. A. Abu-Shanab, "Factors affecting the continuous use of cloud computing services from expert's perspective," in TENCON 2017 - 2017 IEEE Region 10 Conference, 2017, pp. 986-991.

[8] V. Arkorful and N. Abaidoo, "The role of e-learning, advantages and disadvantages of its adoption in higher education," International Journal of Instructional Technology and Distance Learning, vol. 12, no. 1, pp. 29-42, 2015.

[9] T. Sivakumaran, G. Holland, W. Wishart, K. Heyning, and B. Flowers-Gibson, "Electronic Assessment Systems: Implementation, Maintenance and Support," FOCUS on Colleges, Universities & Schools, vol. 6, no. 1, 2011.

[10] I. Shah and N. Baporikar, "The suitability of imported curricula for learning in the Gulf states: An Oman perspective," Teaching and learning in the Arab world, pp. 275-292, 2011.

[11] H. PATEL, A. PATEL, and P. SHAH, "IMPACT OF E-LEARNING IN THE DEVELOPMENT OF STUDENT LIFE," International Journal of Research in Engineering & Technology, vol. 2, no. 4, pp. 233-238, 2014.

[12] A. Fernández, D. Peralta, J. M. Benítez, and F. Herrera, "E-learning and educational data mining in cloud computing: an overview," International Journal of Learning Technology, vol. 9, no. 1, pp. 25-52, 2014.

[13] S. Hrastinski, C. Keller, and S. A. Carlsson, "Design exemplars for synchronous e-learning: A design theory approach," Computers & Education, vol. 55, no. 2, pp. 652-662, 2010.

[14] T. Ercan, "Effective use of cloud computing in educational institutions," Procedia-Social and Behavioral Sciences, vol. 2, no. 2, pp. 938-942, 2010.

[15] U. J. Bora and M. Ahmed, "E-learning using cloud computing," International Journal of Science and Modern Engineering, vol. 1, no. 2, pp. 9-12, 2013.

[16] A. Abbas, K. Bilal, L. M. Zhang, and S. U. Khan, "A cloud based health insurance plan recommendation system: A user centered approach," (in English), Future Generation Computer Systems-the International Journal of Escience, vol. 43-44, pp. 99-109, Feb 2015.

[17] G. Guest, K. M. MacQueen, and E. E. Namey, Applied thematic analysis. sage, 2011.

[18] M. A. H. Masud, M. R. Islam, and J. Abawajy, "Security Concerns and Remedy in a Cloud Based E-learning System," in International Conference on Security and Privacy in Communication Systems, 2013, pp. 356-366: Springer.

[19] M. Al-Zoube, S. A. El-Seoud, and M. F. Wyne, "Cloud computing based e-learning system," International Journal of Distance Education Technologies (IJDET), vol. 8, no. 2, pp. 58-71, 2010.

[20] W. Horton, E-learning by design. John Wiley & Sons, 2011.

[21] G. C. Moore and I. Benbasat, "Development of an instrument to measure the perceptions of adopting an information technology innovation," Information systems research, vol. 2, no. 3, pp. 192-222, 1991.

[22] N. Barik and S. Karforma, "Risks and remedies in e-learning system," arXiv preprint arXiv:1205.2711, 2012.

[23] D. R. Garrison, E-learning in the 21st century: A framework for research and practice. Taylor & Francis, 2011.

[24] N. Sultan, "Cloud computing for education: A new dawn?," International Journal of Information Management, vol. 30, no. 2, pp. 109-116, 4// 2010.

Appendix A

Interview Questions

- How effective is the implementation of Cloud-Based E-Learning (CBEL) in Higher Education Institutions (HEIs) in Oman?
- What is the level of significance and importance of e-learning for successful undertaking of operations in an educational arena?
- What are the future expectations for upgrading the implementation process and operations of CBEL?
- What is the significance of heavy investments made by educational sector to implement e-learning system?
- How significant is CBEL in promoting flexibility and innovation in routine operations?
- How efficient is the adoption of e-learning over cloud in providing increased and convenient access to resources for students and faculty members?
- How efficient is CBEL in cost reduction for educational institutions?
- What is the level of impact of the risks and threats associated with CBEL in Oman?

Design and Simulation of a Rectangular E-Shaped Microstrip Patch Antenna for RFID based Intelligent Transportation

Asif Ali
Postgraduate student in Department of Telecommunication Engineering
Mehran University of Engineering and Technology
Jamshoro, Pakistan

Muhammad Moazzam Jawaid
Assistant Professor in Department of Computer System Engineering
Mehran University of Engineering and Technology
Jamshoro, Pakistan

Nasrullah Pirzada
Assistant Professor in Department of Telecommunication Engineering
Mehran University of Engineering and Technology
Jamshoro, Pakistan

Sajjad Ali Memon
Assistant Professor in Department of Telecommunication Engineering
Mehran University of Engineering and Technology
Jamshoro, Pakistan

***Abstract*—A low profile, rectangular E-shaped microstrip patch antenna is designed and proposed for radio-frequency identification (RFID) based intelligent transportation system (ITS) in this paper. The proposed antenna design aims to achieve high gain and low return loss at 0.96 GHz as it is suitable for ultra-high frequency (UHF) RFID tags. The proposed antenna composed of a radiating patch on one side of the dielectric substrate and the ground plane on the other side, copper is used to produce the main radiator. The simulation of the proposed antenna is performed employing the high-frequency structure simulator (HFSS). The dielectric substrate used for the suggested antenna is an FR4 substrate with dielectric constant of 4.3 and height 1.5 mm. The performance of the proposed antenna is measured in terms of gain, return loss, voltage standing wave ratio (VSWR), radiation pattern and the bandwidth. The antenna gain and the return loss of the suggested antenna at 0.96 GHz are 7.3 dB and -12.43 dB, respectively.**

Keywords*—*E-shaped microstrip patch antenna; antenna gain; return loss; voltage standing wave ratio (VSWR); high-frequency structure simulator (HFSS)

I. Introduction

Presently, public transportation systems are an essential element of human lives. And it is a known fact that every year thousands of human lives are lost in the road accidents. However, road accidents can be mitigated by controlling the transportation systems intelligently using the technology, as every year millions of new vehicles hit the roads. Only in 2016, 77.31 million cars have been sold and this number has increased in 2017 by 1.28 million and counting [1]. Therefore, it is getting difficult to handle the transportation traffic using manual management systems. Thanks to the intelligent transportation systems (ITS), which proffer effectual solution for many transportation troubles as contrasted to the traditional solutions [2].

Electronic toll collection (ETC) systems are an excellent illustration of ITS applications that are employed right through the world and also brings significant advantages to the toll road users and the people at large, as contrasted to the conventional collection system for instance stop and pay at manned toll booths [3]. These ETC systems use ultra-high frequency (UHF) and radio-frequency identification (RFID) technology due to their fast reading and low-cost features. However, the effectiveness of this technology depends on the passive RFID tag reading by RFID reader antenna [4], [5]. The microstrip patch antenna is able to utilize thick substrates to meet the bandwidth demands of many potential applications [6]. We designed the E-shaped microstrip patch to evade impedance mismatches when our antenna placed at toll collection booth for reading the RFID tag [7]. Avoiding impedance mismatches direct the antenna radiate power along it's in a vehicle direction. At the higher frequency, the wide radiation beam has generated the superposition of the radiated signal and oppositely placed element on other side direction. The wide beam width is more suitable for some particular applications [8].

The organization of paper as follows: Section 2 discusses the existing work related to the antenna design for intelligent transport system based on RFID. Antenna designing configuration elaborated in Section 3. Section 4 defines the results analysis and discussion. At last, Section 5 concludes and recommends some future work.

II. Related Work

The developing of transportation industry prompts to the traffic congestion and insist on proficient intelligent transportation. The intelligent transportation system is one the best solution to alleviate these problems from the society. ITS can be portrayed as an advanced system of computing, telecommunication industry advances the security, viable traffic management, and congestion. ITS is applicable for traffic monitoring and managing, toll taxing, parking, and communication between vehicles. The motivation of the work is getting from socio-economic demands and environmental needs. One major part of ITS is antenna design with appropriate materials and shapes sizes. Though, most of the antennas are designed on an unbending substrate which limits the space to align the antenna on vehicles with effective outcomes [9]-[11] Therefore, many types of the antennas have been proposed and designed for the UHF RFID systems in the literature [12]-[16]. However, these proposed antenna designs are specific to certain applications. The microstrip antenna is one of the popular antennas that is used in many applications. Besides many advantages (such as less complex, lightweight, easy to manufacture and cheap) offered by microstrip antennas, they tend to have a larger size in the ultra high frequency (UHF) range [17], [18]. This paper proposed an antenna design for RFID reader that can be employed at the toll plazas to read the passive RFID tags efficiently. The proposed antenna is compact and low profile, rectangular E-shaped microstrip patch that aims to achieve high gain and low return loss at 0.96 GHz. The RFID tag comprises an antenna and semiconductor chip containing some information related the vehicle identification. The chip transmits the radio signal through RFID reader, the tag need the active power support all the time to read the information [19]. Unfortunately, authors fail to activate the RFID reader all the time. The innovative expertise will help to find the fresh solutions to the existing system problems like traffic congestion, vehicle identification, monitoring of observance to traffic rules and aware about the accident for instant management [11]. In this paper, the proposed antenna, wide axial ratio bandwidth is achieved by using coupling feeding mechanism. It is observed that the back lobes are dominant to reduce the antenna gain and performance [20]. The phased array radar with imaging system used here, the millimeter wave antenna is employed for the short-range communication of the application of the vehicle [21]. The literature review presents some limited range of applications of the antenna for ITS in tolling. It is noted that the existing work has discussed the antennas for ITS with larger in size and provided results are not up to mark can be utilized for tolling.

The aim of this paper is more especially to design a compact size antenna with better and efficient results. We have designed the E-shaped microstrip patch antenna for the intelligent transportation systems application.

III. Proposed Antenna Design

Transmission-line model is used for analyzing the suggested E-shaped microstrip patch antenna design [22]. The design of the suggested E-shaped microstrip patch antenna is illustrated in Fig. 1, where W_p and L_p represents the patch width and patch length respectively, therefore, the size of the metallic radiating patch of the proposed antenna is $W_p \times L_p$. Furthermore, W_s, L_s and h represents the width, length, and height of dielectric substrate respectively, therefore, the size of the dielectric substrate of the suggested antenna is $W_s \times L_s \times h$. Moreover, W_f, L_f, g and y_o represents the feed line width, feedline length, inset gap and inset length, respectively. The proposed antenna operating frequency f_o, the dielectric constant of the substrate $\in_r$ and the height (thickness) of the dielectric substrate are 960 MHZ, 4.3 and 1.5 mm respectively. As the proposed antenna is analyzed using transmission-line model, so W_p and L_p can be calculated using following equations, respectively [5].

$$W_p = \frac{c}{2f_o}\left(\frac{2}{\in_r + 1}\right)^{\frac{1}{2}} \tag{1}$$

$$L_p = \frac{1}{2f_o\sqrt{\in_{reff}\ \mu_o \in_o}} - 2\Delta L \tag{2}$$

Where c is the speed of the light, $\in_{reff}$ is the effective dielectric constant, μ_o is permeability of the free space and ΔL is extension length. The extension length ΔL and the effective dielectric constant $\in_{reff}$ can be calculated using following equations, respectively.

$$\Delta L = 0.412h\left[\left(\frac{\in_{reff} + 0.3}{\in_{reff} - 0.258}\right)\left(\frac{\frac{W_p}{h} + 0.264}{\frac{W_p}{h} + 0.813}\right)\right] \tag{3}$$

$$\in_{reff} = \frac{\in_r + 1}{2} + \frac{\in_r - 1}{2}\left(1 + \frac{12h}{W_p}\right)^{\frac{-1}{2}} \tag{4}$$

The proposed E-shaped microstrip patch antenna is simulated via high-frequency structure simulator (HFSS) software and the FR4 substrate is used as the dielectric material having a loss tangent of 0.025. The suggested E-shaped microstrip patch antenna is fed by the input impedance Z_o of 50 Ω.

IV. Methodology

As mentioned in Section III that the proposed antenna is simulated in HFSS, for that we need to design the proposed antenna with the specific values in the said simulator. The specific values can be referred as simulation parameters and some of these values can be calculated from the equations discussed in Section III. After calculating the specific values, the design of the proposed antenna can be modeled in the HFSS simulator that may look like as shown in Fig. 2.

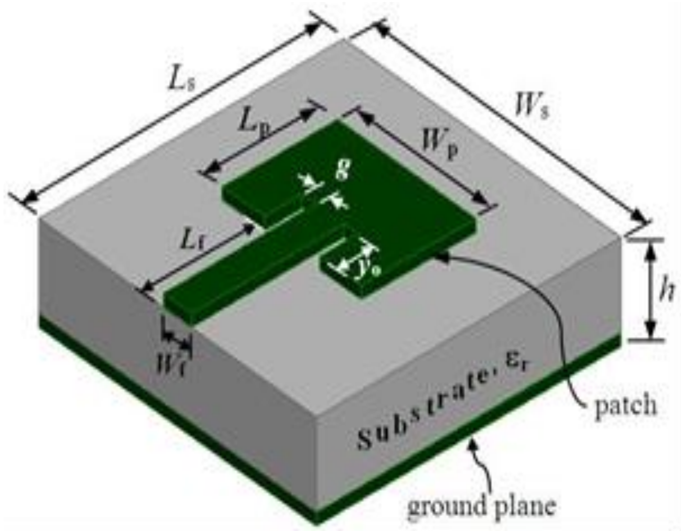


Fig. 1. Proposed design of the E-shaped microstrip patch antenna.

V. Simulation Results and Discussion

The proposed E-shaped microstrip patch antenna is simulated in HFSS simulator using simulation parameters listed in Table I. The configuration of the suggested antenna in HFSS simulator is illustrated in Fig. 2.

TABLE I. SIMULATION PARAMETERS USED IN HFSS SIMULATION

Parameter	**Symbol**	**Value**
The dielectric constant of the substrate	$\in_r$	4.3
Inset Gap	g	3 mm
Inset length	y_o	10 mm
Substrate Thickness	h	1.5 mm
Operating frequency	f_o	0.96 GHz
Patch Dimension Along x	W_p	39 mm
Patch Dimension Along y	L_p	30 mm
Feed Width	W_f	5 mm
Feed Length	L_f	20 mm
Substrate Dimension Along x	W_s	65 mm
Substrate Dimension Along y	L_s	56 mm
Input impedance	Z_o	50 Ω

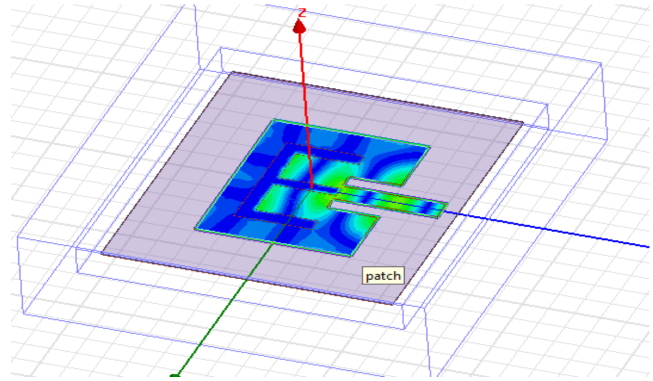


Fig. 2. Proposed E-shaped microstrip antenna in HFSS simulator.

The antenna is characterized by its parameters mentioned already in the previous section. Simulation outcomes are presented in this segment. The simulated return loss of the suggested microstrip antenna is illustrated in Fig. 3. The proposed antenna has a minimum value of -12.38 dB of return loss obtained at 0.96 GHz. Fig. 4 illustrates the 3D gain of the suggested microstrip antenna where the peak gain of 7.3 dB is achieved at the frequency of 0.96 GHz. The directivity of the antenna is in the X-axis direction and the gain is 5.5 dB. The directivity of the suggested E-shaped microstrip patch antenna is illustrated in Fig. 5. The values obtained from the simulation are suitable for the ITS application. Fig. 6 illustrates the radiation pattern the of suggested microstrip patch antenna. It shows the horizontal and vertical half power-beam width. Voltage standing wave ratio (VSWR) value for the suggested microstrip antenna is found to be 0.67, and the simulated input impedance is 48.5 **Ω**, which is not much less than the set input impedance. That translates that the proposed microstrip antenna has low losses.

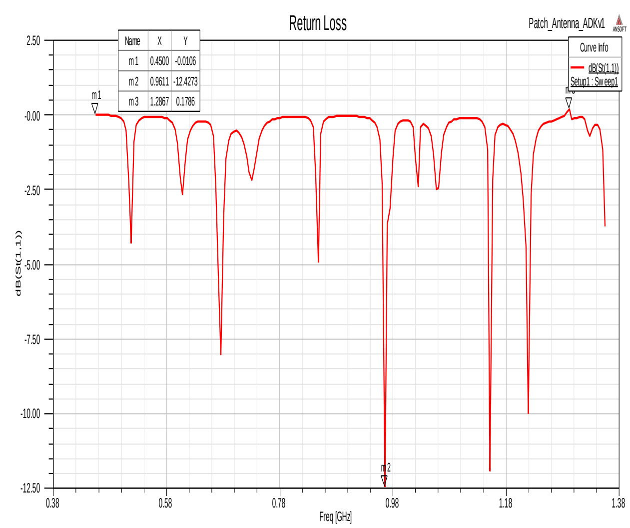


Fig. 3. Proposed antenna simulated return loss.

TABLE II. COMPARING THE RESULTS OF DIFFERENT WORKS

Papers	**Operating frequency**	**Size**	**Return loss**	**Gain in dB**	**VSWR**	**Impedance**
Our paper	0.96 GHz	65X56mm	-12.38dB	7.3 dB	0.67	48.5Ω
[8]	15GHz	6.4X8.5mm	-10dB	9.6dB	0.57	50 Ω
[6]	2.25GHz	55.1X49mm	-12.5dB	9.76dB	0.98	50 Ω
[20]	2.49GHz	30X30mm	-15dB	4.6dB	1.8	49 Ω

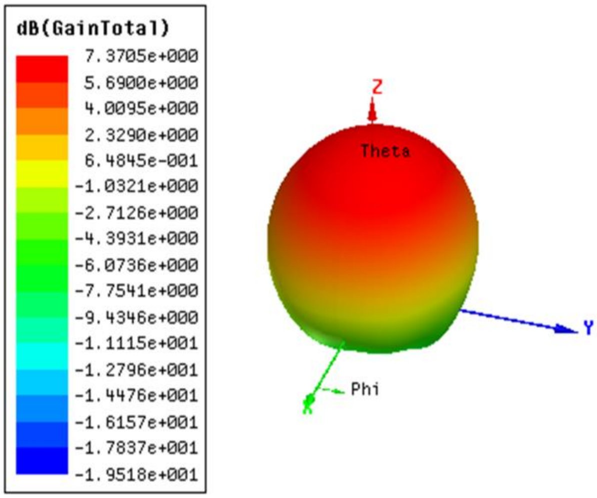


Fig. 4. Proposed antenna simulated gain.

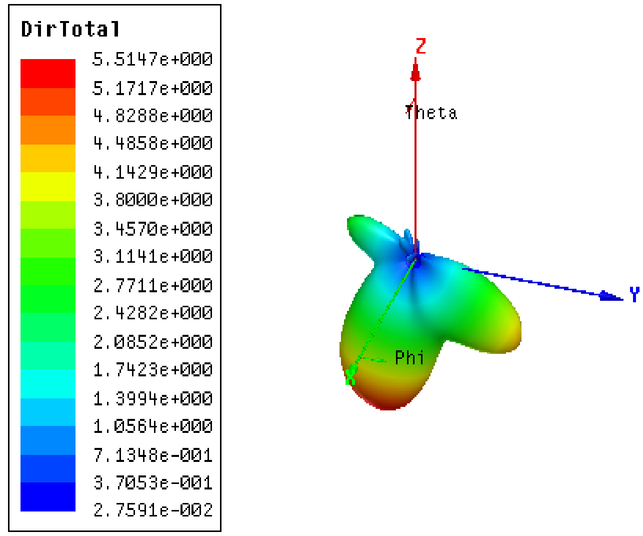


Fig. 5. Proposed antenna simulated directivity.

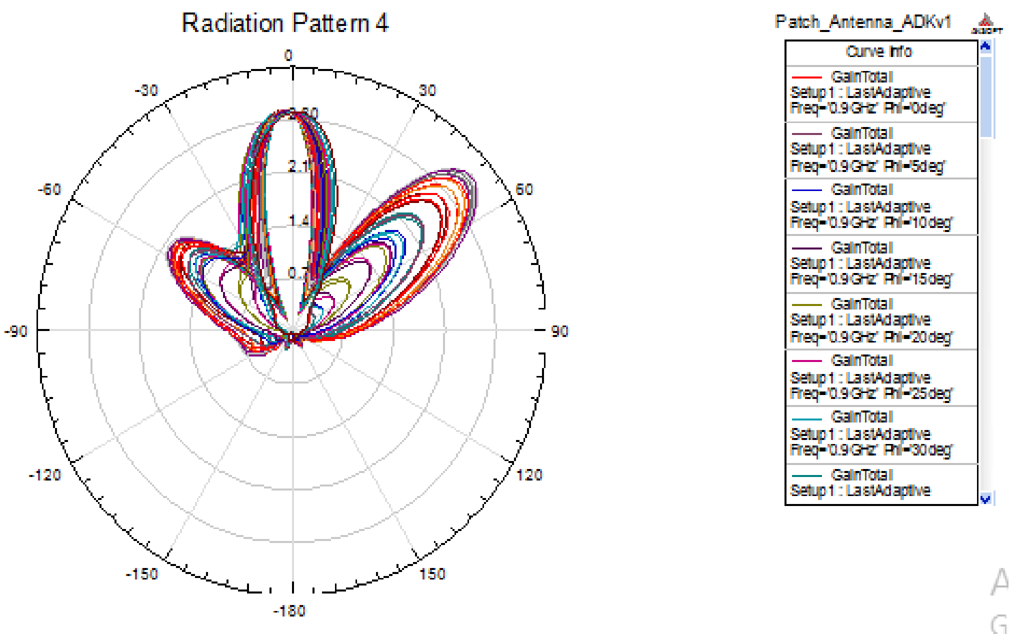


Fig. 6. Proposed antenna simulated radiation pattern.

Table II compares the results of the proposed antennas with other works [8], [6], [20]. As the operating frequency increases the antenna size will decrease as can be seen from the table. Therefore, the antenna size of the proposed antenna is relatively small compared to [8], [6] and [20] as the operating frequency of the proposed antenna is comparatively higher than the other works. Similarly, the return loss of the proposed antenna is better than the [6] and [20], while slightly higher than [8]. The gain of an antenna will increase as the frequency increase [20]. Therefore, the gain of the proposed antenna is relatively higher than [8], [6] and [20]. Similarly, the VSWR and impedance matching of the proposed antenna is better than the [8], [6] and [20] as seen in Table II.

VI. Conclusion

In this paper, the proposed antenna of 65×56 mm^2 size simulated in HFSS v13 on FR4 substrate 4.3 dielectric constants with operating frequency of 0.96GHz is discussed with accomplishing the antenna suite for intelligent transport systems. This proposed antenna suitably improves the antenna with high gain and directivity. In vehicles, antenna orientation is not dependent on anything, the radiation pattern of E-Slotted patch antenna provides omnidirectional pattern can accommodate the tolling traffic in an effective manner. This model yields the effective results in terms of gain 7.3 dB; directivity 5.5 dB with better impedance matching, VSWR of 0.67 was achieved. The simulated radiation efficiency of 70%

along with impedance bandwidth of 33MHz has been revealed. This antenna can be investigated on a different range of wireless applications and vehicular applications. It is observed that due to its higher beam width at 70^o with a circularly polarized response. The benefit of the substrate of low dielectric is that it is favored for maximum radiation. For the reason that of directivity and wide radiation pattern, the rectangular patch antenna is preferred. Proposed configurations for E-shaped microstrip patch antenna meet specifications for highway toll collection or vehicle fleet identification applications.

References

[1] A. J. Al-Dweik, M. Mayhew, R. Muresan, S. M. Ali, and A. Shami, “Using Technology to Make Roads Safer: Adaptive Speed Limits for an Intelligent Transportation System,” IEEE Veh. Technol. Mag., vol. 12, no. 1, pp. 39–47, 2017.

[2] K. Sjoberg, P. Andres, T. Buburuzan, and A. Brakemeier, “Cooperative Intelligent Transport Systems in Europe: Current Deployment Status and Outlook,” IEEE Veh. Technol. Mag., vol. 12, no. 2, pp. 89–97, 2017.

[3] J. Odeck and M. Welde, “Economic evaluation of intelligent transportation systems strategies: the case of the Oslo toll cordon,” IET Intell. Transp. Syst., vol. 4, no. 3, p. 221, 2010.

[4] Y. Yao, Y. Liang, J. Yu, and X. Chen, “Design of a multi-polarized RFID reader antenna for UHF near-field applications,” IEEE Trans. Antennas Propag., vol. 65, no. 7, pp. 3344–3351, 2017.

[5] A. Farswan, A. K. Gautam, B. K. Kanaujia, and K. Rambabu, “Design of Koch Fractal Circularly Polarized Antenna for Handheld UHF RFID Reader Applications,” IEEE Trans. Antennas Propag., vol. 64, no. 2, pp. 771–775, 2016.

[6] J. M. Kovitz, H. Rajagopalan, and Y. Rahmat-Samii, “Circularly polarised half E-shaped patch antenna : a compact and fabrication-friendly design,” vol. 10, pp. 932–938, 2016.

[7] C. Qi, P. Hillyard, A. Al-Hussein, N. Patwari, G. D. Durgin, and A. Tagless, “On-Wall, Wide Bandwidth E-Shaped Patch Antenna for Improved Whole-Home Radio Tomography,” vol. 1, no. 1, pp. 22–31, 2017.

[8] H. Jin, K. Chin, W. Che, S. Member, and C. Chang, “A Broadband Patch Antenna Array With Planar Differential L-Shaped Feeding Structures,” vol. 14, pp. 127–130, 2015.

[9] V. Van Yem et al., “Novel High Gain and Broadband CPW-Fed Antennas With EBG for ITS Applications,” pp. 451–456, 2013.

[10] T. Mondal, S. Samanta, R. Ghatak, and S. R. Bhadra Chaudhuri, "A novel circularly polarized dsrc band square microstrip antenna using the minkowski fractal structure for vehicular communication," 2014 IEEE International Conference on Vehicular Electronics and Safety, Hyderabad, 2014, pp. 141-146.

[11] D. A. Jagtap and V. U. Deshmukh, “Review Paper on Design of a Coplanar Integrated Microstrip Antenna for ITS Applications,” pp. 1261–1264, 2017.

[12] A. Larionov, R. Ivanov, and V. Vishnevsky, “UHF RFID in Automatic Vehicle Identification: Analysis and Simulation,” IEEE J. Radio Freq. Identif., vol. 1, no. 1, pp. 1–1, 2017.

[13] H. Salah, H. A. Ahmed, J. Robert, and A. Heuberger, “Multi-Antenna UHF RFID Reader Utilizing Stimulated Rate Tolerance,” vol. 1, no. 2, pp. 124–134, 2017.

[14] P. Forouzannezhad, A. Jafargholi, and A. Jahanbakhshi, “Multiband compact antenna for near-field and far-field RFID and wireless portable applications,” IET Microwaves, Antennas Propag., vol. 11, no. 4, pp. 535–541, 2017.

[15] S. Maddio, “A Compact Circularly Polarized Antenna for 5 . 8 GHz Intelligent Transportation System,” vol. 1225, no. c, pp. 8–11, 2016.

[16] R. Caso, A. Michel, M. Rodriguez-Pino, and P. Nepa, “Dual-band UHF-RFID/WLAN circularly polarized antenna for portable RFID readers,” IEEE Trans. Antennas Propag., vol. 62, no. 5, pp. 2822–2826, 2014.

[17] K. Fan, Z. Hao, S. Member, and Q. Yuan, “A Low-Profile Wideband Substrate-Integrated Antenna for the Q-LINKPAN Applications,” vol. 65, no. 11, pp. 5667–5676, 2017.

[18] Dahri, F. A, Soomro R. A, Memon Z. A “Design of Wearable Microstrip Yagi Array Antenna aimed for Telemedicine Applications”, Academic Journal of Management Science (AJMS), Vol. 5, No.2, pp 45-52, 2017.

[19] I. Tabakh, M. Jorio, N. El Amrani El Idrissi, and T. Mazri, "Design and simulation of an optimized slotted patch antenna for RFID reader applications," 2016 International Conference on Information Technology for Organizations Development (IT4OD), Fez, 2016, pp. 1-6.

[20] T. Mondal, S. Maity, R. Ghatak, S. Ranjan, and B. Chaudhuri, “Compact Circularly Polarized Wide Beam Width Fern Fractal Shaped Microstrip Antenna for Vehicular Communication,” vol. xx, no. xx, pp. 1–9, 2018.

[21] N. Kishore, G. Upadhyay, A. Prakash, and V. S. Tripathi, “Millimeter Wave Antenna for Intelligent Transportation Systems Application,” vol. 17, no. 1, pp. 171–178, 2018.

[22] H. Pues and a. van de Capelle, “Accurate transmission-line model for the rectangular microstrip antenna,” IEE Proc. H Microwaves, Opt. Antennas, vol. 131, no. June, p. 334, 1984.

Method for Thermal Pain Level Prediction with Eye Motion using SVM

Kohei Arai
Department of Information Science
Saga University
Saga City, Japan

***Abstract*—Method for thermal pain level prediction with eye motion using SVM is proposed. Through experiments, it is found that thermal pain level is much sensitive to the change rate of pupil size rather than pupil size itself. Also, it is found that the number of blinks shows better classification performance than the other features. Furthermore, the eye size is not a good indicator for thermal pain. Moreover, it is also found that user respond to the thermal stimulus so quickly (0 to 3 sec.) while the thermal pain is remaining for a while (4 to 17 sec.) after the thermal stimulus is removed.**

***Keywords*—Eye motion; thermal pain; support vector machine; thermal stimulus; classification**

I. Introduction

The locus coeruleus (LC) signals salience to sensory stimuli and these responses can modulate the experience of pain stimuli. The pupil dilation response (PDR) to noxious stimuli is thought to be a surrogate for LC responses, but PDR response to Peltier-controlled noxious heat stimuli, the most commonly used method in experimental pain research, has not been described [1].

A bedridden person is a person who has become a plant state, a person who cannot speak words well. Such people are difficult to convey their feelings. Currently, researches on gaze input, etc. are being conducted as a way for people who feel difficult to convey feelings to emotion. However, there are times when I have something I want to convey without having time to spare time of line of sight. Therefore, there is a strong demand to create methods and systems which allow pain level prediction and representation for disable persons.

There are not so small number of methods of thermal pain level prediction, such as MEG, MRI, fMRI, EOG, EEG, etc. These, however, are not so easy to measure and prediction of thermal pain due to the fact that it is difficult to recognize thermal pain with measured data and also it does cost for the measuring instruments. Also, these are not accurate enough (around 75 % of accuracy with some time delay) [1].

Pupillary responses to thermal pain stimulation in healthy volunteers (Lauren A. Banker)[1]. Pupil responses and pain ratings to heat stimuli: Reliability and effects of expectations and a conditioning pain stimulus (James C. Eisenach et al.) [1]. Although these studies are based on only the pupillary response, in this research I add elements of other eye motion. Towards a physiology-based measure of pain: patterns of human brain activity distinguish painful from non-painful thermal stimulation (Justin E. Brown et al.[2]) [2]. Although this study is a study of brain activity by fMRI data using classification method of Support Vector Machine: SVM, this research is a study using eye motion data using SVM.

With regard to the psychological status monitoring with eye motion, rescue system for elderly and disabled persons using wearable physical and psychological monitoring system is proposed [3]. Also, a method for psychological status estimation by gaze location monitoring using Eye-Based Human-Computer Interaction: EBHCI is proposed and validated [4]. On the other hand, a method for psychological status monitoring with line of sight vector changes (Human eyes movements) detected with wearing glass is developed [5].

Frequent physical health monitoring as vital sign with psychological status monitoring for search and rescue of handicapped, disabled and elderly persons is developed [6]. Meanwhile, psychological status monitoring with cerebral blood flow: CBF, Electroencephalogram: EEG and Electro-Oculogram: EOG measurements is proposed and validated [7]. Relations between psychological status and eye movements are investigated [8] together with psychological status monitoring with cerebral blood flow, electroencephalogram and electro-oculogram measurements [9]. Then, rescue system with sensor network for physical and psychological health monitoring is summarized [10].

In this paper, a new approach for thermal pain level prediction is proposed. Namely, a method for thermal pain level prediction with human eye motions, in particular, with pupil size change rate using SVM is proposed. There are previously proposed methods which use pupil size [11]. It is found that thermal pain level is much sensitive to pupil size change rate than pupil size. Such this sensitivity analysis is conducted.

The next section describes the proposed method for prediction of thermal pain level followed by experimental method and results. Then, the results are discussed followed by conclusion. Finally, future research work is described.

[1] https://steinhardt.nyu.edu/appsych/psych_cafe

[2] http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0024124

II. Method for Thermal Pain Prediction with Eye Motion using SVM

A. System Configuration

Near Infrared: NIR camera which is mounted on a glass is used for acquisition of eye images. Fig. 1 shows the outlook of the proposed system.

Fig. 1. Outlook of the system configuration.

NIR camera imagery data is acquired through USB interface to PC. The outlook of the camera is shown in Fig. 2 while major specification of the camera is shown in Table I.

Fig. 2. Outlook of the NIR camera used.

TABLE I. Major Specification of the NIR Camera Used

Weight	200 g
Size (dimension)	21.8 x 14.8 x 7.6 cm
Camera Name	DC-NCR13U

B. Thermal Pain Level Prediction Method

From the acquired eye image 1) eye size, eyelid width; 2) pupil size; 3) the number of blink for a time interval (blink frequency); 4) pupil size change rate; 5) pupil center location changes; and 6) pupil center location change rate are selected as feature components for thermal pain level prediction. Using relations between these features and subjective evaluation results of thermal pain, thermal pain is evaluated with the features through regressive analysis.

C. SVM

SVM is well known classification method which allows classify the data into the previously designated class categories. In the proposed method, two class categories, pain and not pain are designated. By using the feature components, SVM based classification is applied.

III. Experiments

A. Experimental Configuration

Fig. 3 shows the experimental configuration which consists of the hot water pot with temperature measuring instrument and NIR camera mounted glass.

Fig. 3. Experimental configuration.

B. Experimental Configuration

Without any thermal stimulus, 50 seconds of eye images are acquired with 30 frame/second. After that, 10 seconds of eye images are acquired with 40 degree Celsius of the pot attached to the user’s arm. These processes are repeated for six times, 600 seconds of eye images are acquired in total as shown in Fig. 4.

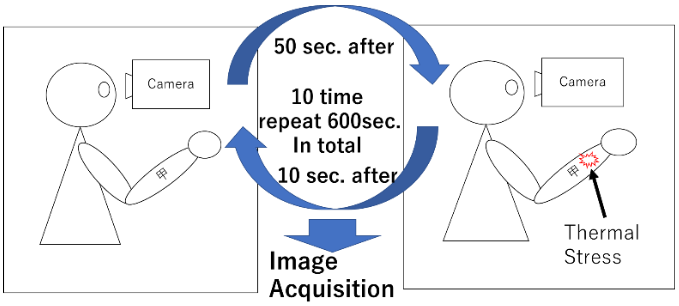
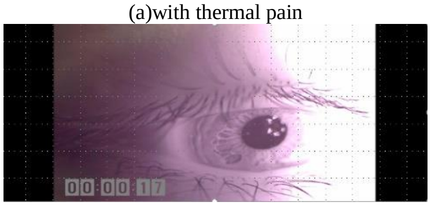


Fig. 4. Procedure of the experiment.

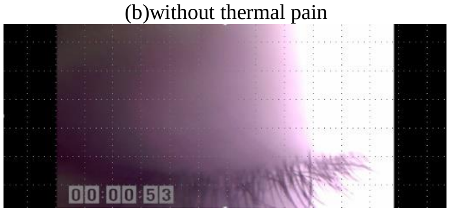
C. Preliminary Results

Fig. 5(a) shows eye images when the user is having thermal stimulus while Fig. 5(b) shows those as the user is not having any thermal stimulus. It is quite obvious that pupil size as for having thermal stimulus is larger than that for having no thermal pain. Also, Fig. 5(c) shows eye image when the user is blinking.

(a)with thermal pain



(b)without thermal pain

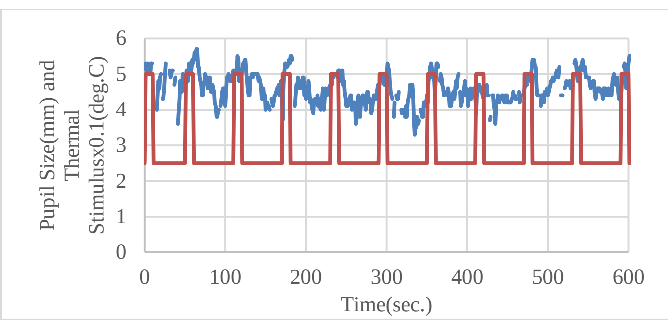


(c)Blinking

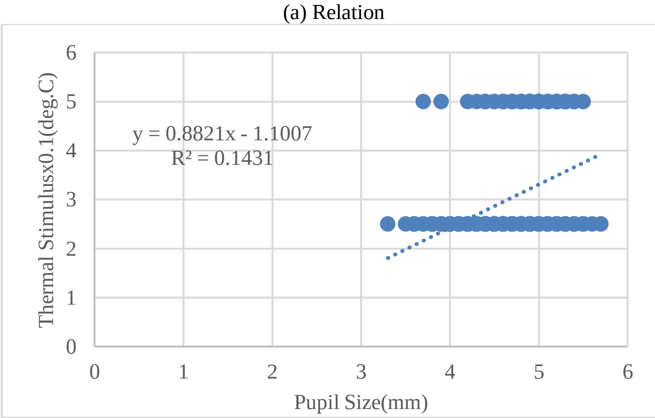
Fig. 5. Examples of eye images when user is having thermal pain and no thermal pain as well as blinking.

D. Pupil Size

Fig. 6(a) shows relation between thermal stimulus and pupil size. Meanwhile, Fig. 6(b) shows the result from linear regressive analysis between thermal stimulus and pupil size.



(a) Relation



(b) Regressive analysis

Fig. 6. Relation between thermal stimulus and pupil size.

Although there is definite relation between thermal stimulus and pupil size, correlation coefficient between both is not so high, around 0.378. On the other hand, response time of pupil size changes against thermal stimulus is shown in Fig. 7. Leading edges of the response are so quick in comparison to the trading edges. Namely, user respond to the thermal stimulus so quickly (0 to 3 sec.) while the thermal pain is remaining for a while (4 to 17 sec.) after the thermal stimulus is removed.

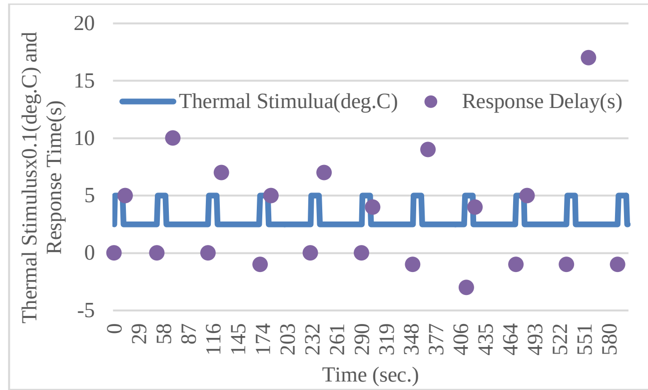
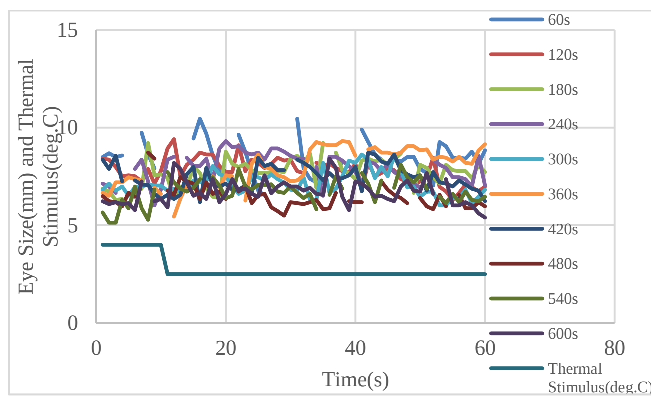


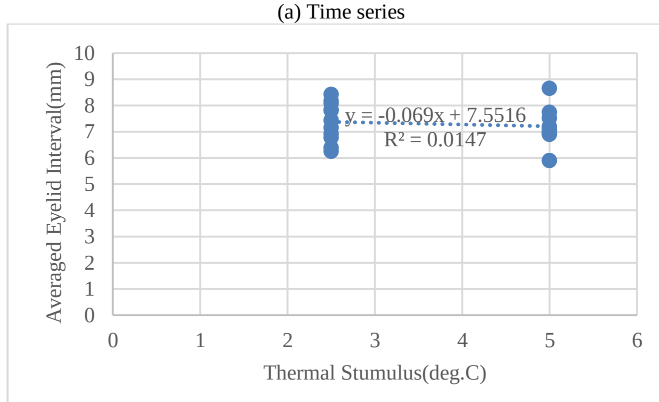
Fig. 7. Response time of pupil size changes against thermal stimulus.

E. Eye Size

Eye size is defined as the distance between the upper and lower eyelids. Usually, human eye size is getting large when users are surprising. Therefore, it is considered that eye size is one of the indicators for thermal pain. Fig. 8(a) shows the time series of eye size data against thermal stimulus while Fig. 8(b) shows the relation between both.



(a) Time series



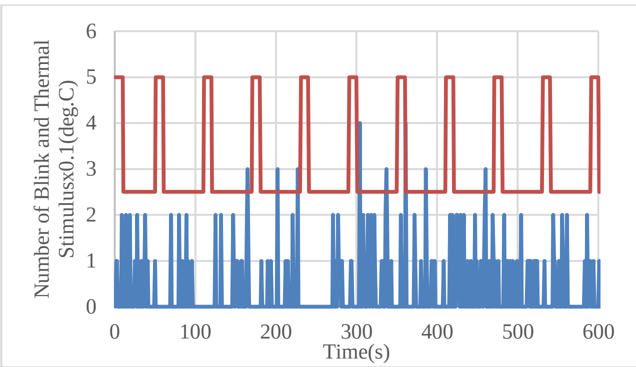
(b) Relation

Fig. 8. Relation between eye size and thermal stimulus.

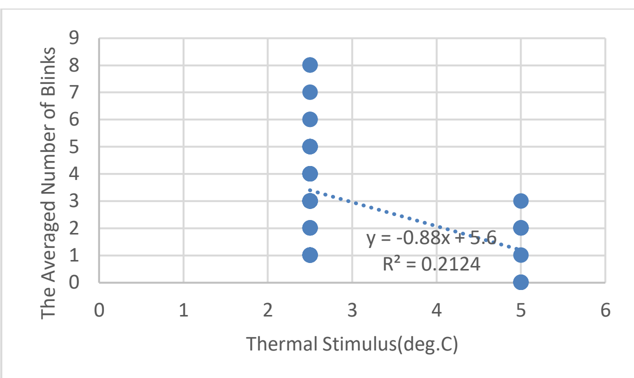
As shown in Fig. 8, relation between eye size and thermal stimulus is not high (correlation coefficient is around 0.1212). Therefore, eye size is not changed so much for thermal stimulus and no thermal stimulus period of time.

F. *Blink*

The number of blinks during thermal stimulus and no thermal stimulus are applied to user is measured. Fig. 9(a) shows the time series of the number of blinks data while Fig. 9(b) shows the result from linear regressive analysis between the number of blinks and thermal stimulus. Also, Table II shows the number of blinks for thermal stimulus and no thermal stimulus period of time.



(a) Time series of data



(b) Linear regression

Fig. 9. Relation between thermal stimulus and the number of blinks.

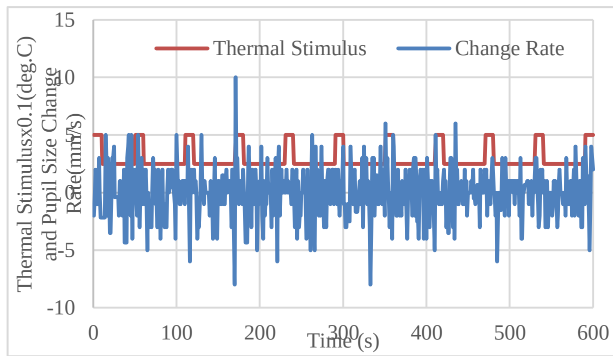
As shown in Fig. 9(b), the relation between the number of blinks and thermal stimulus is not high (correlation coefficient is around 0.113). Therefore, the number of blinks is not changed so much for thermal stimulus and no thermal stimulus period of time.

G. *Pupil Size Change Rate*

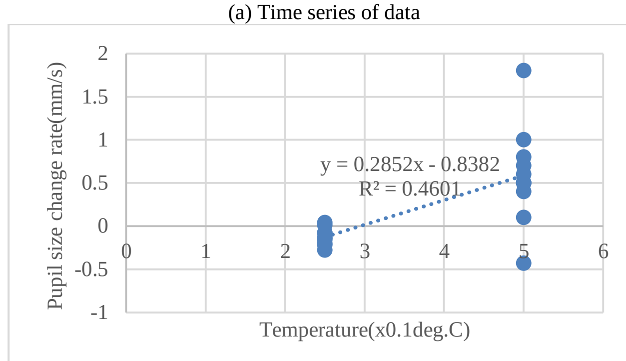
Meanwhile, pupil size change rate is investigated with the pupil size measured data. Human eye pupil size maybe rapidly changes due to thermal stimulus while pupil size is changed slowly for no thermal stimulus. Fig. 10(a) shows time series of pupil size change rate data while Fig. 10(b) shows the result from linear regression between thermal stimulus and the pupil size change rate. The correlation coefficient between both is approximately 0.678. Therefore, the pupil size change rate can be a good indicator of the thermal pain.

TABLE II. THE NUMBER OF BLINKS

Time Duration(s)	The Number of Blink/min.
1～60	35
61～120	16
121～180	17
181～240	6
241～300	6
301～360	8
361～420	4
421～480	2
481～540	5
541～600	7



(a) Time series of data



(b) Linear regression

Fig. 10. Relation between thermal stimulus and pupil size change rate.

H. *Pupil Center Location and Its Change Rate*

It is used to be occurred that pupil center location is changed when users feel some strange actions including thermal stimulus. Therefore, it may be possible to use these two features, pupil center location change and its change rate for detection of thermal pain. Fig. 11 shows time series of the features, pupil center location change and its change rate. In the figure, short and long radius of the ellipsoid of pupil shape, pupil center location of x and y coordinates, as well as short radius with 0 data and the difference between the short radius and short radius with 0 data are shown.

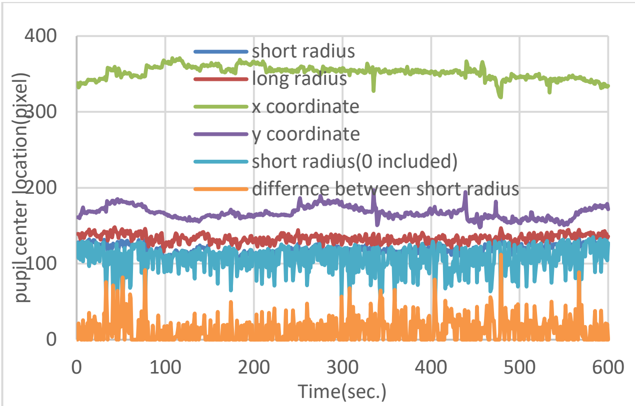
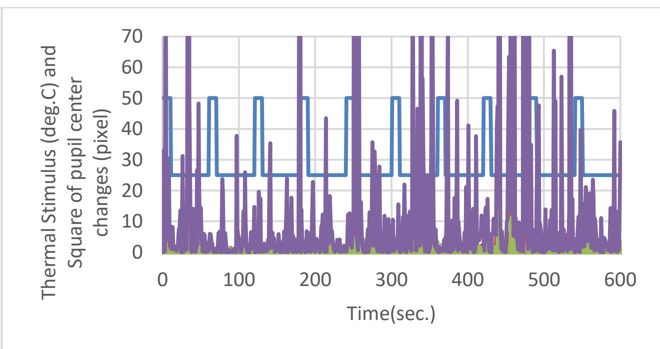
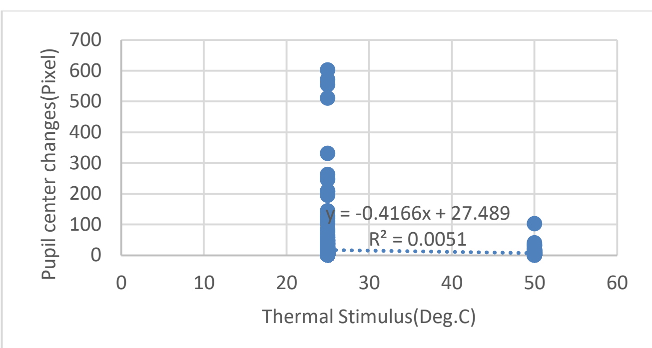


Fig. 11. Time series of features of short and long radius of the ellipsoid of pupil shape, pupil center location of x and y coordinates, as well as short radius with 0 data and the difference between the short radius and short radius with 0 data.

From the features, square of the pupil center location changes in unit of pixel and its change rate are calculated. Fig. 12(a) shows time series of pupil center location changes and thermal stimulus and Fig. 12(b) also shows the relation between thermal stimulus and the pupil center location changes.



(a) Time series



(b) Relation

Fig. 12. Time series of the pupil center location changes and thermal stimulus as well as its relation between both.

Correlation between both is quite low, 0.0051. Therefore, it is difficult to get a good classification performance with SVM. On the other hand, correlation coefficient between pupil center location change rate in unit of pixel/sec. and thermal stimulus is around 0.5 as shown in Fig. 13. Therefore, it is better than the feature of pupil center location changes and may be possible to get a good classification performance.

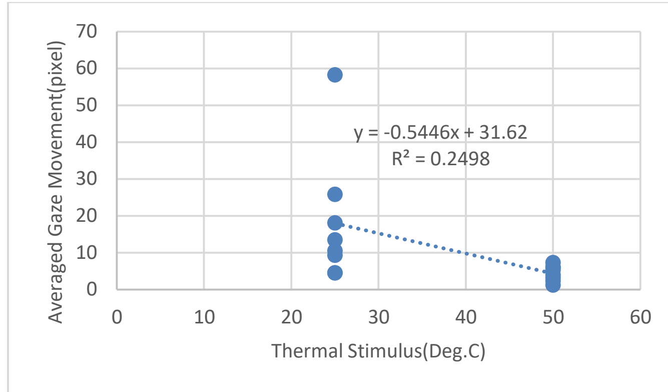


Fig. 13. Relation between the pupil center location change rate and thermal stimulus.

In summary, although the conventional method uses the feature of pupil size, pupil size change rate is much better than pupil size for the reason that correlation coefficients between thermal stimulus and the above two features are quite different.

I. Classification Performance

SVM based classification performances with the aforementioned four features, pupil size (P), the number of blinks (B), eye size (E) and pupil size change rate (S) and their possible combinations are evaluated. Table III(a) shows classification performance with four single features while Table III(b) shows that with six two feature combination. Meanwhile, Table III(c) shows the classification performance with five possible three feature combinations.

TABLE III. SVM BASED CLASSIFICATION PERFORMANCE (PAIN AND NO-PAIN)

(a) Single feature

P	B	E	S
60.97%	75.82%	51.3%	62.46%

(b) Two feature combination

P+B	P+E	P+S	B+E	B+S	E+S
73.71%	57.92%	59.34%	70.88%	73.82%	50%

(c) Three feature combination

P+B+E	P+B+S	P+E+S	B+E+S	P+B+E+S
69.86%	74.02%	56.94%	69.95%	67.48%

It is concluded that the number of blinks shows the best classification performance followed by the three feature combination among pupil size, the number of blinks and pupil size change rate.

IV. CONCLUSION

Method for thermal pain level prediction with human eye motion is proposed. Through experiments, it is found that thermal pain level is much sensitive to the change rate of pupil size rather than pupil size itself. Moreover, the number of blinks shows better classification performance than the other features.

Also, it is found that the number of blinks and the eye size are not good indicator for thermal pain. Furthermore, it is also found that user respond to the thermal stimulus so quickly (0 to 3 sec.) while the thermal pain is remaining for a while (4 to 17 sec.) after the thermal stimulus is removed.

V. Future Study

Further study is required for more experimental data with a variety of examiners. In order to validate the proposed method, more experiments with a variety of examiners are highly required.

Acknowledgment

The author would like to tank Mr. Takuya Kumasaki of Saga University for his effort to conduct the experiment.

References

[1] James C. Eisenach, Regina Curry, Carol A. Aschenbrenner, Robert C. Coghill Timothy, T.Houle, Pupil responses and pain ratings to heat stimuli: Reliability and effects of expectations and a conditioning pain stimulus, Journal of Neuroscience Method, 279, 1, 52-59, 2017.

[2] Justin E. Brown, Neil Chatterjee, Jarred Younger, Sean Mackey, Towards a Physiology-Based Measure of Pain: Patterns of Human Brain Activity Distinguish Painful from Non-Painful Thermal Stimulation, PLoS ONE 6(9): e24124, 2011.

[3] J. Kaprzy Edit., Kohei Arai, Rescue System for Elderly and Disabled Persons Using Wearable Physical and Psychological Monitoring System, Studies in Computer Intelligence, 542, 45-64, Springer Publishing Co. Ltd., 2014.

[4]

[5] Kohei Arai, Ronny Mardiaynto, Method for psychological status estimation by gaze location monitoring using eye-based Human-Computer Interaction, International Journal of Advanced Computer Science and Applications, 4, 3, 199-206, 2013.

[6] Kohei Arai, Kiyoshi Hasegawa, Method for psychological status monitoring with line of sight vector changes (Human eyes movements) detected with wearing glass, International Journal of Advanced Research in Artificial Intelligence, 2, 6, 65-70, 2013.

[7] Kohei Arai, Frequent physical health monitoring as vital sign with psychological status monitoring for search and rescue of handicapped, disabled and elderly persons, International Journal of Advanced Research in Artificial Intelligence, 2, 11, 25-31, 2013.

[8] Kohei Arai, Psychological status monitoring with cerebral blood flow: CBF, Electroencephalogram: EEG and Electro-Oculogram: EOG measurements, International Journal of Advanced Research on Artificial Intelligence, 4, 6, 8-15, 2015.

[9] Kohei Arai, Relations between psychological status and eye movements, International Journal of Advanced Research on Artificial Intelligence, 4, 6, 16-22, 2015.

[10] Kohei Arai, Psychological status monitoring with cerebral blood flow, electroencephalogram and electrooculogram measurements, Journal of Information Technology and Computer Science, 7, 8, 1-7, 2015

[11] Kohei Arai, Rescue system with sensor network for physical and psychological health monitoring, Proceedings of the International Seminar on Intelligent Technology and Its Applications: ISITIA 2015, (Keynote Speech), 2015.

[12] D. Alnaes, M.H. Sneve, T. Espeseth, T. Endestad, S.H. van de Pavert, B. Laeng, Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus, Journal of Visual., 14, 1-20, 2014.

Authors' Profile

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

Recommendation using Rule based Implicative Rating Measure

Lan Phuong Phan
College of Information and Communication Technology
Can Tho University, Can Tho, Viet Nam

Hung Huu Huynh
University of Science and Technology, University of Da Nang
Da Nang, Viet Nam

Hiep Xuan Huynh
College of Information and Communication Technology
Can Tho University, Can Tho, Viet Nam

***Abstract*—The paper presents a rule based implicative rating measure to calculate the ratings of users on items. The paper also presents a new model using the ruleset with the rule length of 2 and the proposed measure to suggest to users the list of items with the highest ratings. The new model is compared to the three existing models that use items (such as the popular items, the items with highest similarities, and the items with strong relationships) to make the suggestion. The experiments on the MSWeb dataset and the MovieLens dataset indicate that the proposed recommendation model has the higher performace (via the Precision - Recall and the ROC curves) than the compared models for most of the given.**

Keywords*—*Model evaluation; recommendation model; rule based implicative rating measure; ruleset

I. Introduction

Recommendation systems (or recommender systems - RSs) [1] are used to predict the ratings of users for products; and then suggest to users the products that can be preferred by those users. Therefore, RSs help to reduce the information overload; and used in many fields of life [2]. There are various kinds of RSs [2]-[4], such as content based RSs, collaborative filtering RSs, context based RSs, hybrid RSs, etc. RSs always use the measures such as finding the nearest neighbors, identifying the strong relationships among items, etc. to make the suggestion. In recommendation techniques, the association rule based approach can provide the deep explanation on recommendation to users [5]. This approach uses the support and confidence measures and the maximum rule length to find the ruleset and make the suggestion. The recommendation performance (via e.g. the precision, the recall) will be high if the maximum rule length is greater than 2. However, if the length is increased, the number of rules will increase; and consequently, the recommendation time will also increase or the computer may be overloaded. For that reason, in order to maintain the high performance, another measure applied on the ruleset where the length of each rule equals to 2 should be used.

The statistical implicative analysis [6], [7] is applied in knowledge management, bioinformatics, etc. to study the trends (the strong implicative relationships) among data attributes. This method proposed the measures such as the implicative intensity, the typicality to detect those trends. Therefore, the statistical implicative measures can be considered for building RSs. Author in [8] proposed an approach to make a RS using the implicative intensity measure, but did not compare the performance of the approach with that of others. In [9], we proposed a recommendation model based on the important statistical implicative measures and association rules, and then conducted the internal evaluation on the performance of model. Although that model shows the high performance when compared to some existing models, it is still inherited the drawback of the association rule-based approach mentioned above. Besides, there is the lack of consistency in comparing models because we used the random values for a few compared models instead of selecting the best suitable values.

This paper proposes a new measure called as the rule based implicative rating measure to predict the users' ratings and a new recommendation model to present to users the top *N* items (e.g. movies, songs, products, etc.). The new model is based on the rule mining approach and the proposed measure. The model uses the binary rating matrix of as the input; the measures including the confidence measure and the support measure for mining the rules with the rule length of 2, the statistical implicative intensity for calculating the implicative value of each rule, and the proposed measure for predicting the ratings of users on items. The new measure uses the confidence and the statistical implicative values of rules and the items that were rated by those users.

The remaining of paper is organized as follows. Section II proposes not only the rule based implicative rating measure, but also the recommendation model using the ruleset and that measure. Section III presents the experiments. Section IV is the conclusion.

II. Recommendation Model using Rule based Implicative Rating Measure

The proposed recommendation model consists of *n* users $u_1, u_2, \dots, u_n$ (*U*); *m* items $i_1, i_2, \dots, i_m$ (*I*); and the ratings of users on items stored in a binary matrix of size $n \times m$ (*R*). Each cell of matrix r_{jk} is set to 0 if the item i_k is not liked (rated or known) by the user u_j, and 1 otherwise.

Fig. 1 shows the sketch of proposed recommendation model. The model uses the rule mining approach and the proposed measure *ArIR1* (Rule based Implicative Rating). However, the model just uses the ruleset with the rule length of 2 for reducing the recommendation time or avoiding the computers' overload. The model generates the ruleset using the confidence measure and the support measure, calculates the implicative value of each rule using the statistical implicative intensity, creates an item matrix of *m* items x *m* items basing on the confidence and implicative intensity values of rules, predicts the ratings *R'* that can be given by a user who needs the recommendation, and then presents to that user the items whose predicted ratings are highest (the top *N* items). The general objective of proposed recommendation model is that the performance of model (via the Precision-Recall and the ROC curves) is higher when compared to those of some existing models.

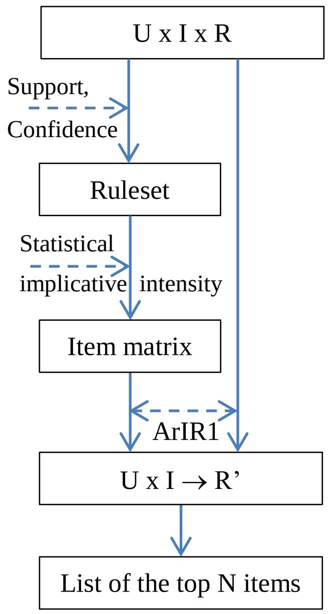


Fig. 1. The recommendation model using the rule based implicative rating measure.

A. Rule based Implicative Rating

The proposed measure *ArIR1* is built on the confidence value and the statistical implicative intensity value of rules, and the items that were rated by the user needing the recommendation. The purpose of *ArIR1* is to predict the rating of a user on an item.

1) Statistical Implicative Intensity

The statistical implicative intensity measure was proposed by Gras [6] to calculate the implicative value between the antecedent and consequent of a rule $a \rightarrow b$. The formula of this measure is shown in (1).

$$\varphi(a,b)=1-\sum_{s=0}^{nab\_}(\lambda^s/s!)e^{-\lambda} \text{ if } n_b \neq n; \text{ and 0 otherwise} \quad (1)$$

Where n is the number of users; n_a (and n_b) is the number of users rating the item a (and the item b); $n_{ab\_}$ is the number of users who rate the item a but do not rate (or know) the item b; and $\lambda=n_a(1-n_b)/n$.

2) Rule based Implicative Rating

The measure *ArIR1* predicts the final ratings given by a user u_a for an item $i \in I$. The value of $ArIR1(u_a,i)$ is in range [0,1]; and is defined by (2) where $AIR1(u_a,i)$ (Atom Implicative Rating) is the original implicative rating value of u_a for i. The formula of *AIR1* is shown in (3).

$$ArIR1(u_a,i)=AIR1(u_a,i)/max_{l \in I}AIR1(u_a,l) \quad (2)$$

$$AIR1(u_a,i)=\sum_{j=1}^{k}IIntConf(i_j,i) \quad (3)$$

In (3), k is the number of items that were rated (liked) by u_a; i_j is one of k items rated by u_a; and $IIntConf(i_j,i)$ is the combination of two values as shown in (4). $\varphi(i_j,i)$ is the statistical implicative intensity value of the rule $i_j \rightarrow i$; and $conf(i_j,i)$ is the confidence value of the rule $i_j \rightarrow i$. The difference among rules will increase quickly if we use the *IIntConf* instead of φ or *conf*.

$$IIntConf(i_j,i)= \varphi(i_j,i)*conf(i_j,i) \quad (4)$$

The value of $AIR1(u_a,i)$ is calculated from the confidence value and the implicative intensity of rules where the right side of those rules are the same item (i) and the left side are the items rated by u_a.

B. Top N Recommendation

1) Creating a Matrix of Items

A matrix of items *IIntConf* is a m x m square matrix where each element $IIntConf[i_j,i]$ is the value of the rule $i_j \rightarrow i$ according to (4) and i_j as well as i is an item of the itemset I. The algorithm named *CreateItemMatrix* is used for building this matrix. The inputs of algorithm are: the rating matrix R; the thresholds of the support measure and the confidence measure to be s and c respectively. The processing steps of *CreateItemMatrix* are:

- Generating a ruleset by the algorithm *Apriorio* [10] using the thresholds s and c.
- Presenting each rule of ruleset by four values $\{n, n_a, n_b, n_{ab\_}\}$, then calculating the implicative intensity value of that rule according to the algorithm named as *ImplicativeIntensity*. We proposed this algorithm in [11].
- Building a matrix of items *IIntConf* according to (4).

```
CreateItemMatrix(ratingmatrix R; float s,c) {
    Ruleset = Apriorio(R,s,c);
    Ruleset = ImplicativeIntensity(Ruleset,R);
    for each rule r_k of form i_j → i belongs to Ruleset {
        IInt[i_j,i] = GetImplicativeIntensity(Ruleset,r_k)
        Conf[i_j,i] = GetConfidence(Ruleset,r_k);
        IIntConf[i_j,i] = IInt[i_j,i] * Conf[i_j,i];
    }
    return IIntConf;
}
```

2) Predicting Ratings

The algorithm named as *PredictRatings* is used for predicting the ratings of a user for each item in the itemset I. The inputs of *PredictRatings* are: the vector A of size m, $A[i_j]=0$ if u_a did not rate (or know) the item i_j and 1 otherwise,

u_a is a user needing the recommendation; the square matrix of items *IIntConf*. The steps of *PredictRatings* are as follows:

```
PredictRatings(vector A; matrix IIntConf) {

AIR1 = InitializeMatrix(0,1,m);
ArIR1 = InitializeMatrix(0,1,m);
for each item i_j ∈I
    if (A[i_j]==1)
        for each item i∈I
            AIR1[u_a,i] += IIntConf[i_j,i];
ArIR1[u_a, ] = AIR1[u_a, ]/max(AIR1[u_a, ])
return ArIR1;
}
```

3) Recommendating the Top N Items

The proposed model uses the algorithm named as *IRRS* to present the top *N* items to a user u_a who needs the recommendation. Its inputs are: the vector *A* as mentioned above; the rating matrix *R*; the thresholds of two measures (support, confidence) to be *s* and *c* respectively. *IRRS* conducts the following steps:

- Building a matrix of items *IIntConf* by the algorithm named as *CreateItemMatrix*.
- Predicting the ratings of the user u_a for each item $i \in I$ using the ruleset and the measure *ArIR1* according to the algorithm named as *PredictRatings*.
- Filtering the predicted result by removing *given* items that were rated by u_a.
- Suggesting *N* items (of the filtered list) with the highest predicted implicative ratings to u_a.

```
IRRS(vector A; ratingmatrix R; float s,c){
    IIntConf = CreateItemMatrix(R,s,c);
    ArIR1 = PredictRatings(A,IIntConf);
    Ratings = RemoveKnownGiven(A, ArIR1);
    Reclist = TopNItems(Ratings);
    return Reclist;
}
```

III. Experiments

A. Experimental Setup

1) Experimental Data

MovieLens[1] and MSWeb[2] are the datasets to be used in our experiments. The former dataset stores the ratings of users on movies while the latter dataset stores all Vroots (the areas of the website) visited by users. The number of users, the number of movies or Vroots, and the number of ratings of these two dataset are (943, 1,664, 99,392) and (32,710, 285, 98,653) respectively. For the MovieLens dataset, the value of rating is from 1 to 5. Therefore, the dataset has to be binarized in order to use the proposed recommendation model. A rating is set to 1 if the user rated the movie and 0 otherwise.

To avoid bias because of lack of data, the MovieLens dataset and the MSWeb dataset have to be preprocessed by removing the items rated only a few times as well as the users rating only a few items. The minimum number of items per user and the minimum number of users per item to be selected for extracting these datasets are (50, 100) for MovieLens and (10, 50) for MSWeb. Table I shows the information of these two datasets after preprocessing.

TABLE I. Information of Two Datasets after Preprocessing

Dataset	The number of users	The number of items	The number of ratings	The maximum number of given*
MovieLens	565	202	41,245	10
MSWeb	875	135	10,487	7

*: For evaluating the models, if a user needing the recommendation were rated *t* items (i.e. *t* ratings), then *given* ratings (also called as *given* known ratings) is used for making the recommendations and the remaining ratings (*t*-*given* ratings) is used for evaluating the recommended result. The maximum number of given is identified by analyzing the percentiles of datasets.

2) Experimental Tool

We developed the recommendation model using rule based implicative measure in the R language, and integrated it in the Interestingnesslab tool [12]. Besides, the three existing models of recommenderlab package [13] are also used for the comparison purpose. Those models are named as IBCF - the item based collaborative filtering model, POPULAR - the popular model, and AR - the association rule based model. The reason for selecting three models is that they all are based on the items to suggest the top *N* items to users. The first one uses the items with highest similarities. The last two ones use the popular items and the items with strong relationships (rules), respectively.

3) Recommendation System Evaluation

We use the evaluation method presented in [14]. In that way, the k-folds cross validation method [15], the Precision - Recall curves and the Receiver Operating Characteristic (ROC) curves [16] are used for evaluating the performance of RSs. The ROC curve is plotted from the values of True Positive Rate (TPR or recall) and False Positive Rate (FPR) while the Precision - Recall curve is plotted from the values of precision and recall. In the experiments of this research, the value of k-folds is set to 4.

B. Experimental Results

We develop four RSs named as: Popular RS, IBCFJaccard RS, ArIR1 RS, and ARConfidence RS.

- ArIR1 RS uses the proposed recommendation model and the rule based implicative rating measure.
- Popular RS suggests to users the top *N* popular items without known items. The POPULAR model is used by this RS.
- IBCFJaccard RS uses the IBCF model and the Jaccard measure. The Jaccard measure is used because it only focuses on matching ones and thus prevents the problem with zeroes. IBCFJaccard RS recommends the top *N* items based on the items with highest similarities.
- ARConfidence RS finds the strong rules (the strong relationships among items) and then suggest to users the

[1]http://www.grouplens.org/node/73
[2]http://www.ics.uci.edu/~mlearn/MLRepository.html

top *N* items. This RS uses the AR model and the confidence measure.

All of these RSs use the following parameters:

- The number of items to be presented to the users (*recs*) is: 1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 for the MovieLens dataset; and 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12 for the MSWeb dataset.
- The number of known ratings of each user used for making the recommendation (*given*) is: from 1 to 10 for the MovieLens dataset; and from 1 to 7 for the MSWeb dataset.
- The number of times of evaluation (*times*) is varied. For each times of evaluation, the k-folds cross validation method is used.

1) Selected Values of Important Parameters

The important parameters of IBCFJaccard RS and ARConfidence RS are: *k* items to be used for finding the items with highest similarities; and the support value *s*, the confidence value *c*, and the maximum length of rule *l*, respectively. To find the suitable values of these parameters, we vary their values as well as the values of *times*, *given*; then observe the ROC curves and the Precision - Recall curves. *k* is varied from 15 to 202 for the MovieLens dataset, from 20 to 135 for the MSWeb dataset with the difference between two adjacent values to be about 10. To do not omit the association rules with high quality, *s* and *c* should be assigned to small values: from 0.01 and 0.2 respectively for the MovieLens dataset; and from 0.01 and 0.02 respectively for the MSWeb dataset. As a result, Table II shows the selected values of important parameters.

TABLE II. SELECTED VALUES OF IMPORTANT PARAMETERS

Dataset	**k**	**s, c, l**
MovieLens	202 for all givens	s=0.01,c=0.3, and l=3 for given from 1 to 5. s=0.04,c=0.5, and l=3 for given from 6 to 10**
MSWeb	135 for all givens	s=0.01, c=0.1, l=7 for all given

**: The larger the value of *l* is, the bigger the size of ruleset will be. The computer used in our experiments is overloaded if the size of ruleset is too big. For example, if s=0.01, c=0.3 and l=3, the size of ruleset is more than 3.5 million rules. Therefore, on the MovieLens dataset, the AR model used the above listed values to compare the models.

The important parameters of ArIR1 RS are the support value *s*, the confidence value *c*. They are used for generating a ruleset where each rule has length of 2. In this paper, *s* and *c* to be selected are equal to ones of ARConfidence RS.

2) Comparison of Recommendation Systems

On the MSWeb dataset, the evaluation of four RSs (Popular RS, IBCFJaccard RS, ArIR1 RS, and ARConfidence RS) using the parameters as mentioned above is conducted.

Fig. 2 displays the ROC curves of four RSs with *times*=10, *given*=4, and the selected values (*recs* from 1 to 12, *k*=135, *s*=0.01, *c*=0.1 and *l*=7) as shown in Table II. The result shows that the ROC curve of RS using the proposed recommendation model is above the other curves; the probability of false alarm (i.e. FPR) of that RS is lowest when compared to that of others for each value of *recs*. When varying the value of *times* and using the selected values of important parameters, we obtain the result similar to Fig. 2 for *given* from 3 to 7.

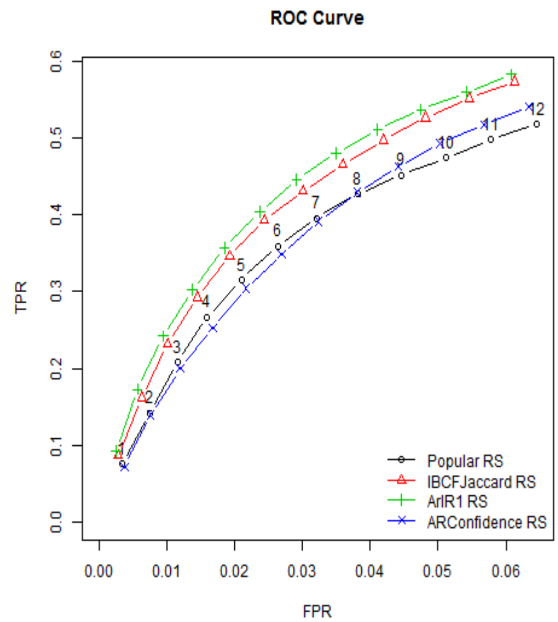


Fig. 2. The ROC curves of four RSs on the MSWeb dataset with times=10, given=4, recs from 1 to 12, k=135, s=0.01, c=0.1, and l=7.

Fig. 3 shows the Precision - Recall curves of four RSs with *times*=2, *given*=2, *recs* from 1 to 12, *k*=135, *s*=0.01, *c*=0.1 and *l*=7. The result indicates that the precision value and the recall value of ArIR1 RS are highest if *recs* is from 1 item to 8 items; and higher than those of Popular RS and IBCFJaccard RS if *recs* is from 9 item to 12 items. For *given*=1, the precision and the recall of ArIR1 RS are just higher than those of IBCFJaccard RS if *recs* is from 1 item to 12 items. When varying the value of *times* and using the selected values, we also achieve the same results as described for *given* less than 3.

On the MovieLens dataset, the performances of four RSs are also compared one another.

Fig. 4 displays the Precision - Recall curves of four RSs with *times*=4, *given*=3, and the selected values shown in Table II (*recs* from 1 to 100, *k*=202, *s*=0.01, *c*=0.3, and *l*=3). The precision value and the recall value of system using the proposed recommendation model are: highest when the number of items to be presented to users (*recs*) is less than 70; and lower than those of IBCFJaccard RS, yet still higher than those of remaining RSs when *recs* is greater than 70. Fig. 4 also shows that the difference between corresponding values of ArIR1 RS and IBCFJaccard RS is quite small for *recs* greater than 30 (the curve portion from 30 items to 100 items). When we vary the value of *times*, use the selected values as shown in Table II, and change *given* from 4 to 10, the achieved result is similar to that obtained for *given*=3. Besides, the length of curve portion with small difference (of ArIR1 RS and IBCFJaccard RS) is increased while *given* is increased.

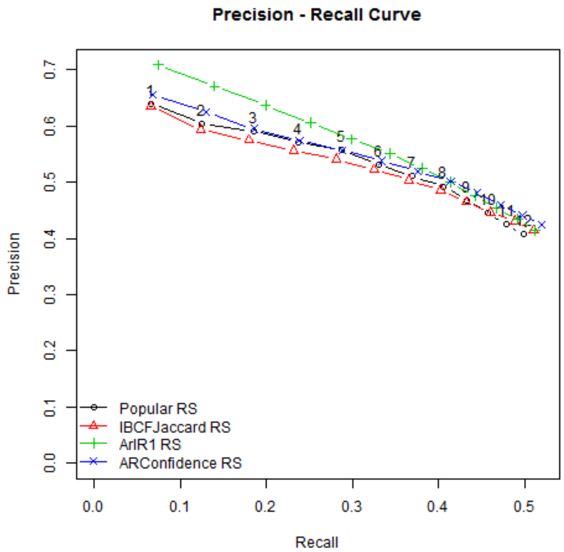


Fig. 3. The Precision - Recall curves of four RSs using the MSWeb dataset, times=2, given=2, recs from 1 to 12, k=135, s=0.01, c=0.1, and l=7.

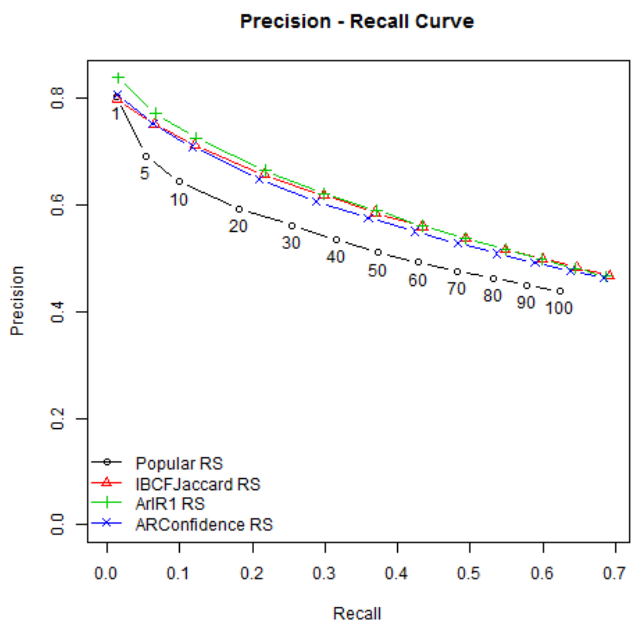


Fig. 4. The Precision - Recall curves of four RSs using the MovieLens dataset, times=4, given=3, recs from 1 to 100, k=202, s=0.01, c=0.3, and l=3.

Fig. 5 shows the ROC curves of four RSs with *times*=2, *given*=1, and the selected values. The ROC curve of ARConfidence RS dominates that of others. The curve of ArIR1 RS is the next one almost dominating the remaining curves. The difference between curves (of ARConfidence RS and ArIR1 RS) is quite small on the portion from 1 item to 30 items. Besides, when observing the values of FPR (the probability of false alarm), we see that the FPR values of ArIR1 RS are almost lowest at all *recs*, especially for the curve portion from 1 to 70 items. We also obtain the result as shown in Fig. 5 when varying the values of *times*.

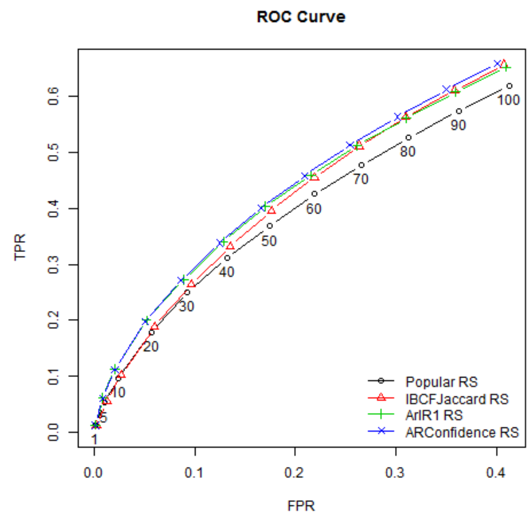


Fig. 5. The ROC curves of four RSs using the MovieLens dataset, times=2, given=1, recs from 1 to 100, k=202, s=0.01, c=0.3, and l=3.

Fig. 6 shows the Precision-Recall curves of four RSs using *times*=5, *given*=2, and the selected values. The figure shows that the precision and recall values of ArIR1 RS are highest on the curve portion from 1 to 70 items; and nearly equal to those of IBCFJaccard RS as well as higher than those of remaining RSs on the portion greater than 70. When varying the values of *times*, the achieved result is similar to that shown in Fig. 6.

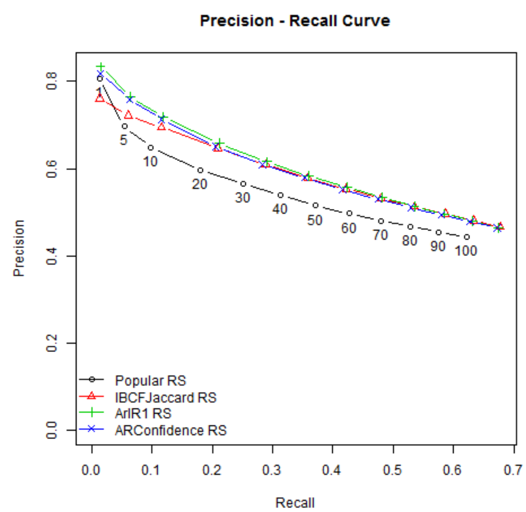


Fig. 6. The ROC curves of four RSs using the MovieLens dataset, times=5, given=2, recs from 1 to 100, k=202, s=0.01, c=0.3, and l=3.

From the experimental results, it is found that: 1) the AR model should only be used for developing RS if *given* equals to 1, its performance is decreased while *given* is increased and the number of generated rules is big if the maximum length of rule is increased; 2) the proposed model should be used for most of *given* (such as: *given* greater than 2 for the MSWeb dataset, and *given* greater than 1 and the number of items to be presented to users less than and equal to 70 for the MovieLens dataset). Besides, either IBCF model or the proposed model can be used for most *given* greater than 1 and the number of items to be presented to users greater than 70 for the MovieLens dataset because the difference between corresponding values of these two models is quite small.

IV. Conclusion

RSs help to reduce the information overload because they can suggest the valuable products to users from the rating matrices. The paper proposes the measure named as *ArIR1* for predicting the users' ratings on items. *ArIR1* is built on the implicative intensity and the confidence value of rules, and the items rated by users needing the recommendation. Besides, the paper also presents the new model based on the ruleset with the rule length of 2 and the proposed measure. This model with the binary rating matrix will suggest to users the top *N* items. The performance of the proposed recommendation model (via the Precision - Recall and ROC curves) is compared with that of some existing models integrated in the recommenderlab package (IBCF, POPULAR, and AR) on two datasets (MovieLens and MSWeb). The results indicate that the performance of proposed recommendation model is higher than that of others for most of *given*.

However, the proposed model has not yet been effective when the number of items that were rated by a user needing the recommendation is small (given=1 or 2 for the MSWeb dataset, and given=1 for the MovieLens dataset). Therefore, in the future direction, we will improve the performance of proposed model by using the hybrid approach such as using the AR model for the small *given* but the creation of a ruleset and a matrix of items will be performed offline; and switching to use the proposed model for the remaining *given*. Moreover, we will also observe the performance of proposed model by varying the threshold value of *IIntConf* to identify the suitable value.

References

[1] F. Ricci, L. Rokach, B. Shapira, and P.B. Kantor, Recommender Systems Handbook, Springer US, 2011.

[2] J. Lu, and et.al, "Recommender system application developments: a survey", Decision Support Systems, 74, pp. 12-32, 2015.

[3] C. Aggarwal, Recommender Systems: The Textbook, Springer International Publishing Switzerland, 2016, ISBN 978-3-319-29657-9.

[4] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, An introduction to recommender systems, Cambridge University Press, 2011.

[5] A. Felfernig, M. Jeran, G. Ninaus, F. Reinfrank, S. Reiterer, and M. Stettinger, "Basic Approaches in Recommendation Systems", Recommendation Systems in Software Engineering, pp. 15-38, Springer-Verlag Berlin Heidelberg, 2014.

[6] R. Gras, E. Suzuki, F. Guillet and F. Spagnolo, Statistical Implicative Analysis, Springer-Verlag, 2008.

[7] J. Régnier, Y. Slimani, and R. Gras, Proceedings of Conference: Analyse Statistique Implicative, des sciences dures aux sciences humaines et sociales, ASI 8, Tunisie, 2015, ISBN : 978-9973-9819-0-5.

[8] R. Couturier, "Un système de recommandation basé sur l'A.S.I.", L'Analyse Statistique Implicative. Des Sciences dures aux Sciences Humaines et Sociales, dir. R.Gras, Cépaduès Ed. Toulouse, 2017, pp. 447-452.

[9] L.P. Phan, H.H. Huynh, H.X. Huynh, and R. Gras, "Recommender System based on Important Implicative Measures", In Proceedings of the 9th International Conference A.S.I. Analyse Statistique Implicative, Belfort, France, pp.508-531, 2017.

[10] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases", In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 207–216, 1993.

[11] L.P. Phan, K.M. Nguyen, H.H. Huynh, and H.X. Huynh, "Association-based recommender system using statistical implicative cohesion measure", In Proceedings of the 8th International Conference on Knowledge and Systems Engineering - KSE, Hanoi, pp. 144-149, 2016.

[12] L.P. Phan, N.Q. Phan, K.M. Nguyen, H.H. Huynh, H.X. Huynh, F. Guillet, "Interestingnesslab: A Framework for Developing and Using Objective Interestingness Measures", In Advances in Intelligent Systems and Computing, vol 538. Springer, Cham, pp. 302-311, 2017

[13] M. Hahsler, "recommenderlab: A Framework for Developing and Testing Recommendation Algorithms", SMU, 2011.

[14] L.P. Phan, H.H. Huynh, and H.X. Huynh, "User based Recommender Systems using Implicative Rating Measure", International Journal of Advanced Computer Science and Applications, pp. 37-42, 2017.

[15] R.Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection.", In Proceedings of the14th International Joint Conference on Artificial Intelligence, pp.1137-1143, 1995.

[16] A. Gunawardana and G. Shani, "A Survey of Accuracy Evaluation Metrics of Recommendation Tasks", Journal of Machine Learning Research, 10, 2009, pp. 2935–2962.

Distributed QoS Constraint based Resource Adaptation Strategy for Cognitive Radio Networks

Yakubu S. Baguda
Information Systems Department,
Faculty of Computing and Information Technology,
King Abdulaziz University, Saudi Arabia.

Abstract—This paper primarily investigates and addresses the optimal power adaptation strategy problem for multi-user cognitive radio network. The need to determine the optimal power transmission for the secondary user (SU) in a distributed network environment is challenging and important. This requires an efficient adaptive strategy with high convergence capability in order to meet up with the multi-users quality of service (QoS) requirements in a cognitive network and ultimately ensure more efficient resource utilization amongst users. In this paper, a distributed QoS contraint scheme which considered both the transmission power and outage probability has been proposed to maximize the network performance and to control the SUs transmission power. Firstly, the QoS constraint optimization problem for distributed secondary users is formulated and solved in order to adapt with the network dynamics of the cognitive networks. Subsequently, the solution has been used to dynamically adjust the radio power of the SUs to conform with the stringent network constraint and ensures coexistence amongst the users. More importantly, the simulation result shows that the scheme increases the overall network utility when compared to the scheme without adaptation.

Keywords—Cognitive radio network; outage probability; quality of service (QoS); network utility; power adaptation; primary user; secondary user

I. Introduction

There has been explosive growth in mobile devices market and high bandwidth demand which eventually leads to more congestion and traffic within the frequency spectrum. It is known fact that the spectrum is primarily allocated to primary users (PU), but has not been utilized effectively. Hence, making usage of the available resources in effective manner requires an efficient resource allocation and transmission strategy to dynamically adapt with the network condition. Therefore, there is dramatic need for highly adaptive schemes to efficiently manage the spectrum utilization. This eventually led to development of cognitive radio technology in order to efficiently and effectively manages the spectrum [1]. Sensing for free spectrum is challenging and it is very important in cognitive radio as such the decision depends on the precision to determine the spectrum availability or presence of the primary user. The performance of the secondary user can be significantly improved by determining the available spectrum that can be used regularly by the secondary user if condition warrant. When more users are involved in the decision, the probability of false alarm is higher, but at the expense of increasing the latency.

As it has previously been mentioned, making effective decision is essential which can lead to higher performance and reliable communication in cognitive radio network. More importantly, efficient power level sensing mechanism will improves the performance and QoS provision by exactly determining the presence of PU and the network utilization. For instance, the network condition rapidly changes and hence the decision time should be negligible and precise as well to ensure effective utilization of the network. The ability to determine the probability of false alarm precisely matters lot while taking decision. It is challenging to reduce the decision delay time especially in a network with large number of users and at the same time achieving optimal performance. Hence, there is need for high effective optimization scheme in order to achieve high performance and conforming to the power constraint in which every SU is expected to abide by. Managing and utilizing the available limited resources will increase the network performance, capacity and reduces the complexity. In most of the literature works, the users were assumed to have the same features and setting or uniform parameters. In real world scenario, the devices have different sensing parameter and requirements. This should be considered while designing such systems. Designing schemes to combine the collective sensing information from different devices and ensuring effective decision with sophistication and precision is extremely important for effective communication to harness.

Different research works related to dynamic spectrum sensing have been conducted in order to improve the throughput of the SU and at the same time protecting the PU [2], [3]. It is known fact that the probability of false alarm depends greatly on the sensing parameters and throughput as well. In [4], the resource allocation by maximizing the total throughput of the secondary user has been exploited. More importantly, both sensing time and power were considered for enhancing the cognitive radio performance. Some researches related to spectrum sharing have been focusing on how to used power constraint on secondary users not exceeding the maximum transmission power level so that not to interrupt with the primary user operation [5]-[7]. Consequently, this will reduce the intereference in the cognitive radio network and enhence performance.

Section 2 will cover the related works. Section 3 mainly focuses on the system model and problem formulation. Section 4 discusses the developed distributed QoS constraint based power adaptation scheme. Numerical results and discussion for experimental simulation is presented in Section 5. Finally, the conclusions for the research are drawn in Section 6.

II. Related Work

Distributed power control has been studied in [8] and the scheme mainly ensures that the feasible solution meet up with the requirement of the users. A joint optimization scheme for cell selection and power control was proposed [9]. Huang et al. [10] used power price and interference to adjust power transmitted by the users with the multi-hop wireless network. In [11], both flow rate and transmitted power were used in order to achieve high network utilization. It assumed static channel condition in which is not possible in real world scenario. A centralized power control scheme using outage probability in [12] has been proposed and it considers the channel fading. In [13], it has formulated the QoS optimization problem in a multi-user environment.

An auction based power allocation scheme was developed in [14] to share spectrum amongst distributed users but it has taken the interference temperature into consideration. Srinivasa et al. [15] studied joint beam forming and power control scheme for different users with multiple antennas.

Wang et al. [16] considered the interference amongst the secondary users and used an exponential pricing function to protect the primary users from interference. More importantly, it considered many secondary and primary users which use the same band for their operation. In [17], multiple users and single PU were considered in the distributed allocation problem scenario in order to minimize the transmitted power. Mei et al. [18] proposed a fast convergence distributed power control algorithm which uses round robin. In [19], it uses game theory to minimize the sum throughput of the SUs.

Hence, it is very important to note that the developed scheme is different from other power control scheme because it primarily considers the maximum transmission power and the QoS needed by PUs and SUs for reliable and effective communication. More importantly, the developed scheme maximize the overall cognitive radio network spectral utility and at the same time conforming to the general coexistence policy of the network in such a way that the primary users are not interrupted or interfered by the secondary users.

III. System Model and Problem Formulation

In this section, the system model used for the distributed QoS constraint based power adaptation scheme has been described. The scenario setting shown in Fig. 1 has been used for the multi-user cognitive radio network. Firstly, let's assume that there are N primary and M secondary users within the network. Also, the basic assumption for the development of the model are introduced in order to have clear understanding of the general assumptions and setting used for the development of the proposed scheme.

In order to model the QoS constraint power adaptive scheme for the above cognitive radio network, it is assumed that there secondary transmitter (SUBs) which is within the coverage area of the cognitive network. Also, it is assumed that the transmitted power by N secondary users can be represented mathematically as follows:

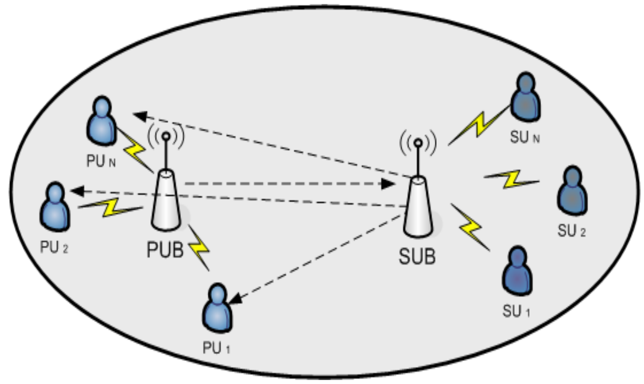


Fig. 1. A typical multi-user cognitive radio network scenario.

$$P_i = [P_1, \cdot \cdot \cdot, P_N] \tag{1}$$

The channel utilization is assumed to be slotted and the secondary users synchronize their spectrum access based on the time slot. In order to avert collision, every node ensures that it listen to the channel before transmit. In this work, the spectral utility of secondary users is determined using the mathematical equation shown as follows:

$$U_i = \log(1+h_iP_i) \tag{2}$$

Where h_i represent the constant taking care of the parameters such as coverage, intereference for the secondary users (i.........N). The utility is connected or related to the information rate which can be effectively conveyed over the secondary user link at a particular time. Ultimately, it varies with the change in network condition as in the real world scenario. To ensure coexistnce amongst the users within the network, the power transmitted by users P_i should be less than the maximum power P_{max}. So all the users have to abide with the maximum power threshold in order to prevent other users interefering the primary users.

Lets assumed that No is the gussian noise at PUB and hence the SINR due to channel can be computed using the equation below:

$$\beta = \frac{P_0 G_0 F_0}{N_0 + \sum_{i=1}^{N} P_i G_i F_i} \tag{3}$$

Assuming F_0 is exponentially distributed and the outage probability of the PUB is given by

$$\emptyset_0 = 1 - \exp\left[-\frac{N_0 \beta_{th}}{P_0 G_0}\right] \tag{4}$$

β_{th} present the required SINR and it ultimately ensures that both the QoS requirements of the PU and SU have been achieved.

In order to develop the optimization problem needed to meet up the power and outage probability. The primary objective is to maximize the secondary user utility function given in (5) and subsequently taking into consideration the power and outage probability constraint as well. The optimization problem can be mathematically described as follows:

$$Maximize\ U = \sum_{i=1}^{M} \log(1 + h_i P_i)$$

Subject to:

$$\sum p_i \leq p_{max} \quad \forall i$$
$$\sum \emptyset_i \leq \emptyset_{max} \quad \forall i \tag{5}$$

In the optimization problem presented in (5), the outage probability and transmission power have been considered as constraints need to be fulfilled by the optimization problem. The SU utility which satisfied the constraint is selected as the utility values to conform to the PU and SU QoS requirement. The value of p and $\emptyset$ for SUs $(i = 1 \ldots\ldots\ldots. M)$ is compared with the maximum value in order to ensure that the PU and SU coexist within the same coverage without interference amongst the users. In order to achieve optimal utility while considering the QoS constraint, (5) has been used in optimizing the SU performance. In the next section, a detail about the QoS constraint power adaptive algorithm has been described.

IV. QoS Constraint Power Adaptive Algorithm

The main primary objective is to reduce the outage probability by maximizing the utility of the secondary users within the cognitive radio network. Initially, the problem has been formulated as convex optimization and its closed solution is derived. The algorithm is used to approximate the optimal solution based on the transmission power and outage probability of the secondary users in the cognitive radio network. The dynamics of the secondary users with the cognitive radio network can be handle using heuristics – that is to learn and improve with time. Regularly, the SU's estimate the outage probability and transmission power in order to tackle the rapid change in the network condition.

The distributed QoS constraint based power adaptation algorithm is as shown below and the parameters used by the algorithm are initialized. Having initialized the constant parameters, others parameters such as p, q are computed by users as the time increase from 0 to t $(0 \rightarrow t)$. The main key issue of concern is the fact that how the distributed QoS constraint based power algorithm will meet up with both power and outage probability constraints. The SUs receives information from the SUB about the network dynamics and based on that it decides on either to utilize the PU spectrum.

Algorithm 1 *Distributed QoS Constraint Based Power Adaptation Scheme*

01. Set value for N_0 , p_{max}, $\emptyset_{max}$, number of SUs
02. Calculate p_i & $\emptyset_i$ for each SU
03. **while (i < M) do**
04. If $((\emptyset_i \geq \emptyset_{max})\ AND\ (0 \leq p_i \leq p_{max}))$
05. $\mathbf{U_i}$ = max $(\sum_{i=1}^{M} \log(1 + h_i P_i))$
    else if $(p_i \geq p_{max})$
06. $p_i = p_{max}$
07. Update U_i^{n+1} and P_i^{n+1}
08. Go to step 5
09. end if
10. end if
11. *i =i+1*
12. end while

Initially, the values for p_{max} and $\emptyset_{max}$ are set as the constraint need to be satisfied before determining the network utility U_i for any SU. They represent the power and outage probability threshold in which all the users have to conform with. In a network with M secondary users, each user while updating its power has to compare with the maximum power threshold. The transmitted power for SUs should always be less than p_{max} in order not to interfere with other users. The values p_i & $\emptyset_i$ for each SU are computed and tested to ensure that they meet up with the constraint. If the condition has been satisfied, the network utility is computed based on (5). If $p_i \geq p_{max}$, the maximum power threshold value is chosen as the value the SU should use as its transmission power level. The SU decides by juxtaposing the P_i with the P_{max}. When $P_i < P_{max}$, the SU will automatically access the spectrum and it transmit with the power P_i. Also, it considers the outage probability threshold as well. Otherwise, the SU does not engage in to any activity with the PU spectrum. In Step 7, both network utility and power for the secondary user are updated. By repeating this procedure over time, the algorithm converges to the peak power level and this dramatically improves the network utility since P_i is directly related to it.

Since the develop algorithm has restriction on SUs knowing about other SUs activities, they only intermingle with the SUB unit. In a very large network which has many SUB and PUB, each SUB has information about all the SUs and PUBs connected to it. This requires more strategic approach to effectively sense the condition of the network and eventually fulfill the constraints thresholds under such situation. Each SU decide independently whether to utilize the spectrum opportunity or not. Once it has been decided, the SU will access or share information via the available medium.

Based on (5), the problem can be further expressed and represented in Lagrange form by

$$\begin{aligned} \boldsymbol{L}(p, \emptyset, \lambda) &= \sum_i \log(1 + h_i p_i) \\ &\quad - \sum_i \lambda_i \left[\left(\sum_i p_{max} - p_i \right) + \left(\sum_i \emptyset_i - \emptyset_{max} \right) \right] \\ &= \sum_i \log(1 + h_i p_i) + \sum_i \lambda_i \left(\sum_i p_i - p_{max} \right) \\ &\quad + \sum_i \lambda_i \left(\sum_i \emptyset_i - \emptyset_{max} \right) \end{aligned} \tag{6}$$

$$= \sum_i \log(1 + h_i p_i) + \sum_i \lambda_i \sum_i p_i - \sum_i \lambda_i p_{max} + \sum_i \lambda_i \sum_i \emptyset_i - \sum_i \lambda_i \emptyset_{max}$$

$$= \sum_i \log(1 + h_i p_i) + \sum_i \lambda_i \sum_i p_i + \sum_i \lambda_i \sum_i \emptyset_i - \sum_i \lambda_i p_{max} - \sum_i \lambda_i \emptyset_{max}$$

$$= \sum_i \left(\log(1 + h_i p_i) + \lambda_i \sum_i p_i + \lambda_i \sum_i \emptyset_i \right) - \sum_i \lambda_i p_{max} - \sum_i \lambda_i \emptyset_{max}$$

If $q_i = \lambda_i \sum_i p_i + \lambda_i \sum_i \emptyset_i$, the equation above can be re-written as

$$= \sum_i (\log(1 + h_i p_i) + q_i) - \sum_i \lambda_i p_{max} - \sum_i \lambda_i \emptyset_{max} \tag{7}$$

Where q = [q_1, q_2,…….q_l] and hence the Lagrange of the dual function to the optimization problem can be represented mathematically as follows:

$$G(p, \emptyset, \lambda) = \min(L(p, \emptyset, \lambda)) \tag{8}$$

$$G(I, x, \lambda) = \min(L(I, x, \lambda))$$

$$=\min\left(\sum_i (\log(1 + h_i p_i) + q_i) - \sum_i \lambda_i p_{max} - \sum_i \lambda_i \emptyset_{max}\right) \tag{9}$$

By setting (9) to zero and re-arranging the terms in the equation, it eventually yields

$$\sum_i (\log(1 + h_i p_i) + q_i) = \sum_i \lambda_i p_{max} + \sum_i \lambda_i \emptyset_{max} \tag{10}$$

The update of the dual variables is achieved using (11) and (12) as shown below:

$$p_i^{(n+1)} = min\left\{p_i^n - \theta^n (p_{max} - \sum_i p_i)\right\}, \quad \forall i \in I \tag{11}$$

$$\emptyset_i^{(n+1)} = \min\left\{\emptyset_i^n - \theta^n (\emptyset_{max} - \sum_i \emptyset_i)\right\}, \quad \forall i \in I \tag{12}$$

A. Optimal Utility

The set of optimal utility for each SU at a particular time t is determined and the best utility is selected based on the constraints. The set of utility for the secondary users can be represented mathematically as follows:

$$U_i = [U_1, \cdot \ \cdot \ \cdot, U_N] \tag{13}$$

More importantly, it has been assumed that there are N secondary users within the cognitive radio network. For each secondary user, it optimal transmission and its required QoS is considered while selecting the best utility. Hence, equation can be further represented in simple form as follows:

$$U_i = [U_1(P_1, Q_1) \cdot \ \cdot \ \cdot, U_N(P_N, Q_N)] \tag{14}$$

Similarly, the outage probability for the secondary users is represented by equation (15) as follows:

$$\emptyset_i = [\emptyset_1, \cdot \ \cdot \ \cdot, \emptyset_N] \tag{15}$$

For every SU, the outage probability varies over period of time t. It can be noticed that $\emptyset_i$ vary randomly below and above the threshold red line as shown in Fig. 2. The outage probability should not be fixed, but changes as the network condition changes. This is primarily to ensure that the simulation scenario mimic the real network scenario as well. In addition, the outage probability is affected by the distance, signal strength, noise and any other factor which can change the network condition. Each SU compare its outage probability with the threshold value in order to decide on whether to increase or decrease its transmission power and also by taking into consideration the maximum power threshold.

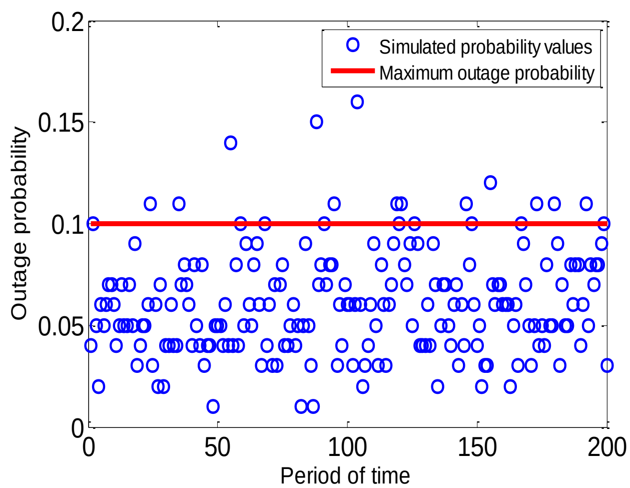


Fig. 2. Constraint outage and simulated outage probabilities.

As can be seen from Fig. 2, it shows the SUs outage probabilities variation over a period of time t. The developed scheme adapts with the change in outage probability and SU transmission power. In a nutshell, the primary motive for setting the outage probability threshold is to ensure that the SU QoS requirements have been accomplished for effective utilization of the network. For every SU, its outage probability is different from other SU as in real world scenario - this is mainly due to the aforementioned reason. The SU will decide to increase its transmission power whenever its outage probability is below the threshold value. More importantly, the SUs update their outage probability based on (12). This strategy helps a lot in making effective utilization of the network amongst the secondary users within the cognitive radio network.

V. Simulation Results and Analysis

This section primarily focus on the simulation results obtained from the proposed distributed QoS constraint based power control algorithm are presented and discussed in more details about the findings. As it has been mentioned that a multiple SUs were deployed within the cognitive radio network as shown in Fig. 1. The parameter settings used for the simulation are represented in Table I as shown below.

TABLE I. Simulation Parameter Settings

Parameter	Value
Path loss exponent	3.5
PU transmission power	2000 mw
SU maximum transmission power	1000 mw
Number of SUs	5
Outage probability	10%

More importantly, the outage probability threshold for the PU has been set to 10% as constraint and it has ultimately been maintained constant throughout the period of time. The main primary key objective is to maximize the SU utility and enhance the overall network performance. Initially, the proposed distributed QoS constraint based power control scheme has been tested to verify that the outage probability does not remain constant but it varies over period of time as in real time scenario. Also, it is very clear that the SU’s outage probability can above or below the threshold value and hence if it has not been properly managed, the QoS of the SU and PU can be significantly affected since there is every tendency that the users would interfere with one another.

A. SUs Power Convergence

Interestingly, the transmitted power by each SU has been a major issue of concern in the distributed QoS constraint based power control algorithm. Firstly, the power transmitted by each is critically analyze in order to ensure that all the SUs operated within the allowable power limit. As it can be seen in Fig. 3, the transmitted power by each user is far below the allowable threshold power level. More importantly, it can be noticed that all the users have different transmission power which is similar to the real world scenario. The power threshold for the SUs is 50% less than that of the PU. This is mainly to ensure that the SUs do not interfere with the PU while operating. For instance, the SU with the highest power is the secondary user 3 and then followed by 2, 1, 4, and 5 as shown in Fig. 3.

Even though that the distributed QoS constraint based power control scheme is decentralized system, coordinating and managing such scheme is challenging and it require thorough testing and evaluation to verify whether the performance of the developed scheme conform the requirements needed to be fulfilled. This requires systematic approach in order to efficiently test the scheme performance and its decision making precision. In order to critically analyze the performance of the developed power control scheme, the performance of each secondary user is evaluated since it is decentralized power control scheme. All the five (5) secondary users were considered so that each user conforms to the maximum power threshold. Based on the numerical experimentation, it can be seen from Fig. 2 that as time t increases; the SUs adjusted their respective transmission power in order to achieve the optimal transmission power.

As it has been explained, each SU has different proximity from both the PUB and SUB. This has had dramatic impact on the SUs transmission power and coverage. As the time varies from 0 to any value of t, the transmitted power for SUs is determined and subsequently each SU updated its power level. As time increases, the power transmitted by each SU becomes relatively constant when the algorithm completely converges. This approach has been used to keep in line with the variation in network condition. Consequently, this improves the network utilization and it ensures fairness amongst the SUs.

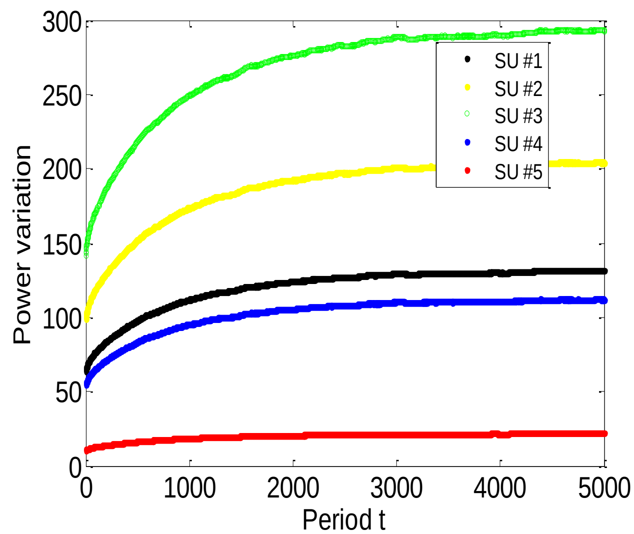


Fig. 3. SUs estimated power over period t.

The power transmitted by the SUs have been observed and evaluated. As can be seen from Fig. 3, it shows that the SUs are capable of maintaining their power level within the acceptable range. None of the SUs has exceeded the power threshold level and it proves how the developed scheme has been able to meet with the network constraint in order to achieve high performance and at the same keep in line with the constraints. The maximum achievable power transmitted by each SU after the network has completely converged is presented in Fig. 4.

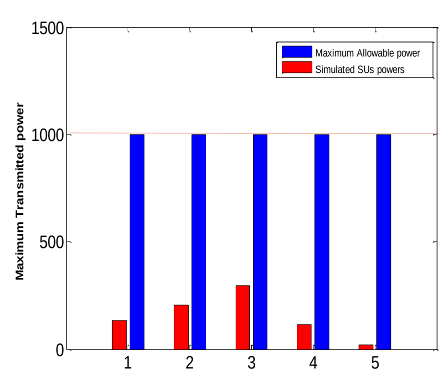


Fig. 4. Maximum allowable power limit for each user and SUs respective transmitted power.

B. Secondary Users Outage Probability Convergence

In order to determine the respective outage probability of each SU, the outage probability was observed and represented in graphical form. As it can be seen from Fig. 5, the SU's have different outage probability as in real world scenario. For instance, SU3 has the highest outage probability and SU5 has the lowest outage probability when compared to the other SUs.

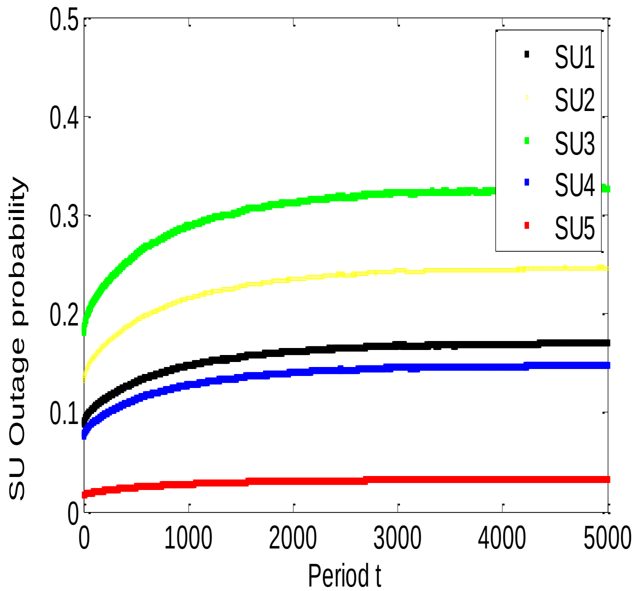


Fig. 5. Secondary users outage probability.

The main primary motive of considering the outage probability is to ensure that the QoS requirements for both the PU and SU have been achieved. This clearly shows the dramatic need to select best possible transmission power and outage probability. In a nutshell, it is very important to note that only the outage probability of SU5 is below the PU outage probability. This is due to the fact that there is need to provide QoS to SUs in order to ensure effective spectrum ulitization by the secondary user within the cognitive radio network. The value of the PU outage probability has stragically been chosen to ensure that the required QoS by both the primary and secondary users to accomplish the targeted network performance.

C. Overall SUs Outage Probability Convergence

In Fig. 6, the developed distributed QoS constraint based power control scheme has been tested based on the maximum number of SUs within the network which implies that the QoS of the PU and constraints have been reached. It clearly shows that the developed scheme can conveniently can be able to determine the outage probability precisely when all the SUs estimated their respective outage probability precisely. Once this has been accomplished, the scheme is said to have had converged to optimal. The overall SUs outage probability steady converges toward the threshold value which has been set to 10%. This indicates that the proposed scheme can be able to adapt with time. It starts to converges from 0 and increase exponentially with time t as shown in Fig. 6.

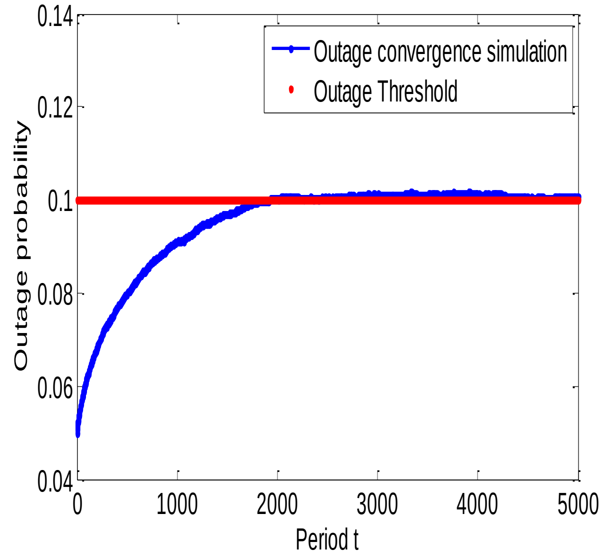


Fig. 6. Distributed QoS constraint based power control scheme outage probability convergence.

Interestingly, it is clear that the ability to utilized the spectrum effectively depend on the precision of the information received by SUs, but in our case no any information is receives from other SUs. Relying more on other SUs would not assist in situation where no any other SU is present within the coverage of the cognitive radio network which makes it difficult for the SU to utilize the network. Hence forth, the need for decentralize scheme is very crucial and extremely important under such conditions. As it can be noticed that the network performance increases rapidly at the initial stage and it become steady as the distributed QoS constraint based power adaptation algorithm converge. It is mainly due to the constraint put in place on the scheme. It has been observed that the proposed decentralize power control algorithm is more effective and reliable since the resources are distributed across the network.

D. Optimal Network Utility Performance Comparison

In this section, the SUs utility performance comparison test was conducted to verify the impact of adaptation in enhancing the network utilization in multi-user cognitive radio network. As can be seen from Fig. 7, the proposed distributed QoS constraint based power control scheme achived high network utilization when compared to the situation when no adaptation strategy has been used. Hence, it can be clearly seen that the proposed adaptation strategy improves the utility performance. Interestingly, the proposed distributed QoS constraint based power adaptation scheme converges faster toward theoptimal network utility when compared to without adaptation scheme. It is very obvious that the proposed adaptation scheme has consistently improves the utility over the period of time t.

The efficiency of the proposed scheme increases with increase in time and it converges faster. After the scheme converged, the network utilization remains relatively constant as it can be seen in Fig. 7. However, the power at which every secondary user transmit is fully controlled in order to ensure that the users coexist together - without affecting the performance of other SUs. Moreover, there is need for effective power management as the network condition dynamically changes in order to ensure fairness to both the primary and secondary users. The proposed scheme achieve more better performance with the update, but the power consumption of wireless applications is growing exponentially which requires highly efficient techniques for power control. The fundamental basic findings of this research will be used in developing more better power control strategies which control the transmitted power, and it will improve the network capacity and reduces network cost. Therefore, the use of nature-inspired approaches for energy-efficient resource allocation in cognitive radio network will be employ and fully exploited.

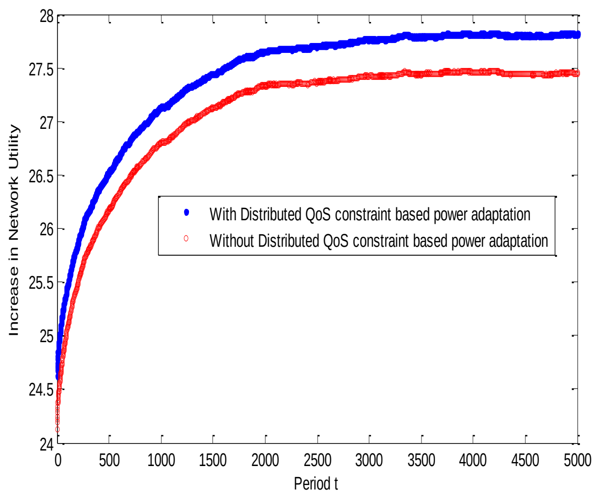


Fig. 7. Distributed QoS constraint based power control scheme with and without adaptation.

On a final note, both high convergence and network utility have been achieved using the proposed distributed QoS constraint based power control scheme. Undoubtedly, as the number SU increases, the network complexity will increase which eventually makes the decision more difficult and challenging. More intensive research work need to be conducted on how to enhance the performance of the proposed scheme further using energy-efficient nature-inspired optimization algorithms for effective power control and management in cognitive radio network.

VI. Conclusion

In this paper, it primarily focuses on the development of distributed QoS constraint based power control scheme in order to maximize network utility in multi-user cognitive radio network. The optimization problem has been formulated and solved to ensure coexistence between the primary and secondary users within the network and subsequently it enhances the secondary user utility. This work has considered the power constraint from the PU so that the SU will sense and utilize the spectrum within the thresholds limits in order to comply with the QoS requirements. Numerically, it has been shown that the proposed scheme effectively improves the performance of the multi-user cognitive radio network. This eventually helps in effective and efficient sensing and management of the SUs with relative ease and sophistication. The distributed QoS constraint based power control scheme has been proposed and its performance is validated numerically. Our future will primarily focus on the usage nature-inspired approach to cooperatively enhance the performance of the network. The behavior of swarms will be mimic and adapted in order to effectively coordinate and manage the multi-users cognitive radio network to achieve higher performance and to reduce the network complexity.

Acknowledgment

The author would like to thank all those who contributed toward making this research successful. Also, I would like to thanks the reviewers for their insightful and valuable comment. This work is supported by the Deanship of Scientific Research (DSR), King Abdulaziz University, Saudi Arabia, under grant No. 830-129-D1437. The author is grateful to the DSR for their technical and financial support.

References

[1] J. Mitola and G. Q. Maguire, "Cognitive radio: Making software radios more personal," IEEE Personal Communications, vol. 6, 1999.

[2] E. C. Y. Peh, Y.-C. Liang, Y. L. Guan, and Y. Zeng, "Optimization of cooperative sensing in cognitive radio networks: A sensing-throughput tradeoff view," IEEE Transactions on Vehicular Technology, vol. 58, 2009.

[3] Z. Quan, S. Cui, A. H. Sayed, and H. V. Poor, "Optimal multiband joint detection for spectrum sensing in cognitive radio networks," IEEE Transactions on Signal Processing, vol. 57, 2009.

[4] H. Zhang, C. Jiang, N. C. Beaulieu, X. Chu, X. Wen, and M. Tao, "Resource allocation in spectrum-sharing OFDMA femtocells with heterogeneous services," IEEE Transactions on Communications, vol. 62, 2014.

[5] J. C. F. Li, W. Zhang, and J. Yuan, "Opportunistic spectrum sharing in cognitive radio networks based on primary limited feedback," IEEE Trans. Commun., vol. 59, 2011.

[6] R. Zhang, "On active learning and supervised transmission of spectrum sharing based cognitive radios by exploiting hidden primary radio feedback," IEEE Trans. Commun., vol. 58, 2010.

[7] Y. He and S. Dey, "Power allocation in spectrum sharing cognitive radio networks with quantized channel information," IEEE Trans. Commun., vol. 59, 2011.

[8] G. Foschini and Z. Miljanic, "A simple distributed autonomous power control algorithm and its convergence," IEEE Transactions on Vehicular Technology, vol. 42, 1993.

[9] R. D. Yates, "A framework for uplink power control in cellular radio systems," IEEE Journal on Selected Areas in Communications, vol. 13, 1996.

[10] J. Huang, R. Berry, and M. Honig, "Distributed interference compensation for wireless networks," IEEE Journal on Selected Areas in Communications, vol. 24, 2006.

[11] M. Chiang, "Balancing transport and physical layers in wireless multihop networks: jointly optimal congestion control and power control," IEEE Journal on Selected Areas in Communications, vol. 23, 2005.

[12] S. Kandukuri and S. Boyd, "Optimal power control in interference limited fading wireless channels with outage-probability specifications," IEEE Transactions on Wireless Communications, vol. 1, 2002.

[13] M. Chiang, C. W. Tan, D. Palomar, D. O'Neill, and D. Julian, "Power control by geometric programming," IEEE Transactions on Wireless Communications, vol. 6, 2007.

[14] W. Ren, Q. Zhao, and A. Swami, "Power control in cognitive radio networks: how to cross a multi-lane highway," IEEE Journal on selected area in communications, vol. 27, 2009.

[15] S. Srinivasa and S. Jafar, “Soft sensing and optimal power control for cognitive radio,” IEEE Global Telecommunications Conference(GLOBECOM), 2007.

[16] W. Wang, Y. Cui, T. Peng, and W. Wang, “Noncooperative power control game with exponential pricing for cognitive radio network,” in Proc. 2007 IEEE Vehicular Technology, 2007.

[17] S Ni.Q. Sun, J.X. Di, and W.M.. “Distributed power control based on convex optimization in cognitive radio networks,” In International Conference onWireless Communications and Signal Processing (WCSP), Suzhou, 2010.

[18] Y.N. Mei, Y.H. Lu, X.M. Mu, and X. Liu. “Distributed fast convergent power allocation algorithm in underlay cognitive radio networks,” Wireless Internet, vol 98, 2012.

[19] C.G. Yang, J.D. Li, and Z. Tian. “Optimal power control for cognitive radio networks under coupled interference constraints: a cooperative game-theoretic perspective,” IEEE Transactions on Vehicular Technology, vol 59, 2010.

Robust Modeling and Linearization of MIMO RF Power Amplifiers for 4G and 5G Applications

Imene ZEMZEMI, Souhir LEJNEF
Dept. of Physics, Faculty of Sciences of Tunis,
University of Tunis El Manar,
Tunis, Tunisia

Noureddine BOULEJFEN
Research Center for Microelectronics and Nanotechnology,
Technopole Sousse,
Sousse, Tunisia

M.Fadhel GHANNOUCHI
iRadio Laboratory, Electrical Engineering Department, Schulich Engineering School,
University of Calgary 2500 University Drive NW, Calgary, Alberta 2TN 1N4, Canada

***Abstract*—In this paper, a novel set of orthogonal crossover polynomials for the baseband modelling and linearization of MIMO RF Pas is presented. The proposed solution is applicable to WCDMA and LTE applications. The new modelling approach has considerably reduced the numerical instability problem associated with the conventional polynomial model identification. In order to validate the efficiency and the robustness of the proposed solution, a 2x2 MIMO LDMOS RF power amplifier has been measured modelled and linearized. A comparison with the conventional polynomial MIMO models showed clearly the superiority of the proposed solution in a fixed-point calculation environment such as DSP and FPGA boards.**

Keywords—MIMO transmitter; RF power amplifiers; orthogonal polynomials; nonlinear transmitters; digital predistortion

I. Introduction

Power amplifiers (PAs) are the major source of nonlinearity in communication system that causes spectral regrowth as well as in-band distortion. Accurate modeling of the RF PAs is required which increases the problem size and reduces the numerical stability of the model identification procedure. This problem is more pronounced in multiple input systems. Multiple-input multiple-output (MIMO) transceivers allow high service quality and increase the capacity range of wireless transmission requiring very high-speed data transfer [1]–[3]. In fact, the capacity of MIMO transceivers M times the capacity of a single-input-single-output (SISO) equivalent system. Volterra series were developed in [4]–[7] and were intensively used in modelling power amplifiers and DPD. However, the Volterra series involves a great number of coefficients, which increases the complexity of the problem in the case of MIMO systems.

In the literature, several attempts reduce the number of coefficients for PA models. In [8] a dynamic deviation reduction model has been suggested for modelling single input PAs. In [9]–[13], reduced polynomial models have been proposed for modelling MIMO PAs. Saffar et al. [14], addressed the joint mitigation of I/Q modulator impairment and PA nonlinearity in MIMO transmitters through an optimized memory polynomial model. However, all these published models suffer from the increasing numerical instability as the problem size increases. This generally happens when a high nonlinearity order, a high memory depth, or a big number of inputs were considered.

Radio frequency (RF) PAs presents a challenge to the transceiver designers. In fact, designers need to boost two contradictory parameters of the PA such as power efficiency and linearity. Several linearization techniques are proposed in the literature to improve PA's linearity. Due to its simplicity and efficiency, the digital pre-distortion (DPD) is considered as the most popular linearization technique. Hence, the DPD approach has been intensively used to compensate for the transmitter nonlinearities as in [15]–[17]. However, the DPD technique requires an accurate modelling of the PA. In fact, high polynomial order leads generally to an accurate model with high complexity. This generally increases the numerical instability and vulnerability of the model identification procedure [18], [19]. Therefore, a tradeoff is always required between the model accuracy and complexity. A good metric for measuring the model accuracy is the time domain normalized mean squared error (NMSE) while the DPD performance can be measured using the frequency domain adjacent channel power ratio (ACPR) of the compensated system [20], [21]. However, the model identification procedure is based on the inversion of an observation matrix that has to be well conditioned in order to avoid numerical instability. The condition number is a metric to test the conditioning of this matrix by measuring the linear dependence of its columns [18], and is generally related to the distribution of the input signal envelope, as well as the dimensionality of the problem. In MIMO PAs, the problem size increases drastically with the number of input signals, making the instability issue more pronounced. This is mainly due to the high correlation that can exist between the input data resulting in ill-conditioned observation matrix with a high condition number. In this work, the numerical instability of the model identification procedure is addressed through the development of a complex multi-input orthogonal polynomial model.

The organization of the paper is as follows. In Section. 2, we introduce the reference crossover conventional polynomials for modelling the MIMO PAs. In Section 3, we present the derived novel robust orthogonal polynomial model. The experimental validation and performance assessment of the proposed model is given in Section 4. Finally, the conclusion is drawn in Section 5.

II. Modelling MIMO Systems using Conventional Polynomials

In this section, the crossover polynomial model [13] is detailed, and its limitations are discussed for a 2×2 MIMO PA case. The conventional polynomial model is a base band model that has been developed to characterizing nonlinear radio frequency power amplifiers with and without memory effects. The closed form expression that relates the input and output complex envelope signals of the PA is given in [13] such that

$$y_i(n) = \sum_{j=1}^{2}\sum_{q=0}^{Q}\sum_{p=0}^{P}\sum_{k=0}^{K} h_{k,p,q}^{j,i}\, x_j(n-p)\left|x_j(n-p-q)\right|^{2k} \tag{1}$$

where x_j and y_i are the complex envelope input and output signals respectively and $h_{k,p,q}^{i,j}$ are the model coefficients with $i, j \in \{1,2\}$. K is the model nonlinearity and P and Q are the memory depths.

Equations (1) can be expressed in a matrix form such that

$$\begin{bmatrix}\overrightarrow{y_1} & \overrightarrow{y_2}\end{bmatrix} = \begin{bmatrix}\Phi_{\overrightarrow{x_1}} & \Phi_{\overrightarrow{x_2}}\end{bmatrix}\begin{bmatrix}\overrightarrow{H^{1,1}} & \overrightarrow{H^{1,2}} \\ \overrightarrow{H^{2,1}} & \overrightarrow{H^{2,2}}\end{bmatrix} \tag{2}$$

For a two-input single output memory less polynomial conventional model, the above expression becomes.

$$\overrightarrow{y_1} = \begin{bmatrix}\Phi_{\overrightarrow{x_1}} & \Phi_{\overrightarrow{x_2}}\end{bmatrix}\begin{bmatrix}\overrightarrow{H^{1,1}} \\ \overrightarrow{H^{2,1}}\end{bmatrix} \tag{3}$$

where $\Phi_{\overrightarrow{x_1}}$ and $\Phi_{\overrightarrow{x_2}}$ are the data matrices defined for inputs $\{x_1, x_2\}$ and $\overrightarrow{H^{1,1}}$, $\overrightarrow{H^{2,1}}$ are the vectors of complex-valued model coefficients between the inputs and the output $\overrightarrow{y_1}$ such that

$$\overrightarrow{y_1} = \left[y_1(1), y_1(2),, y_1(3), \ldots, y_1(S)\right] \tag{4}$$

where S is the number of input samples.

$$H = \begin{bmatrix}\overrightarrow{H^{1,1}} \\ \overrightarrow{H^{2,1}}\end{bmatrix} = \left[h_0^{1,1}, h_1^{1,1}, \ldots, h_k^{1,1}, h_0^{2,1}, h_1^{2,1}, \ldots, h_k^{2,1}\right]^T \tag{5}$$

is a vector of the model coefficients.

$\Phi = \begin{bmatrix}\Phi_{\overrightarrow{x_1}} & \Phi_{\overrightarrow{x_2}}\end{bmatrix}$ is a matrix such that:

$$\Phi_{\overrightarrow{x_j}} = \begin{bmatrix}\phi_0(x_j(1)) & \cdots & \phi_K(x_j(1)) \\ \vdots & \ddots & \vdots \\ \phi_0(x_j(S)) & \cdots & \phi_K(x_j(S))\end{bmatrix} \tag{6}$$

with $\phi_K(x)$ is given by $\phi_K(x) = x|x|^{2K}$.

The least-squares estimate (LSE) of H for MIMO conventional models can be obtained using the Moore-Penrose pseudo inversion such that

$$\overrightarrow{\mathrm{H}}_{\mathrm{LS}} = (\Phi^{\mathrm{H}}\Phi)^{-1}\Phi^{\mathrm{H}}\vec{y}_1 \tag{7}$$

where $[\cdot]^H$ denotes the Hermitian transpose.

The inversion of the observation matrix $\mathrm{M} = \left[\Phi^{\mathrm{H}}\Phi\right]$ is often imperfectly conditioned. Thus, the inversion of such matrix will undergo numerical errors. This problem is more pronounced and leads to erroneous results when the finite precision calculation is used. Define the condition number of a matrix M as:

$$\rho(M) = \left(\left|\frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}\right|\right) \tag{8}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and minimum singular values of M.

In order to estimate the condition number ρ of the observation matrix $\left[\Phi^{\mathrm{H}}\Phi\right]$, two independent sets $\{x_1, x_2\}$ of 184,239 data samples with normalized Gaussian distribution are used.

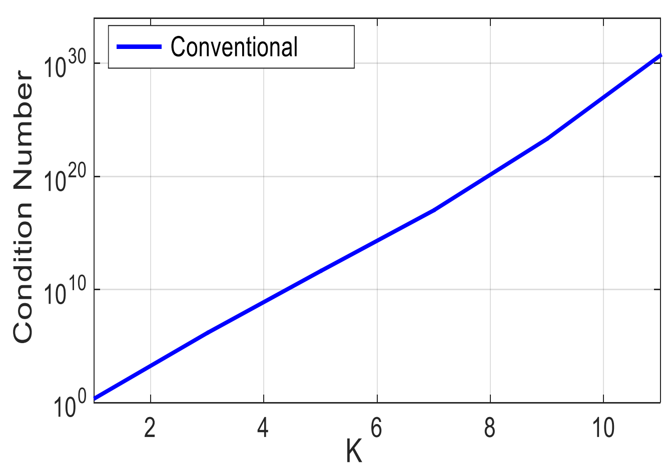


Fig. 1. The condition number of $\left[\Phi^{\mathrm{H}}\Phi\right]$ as a function of K.

Fig. 1 shows clearly that the condition number ρ increases exponentially as a function of the model nonlinearity order. This implies that, in practice, the model can be unstable. Therefore, reduction of the condition number is highly required to ensure the numerical stability of the model coefficient identification procedure. This can be achieved by substituting the set of conventional basis functions $\phi_K(x)$ with an orthogonal set $\psi_K(x)$ leading to an observation matrix with lower ρ.

III. Orthogonal Polynomials Based Model for MIMO RF Power Amplifiers

In this section, we propose a new set of orthogonal polynomial basis functions for modelling PAs excited with two independent signals. To adhere to the statistics of the widely used communication signals such as WCDMA and LTE, Gaussian distributions have been considered for the complex envelops of the RF inputs. The proposed solution is expected to reduce the condition number of the observation matrix and hence ensure the numerical stability of the model identification procedure.

Considering the memory less case and in an attempt to alleviate the numerical instability problem associated with the inversion of the observation matrix in (6) let's consider

$$y_i(n) = \sum_{j=1}^{2}\sum_{k=0}^{K} b_k^{j,i} \sum_{l=0}^{k} u_{j,lk}\, x_j(n) \left|x_j(n)\right|^{2k} \tag{9}$$

where x_j and y_i are the complex envelope input and output signals respectively and $b_k^{i,j}$ are the model coefficients, $u_{j,lk}$ are the coefficients of the orthogonal model, P and Q are the memory depths, K is the model nonlinearity, with $i, j \in \{1,2\}$.

The new data matrix is given by

$$\Psi = \begin{bmatrix} \Psi_{\vec{x}_1} & \Psi_{\vec{x}_2} \end{bmatrix} = \begin{bmatrix} \Phi_{\vec{x}_1} & \Phi_{\vec{x}_2} \end{bmatrix} \mathrm{U} \tag{10}$$

with U is a $(2k+2)\times(2k+2)$ matrix.

We consider the following requirements for the orthogonality of the suggested basis functions: $\Psi_k(x)$ and $\Psi_\ell(x)$ to be orthogonal, the following condition has to be satisfied:

$$\begin{cases} \mathrm{E}\left[\Psi_k^*(x)\ \Psi_\ell(x)\right] = 0 & if\ \ k \neq \ell \\ \mathrm{E}\left[\Psi_k^*(x)\ \Psi_\ell(x)\right] = 1 & if\ \ k = \ell \end{cases} \tag{11}$$

Where, E[.] denotes statistical expectation, and (*) stands for complex conjugation. Therefore, finding the appropriate orthogonal polynomial basis functions returns to finding the U matrix such that

$$\mathrm{E}\left[\Psi^H \Psi\right] = \mathrm{U}^H \mathrm{E}\left[\Phi^H \Phi\right] \mathrm{U} = S\mathrm{U}^H \mathrm{M}\ \mathrm{U} \tag{12}$$

is diagonal, with $\mathrm{M} = \mathrm{E}\left[\Phi^H \Phi\right]$ is a $(2K+2)\times(2K+2)$ matrix. The resolution of the problem returns to find the elements of the matrix U such that $\mathrm{U}^H \mathrm{M}\ \mathrm{U} = I_d$.

To find the U matrix elements for the proposed 2×2 MIMO model consider x_1 and x_2 two independent complexes Gaussian input signals with zero means and variances σ_{x1} and σ_{x2} respectively. The $\mathrm{E}\left[\Phi^H \Phi\right]$ matrix can then be given by

$$\mathrm{E}\left[\Phi^H\ \Phi\right] = \begin{bmatrix} \mathrm{R}_1 & 0_{k+1} \\ 0_{k+1} & \mathrm{R}_2 \end{bmatrix} \tag{13}$$

where 0_{k+1} is a $(k+1)\times(k+1)$ sub matrix with zero elements. R_1 and R_2 are two $(k+1)\times(k+1)$ symmetric sub-matrices such that for i={1,2} we have

$$\mathrm{R}_i = \mathrm{E}\begin{bmatrix} |x_i|^2 & \cdots & |x_i|^{2k+2} \\ \vdots & \ddots & \vdots \\ |x_i|^{2k+2} & \cdots & |x_i|^{4k+2} \end{bmatrix} \tag{14}$$

To construct the required set of basis functions the U matrix is proposed to be a $(2k+2)\times(2k+2)$ matrix such that

$$\mathrm{U} = \begin{bmatrix} \mathrm{U}_1 & 0_{k+1} \\ 0_{k+1} & \mathrm{U}_2 \end{bmatrix} \tag{15}$$

Where, U_1 and U_2 are $(k+1)\times(k+1)$ upper triangular sub-matrices such that

$$\mathrm{U}_i = \begin{bmatrix} u_{\mathrm{i},00} & \cdots & u_{i,0k} \\ 0 & \ddots & \vdots \\ 0 & 0 & u_{i,lk} \end{bmatrix} \tag{16}$$

and 0_{k+1} is a null $(k+1)\times(k+1)$ sub matrix. For the proposed 2×2 MIMO model with two independent inputs x_1 and x_2 with zero mean and σ_{x_i} variances. The $u_{i,lk}$ elements of the matrix U_i can be given by [19].

$$\mathrm{U}_i = \begin{cases} \dfrac{(-1)^{k-\ell}}{\sigma_{x_i}^{\,2k+1}} \dfrac{\sqrt{k+1}}{(\ell+1)!} \begin{pmatrix} \ell \\ k \end{pmatrix} & \ell \leq k \\ 0 & \ell > k \end{cases} \tag{17}$$

with i={1,2} refers to the inputs of the PA.

A. N×N MIMO Orthogonal Crossover Model

The suggested Crossover orthogonal model can be extended to $N\times N$ MIMO transmitters such that:

$$\left[\vec{y_1} \quad \vec{y_2} \cdots \vec{y_N}\right] = \left[\Phi_{\vec{x_1}} \quad \Phi_{\vec{x_2}} \ldots \Phi_{\vec{x_N}}\right] \begin{bmatrix} \overrightarrow{H^{1,1}} & \cdots & \overrightarrow{H^{N,1}} \\ \vdots & \ddots & \vdots \\ \overrightarrow{H^{1,N}} & \cdots & \overrightarrow{H^{N,N}} \end{bmatrix} \tag{20}$$

with N is the number of inputs and outputs of the MIMO transmitter.

For MISO transmitters with N inputs and a single output, expression (9) leads to the following matrix form:

$$\vec{y_1} = \left[\Phi_{\vec{x_1}} \quad \Phi_{\vec{x_2}} \ldots \Phi_{\vec{x_N}}\right] \begin{bmatrix} \overrightarrow{H^{1,1}} \\ \vdots \\ \overrightarrow{H^{1,N}} \end{bmatrix} \tag{21}$$

In order to convert the observation matrix to a diagonal one that ensures the numerical stability of the model identification procedure, the F matrix is substituted with a Y matrix such that

$$\Psi = \left[\Psi_{\vec{x_1}} \quad \Psi_{\vec{x_2}} \ldots \Psi_{\vec{x_N}}\right] = \left[\Phi_{\vec{x_1}} \quad \Phi_{\vec{x_2}} \ldots \Phi_{\vec{x_N}}\right] \mathrm{U} \tag{22}$$

with U is an $(\mathrm{N}k+\mathrm{N})\times(\mathrm{N}k+\mathrm{N})$ orthogonal matrix. U is composed of sub matrices U_i placed in the diagonal line with $i=\{1,2,\ldots,\mathrm{N}\}$,such that:

$$\mathrm{U} = \begin{bmatrix} \mathrm{U}_1 & 0_{k+1} & \cdots\cdots & 0_{k+1} \\ 0_{k+1} & \mathrm{U}_2 & & \vdots \\ \vdots & & \ddots & 0_{k+1} \\ 0_{k+1} & \cdots & 0_{k+1} & \mathrm{U}_N \end{bmatrix} \tag{23}$$

B. Numerical Validation

In this paper, we proposed a closed-form expression for orthogonal polynomials to model MIMO Pas excited by RF signals with Gaussian complex envelops such as WCDMA and LTE signals. To demonstrate the efficiency of the proposed set of basis functions in reducing the risk of numerical instability, the condition number of the resulting observation matrix has been calculated. To do so 184,239 independent realizations of a four-channel WCDMA1001 and a four-channel WCDA1111 have been generated as input signals x_1 and x_2 respectively and used to calculate the $\Phi^H\Phi$ and $\Psi^H\Psi$ observation matrices for different non-linearity orders K. The '1' refers to an ON channel while the '0' refers to an OFF channel. Fig. 2 shows the condition number for the memory less 2×2 MIMO.

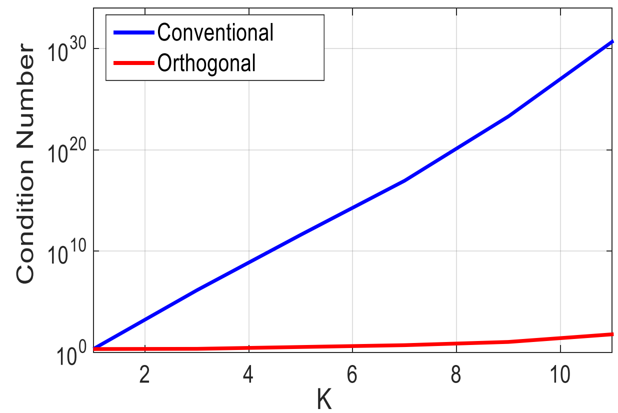


Fig. 2. The condition number of $\Phi^H\Phi$ and $\Psi^H\Psi$ for the 2×2 MIMO case.

Fig. 2 shows clearly that for the conventional model the condition number grows exponentially with the non-linearity order K to reach 10^{30} for K=11. However, for the proposed orthogonal model, the condition number increases at a much lower rate without exceeding 100 for the same range of K.

Hitherto, the above-described numerical simulations have proven the remarkable performance of the proposed orthogonal polynomial basis functions for crossover 2×2 MIMO models. However, an experimental validation is required to verify the impact of the numerical stability on the model accuracy.

IV. Experimental Validation

The performances of the proposed set of orthogonal basis functions have been evaluated by modelling and linearizing a 2x2 RF power amplifier. To do so the experimental setup shown in Fig. 3 has been developed. We generated two signals using two vector signal generators (VSG) ESG1 and ESG2 of type E4438C in order to excite the MIMO transmitter. The latter is equivalent to two drivers followed by two class-AB RF PAs and two attenuators. Two couplers have been used to introduce a non-linear cross talk as shown in the figure. A vector signal analyzer (VSA) of type E4440A is then used to collect and analyze one of the two attenuators' outputs through the RF switch.

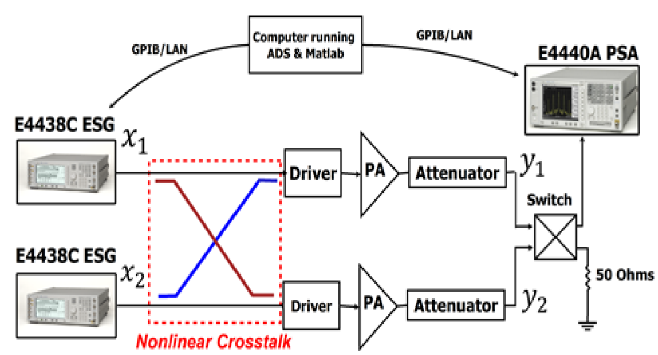


Fig. 3. Block Diagram of a MIMO PA measurement setup.

The measurement is performed by connecting the sources via a (GPIB) bus. The I/Q components of the two multi-carrier wideband code division multiple access (WCDMA) signals (WCDMA1001 and WCDMA1111) are generated using Matlab then downloaded to the source with an average power equal to 9 dBm and a bandwidth of 20 MHz. The output signal is down-converted, sampled and digitized at a sampling frequency of (*fs*= 92.16MHz) for a time window of 2ms leading to a data set of 184,239 samples. The non-linear crosstalk is set to −20 dB

C. Key Performance Metrics

To evaluate the accuracy of the proposed MIMO model in predicting the transmitter output the time domain normalized mean square error (NMSE) has been used. The NMSE [20], [21] is expressed in the logarithmic scale as shown in the following equation:

$$NMSE_{dB} = 10\log_{10}\left[\frac{\sum_{n=1}^{N}\left|y_i(nT_S) - y_m(nT_S)\right|^2}{\sum_{n=1}^{N}\left|y_i(nT_S)\right|^2}\right] \qquad (24)$$

where $y_i(nT_S)$ and $y_m(nT_S)$ are respectively the complex envelopes of the measured and modelled output signals. In the other hand, a frequency domain metric such as the adjacent channel power ration (ACPR) can be used to evaluate the spectral regrowth in the output spectrum as well as the performances of the proposed model based digital predistorter. The ACPR defined as follows [20], [21]:

$$ACPR_{dB} = \frac{\int_{Adj.ch}\left|Y(f)\right|^2 df}{\int_{main.ch}\left|Y(f)\right|^2 df} \qquad (25)$$

The adjacent channel power ration (ACPR) is measured for the adjacent channels below and above the main carrier in dBc.

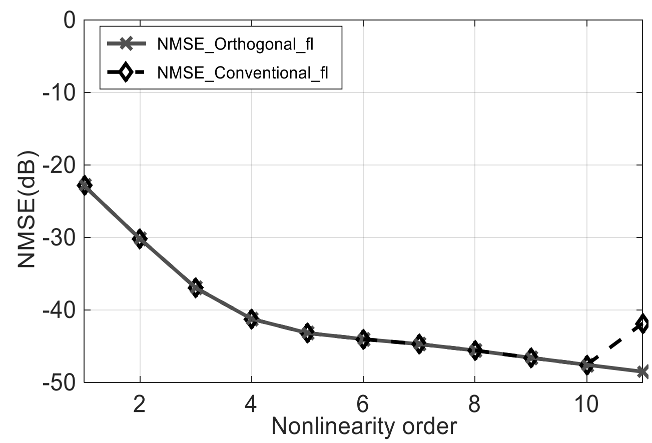
D. Forward Model Experimental Results

The importance of a low condition number resides in the fact that the PA model is generally used in the PAs linearization operation. The linearization algorithm runs on a fixed-point processor like DSP or FPGA with a limited number of bits. In fact, the fixed-point processors are efficient with low computation time, cost and power consumption. However, during the modelling process, the model is generally simulated with a floating-point processor.

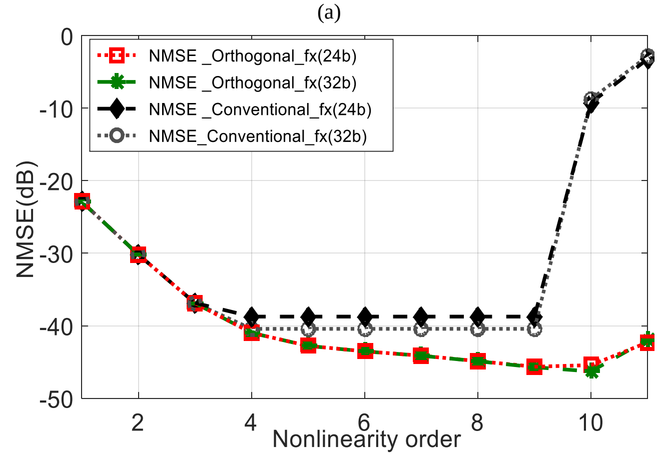
The nonlinearity order is set to 11 and the memory depths *P* and *Q* are set to 2 and 3 respectively. Under these conditions, three different calculation environments have been considered.

- Scenario 1: floating point calculation.
- Scenario 2: fixed point calculation with a fraction length of 32 bits.
- Scenario 3: fixed point calculation with a fraction length of 24bits.

As a first test, the coefficients of the conventional and the proposed orthogonal models for the PA under test have been identified in a floating calculation environment. The models' outputs are then compared to the measured PA output and the NMSE have been calculated. Fig. 4(a) shows the NMSE of the two models. The figure reveals that the two models exhibit a comparable accuracy with a pretty similar NMSEs that reached -47.5dB for K = 10. However, the NMSE of the conventional model bounced back to -42dB for K = 11 while the one of the proposed models continued its fall to -48.53dB.



(a)



(b)

Fig. 4. Measurement of the NMSE of scenario 1(a), scenario 2 and scenario 3(b).

In a second test, the same procedure has been repeated using a 64 bits fixed-point processor with fractional lengths of 32 bits and 24 bits. Fig. 4(b) shows The NMSEs for the two models and for the two fraction lengths. The figure demonstrates clearly that the performance of the conventional model deteriorates for the two cases leading to NMSEs diverging to near -3dB for K = 11. However, the proposed orthogonal model maintained its accuracy with NMSEs below -40dB for the whole range of *K*.

E. Digital Pre-Distorter Experimental Results

Once the coefficients of the model are identified, the PA output can be predicted under different excitation signals. However, in an attempt to linearize the PA and improve its efficiency-linearity compromise, a reverse model can be developed and a digital predistorter (DPD) can be obtained using the indirect learning architecture (ILA). As shown in Fig. 5, the ILA consists of identifying and updating the DPD coefficients using the LS algorithm while exchanging the PA inputs and outputs. The resulting DPD can then linearize the PA and improve the quality of its output spectrum. To validate the effectiveness of the proposed orthogonal model in the linearization of MIMO PAs, a nonlinearity order of 11 and memory depths *P* and *Q* of 2 and 3, respectively were considered. The DPD is then developed and applied using the three different scenarios of calculation environment.

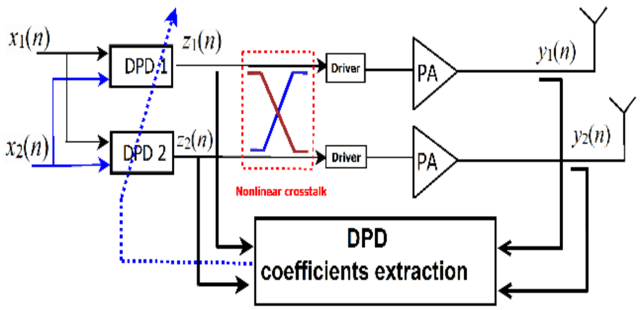


Fig. 5. Block Diagram of a MIMO PAs + DPD measurement setup.

TABLE I. MEASURED ACPR IN DBC FOR THE FIRST SCENARIO

Signal	WCDMA1001		WCDMA1111	
	L	R	L	R
Without DPD	-33.42	-32.76	-34.42	-34.61
With DPD (conventional model)	-42.25	-42.60	-43.56	-43.81
With DPD (orthogonal model)	-49.16	-49.31	-50.21	-50.30

TABLE II. MEASURED ACPR IN DBC FOR THE SECOND SCENARIO

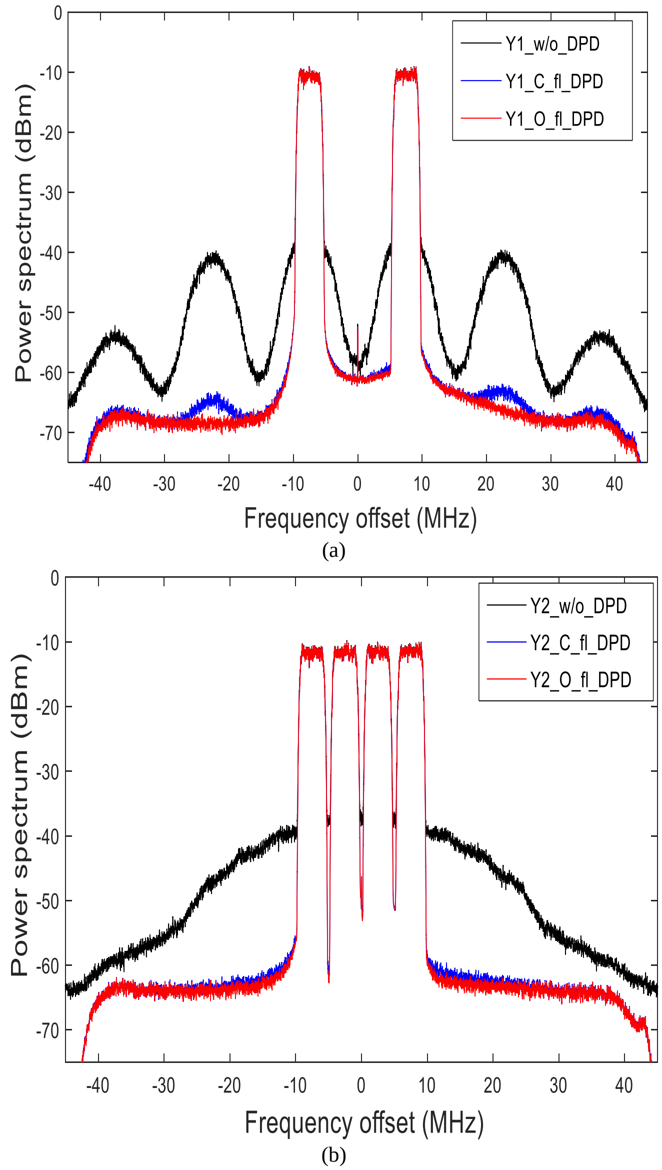
Signal	WCDMA1001		WCDMA1111	
	L	R	L	R
Without DPD	-28.52	-28.92	-28.76	-28.33
With DPD (conventional model)	-20.29	-20.32	-20.34	-20.79
With DPD (orthogonal model)	-42.97	-43.1	-43.52	-43.86

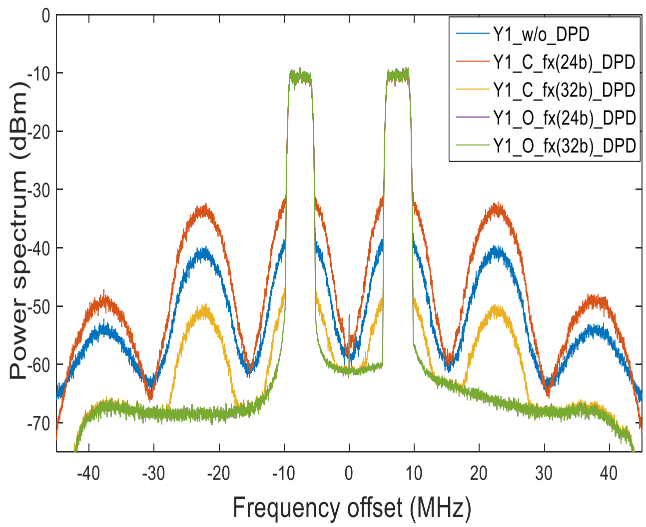
TABLE III. MEASURED ACPR IN DBC FOR THE THIRD SCENARIO

Signal	WCDMA1001		WCDMA1111	
	L	R	L	R
Without DPD	-27.58	-27.92	-27.77	-27.91
With DPD (conventional model)	-18.37	-18.41	-19.21	-19.34
With DPD (orthogonal model)	-41.87	-41.62	-42.64	-42.32

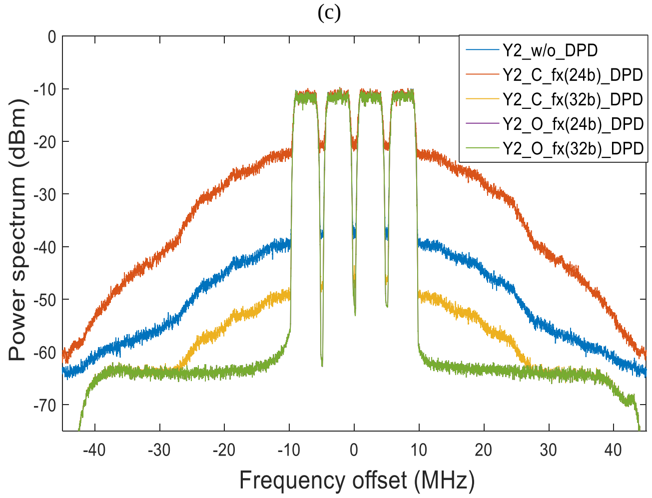
Tables I, II and III show the adjacent channel power ratios (ACPR) of the 2x2 MIMO PA outputs with and without DPD. The table revealed the high values (above -28dBc) of the ACPR for the different scenarios when the DPD is turned OFF. Under the floating-point environment calculation, the conventional and the proposed orthogonal model based DPDs performed pretty well with an ACPR below -42dBc for the two PA outputs. In addition, the results revealed a better performance of the orthogonal model based DPD with an ACPR as low as -49dBc.

Under the fixed-point calculation environment, the performance of the conventional model-based DPD deteriorates significantly, leading to an ACPR exceeding -19dBc. However, the orthogonal model based DPD maintained its good performance with an ACPR below -41dBc for the two PA outputs and for the two fraction lengths of 24 and 32 bits





(c)



(d)

Fig. 6. Frequency domain performance for two channels for 2×2 MIMO transmitter (a) and (b) outputs of DPD for scenario 1, (c) and (d) outputs of DPD for scenario 2 and 3.

In Fig. 6 the power spectral densities of the PA outputs has been plotted for the different scenarios where Y*i*_w/o_DPD are the measured PA outputs when the DPD is turned OFF (Black); Y*i*_C_fl_DPD are the predicted PA outputs when the conventional model based DPD is applied using a floating point calculation environment (Blue); Y*i*_O_fl_DPD are the predicted PA outputs when the orthogonal model based DPD is applied using a floating point calculation environment (Red); Y*i*_C_fx(24b)_DPD are the predicted PA outputs when the conventional model based DPD is applied using a fixed point calculation environment with a fraction length of 24 bits (Orange); Y*i*_C_fx(32b)_DPD are the predicted PA outputs when the conventional model based DPD is applied using a fixed point calculation environment with a fraction length of 32 bits (Yellow); Y*i*_O_fx(24b)_DPD are the predicted PA outputs when the orthogonal model based DPD is applied using a fixed point calculation environment with a fraction length of 24 bits (Violet); Y*i*_O_fx(32b)_DPD are the predicted PA outputs when the orthogonal model based DPD is applied using a fixed point calculation environment with a fraction length of 32 bits (Green), with $i = \{1,2\}$ denotes the number of outputs.

Fig. 6 shows the power spectral densities of the predicted PA outputs for the different scenarios. The measured outputs of the nonlinearity PA, shown in Fig. 6(a)-(b), confirm the spectrum regrowth caused by the PA nonlinearity. In the same time, the figures reveal quite similar and perfect PA linearization when using a floating-point processor (scenario 1) regardless of the DPD model. However, Fig. 6(c)-(d) shows the spectra of the same signals in the cases of scenario 2 and 3. The figures reveal that when using a fixed-point processor, the conventional model-based DPD loses completely its performance and fail to linearize the PA. However, the proposed orthogonal model based DPD maintains its good performances and succeeds to linearize the PA for the 32 bits and 24 bits fraction lengths.

V. Conclusion

In this paper, we proposed closed-form expressions for orthogonal polynomials for MIMO PA modelling under RF signals with Gaussian complex envelopes. The numerical and experimental validations have confirmed the robustness and stability of the proposed model identified in fixed-point calculation environments. The proposed orthogonal model based DPD outperformed the conventional model-based DPD in terms of adjacent channel emission reduction in the presence of a nonlinear coupling in the MIMO PA. Due to its simplicity and closed-form expression, the proposed model can be tuned to fit special cases such as massive MIMO PAs where only the coupling between adjacent inputs needs to be considered.

References

[1] J. Gozalvez, “Samsung Electronics Sets 5G Speed Record at 7.5 Gb/s [Mobile Radio],”*IEEE Vehicular Technology Magazine*,vol.10,no.1, p.12–16, 2015

[2] G.J.Foschini and M. J. Gans, “On limits of wireless communications in a fading environment when using multiple antennas,” *Wireless Pers. Commun*., vol. 6, no. 3, pp. 311–335, Mar. 1998.

[3] N. Doose, P.A. Hoeher, “Massive MIMO Ultra-Wideband Communications Using Multi-Mode Antennas,”*In 10th International ITG Conference on Systems Communications and Coding (SCC), Hambourg (Allemagne)*, pp. 1–6, 2015.

[4] L. Chua and C.-Y. Ng, “Frequency domain analysis of nonlinear systems: general theory*,” IEE Journal on Electronic Circuits and Systems*, vol. 3, no. 4,pp. 165–185, Jul. 1979.

[5] A. K. Swain and S. A. Billings, “Generalized frequency response function matrix for MIMO non-linear systems,” *Int. J.Control*, vol. 74, no. 8, pp. 829–844, 2001.

[6] L. M. Li and S. A. Billings, “Generalized frequency response functions and output response synthesis for MIMO non-linear systems,” *Int.J. Control*, vol. 79, no. 1, pp. 53–62, 2006.

[7] Z. K. Peng, Z. Q. Lang, and S. A. Billings, “Non-linear output frequency response functions for multi-input non-linear volterra systems,”*Int. J. Control*, vol. 80, no. 6, pp. 843–855, 2007.

[8] A. Zhu, J. Pedro, and T. Brazil, “Dynamic deviation reduction-based Volterra behavioral modeling of RF power amplifiers,” *IEEE Trans. Microw. Theory Tech.*, vol. 54, no. 12, pp. 4323–4332, Dec. 2006.

[9] D. Saffar, N. Boulejfen, F.M. Ghannouchi, et al,“Compensation of I/Q Impairments and Nonlinear Distortion in MIMO Wireless Transmitters,”*In 11th IEEE International New Circuits and Systems Conference (NEWCAS), Paris (France)*, pp. 1–4, 2013.

[10] S. A. Bassam, M. Helaoui, F.M. Ghannouchi, “2-D digital predistortion (2-D-DPD) architecture for concurrent dual-band transmitters*,” IEEE*

Transactions on Microwave Theory and Techniques, vol. 59, no. 10, pp. 2547–2553, 2011.

[11] S. A. Bassam, M. Helaoui, F.M. Ghannouchi, “Crossover digital predistorter for the compensation of crosstalk and nonlinearity in MIMO transmitter;”*IEEE Transactions on Microwave Theory and Techniques*, vol. 57, no. 5, pp. 1119–1128, 2009.

[12] E. Zenteno, S. Amin, M. Isaksson, D. Ronnow and P. Handel, “Combating the dimensionality of nonlinear MIMO amplifier predistortion by basis pursuit,”*IEEE EuMC* 2014, Rome, Italy, pp. 822-826, 6-9 Oct. 2014.

[13] S Amin, P. N. Landin, P. Handel, D. Ronnow, “Behavioral Modeling and Linearization of Crosstalk and Memory Effects in RF MIMO Transmitters,”*IEEE Transactions on Microwave Theory and Techniques,* vol. 62, no. 4, pp. 810–823, 2014.

[14] D. Saffar, N. Boulejfen, F.M. Ghannouchi et al, “Behavioral modeling of MIMO nonlinear systems with multivariable polynomials,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 59, no. 11,pp. 2994–3003, 2011.

[15] M. Younes, F.M. Ghannouchi,“Behavioral Modeling of Concurrent Dual-band Transmitters based on Radial-Pruned Volterra Model,”*IEEE Communications Letters*, vol. 19, no. 5, pp.751–754, 2015.

[16] F. Zhu, L. Anttila, M. Abdelaziz, et al., “Frequency-Selective Digital Predistortion for Unwanted Emission Reduction,”*IEEE Transactions on Communications*, vol. 63, no. 1, pp.254–267, 2015.

[17] A. Molina, K. Rajamani, K. Azadet, “Digital Predistortion Using Lookup Tables With Linear Interpolation and Extrapolation: Direct Least Squares Coefficient Adaptation,”*IEEE Transactions on Microwave Theory and Techniques*, vol. 65, no. 3, pp.98–987, 2017.

[18] R. Raich, H. Qian, and G.T. Zhou,“Orthogonal Polynomials for Power Amplifier Modeling and Predistorter Design,”*IEEE Transactions on Vehicular Technology*, vol. 53,no. 5, pp. 1468–1479,Sep. 2004.

[19] R. Raich and G.T. Zhou, “Orthogonal Polynomials for Complex Gaussian Processes”. *IEEE Transactions on Signal Processing*, vol. 52, no. 10, pp. 2788–2797, 2004.

[20] O. Hammi, M.Younes and F. M. Ghannouchi,“Metrics and methods for benchmarking of RF transmitter behavioral models with application to the development of a hybrid memory polynomial model,” *IEEE Transactions on Broadcasting*, vol. 56, no. 3, pp. 350–357, 2010.

[21] C. Nader, P. Landin, W. Van Moer, et al,“Performance evaluation of peak-to-average power ratio reduction anddigital pre-distortion for OFDM based systems,”*IEEE Transactions on Microwave Theory and Techniques*,vol. 59, no. 12, pp. 3504–3511,2011.

Energy Harvesting for Remote Wireless Sensor Network Nodes

Engr. Syed Ashraf Ali
Department of Electrical Engineering
Iqra National University
Peshawar, Pakistan

Engr. Syed Haider Ali
Department of Electrical Engineering
University of Engineering and Technology
Peshawar, Pakistan

Engr. Sajid Nawaz Khan, Engr. Muhammad AAmir Aman
Department of Electrical Engineering
Iqra National University
Peshawar, Pakistan

***Abstract*—Wireless Sensor Network (WSN) technology is widely used for controlling and monitoring purposes. Advancement accomplished in the past era in wireless communications and microsystems have allowed the change of minor degree and least effort sensor nodes outfitted with remote correspondence abilities ready to manufacture a wireless WSN node. Each sensor node is ordinarily outfitted with one or a couple of detecting units, information preparing units, remote correspondence interface and battery. WSNs have discovered application in an extensive variety of various spaces like home and biomedical health monitoring. Providing continuous supply of energy to these nodes at remote locations is a major concern. The aim of this paper is to provide un-interrupted supply of energy to remote WSN nodes. Solar energy is expected to provide the required energy; however, photovoltaic (PV) based system are not able to operate at night. This may influence the operation of WSN nodes, rendering them useless for that instant. Several techniques have been proposed to provide satisfactory energy storage. However, the utilization of a suitable device to provide the required energy storage and operate WSN node for all day long is an open issue. A complete WSN node is developed for flood monitoring with sensing capacity along with energy harvesting using PV system and storage unit, which is able to harvest and store energy for un-interrupted operation of WSN node at remote sites.**

***Keywords*—Wireless sensor network (WSN); photovoltaic (PV)**

I. Introduction

Flood is becoming a major natural disaster in different areas of Khyber Pakhtunkhwa (Pakistan), resulting in significant amount of damage including homes, crops, cattle and human lives. Flood control and monitoring is a major problem in Pakistan, owing for the most part to the consumption and maintenance of the equipment utilized in this regard. At dams and rivers, different water level strategies are executed however tragically; we don't have a warning system for continuous checking and controlling in flood prone zones where human access is constrained. Collecting and processing continuous information has been an interesting subject of installed system design. In any case, to gather information utilizing normal wired sensor systems has dependably been troublesome and costly. The wired system frequently goes excessively too crowded and creates an obstacle to mobility requirements of the applications. In this way, WSN has turned into a head research point in the installed fields is a term used to present a class of implanted specialized communication devices that give dependable remote associations between sensors, processors and actuators. It is developing as an answer for an extensive variety of information assembling and handling applications [1]. A WSN is comprised of nodes that have processing units, sensors, antenna, power source and radio frequency integrated circuit (RFIC) as radio transceivers [2].

WSN typically use battery power as a power source while energy harvesting wireless sensor network (EHWSN) use an energy harvesting system (EHS) as a power source, converting energy from nature to electrical energy using different energy harvesting (EH) techniques. For example electrical energy is converted from solar energy by Photovoltaic cell is most widely recycled in our daily life. Although the output power from the EHS is extremely low (μW-mW) and various over time with the development of low power electronics and energy storage techniques (e.g. low leakage super capacitor), EHWSN becomes reality and attracts more and more researchers attentions [3].

WSN technology is generally utilized for controlling and observing purposes. However, providing continuous supply of energy to these nodes is a major concern. This research aimed to provide un-interrupted supply of energy to remote WSN nodes. Solar energy is expected to provide the required energy with minimum expenses of resources. However PV based system are not able to operate at night. This may influence the operation of WSN nodes, rendering them useless for that instant. A few procedures have been proposed to give satisfactory energy storage. However, the use of a reasonable device to give the required energy storage and work WSN node for throughout the day is an open issue [4].

This paper aims to build up complete WSN nodes that will be fit for sensing (based on specific sensor) and transmitting data, with the assistance of PV based system. The proposed system will also be fit for giving the required energy even at the absence of sun (evening time). This would not just encourage the continuous flood checking process with the use of ease and self-supported assets. In addition, this will likewise help in saving valuable lives by conveying earlier data about any unfortunate flooding events.

II. Energy Harvesting

Batteries are normally used to power the WSN nodes. The node will be dead when their energy is exhausted. Just in exceptionally specific applications batteries can be recharged/replaced. Be that as it may, notwithstanding when this is conceivable, the energizing or replacing operation is moderate, costly and it also drops the network performance. Renewable energy sources are proposed to power WSN nodes at remote sites. Solar energy is a suitable candidate; however, sun is not present at night. Therefore, it is imperative to design a suitable circuit that can harvest and store enough energy to operate WSN node throughout the day. Several energy storage units are proposed to operate with PV system; however, the use of a suitable source for flood monitoring applications is still an open issue.

III. Hardware Design

The main goal of this paper is to create WSN node, which are small devices that collect and transmit data and are often placed in remote areas with the capability to extract energy from ambient sources to last eras of time. The goal of this research is to produce such a WSN node that will make use of a unique energy management scheme to ensure long-lasting operation. A solar cell will be the primary energy source with secondary energy storage devices being a rechargeable battery and super capacitor (Fig. 1).

1) Functions

- Collect on-board electrical data including: solar cell open-circuit voltage, battery voltage and current and super capacitor voltage.
- Transmit and receive the collected data wirelessly.
- Program the microcontroller for testing and final operation.
- Allow the on-board microcontroller to Communicate with each other.
- Recharge the energy storage devices with surplus solar energy.
- Execute a power management program to maximize the operational lifetime of the WSN node.

2) Constraints

The wireless sensor node must adhere to the following constraints:

- Use a super capacitor as an energy storage device.
- Use a rechargeable battery as an energy storage device.
- Operate outdoors during the day under varying lighting conditions and during the night.
- Operate autonomously once deployed.

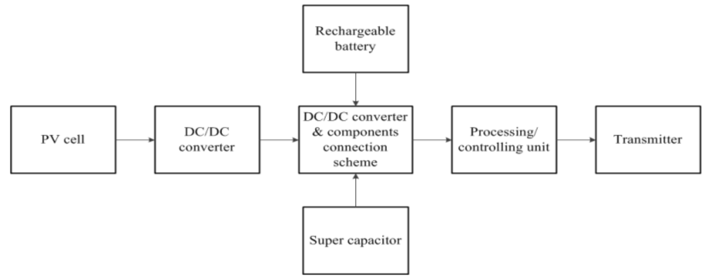


Fig. 1. Higher-level blocks required for the Wireless Sensor Node.

The DC/DC Converter with MPPT Functionality interfaces with the photovoltaic cell to extract solar energy [5]. The component connection scheme contains hardware that will charge and discharge the energy storage devices. The yield of the component connection scheme powers devices, specifically the sensors used to collect data, the transmitter and the controlling units. Data is likewise exchanged among the sensors, transmitter and logic units [6].

IV. Energy Storage and Delivery Module

The Energy Storage and Delivery Module are responsible for:

- Storing input energy from the Solar Boost Converter into either of the onboard energy sources.
- Transferring energy from the Solar Boost Converter or the onboard energy sources to the load on regulated 5V providing signals for the battery voltage, battery current and super capacitor voltage to be used.
- Managing energy viably to guarantee the WSN manages operation once conveyed.

TABLE I. WSN Power Management Operating Modes [7]

Operating Mode	Energy Source Available	Conditions
Normal Day.	Battery, Super Capacitor, PV.	Sufficient ambient light to power the PVs.
Normal Night.	Battery, Super Capacitor.	Insufficient ambient light to lower the PVs.
Charge Battery.	Super Capacitor, PV.	Day operation, PVs Provide enough power to support load and charge both Super Capacitor and Battery.
Charge Super Capacitor.	P.V and /or Battery.	Insufficient power from PVs to support load from PVs and Capacitor only.
Converter Off.	P.V and/ or Super Capacitor.	Super Capacitor has sufficient energy to support load.
Emergency.	P.V and/ Battery or Super Capacitor.	All energy devices unable to support load.

V. Bi-Directional Non-Inverting Buck-Boost Converter

This converter transfers energy between the battery and super capacitor and supports bidirectional power flow. The converter regulates the battery current for charging and discharging purposes (Fig. 2).

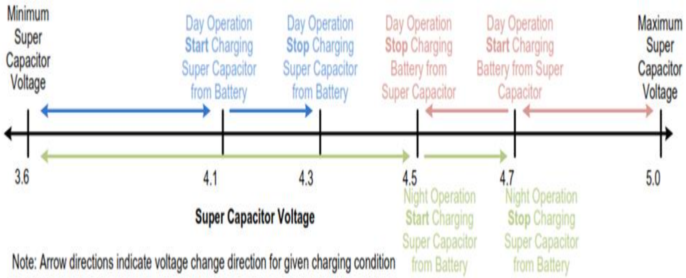


Fig. 2. The voltage conditions that dictate the WSN power management operating mode [11].

VI. Assessment of Proposed Design

The entire design is split into three main sections: energy collection, energy storage and delivery, and energy consumption. Within each of these sections very specific tasks and goals were established to ensure the requirements were met. The energy collection stage needs to efficiently catch solar energy and convey it to the next stage. The Energy Storage and Delivery stage must efficiently and cautiously transfer energy throughout the system. Lastly, the Energy Consumption stage must run the system as proficiently as could be allowed and do its end utilize dependably. A common theme throughout is the careful use of energy: each section is equally responsible for operating efficiently to ensure the WSN node can last for potentially months at a time. Much effort was spent amid the outline/parts-choice stage to secure low power parts that would yield acceptable results. Take note that while the three main sections all depend on each other to work as a whole, each section are modular enough to operate on its own provided suitable input energy is available. To be more particular, there is negligible interconnection between the three modules to ensure successful system integration [8]. Fig. 3 illustrates the relationship between sections: the design leaves little to chance. Each large section can be developed on its own and nearly seamlessly come together. Combining the three large sections was quite trivial relative to the development of every module.

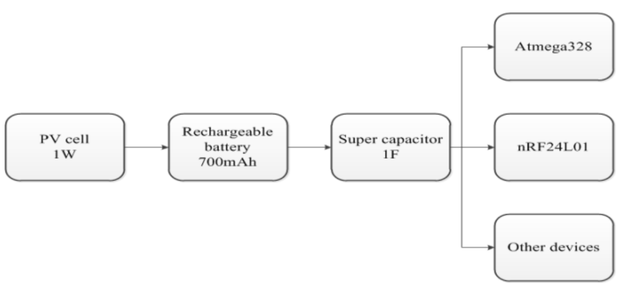


Fig. 3. WSN node design flow.

VII. Power Calculation

To spare restricted power assets, the microcontroller of WSN node remains for huge measure of time in a sleep mode, occasionally changing to an active mode to examine information from sensors or to communicate by means of radio interface. Accordingly, the lifetime of the node for the system without static power supply, would depend also on the node general power utilization and on control source discharge qualities. The power consumption of a node is determined by the consumption of the microcontroller, radio and the peripheral devices for active and sleep modes and on system scheduling. Microcontroller in WSN node is the key part which controls all the work of peripherals and radio communication. Contingent upon application, WSN node is additionally regularly required to make a few information preparing before sending the information to the receiver. In several works it has been shown that in many cases data processing on the node allows reducing the amount of the information, which is required to be transmitted over WSN node. This allows the improvement of the power consumption of the whole WSN node as radio communication usually has higher power consumption than the data processing [9].

The design node has the property that after 1 minute, it goes to sleep mode for 10 minutes. So according to that calculation, the node will be in active mode for 2.4 hours in 24 hours and the remaining 21.6 hours, it will be in sleep mode.

Node has the input voltage=5V.
Current across the node (Active mode) =97.6mA.
Current across the node (Sleep mode) =72.1mA.
For the measurement of power at active mode=97.6mA*5V*2.4h=1,171.2mWh.
For the measurement of power at sleep mode=72.1mA*5V*21.6h=7,786.8mWh.
$Power_{Total}= Power_{Active}+ Power_{Sleep.}$
$Power_{Total}$= 1,171:2mWh + 7,786:8mWh.
$Power_{Total}$= 8,958mWh.

The reading of the node active mode and sleep mode are shown in Fig. 4 and 5.

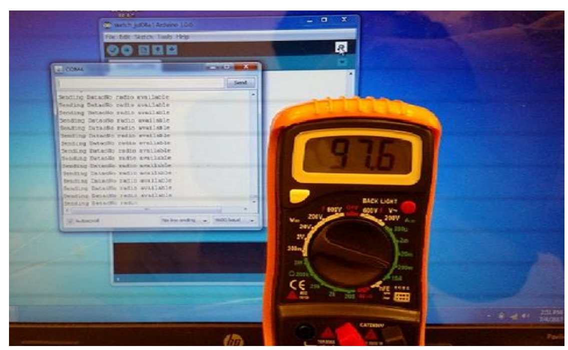


Fig. 4. Active mode.

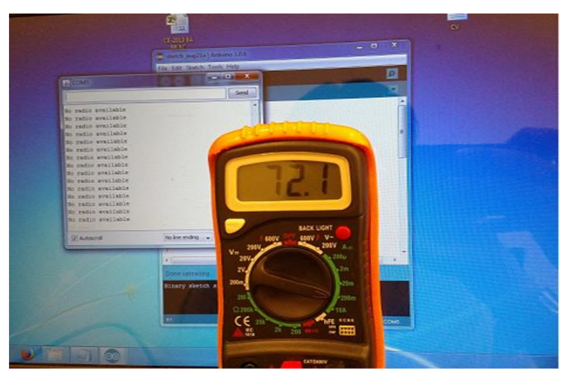


Fig. 5. Sleep mode.

VIII. Design Node

In the design node, the solar cell features open voltage of 6V and short-circuit current at 0.184A (Fig. 6).

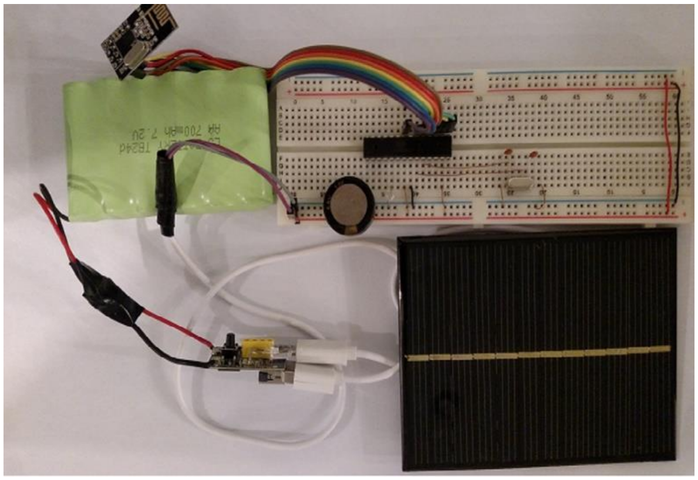


Fig. 6. The implemented WSN node with Solar cell, rechargeable battery and super capacitor.

The individual curves are measured under direct sunlight, on a cloudy day, 1 hour before sunset and under heavily cloudy conditions using different load resistors, which are shown in Fig. 7 and 8.

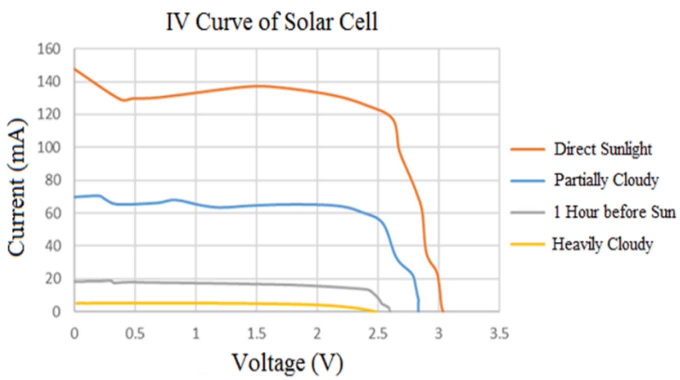


Fig. 7. Current-voltage curve of Solar cell [12].

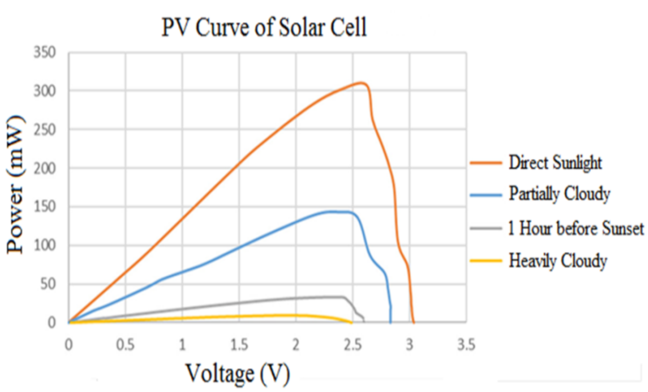


Fig. 8. Power-voltage curve of Solar cell [12].

Fig. 9 shows the Volts-Hours curve of Rechargeable batteries.

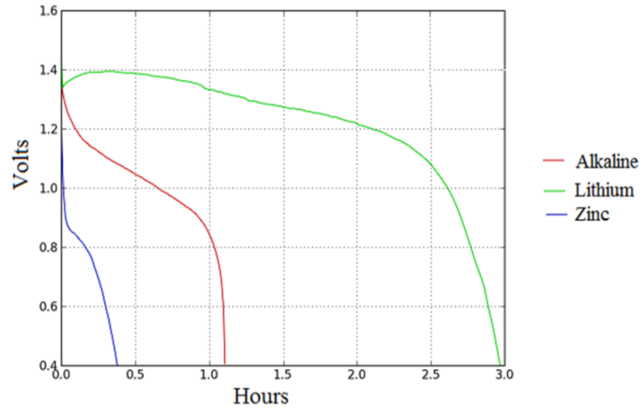


Fig. 9. Volts-hours curve of rechargeable batteries [13].

Fig. 10 shows the measured self-discharge of 1 farad super capacitor.

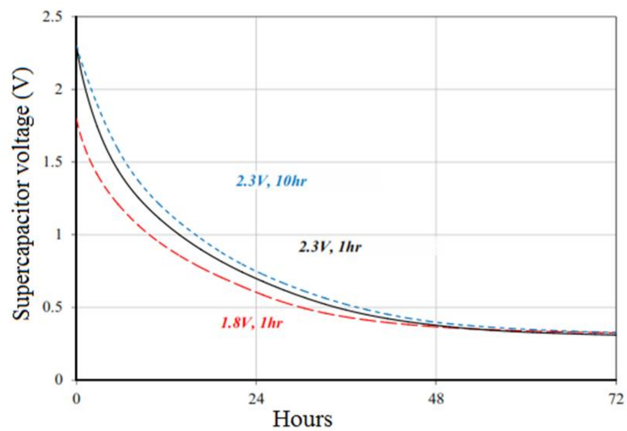


Fig. 10. Measured self-discharge of 1F super capacitor.

Maximum power point tracking circuits, reverse current own protection and charging controller are typically included by the influence of PV cell and energy storages units (rechargeable batteries and super capacitor). The most straightforward of this interface incorporates a turn around current protection diode, which keeps energy stockpiling from compelling current into the PV cell under low brightening conditions yet includes a diode voltage drop between the PV cell and energy storages. This interface can be efficient and even sufficient in the case where the PV cell nominal operating voltage range fall a diode drop above the operating voltage range of energy storage and this arrangement is suitable for systems that are able to select energy storages and PV cell accordingly.

This interface is also the part of maximum power point tracking (MPPT). The main concern of MPPT in case of low-power harvesting is the efficiency of the block and a number of implementations have been proposed. In some cases it is shown that the benefit extended from MPPT is peripheral as given in Fig. 11 matched connection between the maximum power point and PV cell operating time.

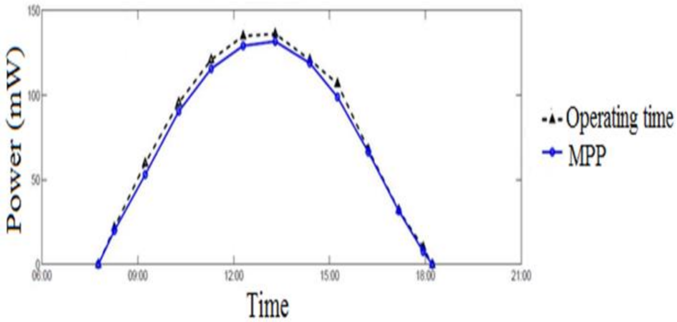


Fig. 11. Comparison of Solar cell output: Operating point verses MPPT.

Then again, when utilizing a super capacitor the subsequent voltage swing as a super capacitor completely releases or charges can make the PV cell go amiss incredibly from its close most extreme power working point as appeared in Fig. 12.

For this situation the MPPT piece can guarantee that PV cell works near its most extreme.

With a completely released super capacitor, an additional issue is cold-booting where the drained super capacitor does not permit the PV cell to energize the super capacitor. A rechargeable based battery charging circuit is recommended that permit the fast charging of a super capacitor contrasted with coordinate charging (Fig. 13).

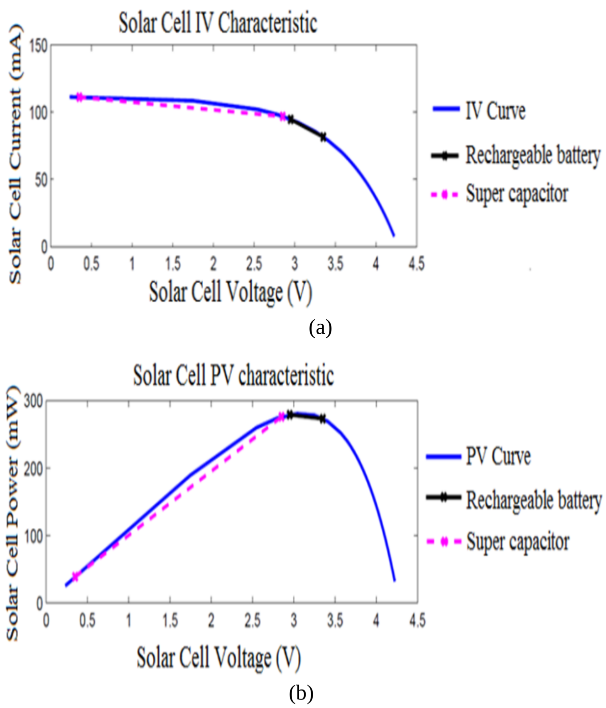


Fig. 12. Operating range of the WSN node.

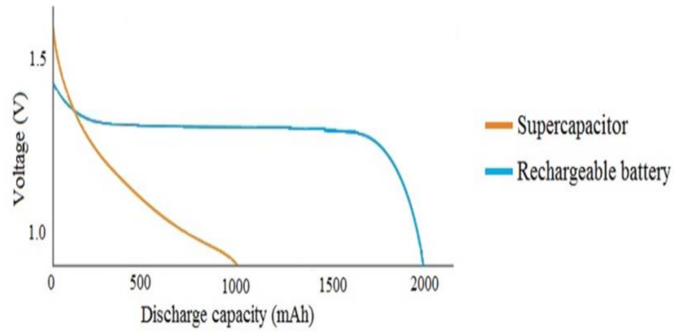


Fig. 13. Voltage-discharge capacity comparison of super capacitor and rechargeable battery [10].

IX. Conclusion

The aids of this paper can be summarized in two phases:

The first phase is improved management of uncertainty of energy supply through improved calculations and using this to achieve low variability in performance of energy harvesting systems.

The second phase is to enhance the energy management procedures that achieve better match of application workload demand with energy supply using system components and practical design.

References

[1] Andreas Willig, Kirsten Matheus, and Adam Wolisz. wireless technology in industrial networks. Proceedings of the IEEE, 93(6):1130-1151, 2005.

[2] Sadok Bdiri, Faouzi Derbel, and Olfa Kanoun. Wireless sensor nodes using energy harvesting and b-mac protocol. In Systems, Signals & Devices (SSD), 2013 10th International Multi-Conference on, pages 1-5. IEEE, 2013.

[3] Prabhakar T Venkata, SN Akshay Uttama Nambi, R Venkatesha Prasad, and Ignas Niemegeers. Bond graph modeling for energy-harvesting wireless sensor networks. Computer, 45(9):31-38, 2012.

[4] Masateru Minami, Takashi Morito, Hiroyuki Morikawa, and Tomonori Aoyama. Solar biscuit: A battery-less wireless sensor network system for environmental monitoring applications. In The 2nd international workshop on networked sensing systems, 2005.

[5] A Dolara, R Faranda, and S Leva. Energy comparison of seven mppt techniques for pv systems. Journal of Electromagnetic Analysis and Applications, 1(03):152, 2009.

[6] Hong Yao. Modeling and design of a current mode control boost converter. PhD thesis, Colorado State University. Libraries, 2012.

[7] Ali Reza Reisi, Mohammad Hassan Moradi, and Shahriar Jamasb. Classification and comparison of maximum power point tracking techniques for photovoltaic system: A review. Renewable and Sustainable Energy Reviews, 19:433-443, 2013.

[8] Chen Xu, Caofeng Pan, Ying Liu, and Zh L Wang. Hybrid cells for simultaneously harvesting multi-type energies for self-powered micro/nanosystems. Nano Energy, 1(2): 259-272, 2012.

[9] Hannes Reinisch, Stefan Gruber, Hartwig Unterassinger, Martin Wiessflecker, Gunter Hofer, Wolfgang Pribyl, and Gerald Holweg. An electro-magnetic energy harvesting system with 190 nw idle mode power consumption for a baw based wireless sensor node. IEEE Journal of solid- state circuits, 46(7):1728-1741, 2011.

[10] C.-Y. Chen and P. H. Chou. Duracap: A Supercapacitor-Based, Power Bootstrapping, Maximum Power Point Tracking Energy- Harvesting System. In Proceedings of the 16th ACM/IEEE international symposium on Low power electronics and design, ISLPED ’10, pages 313–318, New York, NY, USA, 2010.

[11] Christian Renner and Jürgen Jessen and Volker Turau, “Lifetime Prediction for Supercapacitor-powered Wireless Sensor Nodes”, Proceedings of the 8th GI/ITG KuVS Fachgespräch, Drahtlose Sensornetze (FGSN'09), Hamburg, Germany, 2009.

[12] P. Dutta, M. Grimmer, A. Arora, S. Bibyk, and D. Culler. *“Design of a Wireless sensor network platform for detecting rare, random and ephemeral events”*.IEEE IPSN, 2005.

[13] Sakar, S., Adamu: *”A framework for optimal battery management for wireless nodes. IEEE Journal on Selected Areas in Communications”* 21 (2): 179–188, February 2003.

The Impact of Knowledge Management on Organizational Performance

Hayfa.Y. Abuaddous, Abdullah A.M. Al Sokkar
Faculty of Computer Science and Informatics
Amman Arab University (AAU)
Amman, Jordan

Blaqees I. Abualodous
Digital Product Planner
Disrupt Technologies
Amman, Jordan

***Abstract*—In today's business, knowledge is considered as a core asset in any organization, even it can be considered as important as technological capital. It is part of human abilities and thus human capital. Knowledge management (KM) is becoming a fad in an increasing way so many organizations are trying to apply it in order to enhance their organizational performance. In this paper, literatures were investigated critically in order to show the real influence of knowledge management and some of its practices on organizational performance. It has been founded that KM including knowledge process and infrastructure capabilities affect positively in a huge manner on all aspects of organizational performance directly or indirectly. In the same vein, there is a huge need to continuously train and educate the learning organizations' CEOs about the importance of KM through group works and training programs.**

Keywords*—Knowledge management; infrastructure capabilities; process capabilities; organizational performance; learning organizations

I. Introduction

In this dynamic competitive world and with the clearly influence of information technology (IT) on the business environment, the need arises to get a competitive advantage amid the vast amount of competitors. It has been noticed that a strong tendency within organizations to consider knowledge and its management as a core asset and valuable concept , where their competitive advantage -which they all aspire to- lies in [1]. Knowledge is not just facts and numbers that organizations have on spreadsheets or in maps but rather it concerns the whole human experience acquired through education and working skills and experience [2]. Today's businesses are characterized by a high degree of complexity and that indicates for a more accumulation in knowledge, and for sure this leads to a more difficulty in managing and controlling it in terms of storage, organizing and retrieval.. etc. Correspondingly and according to [3], there is an increasing investment in knowledge management every year, where they have cited in their research, Forrester Research Inc. (2010) found that 20 percent of businesses with small and medium size have the intention to construct Customer Relation Management (CRM) or information and knowledge management tools in 2010 or later in North America and Europe. Therefore, we can induce that organizations are strongly starting to believe in Knowledge Management (KM) as a chance to achieve what the majority of them strive for. Logically, when organizations discuss all of the above, it is mainly prompt to improve their performance and the development of decision-making processes and thus to increase profitability and sustainability in today's global markets, and perhaps to get the knowledge capability that enhance the effective management and effective flowing for information and knowledge inside the business [4].

This paper will focus on evaluating the impact of knowledge management and its resources on the organizational performance in order to reach competitive advantages. Besides, this paper will also aim to provide the reader with an overview of knowledge management, and learning organizations.

This paper is organized as follows: the first section of the literature presents an overview of knowledge management; the second section gives an overview to the knowledge management and its impact on organizational performance. The third section will address the learning organizations and how knowledge management will affect them. Finally, future work and conclusion are presented..

II. An Overview of Knowledge Management and its Capabilities

The knowledge that organizations possesses is one of the most important assets that helps them in attaining competitive advantage, even in some cases, it becomes more important than the financial resources and all other tangible assets [2]. With this in mind, knowledge management has become an attractive subject for research and studies in the last twenty years [5] including this research. It can be precisely defined as "the process of capturing information and experience of individuals and the organization –which is available in databases, on paper, or even in people's intellect – and distributing it to wherever it can produce benefit" [6]. Additionally, knowledge management can be attributed as a portfolio of strategies and activities that is related to the process of acquiring, transferring, and sharing knowledge with all organization's people [7]. Researchers noticed that most of their respondents are well educated about knowledge management [7], while others argue that there is unexpected level of misunderstanding to the real knowledge management and its real importance [6]. Above all, the strongly need for an effective implementation to knowledge management arises from the organizations' need to achieve goals such as establishing a competitive advantage to amid the current globalization, also, adapting the organizations with constantly changes and build involved workforce as well as increasing the productivity amid to a highly turnover and downsizing [12].

Author in [12] considered knowledge management as a technique that uses the values of knowledge resources in order to enhance the performance for both organizations and employees. They also focus on its ability in facilitating the process of attaining data and information on the needed time, and in improving both strategic and tactical operational activities needed to attain organizational goals. From the theoretical point of view, the components of the knowledge management are (people, processes, technology, culture and structure) and those are the foundation of the knowledge management system [5], where literatures acknowledged that knowledge management determines the knowledge flow inside firms in four steps inside each defined process, which are knowledge creation, retention, transfer, and utilization [2], [6], [8]. Other research added another step titled with knowledge internalization [9]. All the above mentioned activities can be conducted in the organization with relatedness to knowledge management and they can be defined as knowledge management practices and that may include locating, sharing, creating and encouraging through openness culture [10], [11]. Knowledge management may in some cases be affected by demographic characteristics of the organization's employees in a highly manner [7]. Usually, when talking about the knowledge capability of a firm, we can distinguish it into two types of capabilities including firm's knowledge infrastructure capability (e.g., technology infrastructure, organizational structure, and culture) and the firm's knowledge process capability or knowledge management enablers (e.g., knowledge acquisition, conversion, application and protection) [9]. The first and the second comprise the knowledge management capabilities. While [3] stressed the importance of the nature of knowledge capabilities, so that many organizations can include different levels and combinations of knowledge enablers and processes construct the knowledge capability together. Knowledge process capabilities affected positively by collaboration, culture, the support of both top management and information technology. In addition, it is affecting positively- in combination with creative organizational learning- on the organizational performance [8].

When talking about knowledge management we have to mention knowledge management system that emerged with the revolution of ICT and of course automate and the ease of all what managers strive to do through knowledge management. Knowledge management systems (KMS) can be defined as "the planned workplace linkage of specific process steps or domains within an organization. A KMS allows an organization to systematically manage knowledge in order for its workforce to acquire, create and use knowledge to innovate and compete in the marketplace" [5]. Author in [12] portrayed that KMS is comprised from subcomponents including repositories, collaborative platforms, networks, and culture. Also, they considered that building KMS without a robust design will lead to failures. Researchers address that in order to implement an effective knowledge management or even knowledge management system many initiatives need to be done, for example, it is important to take into account the necessary support services to create a supportive environment for knowledge and learning, and linking operations with technology and culture. More importantly, the support of senior management in consolidation of building knowledge management and its related systems and in spreading the culture of knowledge sharing need to be taken under consideration [2], [5], [9].

III. KM and Organizational Performance

Organizational performance (OP) means to what extent the organizational goals and objectives can be achieved [13]. Many empirical and conceptual studies were conducted to test the relationship between KM and OP which are summarized in Table I. [4] Investigated the organizational impact of knowledge management practices. The study argued that it is expected that knowledge management practices including (communication, the ability to create new knowledge, acquisition, policies and strategies of KM, and training) affect organizational performance dimensions which include (financial performance, new product success, customer satisfaction, market share) in a positive way among all the dimensions. However, it was revealed that the highest impact was on the success of a new product, which is one of the OP dimensions, and the training was the strongest affecting dimension on OP.

In order to help organizations including academic ones (e.g., Universities) to correctly choosing the strategies for investing in knowledge resources, an empirical study was conducted in the Isfahan universities in Iran, presented that knowledge management resources such as organizational structure and knowledge application are positively affecting OP while other resources such as technology and knowledge conversion are not [1]. It is likely to address that KM strategy could include knowledge transfer concentration, open mindedness orientation, skill sharing and integrated value knowledge [13]. The most significant positive relationships from the whole KM processes performance indicators are factor strategy and leadership, among knowledge management enablers [9]. On the other hand, other similar studies were conducted to reveal the influence of KM resources on organizational performance with the same dimensions. The results were supported along with organizational structure and knowledge application and weren't the case with as technology and knowledge conversion [14], [15]. With this in mind, we can induce that not all KM resources contribute directly or positively on OP, in other words, each resource is not linked to performance, rather as a composed. It has been found that a non-expected percentage of employees have no interest in knowledge sharing and retrieving, and most of them prefer to depend on their own knowledge and intuition [6]. Since the culture can be considered as a practice of KM beside processes, human capital, and strategy, there is a strong need to construct the culture to ease the process of sharing knowledge between employees. Other researchers also address that well-constructed culture will lead to support knowledge management process and thus improving the organizational performance [16]. Therefore, they improved the conceptual framework under the name of PICS that aggregate the last enablers for KM effectiveness. Many other factors were found to have heavy effects on organizational performance such as competent competitive advantage, operation improvement, and potential growth. Knowledge management strategy affected the last factors significantly in clothing manufacturing in Thailand and thus it affects OP in a positive way [13]. Although they found that, there is no significant relationship between KM practices and financial performance, researchers has also found that they can generalize another rule; There is an existence of a direct significant relationship between KM practices and Organizational performance which includes the financial performance [10]. Fig. 1 is a similar model illustrating the idea of the study.

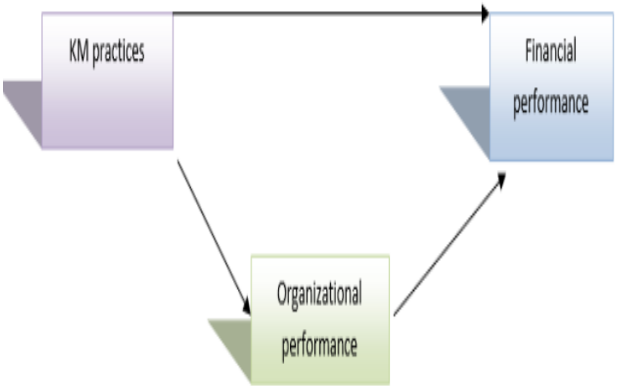


Fig. 1. Research model.

Author in [6] focused on the engineering field; he indicated that knowledge management systems have a significant effect on the effectiveness of strategic decisions that will lead to a better organizational performance. While in [11] focused on banking sector and argued that organizations need to develop their knowledge management processes in order to reach better decisions creations and thus a better organizational performance. They found that there is a significant impact on decision making related to knowledge management practices including what we have mentioned previously (IT infrastructure, HR, shared knowledge and culture).

IV. KM and Learning Organizations

Not all organizations are learning organizations (LO) although every organization can continuously learn. Rather, every learning organization is any organization that uses the managed knowledge systematically in order to continuously aggregate the best knowledge and to construct its long-term memory [17]. Those organizations arrived to this level due to their employees, where learning depends always on them [7].

TABLE I. Summarization of Reviewed Literatures

Article	**The study**	**Methodology and Hypothesis**	**Findings**
Kharabsheh, et al (2012), [4]	KM practices and its impact on OP in Jordanian pharmaceutical firms	Using questionnaire survey. Data gathered from 13 pharmaceutical companies in Jordan. 11 positive hypothesis	All hypothesis were supported.
Fattahiyan, et al (2012), [1]	Study of relationship between KM enablers and processes with OP in Isfahan universities faculty members	Using questionnaire gathered from 1554 randomly selected members. 9 positive hypothesis	Some of them were supported, others were not.
Bhatti, et al (2011), [16]	The affect of KM practices on OP	Conceptual study with 2 general recommended hypothesis	Recommend PICS model will be tested in the future on local service industry
Janepuengporn & Ussahawanitchakit (2011), [13]	The impact of KM strategy on OP in clothing manufacturing business in Thailand	Using questionnaire survey gathered from the clothing manufacturing sample. 15 positive hypothesis	All hypothesis were supported.
Emadzade, et al (2012), [14]	KM capabilities and OP	Using a questionnaire gathered from 245 small size business owners and managers in 86 firms in Isfahan. 9 positive hypothesis	Some of them were supported , others were not.
Mills & Smit (2010), [15]	KM and OP , decomposed view	survey data from 189 managers and structural equation modeling	Some of them were supported, others were not.
Zack, et al (2009), [10]	KM and OP	Exploratory analysis, literature revealed 12 KM practices whose performance impact was assessed by a questionnaire of business firms.	Some of them were supported, others were not.
Ho (2009), [9]	KM enablers and OP	Exploratory analysis, literature reviews, 3 hypothesis	Supported
Mills & Smit (2010), [3]	KM and performance	internet-based survey on 120 responder in company adapting KMS with 7 hypothesis	Supported
Shannak (2010). [6]	KBSs support for strategic decision	Using a questionnaire in engineering offices in Jordan	Supported
Mohammed and Jalal (2011), [11]	KMSs support for decision making process	Using a questionnaire in banking sector,4 main hypothesis	Supported
Frigidian, C. & Harris, [18]	Knowledge management support to product development	Literature review, case study and process modeling using IDEFO technique	Develop knowledge management framework

When employees both understand and apply the captured knowledge, then they are learning and eventually those employees are constructing the learning organizations. Learning organizations know exactly how to adapt itself with changes quickly among all levels including developing a product, how they response to their customers, and how they manage their human capital [19], [20]. Author in [2] discussed that KM activities are strongly linked to LOs, but they presented that private sector organizations have extremely better processes among all dimensions of LO in comparison with public sector organizations. Researchers portrayed the relationship between knowledge management dimensions discussed above and learning organizations in seven sub relations. LO strategy and vision affected positively by KM processes and leadership, work practices in LO affected positively by KM culture, technology and leadership, and also the culture of LO affected positively by KM culture. In addition, the structure of information and knowledge flow influenced positively by KM process, culture and technology, LO's processes improvement affected positively again by KM process, culture, but also measurement. As well as, the KM process, culture, and leadership affecting positively training and development in LO, and finally KM leadership and measurement affecting positively both reward and recognition in learning organization [2]. Table II inserted to show this supported study's details.

TABLE II. CHAWLA & JOSHI [2] STUDY'S SUMMARIZATION

Article	The study	Methodology and Hypothesis	Findings
Chawla and Joshi (2011), [2]	The impact of KM on Los, comparison between private and public sector organizations	Using surveys among 51 executives among 16 private and public firms in India. 7 Hypothesis.	Supported

Author in [7] conducted a survey with 180 managers and engineers to test the impact of knowledge management on learning organizations. The major findings revealed that the higher level of information, knowledge application, and KM processes will clearly help learning organizations in enhancing their performance as a whole including the innovation and the process of decision-making. They advised to conduct this empirical study yearly in each organization in order to evaluate all new practices so that organizations can learn.

V. FUTURE WORK

As a future direction for this research, an empirical study will be conducted to involve a survey of Jordanian manufacturing that implement knowledge management systems. In order to test the most critical success factors (CSFs) that affect implementing knowledge management systems (KMSs) in the manufacturing Jordanian environment. This study may also incorporate user's factors that may affect (users' perceptions towards the use and implementation of knowledge management system in manufacturing organizations, user's satisfaction, and user's training). In order to prepare the final Critical Success Factors in Jordanians manufacturing, this study will be tested with the following three theoretical hypotheses:

H1: There is a significant impact of user's satisfaction on KMSs implementation success.

H2: There is a significant impact of user's perception of ease of use on KMSs implementation success.

H3: There is a significant impact of user's training on KMSs implementation success.

VI. CONCLUSION

Nowadays, knowledge is considered as an essential asset in any organization. Thus, many organizations are trying to apply Knowledge management in order to improve their organizational performance. This article has reviewed the positive impact of knowledge management and some of its practices on organizational performance. Many studies have concluded that KM is the main reason to business growth. Thus, it is good to invest in KM resources to attain organizational performance improvement, since KM resources and practices are related directly and indirectly to it when they are implemented effectively. In order to implement an effective knowledge management system, the support of senior management in consolidation of building knowledge management and in spreading the culture of knowledge sharing need to be taken under consideration. Furthermore, there is a huge need to continuously train and educate the organizations' CEOs about the importance of KM through group works and training programs.

REFERENCES

[1] S. Fattahiyan, R. Hoveida, S. Siadat, and H. Tallebi, "Study Of Relationship Between Knowledge Management Enablers And Processes With Organizational Performance", Interdisciplinary Journal Of Contemporary Research In Business, Vol.4 (4), pp. 36-44, 2012.

[2] D. Chawla, and H. Joshi, "Impact of Knowledge Management on Learning Organization in Indian Organizations—A Comparison", Knowledge and Process Management, Vol. 18(4), pp. 266–277,2011.

[3] A. Mills, and T. Smith," Knowledge Management And Organizational Performance: A Decomposed View", Journal Of Knowledge Management, Vol. 15 (1), pp. 156-171, 2010.

[4] R. Kharabsheh, I. Magableh, and S. Sawadha, "Knowledge Management Practices (KMPs) and Its Impact on Organizational Performance in Pharmaceutical Firms".,European Journal of Economics, Finance and Administrative Sciences, Vol.48, pp. 6-15, 2012.

[5] E. Fibuch, and CW. Vanway, "What Is a Knowledge Management System and Why Should Care?", PEJ, Vol. 37(5), pp.34-39,2011.

[6] R. Shannak, "Knowledge-based Systems Support for Strategic Decisions". European Journal of Economics, Finance and Administrative Sciences, Vol.21 (2010), pp. 7-20, 2010.

[7] U. Aktharsha, and H. Anisa, "Knowledge Management System and Learning Organization: An Empirical Study in an Engineering Organization". The IUP Journal of Knowledge Management, Vol. IX(2), pp.27-43, 2011.

[8] S. Lee, B. Kim, and H. Kim, " An Integrated View Of Knowledge Management For Performance", Journal Of Knowledge Management, Vol.16 (2), pp. 183-203. 2012

[9] C. Ho, "The Relationship Between Knowledge Management Enablers And Performance", Industrial Management & Data Systems. Vol. 109 (1), pp. 98-117, 2009.

[10] M. Zack, J. McKeen, and S. Singh, "Knowledge Management And Organizational Performance: An Exploratory Analysis". Journal Of Knowledge Management, Vol. 13 (6), pp. 392-409, 2009.

[11] W. Mohammed, and A. Jalal, "The Influence of Knowledge Management System (KMS) on Enhancing Decision Making Process (DMP)", International Journal of Business and Management, Vol.6 (8), pp.216-229, 2011.

[12] R. Gopal, and P. Joy. "Creation of Knowledge Management System". Advances In Management, Vol. 4(11), pp.7-14, 2011.

[13] K. Janepuengporn, and P. Ussahawanitchakit, " The Impacts Of Knowledge Management Strategy On Organizational Performance: An Empirical Study Of Clothing Manufacturing Businesses In Thailand".

International Journal Of Business Strategy, Vol. 11 (1), pp. 92-109, 2011.

[14] D. Emadzade, B. Mashayekhi, and E. Abdar, "Knowledge management capabilities and organizational performance", Interdisciplinary Journal Of Contemporary Research In Business, Vol 3 (11), pp. 781-790, 2012.

[15] A. Mills, and T. Smith., "29P. Exploring the Impact of Knowledge Management Capabilities on Organizational Effectiveness", CONF-IRM 2010 Proceedings. 35, 2010.

[16] https://aisel.aisnet.org/confirm2010/35

[17] W. Bahatti, A. Zaheer, and K. Rehman, "The effect of knowledge management practices on organizational performance: A conceptual study", African Journal of Business Management, Vol. 5(7), pp. 2847-2853, 2011.

[18] C. Frigidian, and A. Harris. "Knowledge Management To Support Product Development In Cold Roll Forming Environment", Int J Adv Manuf Technol, Vol.57 (5), pp 585-596 pp.585–596, 2011.

[19] A. Al Sokkar, and E. Law, "Validating an episodic UX model on online shopping decision making: A survey study with B2C e-commerce". In Proceedings of the 5th ACM SIGCHI symposium on Engineering interactive computing systems (pp. 297-306). ACM, 2013.

[20] T. Orehovački, A. Al Sokkar, J. Derboven, A, Khan, "Exploring the Hedonic Quality of Slow Technology". In: CHI 2013 workshop on Changing Perspectives of Time in HCI, 2013.

Wireless Body Area Network Security and Privacy Issue in E-Healthcare

Muhammad Sheraz Arshad Malik, Muhammad Ahmed, Tahir Abdullah, Naila Kousar, Mehak Nigar Shumaila
Department of Information Technology
Government College University
Faisalabad, Pakistan

Muhammad Awais
Department of Software Engineering
Government College University
Faisalabad, Pakistan

***Abstract*—Wireless Body Area Network (WBAN) is a collection of wireless sensor nodes which can be placed within the body or outside the body of a human or a living person which in result observes or monitors the functionality and adjoining situations of the body. Utilizing a Wireless Body Area Network, the patient encounters a greater physical versatility and is never again constrained to remain in the hospital. As the Wireless Body Area Network sensor devices is being utilized for gathering the sensitive data and possibly will run into antagonistic situations, they require complicated and very secure security medium or structure to avoid the vitriolic communications within the system. These devices represent various security and privacy protection of touchy and private patient medical data. The medical records of patients are a significant and an unsolved situation due to which a changing or exploitation to the system is possible. In this research paper, we first present an overview of WBAN, how they utilized for healthcare monitoring, its architecture then highlight major security and privacy requirements and assaults at different network layer in a WBAN and we finally talk about various cryptographic algorithms and laws for providing solution of security and privacy in WBAN.**

***Keywords*—E-Health; privacy; security; wireless body area networks**

I. Introduction

Recent improvement in wireless technologies and ICT (information and communication technology) frameworks are empowering the health care segment to effectively and efficiently control and provide a variety of solution and health services. Progressed frameworks of ICT will have the capacity to convey medicinal services administrations to patients in healing facilities and therapeutic focuses, as well as in their homes and working environments, along these lines offering cost reserve funds and upgrading the individual fulfillment of patients. E-Health administrations can make use of WBAN, which can go about as an empowering innovation, and gaining its popularity day by day due to its benefits. As by using WBAN patient don't need to visit hospital daily, they stay in home, save its time, and data related to patient is examined by doctor any time at any place.

A WBAN make use of litter sensor and display medical related information on screens by means of WIFI transmission. Sensors are set inside or outside one's body.

Sensors are used to collect sensitive and important medical related information of a patient or it can also be used in sports. WBANs communicate with the net and other technologies such as ZigBee technology, WSNs, WI-FI, Bluetooth, cell systems and Wireless Personal Area Network (WPAN) technology. Sensor collect patient related data, transfer it to cloud using different technology, this data or information is used by doctor, etc.

A wireless body area network has generally two types of nodes i.e. wearable and implantable nodes which work at different frequencies. An implantable node is well on the way to work at 400 MHz utilizing the MICS (Medical Implantable Communication Service) band, while the wearable hub/ node could work in ISM/UWB (Instrumentation Scientific Medical/Ultra Wide Band) or some other specific groups [2], [3].

The paper is sorted out into five Sections. Section I give overview of what is being introduced in the paper. Section II introduces a WBAN architecture. Section III presents the related work done by different researchers which provide solution of security and privacy issues. Section IV presents the WBAN security Requirements and conceivable assaults which occur when utilizing Wireless body area network in E-Health. Section V represents biometric solution for securing Wireless Body Area Network. The last section finishes up our work.

II. General Architecture

A. 3 Tier Architecture

This area gives a general layout of WBAN plan as shown in Fig. 1. WBANs are a vital piece of a multi-level telemedicine framework [4]. Tier 1 incorporates various wireless medical sensor nodes. A WBAN screens physiological signs from these little sensors node with remote transmission capacity set either inside or around a man's body, which are utilized to gather vital wellbeing information of a man amid a specific movement medicinal or game or training related activities. Every sensor node can recognize and test as well as process at least one physiological signs. For instance, heart rate can be checked by an electrocardiogram sensor (ECG), Oxygen saturation sensor (SpO2) used to quantify the level of oxygen, and blood pressure is observed by blood pressure. Tier 2 incorporates the personal server (PS) application running on a client PDA or iPod or some other convenient gadgets telephone which goes about as a sink for information of the remote devices [1], [2], [4] and at that point exchange those data to an appropriate PC when a correspondence interface is accessible. Tier 3 contains various remote base-stations that keep patient's therapeutic/non-medicinal records and gives huge (indicative) recommendation.

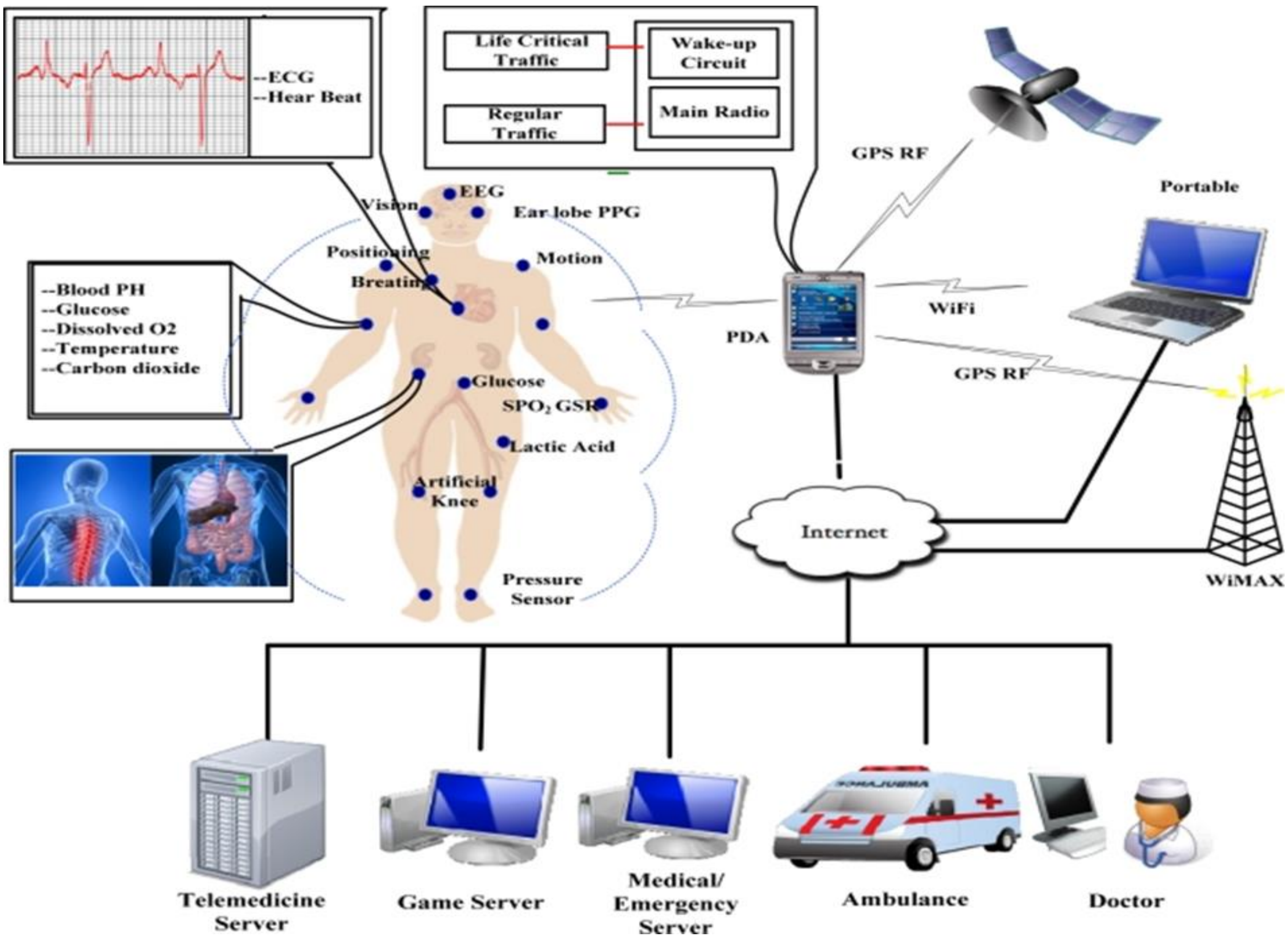


Fig. 1. WBAN architecture [6].

B. Node Classification in a WBAN

In WBAN a node is a free gadget which has communication ability. We can classified nodes into three distinct groups based which will perform a role in the network founded on his functionality, implementation.

a) *Node Type based on functionality*

Personal Device (PD): This gadget is responsible of collaboration with other clients by gathering the data given by the sensors and actuators. At that point it is advised the client over an external gateway by the PD, On the Actuator or on the gadget there is a screen/LED. The gadget is called body-gateway, sink, Body Control Unit (BCU) or PDA in few of the applications.

Sensor: In WBANs Sensors measure very certain amount or parameters in patient's body either inside or outside the body. When a physical stimuli happened the nodes automatically assembles and react to the information, process vital information and the information get a reaction from the wireless. The example of this type of sensors is physiological sensors, and the ambient or the biokinectics [5]. Few of the current sorts of these types of sensors could be utilized as a part of one's patients or a person's wrist watch, portable, or headphone and therefore, permit wireless checking of a man anyplace, whenever and with anyone. A rundown of various sorts of monetarily accessible sensors utilized as a part of WBANs are as per the following: DNA Sensor, Blood pressure, EMG, Spirometer, EEG, ECG, Temperature, Humidity, CO2 Gas sensor and so on.

Actuator: After receiving data from sensor the actuator collaborate with the user. The actuator gives feedback in the system by acting on sensor information, such as directing the right dosage of medication into the patient body in omnipresent medicinal services apps.

b) *IEEE classification for Nodes*

Another classification has been purposed by the IEEE 802.15.6 for node in a WBAN in which they are actualized inside the body, [2] [9] that is provided as follows:

Body Surface Node: In this type the node are placed on the surface or the outer side or the node is keep far away from the patient i.e. 2 cm far away from the patient.

External Node: Node in this category is not contacted to the patient's rather it is placed a couple of cm i.e. 5 meters far away from the patient's body.

Implant Node: Implant Nodes are those which are embedded within the human body, under the skin or situated on or in the inside the body tissue.

c) *Node Type base on Role*

These nodes are classified in the WBANs were based on their role as follows:

Coordinator: The working of a coordinator node is resembles to a gateway, or it resembles with another WBAN, or the access coordinator. All other nodes that want to communicate with each other make use of PDA known as WBAN coordinator.

End Nodes: The devices or end nodes used in WBANs are restricted to perform their inserted application. In any case, they don't have ability for transferring messages between nodes.

Relay: The relays node also known as intermediate nodes. They have two nodes one is called parent and the other node is called child node and they transfer the messages. Fundamentally if one of the node reaches at a limit, at that point any of the information is sent is required to be handed off from the alternate nodes before it reaches the PDA. The relays nodes can be used for detecting information.

III. Related Work for Security and Privacy in WBAN

In this section we talk about the recently published literature on E-Health using WBANs. In [17] Muhammed et al. also introduced BARI+ distributed key management protocol based on biometric, for WBANs. The authors assert that this protocol will help numerous security related services like confidentiality, authentication, and will provide security against different routing attacks and also defense against node compromise.

Huang et al. [18] stated secure access to a hierarchical sensor based healthcare monitoring architecture. Its architecture has three tier i.e. sensor, mobile, and back-end network tier and used for three healthcare apps (in-hospital, in-home, and nursing-home).In the sensor network tier, a wearable sensor system (WSS) make use of Bluetooth along with biomedical sensors to monitor the fundamental signs of individuals. Wireless sensor motes (WSMs) i.e., Mica2 are set inside the building to gather the ecological parameters. WSS and WSM safely broadcast physiological info and environmental parameters to the upper layer. WSS use AES authentication (i.e., CBC-MAC) associated with an encryption scheme, on the other hand WSMs utilize polynomial based encryption scheme to set up secure point-to-point communication between two WSM motes.

A protocol based on public-key is utilized to set up the secure keys. In the mobile computing network tier, mobile computing devices (MCD) like PDAs sorted out as in an ad-hoc network, route the data by means of multi-hops to the local station. MCD has the computational capacities needed to dissect the WSS and WSM data. Besides, the authors guarantee that their secure architecture gives security CIA i.e., confidentiality, integrity and authentication.

Muraleedharan Osadciw [19] introduced a secure and safe health monitoring network against DOS attacks utilizing cognitive intelligence. To prevent Sybil and worm whole attacks in health care app they introduced energy efficient cognitive routing protocol.

Le et al. [20] stated a protocol called MAACE in which only an authorized person can gain access to the patient confidential data. Their scheme gives mutual authentication and access control depends on ECC. Besides, the authors guarantee that purposed technique can shield against replay and DOS assaults. Le et al.'s protocol gives security for patient's private and confidential data that could be hazardous for a patient when compromised. Real-time healthcare applications can't compromise or accept, if patient's data disclosed to illegal users.

Malasri et al. [21] actualized a protected wireless mote-base medical sensing network for health care applications. Their set up defend against spoofing and physical layer attack, and gives CIA. Misic et al. [22] purposed 2 key distribution algorithms for implementing patient privacy in healthcare WSNs by utilizing key distribution algorithms.

Haque et al. [23] projected an effective and efficient security mechanism for patient monitoring systems using WSNs. Their scheme make use of public key primarily based infrastructure which gives data confidentiality.

Hu et al. [24] projected a patient healthcare monitoring system called tele-cardiology sensor network (TSN) that is based on software and hardware. TSN is specially intended for the U.S. healthcare community, it performs ongoing healthcare information gathering for aged patients in large nursing homes. In order to secure the patient privacy, TSN encourages confidentiality and integrity. Intra-cluster security, skipjack block cipher cryptography algorithmic program accustomed secure patient physiological (i.e., ECG data), associated an inter-cluster security make use of pre-distributed session keys. For various patients, cluster based routing is employed to diminish the patient-to-doctor routing overhead, and accomplished effectiveness. Security scheme which they used expends twenty six mJ for data processing, 1,002 mJ for radio communication, and 11 mJ for memory accesses.

Dagtas et al. [25] introduced secure architecture for health monitoring by utilizing ZigBee. Session key in WBAN is setup using secure and reliable key management protocol since it gives cryptographic keys that encourage security services, e.g., confidentiality, authentication, and data integrity. An authentication algorithmic program is utilized between the body sensors and also the handheld gadget of the mobile patient. Nonetheless, the authors give security to physiological data, but they didn't discuss which symmetric cryptosystem they utilized, and didn't analyze the energy potency for security services.

Kang et al. [26] for pervasive healthcare introduced a wearable context-aware system. The context-aware system is made out of wearable sensor systems, wearable PCs and communication modules. The wearable sensors are associated with wearable PCs by means of ZigBee. There are two kinds of sensors which can be used to a wearable. For exp, it is a sensor which is like a watch and there is a sensor which can be wearable on chest like a belt. Examples of wearable PCs, Personal digital assistant, are applied to collect the sensors' information. There is a technology which is called ZigBee which is used to communicate between the two first is the wearable sensors and other is Personal digital assistant. A LAN along with 802.11b WiFi is used to communicate among the Personal digital assistant and the services provided by the healthcare over the Internet.

Lin et al. [27] introduced SAGE is the idea which is good to provide solid security and privacy against the eavesdropping on the E-Health applications or systems. The system achieves content which is concerned with security and related privacy beside a solid worldwide foe. The idea behind SAGE is that when a patient information database (PIDB) acquires the Patient health information (PHI) from patient's body sensors for example, the accelerometer, and the blood pressure, or oxygen saturation, and temperature sensors [27]. It transmits the Patient health information to all of the patient doctors but the access is given to only one doctor or doctors which are applicable.. Authors proposed ECC which is based on privacy and security solution and show a proper or written evidence for the suggested solution which is against the eavesdropping worldwide.

Kumar et al. [28] gave another securing health solution which is known as the SHM stand for the secure health Monitoring via a medical WSNs. Secure Health Monitoring offers security facilities including the confidentiality, and the validity, as well as the integrity towards the patient's information at very little cost. In SHM, the privacy is acquired by Ping-Pong 128 stream cipher, validity and the integrity on the other hand are acquired by Ping-Pong-MAC.

Wu et al. [29] introduced AFTCS idea which is based on faultless connection mechanism for the body sensor networks. As the AFTCS give consistent and secure communication of data for precarious sensors by keeping the bandwidth which is regarding to the quantity of human physiological information, and the exterior environs as well as the system itself.

IV. Security and Privacy Requirements in WBANs

A. *Security and Privacy Issue to E-health*

The data which is created, stored related to patient conveyed by BANs is critical for medical diagnosis and treatment. Importance of securing this data, therefore, cannot be overemphasized. Significance of securing this information, thus, cannot be overemphasized. On the off chance that the information is tainted or appropriated, the impact could extend from ineffectual treatment of patients to abuses debilitating patients' lives.

For across the board acknowledgment by patients, as an essential piece of the healthcare framework, WBANs should likewise address privacy concerns. The conveyed idea of information in WBANs makes implementing security and privacy troublesome. It is imperative to manage these issues notwithstanding when the node is compromised or falls flat. Utilization of encryption and cryptography is picking up cash for upholding access control to secure the privacy of patients. The essential security necessities in WBANs are discussed below

a) *Data Confidentiality*

The confidentiality of data and information can be achieved by protecting data disclosure from illegal and authorized person it also involve data protection from other networks. It is an important issue and proper step should be taken so that we can save patient important information with neighboring or external networks. In a communication channel, we can obtain Data confidentiality of patient and its private or personal data by means of shared secret key and security algorithms.

b) *Scalability*

As in E-Health there are various patient related data, the distributed access control system ought to be versatile with the number of clients. The main important parts of it are low consumption and storage overhead is required. Low management overhead is required to set up and modified easily to obtain better result.

c) *Data Integrity*

The information can be modified or changed when ever data or information is transmitted in an insecure network. To protect information during transmission process where information can be altered, the error control can be guarantee by data integrity amid the transmission time of information. Despite the fact that it is hard to get error free transmission, the process of data integrity are used to confirm that data packets are not exchanged by opponents.

d) *Data Authenticity*

The term Authentication is important for medical as well as non-medical applications. Nodes which take part in communication process proved their identity by authenticity. During transmission process, the coordinator node and the member nodes require affirmation that data is being sent to a guaranteed center and not by intruder for performing some illegal actions.

e) *Data Availability*

Must provide guarantee of data availability in all health care system so that required operation can be carried out any time anywhere during emergency.

It gives an assurance and makes the work easy as they are designed in advance for patient security. Network degradation occurs if there is any problem in the network, switching to other network is important otherwise it leads to loss of life of a patient.

f) *Data Security*

In data security protection of the database from ruinous forces by maintaining secrecy is done. Due to which, an illegal and unauthorized person cannot access as well as cannot alter data. To ensure data is transmitted securely encryption technique is used.

g) *Encryption*

Data Confidentiality is carried out by algorithm such as AES, DES and Advance Data Encryption Standards (ADES).

h) *Data Privacy*

The data or information can be approved to use only by legal and authorized person in data privacy. When data is disclosed to unauthorized entities (persons) this will lead to several risk factor and medical information of a person is very sensitive issue and can be only accessed by authorized person. Non- cryptographic technique is used for designs and protection of data privacy. The privacy of source location which is essential in WBAN can be improved by a protocol called phantom routing. This protocol will remarkably expands the privacy of source location by initializing a phantom source and flooding and privacy of data is provided. The privacy of medical data must be addressed through proper mechanism by all the system to carry out data privacy in the network.

B. *Security and Privacy Related Attacks/Threats*

Comparable to any wireless network, WSNs are experiencing a wide range of assaults. In this section, we discuss threats or attacks related to WSNs.

a) *Physical Layer*

Jamming: It is an attack which meddling with the frequencies of the radio that the nodes of a network are utilizing in jamming [11], [12]. These are the typical barriers which are against the jamming incorporate varieties of wide range communication, there are two examples, and one is frequency hopping and the second is code spreading.

Tampering: It is an attack in which a node is given the physical access to intruder, or an intruder can draw delicate data, for example, cryptographic secret keys or any other confidential information on the node. Then node is controlled by the intruder which can alter or supplanted to make it compromised. To make the physical package of the node with proofing with tamper is the defense for this type of attack.

b) *Data Link Layer*

Collision: A collision happens when more than one nodes endeavor to transmit on the same frequency at the exact time. A common safeguard to prevent the collisions is the utilization of correcting codes for the errors [12].

Exhaustion: An attacker can cause collisions that can repetitively be made utilization of to cause asset exhaustion. A solution which is attainable to force rates which limits the confirmation control of MAC and can disregard the excessive requests by the network, so by repeating transmissions it can prevent energy from draining [12].

Unfairness: As opposed to blocking a service access to outright, It can be corrupted by an attacker to gain the advantage and miss their transmission due date by causing the other nodes within a real time of MAC protocol. Utilizing minor frames lessens the effect of these type attacks just with diminishing amount of the time from which attacker grab and hold the communication of the channel.

c) *Network Layer*

Selective Forwarding: Which is mischievous node endeavors in the network to block or stop packets just by dropping or dismissing messages which is going through them. What’s more, the malevolent node sends the data or information typically to the wrong or opposite path with the aim to put false routing information [13]. Utilizing various ways to send data or information provide defense against selective forwarding attacks on network layer. While the 2nd resistance is to identify the malevolent node and assume that it has fizzled and searched for an alternate route.

Sinkhole Attack: The main concern of the foe is that to draw all of the traffic from a specific region just with using the method of a traded off node, making a metaphorical sinkhole with the enemy at the middle. The attacks of Sinkhole usually succeed just with making a traded off nodes to appear especially it can be very attractive in terms of algorithm routing with neighboring nodes [14]. These are caused by the attacks of selective forwarding making it very easy and simple because almost all the traffic coming from a massive area within the network moves over the enemy's node.

Sybil Attacks: A single or one node which duplicates or multiplies himself and is introduced in multiple location. These attacks goes for the schemes of fault tolerance, for instance, the storage which is distributed, multipath routing, and topology maintenance. In these attacks, the solitary node shows numerous identities to different nodes in the network. The techniques of Authentication and the encryption methods can obstruct an outcast from beginning a Sybil attack on the sensor network.

Worm hole Attacks: In this attack, the assailant acquires some packets at one of the point within network, and then “tunnel” the packets with other points in the network, and then repeats them within the network from that point [15].

HELLO Flood Attacks: This is substantial many protocols using HELLO packets innocently expect that accepting these packets implies that the sender or transmitter is within the range of the radio also it is subsequently the neighbor. An attacker can utilize powerful transmitter which used to trick the large zone of the nodes to believe that they were neighbors of transmitting node. The Cryptography algorithm is the solution to these kinds of assaults.

d) *Transport Layer*

Flooding: When a new connection is made by an attacker then it requests again & again till the point when the assets compulsory by every connection is shattered or achieved a most extreme cutoff [16]. Answer of these types of issues is to necessitate every connecting user to prove its commitment to the connection by comprehending a puzzle.

Desynchronization: An enemy which repeatedly sends the messages which passes arrangement numbers to either at the endpoints. Requires the authentication of communication between all packets which is within hosts is one of the conceivable answers assault.

C. Regulations and Laws in Security and Privacy

Medical security and privacy is a basic prerequisite in E-Health everywhere throughout in world, so here is various rules and regulations that influence healthcare suppliers. Truth be told the regulations and acts fluctuate enormously from nation to nation. Now talk about the American in which Health Insurance Portability and Accountability Act of 1996 (HIPAA) [7] and the Health Information Technology for Economic and Clinical Health Act (HITECH) [8]. The health Insurance Portability and Accountability Act (HIPAA) orders that, in light of the fact that the sensors in WBAN gather the wearer's health knowledge, should be kept in mind to guard it from illegal access and change of state [9], [10]. As per the Act, human services suppliers are subjected with harsh punishments for exp fine of dollars $250,000 or detainment for Ten years, for the individuals which obtain and uncover the health data of the patient data for making money or malicious damage [8].

V. WBAN Security and Privacy Solutions

As we have realized from the previous sections that all the research aim is to secure the healthcare applications in wireless body area network. Clearly more research is required in E-health application, so that we can solve the security and privacy issues we have talked about in above section.

Wireless medical sensor networks make patients' life more relaxing as patient don't need to visit hospital, WBANs also give reasonable solutions to healthcare applications, for example, key sign observing, hospitals, home care, ambulatory care and as well as in the clinical examine.

To keep up solid security in an ongoing E-Health application, security and privacy should to be each stage like in application design, deployment, and implementation. HIPAA managed stringent principles for healthcare provider. In spite of the fact that it might appear sensible to utilize 128 bit AES. Due to longer secret key encryption/decryption time for this algorithm will be more, its might not be appropriate to use this algorithm.

There is also new block cipher which is more secure and suitable for low power consumption and that is known as HIGHT (high security and light weight) block cipher having 64-bit block size with 128 bit key [7].

Steam cipher, Digital signature, MAC, AES-CTR, AES-CBC-MAC, and AES-CCM based technique is more secure as compare to other but its complex, due to complexity they are difficult to implement.

A solid user verification and validation i.e. authentication protocol has not yet been tended properly at the application layer. User authentication with other possible mechanism should be introduce keeping in mind the end goal to prevent illegal user to gain access to confidential medical data of patient because its disclose can result in patient death.

We also purposed the utilization of Biometric strategy for solving security and privacy issue in WBANs since it is more proficient than other techniques used in WBANs of accomplishing all the security and privacy prerequisites. As in Biometric we make use of physical characteristics like fingerprint retina scanning and palm scanning so no one can take this characteristic so security and privacy is achieved.

Biometric is a procedure or technique which is used for providing distinguishing proof or checking of a person by his or her exceptional physiological or behavioral characteristics. WBAN conveys different security and privacy issues, for example, loss of information, authentication and access control. We recommended that biometric characteristics will be utilized which will not only increase security and privacy but also provide effectiveness in WBAN.

It is more secure, because in biometric provide defense against attack and risk influences, small key utilized, and it's more effective to implement.

A. Heart Rate Variability (HRV)

HRV is a physiological phenomenon where the time interval between heartbeats changes randomly. HRV have unique characteristics and we can use it in secure communications. HRV is measured by calculating the time between the spikes. HRV can be measured by any heart related signal; however Electrocardiogram (ECG) is the most preferred.

We have purposed the use biometric based security technique for the information verification and validation within WBANs. In particular, we use client's ECG feature as a key which is biometric based for information verification in Wireless body area network system. So, the data or the information of a patient can be detected and obtained individually from a patients assigned Wireless body area network scheme and one patient data or record is not mixed with other patients due to different physical and behavior characteristics in biometric system.

The security system implemented in biometric make use of low computational complexity and is more efficient instead of other cryptographic key distribution.

VI. Conclusion

Wireless Body Area Networks supporting healthcare applications are in early development stage yet offer significant commitments at monitoring, diagnostic, or therapeutic levels. As the WBAN sensor devices is being utilized for gathering the sensitive data and possibly will run into antagonistic situations, they require complicated and very secure security medium or structure to avoid the vitriolic communications within the system. These devices represent various security and privacy protection of touchy and private patient medical data.

We have purposed the use of Biometric technique as it's more efficient for acquiring security than other cryptographic procedures and algorithm. It is more secure, because in biometric framework there is no chance of replay eavesdropping and other such attack and threat which compromised the system security when we make use of algorithms having small key.

When we develop a security solution for WBANs we should keep in mind that it conforms to every side of WSN such as data privacy, integrity, data freshness, identity authentication, and availability which make WBANs secure. As it's necessary to recommend a new policy adaptation in emergency healthcare, a future direction is to develop better, flexible, more secure, cryptographic imposed, and attribute based access control mechanism for wireless body area network because biometric implementation is somehow complex and costly.

References

[1] M. Chen, S. Gonzalez, A. Vasilakos, H. Cao, and V. Leung, "Body area networks: A survey," Mobile Networks and Applications,vol.16,pp. 171–193, 2011.

[2] J. Xing and Y. Zhu, "A survey on body area network," in 5th Int.Conf.on Wireless Communications, Networking and Mobile Computing(WiCom '09), pp. 1 –4, Sept. 2009.

[3] Khan, J.Y., and Yuce, M.R.: 'Wireless body area networks: technology, implementation, and applications' (Pan Stanford Pub., 2012. 2012).

[4] Milenković, A., Otto, C., and Jovanov, E.: 'Wireless sensor networks for personal health monitoring: Issues and an implementation', Computer Communications, 2006, 29, (13), pp. 2521-2533.

[5] Ullah, N., Khan, P., and Kwak, K.S.: 'A Very Low Power MAC (VLPM) Protocol for Wireless Body Area Networks', Sensors, 2011,11,(4).

[6] K. Y. Yazdandoost and K. Sayrafian-Pour, "Channel model for body area network (BAN)," Networks, p. 91, 2009.

[7] Deuko Hong, et al, "HIGHT: A New Block Cipher Suitable for Low-Resource Device," CHES'06, LNCS 4249, 2006.

[8] http://waysandmeans.house.gov/media/pdf/110/hit2.pdf ,accessed on 22 December 2017.

[9] Venkatasubramanian, K. K., Banerjee, A., & Gupta, S. K. S.. "PSKA: usable and secure key agreement scheme for body area networks," IEEE transactions on information technology in biomedicine a publication of the IEEE Engineering in Medicine and Biology Society, vol. 14, 2010, pp. 60-68.

[10] Li, M., Lou, W., & Ren, K.. "Data security and privacy in wireless body area networks," IEEE Wireless Communications, IEEE Press, vol. 17, Feb. 2010, pp. 51-58, doi:10.1109/MWC.2010.5416350.

[11] CC2420 DataSheet, "C2420, 2.4GHz IEEEE 802.15. 4/ZigBee-ready RF Transceiver," Chipcon, 2006.

[12] J. Sen, "Security in wireless sensor networks," inWireless Sensor Networks: Current Status and Future Trends,2012.

[13] K. Venkatraman, J. VijayDaniel, and G.Murugaboopathi, "Various attacks in wireless sensor network: survey," International Journal of Sot Computing and Engineering,vol.3,no.1,2013.

[14] V. Soni, P. Modi, and V. Chaudhri, "Detecting Sinkhole attack in wireless sensor network," International Journal of Application or Innovation in Engineering & Management,vol.2,no.2,2013.

[15] T. K. Rao, M. Sharma, and M. V. Saradhi, "Wormhole attacks in Ad-Hoc networks," International Journal of Latest Trend in Computing,vol.4,no.2,2013.

[16] H. C. Chaudhari and L. U. Kadam, "Wireless sensor networks: security, attacks and challenges," International Journal of Networking, vol. 1, no. 1, pp. 4–16, 2011.

[17] Muhammad, K.R.R.S.; Lee, H.; Lee, S.; Lee, Y.K. BARI+: A Biometric Based Distributed Key Management Approach for Wireless Body Area Networks. Sensors 2010, 10, 3911-3933.

[18] Huang, Y.M.; Hsieh, M.Y.; Hung, H.C.; Park, J.H. Pervasive, Secure Access to a Hierarchical Sensor-Based Healthcare Monitoring Architecture in Wireless Heterogeneous Networks. IEEE J. Select. Areas Commun. 2009, 27, 400-411.

[19] Muraleedharan, R.; Osadciw, L.A. Secure Healthcare Monitoring Network Against Denial-of-Service Attacks Using Cognitive Intelligence. In Proceedings of Communication Networks and Services Research Conference, Halifax, NS, Canada, 5–8 May 2008; pp 165-170.

[20] Le, X.H.; Khalid, M.; Sankar, R.; Lee, S. An Efficient Mutual Authentication and Access Control Scheme for Wireless Sensor Network in Healthcare. J. Networks 2011, 27, 355-364.

[21] Malasri, K.; Wang, L. Design and Implementation of Secure Wireless Mote-Based Medical Sensor Network. Sensors 2009, 9, 6273-6297.

[22] Misic, J.; Misic, V. Enforcing Patient Privacy in Healthcare WSNs Through Key Distribution Algorithms. Secur. Commun. Network 2008, 1, 417-429.

[23] Haque, M.M.; Pathan, A.S.K.; Hong, C.S. Securing u-Healthcare Sensor Networks Using Public Key Based Scheme. In Proceedings of 10th International Conference of Advance Communication Technology, Pyeongchang, Korea, 19–22 February 2008; pp. 1108-1111.

[24] Hu, F.; Jiang, M.; Wagner, M.; Dong, D.C. Privacy-Preserving Tele cardiology Sensor Networks: Toward a Low-Cost Portable Wireless Hardware/Software Codesign. IEEE Trans. Inform. Tech. Biomed. 2007, 11,619-627.

[25] Dagtas, S.; Pekhteryev, G.; Sahinoglu, Z.; Cam, H.; Challa, N. Real-Time and Secure Wireless Health Monitoring. Int. J. Telemed. Appl. 2008, doi: 10.1155/2008/135808.

[26] Kang, D.O.; Lee, H.J.; Ko, E.J.; Kang, K.; Lee, J. A Wearable Context Aware System for Ubiquitous Healthcare. In Proceedings of 28th IEEE EMBS Annual International Conference, New York, NY, USA, 30 August–3 September 2006; pp. 5192-5195.

[27] Lin, X.; Lu, R.; Shen, X.; Nemoto, Y.; Kato, N. SAGE: A Strong Privacy-Preserving Scheme Against Global Eavesdropping for eHealth System. IEEE J. Select. Area Commun. 2009, 27, 365-378.

[28] Kumar, P.; Lee, Y.-D.; Lee, H.-J. Secure Health Monitoring Using Medical Wireless Sensor Networks. In Proceedings of 6th International Conference on Networked Computing and Advanced Information Management, Seoul, Korea, 16–18 August 2010; pp. 491-494.

[29] Wu, G.; Ren, J.; Xia, F.; Xu, Z. An Adaptive Fault-Tolerant Communication Scheme for Body Sensor Networks. Sensors 2010, 10, 9590-9608.

A Novel Stable Clustering Approach based on Gaussian Distribution and Relative Velocity in VANETs

Mohammed Saad Talib
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
College of Administration and Economics
University of Babylon, Babel – Iraq

Zaniab Saad Talib
Faculty of Science, University of Kerbala
Kerbala – Iraq

Aslinda Hassan, Burairah Hussin, Z.A. Abas
Center for Advanced Computing Technology
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Melaka- Malaysia

Zainab Sabah Rasoul
Faculty of Manufacturing Engineering
Universiti Teknikal Malaysia Melaka
Melaka- Malaysia

Abstract*—Vehicles in Vehicular Ad-hoc Networks (VANETs) are characterized by their high dynamic mobility (velocity). Changing in VANET topology is happened frequently which caused continuous network communication failures. Clustering is one of the solutions applied to reduce the VANET topology changes. Stable clusters are required and Indispensable to control, improve and analyze VANET. In this paper, we introduce a new analytical VANET's clustering approach. This approach aims to enhance the network stability. The new proposed grouping process in this study depends on the vehicles velocities mean and standard deviation. The principle of the normal (Gaussian) distribution is utilized and emerged with the relative velocity to propose two clustering levels. The staying duration of vehicles in a cluster is also calculated and used as an indication. The first level represents a very high stabile cluster. To form this cluster, only the vehicles having velocities within the range of mean $\pm$ standard deviation, collected in one cluster (i.e. only 68% of the vehicles allowed to compose this cluster). The cluster head is selected from the vehicles having velocities close to the average cluster velocity. The second level is to create a stable cluster by grouping about 95% of the vehicles. Only the vehicles having velocities within the range of mean $\pm$ 2 standard deviation are collected in one cluster. This type of clustering is less stable than the first one. The analytical analysis shows that the stability and the staying duration of vehicles in the first clustering approach are better than their values in the second clustering approach.

Keywords—VANET; clustering; cluster stability; Gaussian; relative velocity; staying duration

I. INTRODUCTION

Vehicular ad hoc networks (VANETs) represent a crucial view of the Intelligent Transportation System (ITS). Increasing the daily traffic represents a great challenge for the citizen of all urban and rural places. The street traffic safety and management principles continuously affect the individual's life. Focusing must be made on street safety to improve its traffic efficiency. Efficient street traffic safety and management can be achieved with the huge advancement in Information Technology (IT) by developing reliable communication among vehicles [1]. A vast development in manufacturing, wireless communication and intelligent technologies helps in equipping modern vehicles with onboard unit (OBU) wireless communication to enable vehicles communicate with their neighbors through a Vehicle-to-Vehicle (V2V) and with Road Side Units (RSUs) as a Vehicle-to-Infrastructure (V2I) or Infrastructure-to-Vehicle (I2V) communication style [2]. VANET is characterized by its high moving speed on streets. The vehicles high mobility is significantly affecting the connectivity, throughput and the process of exchanging safety or non-safety traffic information among vehicles [3], [4] VANET aims to prevent collision, accidents and to offer comfortable for passengers by disseminating emergency information in advance [5].

Dynamic and dense network topology characteristics in VANETs cause problems, such as congestion, rerouting and the hidden terminal problem. Clustering techniques have been proposed to improve the communication in VANETs by grouping nodes in geographical vicinity together to enhance communication efficiency. It can also be used for frequency reuse [6], to reduce data congestion and to support QoS requirements. By grouping the vehicles into sets of similar mobility, the relative mobility between communicating neighbor nodes can be reduced [7]. Clustering structures for VANETs are well studied and discussed in the related literatures. Various clustering approaches have been proposed in VANETs [8], [4], [9]. Grouping of vehicles in clusters may deliver certain level of routing scalability, improving resource sharing and developing communication efficiency [10].

Clustering performance is commonly evaluated by cluster stability. Clustering forms must ensure good stability with less overhead, longer Cluster Head (CH) and Cluster Member

(CM) lifetime duration, fewer numbers of changes in vehicles states and minimum CH changes [11], [12]. Most of the VANETs proposed clustering approaches are based on velocity, position, lanes, vehicles density, movement pattern and degree of connectivity [10], [9]. Ensuring stability represents the essential challenge for clustering procedures in dynamic environments. The well-organized clustering procedures must have the ability to maintain the existing cluster formation and reducing its overhead [13], [11], [14].

In this work, we present a new approach to form vehicles clusters. This approach aims to increase the network stability and to decrease its topology dynamics. The principles of the normal (Gaussian) distribution are utilized for the first time in accepting or rejecting the vehicles in the process of cluster formation. Relative velocity and staying duration are also emerged in this study as crucial decision variables. The velocities method and standard deviation are also represents the corner stone in this suggested clustering approach.

II. Clustering Process

Clustering is the process of grouping the vehicles into different sets. These sets are called clusters. Every cluster must have one vehicle that plays the role of leader or CH. CH is responsible for the communication with the CMs and with other neighbor clusters. The clustering process is used to reduce the routing overhead, suitable use of the network bandwidth and enhance the message delivery. In order to make the clustering processes more efficient, different clustering algorithms are suggested and designed. These algorithms are developed for the cluster formation and cluster maintenances. In cluster formation step, the algorithm is used to make a cluster, elect the CH, assign members, gateway and assist communication. In maintenance step, the algorithm is used to improve the links and to describe the leaving vehicles, the joining vehicles and the merging clusters [15], [11].

III. Normal Distribution

The Normal distribution is also known as Gaussian distribution is usually formed as a bell- curve. It represents an organization of a data set in which most values are groups in the central of the sequence and the others are distributed symmetrically toward both ends. Normal distribution are widely used to approximated and model several actual random phenomena in a well manner. There is a very robust association between the size of a sample n and the possibility to which a sampling distribution tends to the Gaussian shape. Several sampling distributions created on big n can be approximated by the Gaussian distribution even though the population distribution itself is definitely not Gaussian. Gaussian is a continuous time probability distribution to represent all x values from -∞ to +∞. So each possible interval of actual quantities has a probability other than zero. The Gaussian (Normal) distributed is usually denoted by N (μ, σ^2) (where μ is the data mean and σ^2 is its variance). Equation 1 represents the probability density function (f(x)) of the normal distribution. Mode, median and method have equal values in Gaussian distribution. Literatures show the percentage value of the data falls within any value of standard deviations from the mean. Their limits and percentages were empirically calculated and listed in tables. Fig. 1 shows some of these percentage values under the normal curve [16], [17].

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{1}$$

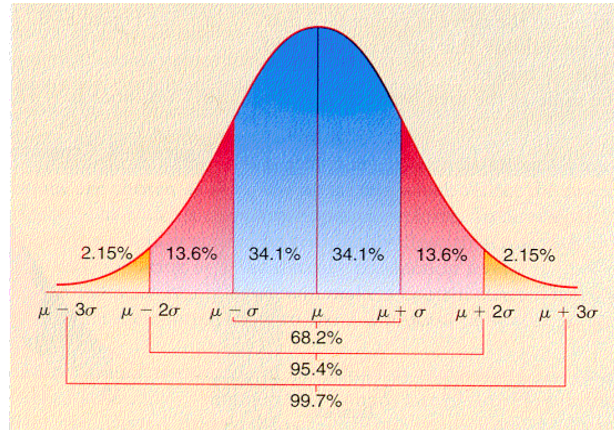


Fig. 1. Percentage values under the normal curve.

The following equations [16] are used to estimate the sample mean ($\bar{X}$,**)** and the sample variance (s^2) or the standard deviation (s).

$$\mu = \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{2}$$

$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \tag{3}$$

and;

$$\sigma = \sqrt{\sigma^2}$$

In this study, μ is used to represent the sample mean ($\bar{x}$**)** and σ is used to represent the sample standard deviation (s).

IV. Vehicles Relative Velocity

The relative velocity between any two moving vehicles can be calculated as the absolute difference between these two vehicles velocities as shown in (4) [18].

$$Relative - velocity = \left|v_j - v_i\right| \tag{4}$$

Where v_j represent the velocity of the vehicle j and v_i is the velocity of the vehicle i. The impact of the relative vehicles velocities will result in two events affecting the cluster structure. These events are some vehicles leaving the cluster and/or some vehicles entering the cluster [19]. Relative speed level among vehicles can helps in keeping the vehicles in a cluster and reducing the probability of moving out of that cluster [20], [13], tried to create new reliable cluster. They depended on the velocity difference among vehicles, vehicles direction and vehicles location in their clustering formation process. Their approach generates two groups of vehicles whose relative velocity is less and greater than each vehicle [21].

All the related literatures aim to create relatively stable cluster in VANET although the CMs and CHs are high velocity vehicles in order to extend the clusters lifetime. When the relative velocity between CMs and their CH is small then the cluster will be more stable. The vehicle having relatively small velocity difference represents the optimal choice to play the role of the CH [22]. In VANETs, the connection duration between any two vehicles depends on their relative velocity. The connectivity duration between any two vehicles is a function of the communication range and their relative velocity. The relative velocity distribution can be used to estimate the connectivity duration. Velocities Gaussian distribution models can be used to model the connectivity duration. The connectivity duration (t_c) between any two vehicles estimated using (5) [23].

$$t_c = \frac{2R}{\Delta v} \qquad (5)$$

Where *R* is the vehicle communication range and *Δv* is the relative velocity between the two vehicles.

Merging the velocity and location of the adjacent vehicles represents one of the best approaches to measure the link duration. Most of the suggested cluster merging approaches are motivated either when the distance between two adjacent CHs is less than a specific threshold [13]. Or when the CHs stay associated for a time duration larger than a predetermined value [24].

The applied performance metrics to evaluate the cluster stability are the CM and CH duration, CH change rate, clustering overhead, and number of vehicles that needs to make a decision about the next state. The average relative speed is preferred to be low which means it has less moveable compared with their adjacent. It can be elected as a CH. To maximize the connection probability, the relative velocity of the vehicles must be minimized. When the relative vehicle velocities are increasing, the probability of breaking the connection will increase [25]. Therefore, the relative velocity between two vehicles can vary from 0 to 260 km/h as indicated in certain highways zones [26].

V. Clustering Management

In this paper, the proposed model is heavily based on the clustering process. The instantaneous velocity of each vehicle represents the pivot stone in this model. The moving vehicles in the same direction are considered in this study. Each vehicle must have a unique ID number and equipped with an on-board-unit (OBU) to be able to deal with the IEEE802.11 as a Dedicated Short Range Communications (DSRC) system. The vehicles are also equipped with a Global Positioning System (GPS) device to provide information about real-time vehicle speed, direction, and land attributes. The vehicles can also proactively and periodically collect real-time traffic information.

The cluster formation process must be proposed with the purposes of reducing the number of cluster heads in the network, maximizing the duration of CH and CMs to provide certain stability level and reducing the overhead [27].

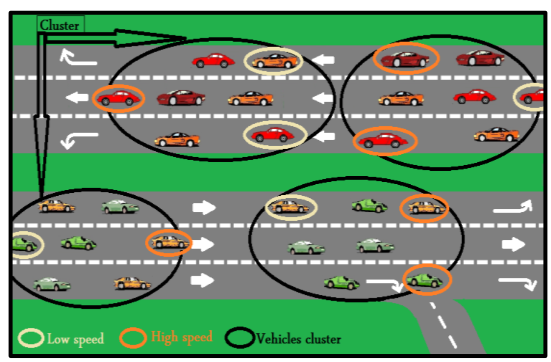
Most literatures stated that the vehicles velocities are normally distributed [28]. So the fitted normal curve can be applied and used to estimate the required parameters. These parameters are the numbers of vehicles, velocities median, mean, mode and standard deviation.

VI. Proposed Solutions

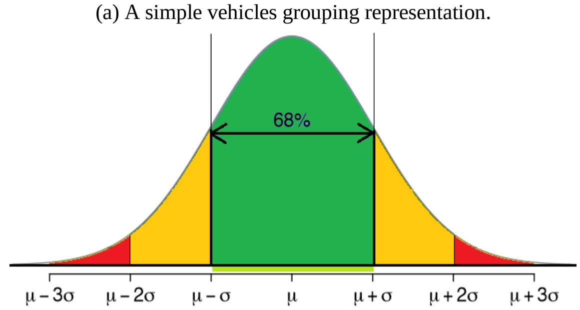
The suggested approach in this analytical study is to design new reliable stable vehicles clusters. Merging the fundamentals of the normal probability distribution function, vehicles mean velocity; velocities standard deviation, relative velocity and vehicles transmission range can be utilized in creating a novel real time clustering approach.

Due to the communication facilities, the number of vehicles in certain street segment can be calculated and controlled. There velocities and locations are changeable with time. The velocities mean (μ) and standard deviation (σ) can be calculated easily using equations (2 and 3). The following percentages and numbers can be developed:

- Area under the curve from (μ - σ to μ + σ) represents 68.2% of the total area. We can use this value to select about (68%) of the vehicles to compose a very high stability cluster and exclude about 32% of the vehicles. These excepted vehicles represent the vehicle affecting the clusters maintain. Some of these vehicles will leaving the cluster after certain time due to their very high or very low velocity. Such cluster will have long life time cluster. All the vehicles in this cluster are within a very close level of velocity. The relative velocity among them is very small. Fig. 2 shows in (a) a simple representation of the vehicles grouping and in (b) the suggested area under the normal curve with its percentage.



(a) A simple vehicles grouping representation.



(b) The suggested area under the normal curve and its percentage.

Fig. 2. A simple representation and the suggested area under the curve and its percentages.

- The area under the curve from (μ - 2σ to μ + 2σ) represents 95.4% of the total area as indicated in Fig. 1. Fig. 3 represents the limitations and the (95%) of the area under the normal curve. We can use this value to select about (95%) of the vehicles to create a stabile cluster and exclude about 5% of the vehicles. These excluded vehicles represent the vehicle affecting the cluster maintains. Some of these vehicles will leaving the cluster after certain time due to their very high or very low velocity. Such cluster has less life time compared with the previous one. Most of the vehicles in this cluster are within a close level of velocities.

There is a tangible difference among the vehicles velocities in this cluster. The relative velocity in this cluster is greater than that in the first clustering approach.

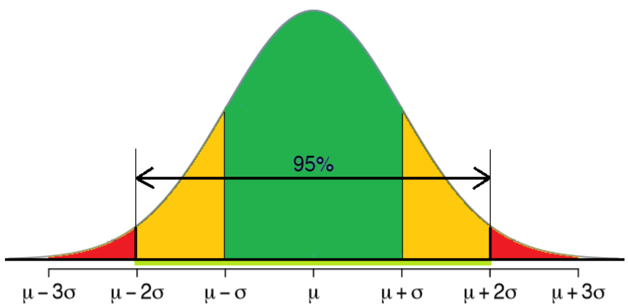


Fig. 3. The limitations and the area that represents 95% under the normal curve.

VII. Case Example

In order to implement and apply this idea, the following example is suggested. The data sample in Table I represents a suggested example of 21 vehicles (ID) and their velocities in (km/hr.) in certain street.

TABLE I. Sample of 21 Vehicles ID and Their Velocities

Vehicle ID	**Velocity Km/hr.**	**Vehicle ID**	**Velocity Km/hr.**	**Vehicle ID**	**Velocity Km/hr.**
1	70	8	95	15	70
2	90	9	100	16	60
3	60	10	90	17	120
4	75	11	80	18	100
5	100	12	75	19	95
6	120	13	110	20	85
7	110	14	90	21	90

According to equations (2 and 3):
The sample Mean μ = 89.7619
Standard deviation σ =17.4983

The first approach is to collect the vehicles within the percentage range (μ - σ to μ + σ) in one cluster. About 68% of the vehicles (13 vehicles out of 21) will be grouped to form this cluster. The following indications can be reached after analyzing the collected data with the Gaussian fundamentals:

Maximum velocity allowed = 107.2602 km/hr
Real data Maximum velocity =100 km/hr
Minimum velocity allowed = 72.2636 km/hr
Real data Minimum velocity = 75 km/hr
Velocities Range (faster - slower) = 34.9966 km/hr
Real velocities Range = 25 km/hr
Number of vehicles in this cluster = 13 vehicles
Cluster head velocity = 90 km/hr
Maximum relative velocity = 100-75 = 25 km/hr

The relative velocity between the CH and the slowest vehicle in this cluster is 15 km/hr.

The relative velocity is ranged from 0 to 25 km/hr.

The expected duration for this cluster can be estimated using equation (5) after suggesting the transmission range to be 1000 m (1km).

t_c= 2/15 = 0.1333hr. (This means that the slowest vehicle will stay about 8 min in this cluster).

t_c= 2/10 = 0.2 hr = 12 min ((this means that at least 11 vehicles (85%) will stay after 12 min in this cluster).

Table II shows the IDs and velocities for the vehicles forming this high stable cluster.

TABLE II. Vehicles ID and Velocity

ID	2	4	5	8	9	10	11	12	14	18	19	20	21
Velocity	90	75	100	95	100	90	80	75	90	100	95	85	90

The second approach is to collect the vehicles within the percentage range of (μ - 2σ to μ + 2σ) in a cluster. About 95% of the vehicles (20 vehicles) will be grouped to form this cluster. Practically, according to the velocities data range in Table I, 20 vehicles will be collected in one cluster.

Maximum velocity allowed in this cluster is = 142.2568 km/hr
Real Maximum velocity =120 km/hr
Minimum velocity allowed is = 54.7653 km/hr
Real Minimum velocity = 60 km/hr
Velocities Range (faster - slower) = 60 km/hr
Real velocities Range = 60 km/hr
Number of vehicles in this cluster = 20 vehicles
Cluster head velocity = 90 km/hr

Maximum relative velocity = 90-60 = 30 km/hr (between the CH and the slowest vehicle in this cluster). The relative velocity is ranged from 0 to 30 km/hr.

The expected duration for this cluster can be estimated using (5) after suggesting the transmission range to be 1000 m (1km).

t_c= 2/30 = 0.066 hr. (this means that the slowest vehicle will stay about 4 min in this cluster).

t_c= 2/10 = 0.2 hr = 12 min ((this means that at least 11 vehicles will stay after 3.6 min in this cluster).

The same data in Table I represents the vehicles forming this stable cluster.

VIII. Conclusion

The dynamic features of VANET require new developed approaches to form stable clusters. In this paper, we presented a new VANET cluster formation analytical approach.

According to vehicles velocities, the average (mean) and standard deviation are estimated and utilized to establish the cluster velocity limits. These velocity limits are imported from the normal (Gaussian) distribution fundamentals using the mean and standard deviation. The upper limit is estimated by adding the (k) standard deviation to the mean and the lower velocity limit by subtracting (k) standard deviation from the mean (where k = 1, 2, and 3). So, one stability level can be selected to form a cluster out of these two possible clustering levels. These limits are used to create two stable clustering levels. Any vehicle in these clusters having velocity equal or close to the average value can be elected to play the role of the CH. Each vehicle having velocity within the velocity limits registered as CM. The staying duration and relative velocity are also used in this study to evaluate and analyze the cluster maintain step. The cluster stability will be increased with the decreasing in the suggested velocity limit. The resulted relative velocity is decreased with the decreasing in velocity limit while the staying duration is increased. Developing equations to estimate the vehicles leaving and joining probabilities are proposed for the future work.

Acknowledgment

This research is fully sponsored by Universiti Teknikal Malaysia Melaka (UTeM) Zamalah Scheme.

References

[1] I. Hussain and C. Bingcai, “Cluster Formation and Cluster Head Selection approach for Vehicle Ad-Hoc Network (VANETs) using K-Means and Floyd-Warshall Technique,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 12, pp. 11–15, 2017.

[2] I. T. Abdel-Halim and H. M. A. Fahmy, “Prediction-based protocols for vehicular Ad Hoc Networks: Survey and taxonomy,” Computer Networks, vol. 130. Elsevier B.V., pp. 34–50, 2018.

[3] K. Liu, M. Li, Y. Liu, M. Li, Z. Guo, and F. Hong, “Passive Diagnosis for Wireless Sensor Networks,” 6th ACM Conf. Embed. Networked Sens. Syst., vol. 18, no. 4, pp. 1132–1144, 2010.

[4] R. S. Bali, N. Kumar, and J. J. P. C. Rodrigues, “Clustering in vehicular ad hoc networks: Taxonomy, challenges and solutions,” Veh. Commun., vol. 1, no. 3, pp. 134–152, Jul. 2014.

[5] C. Wu, S. Wang, D. Ben, X. Wang, and M. Guizani, “Dynamic cluster based price control and gateway management for VANETs,” in IEEE International Conference on Communications, 2012, pp. 5026–5030.

[6] M. Azizian, S. Cherkaoui, and A. S. Hafid, “A distributed D-hop cluster formation for VANET,” IEEE Wirel. Commun. Netw. Conf. WCNC, vol. 2016–Septe, no. Wcnc, pp. 1–6, 2016.

[7] C. Shea, B. Hassanabadi, and S. Valaee, “Mobility-based clustering in VANETs using affinity propagation,” in GLOBECOM - IEEE Global Telecommunications Conference, 2009, pp. 1–6.

[8] A. Ahizoune and A. Hafid, “A new stability based clustering algorithm (SBCA) for VANETs,” in Proceedings - Conference on Local Computer Networks, LCN, 2012, pp. 843–847.

[9] C. Cooper, D. Franklin, M. Ros, F. Safaei, and M. Abolhasan, “A Comparative Survey of VANET Clustering Techniques,” IEEE Commun. Surv. Tutorials, vol. 19, no. 1, pp. 657–681, 2017.

[10] M. Azizian, S. Cherkaoui, and A. S. Hafid, “DCEV: A distributed cluster formation for VANET based on end-to-end realtive mobility,” in 2016 International Wireless Communications and Mobile Computing Conference, IWCMC 2016, 2016, pp. 287–291.

[11] M. Ren, L. Khoukhi, H. Labiod, J. Zhang, and V. Vèque, “A mobility-based scheme for dynamic clustering in vehicular ad-hoc networks (VANETs),” Veh. Commun., vol. 9, pp. 233–241, 2017.

[12] E. Daknou, M. Thaalbi, and N. Tabbane, “Clustering enhancement for VANETs in highway scenarios,” in 2015 5th International Conference on Communications and Networking, COMNET 2015 - Proceedings, 2016, pp. 1–5.

[13] Z. Y. Rawashdeh and S. Mahmud, “A novel algorithm to form stable clusters in vehicular ad hoc networks on highways,” EURASIP J. Wirel. Commun. Netw., vol. 2012, no. 1, p. 15, 2012.

[14] S. Y. Kannekanti and G. S. P. Nunna, “An Efficient Clustering Scheme in Vehicular Ad-Hoc Networks,” in Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), 2017 IEEE 8th Annual, 2017, pp. 282–287.

[15] B. Sivakumar and K. Srilatha, “A Review on Clustering in VANET,” Int. J. Innov. Res. Comput. Commun. Eng., vol. 4, no. 5, pp. 365–373, 2016.

[16] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, Probability and Statistics for Engineers and Scientists, vol. 3rd. 2012.

[17] P. Sahoo, Probability and Mathematical Statistics. 2013.

[18] S.-C. Lo, Y.-J. Lin, and J.-S. Gao, “A Multi-Head Clustering Algorithm in Vehicular Ad Hoc Networks,” Int. J. Comput. Theory Eng., vol. 5, no. 2, pp. 242–247, 2013.

[19] K. Abboud and W. Zhuang, “Stochastic modeling of single-hop cluster stability in vehicular ad hoc networks,” IEEE Trans. Veh. Technol., vol. 65, no. 1, pp. 226–240, 2016.

[20] Z. Y. Rawshdeh and S. M. Mahmud, “Toward strongley connected clustering structure in vehicular ad hoc networks,” in IEEE Vehicular Technology Conference, 2009, pp. 0–4.

[21] I. Kim and J. Song, “Classification of Clustering Schemes in VANETs,” in Proc. of the Intl. Conf. on Advances in Computing and Information Technology-- ACIT, 2014, pp. 102–105.

[22] R. Chai, B. Yang, L. Li, X. Sun, and Q. Chen, “Clustering-based data transmission algorithms for VANET,” in 2013 International Conference on Wireless Communications and Signal Processing, WCSP 2013, 2013, no. 61102063.

[23] S. M. Abuelenin and A. Y. Abul-Magd, “Empirical study of traffic velocity distribution and its effect on VANETs connectivity,” in 2014 International Conference on Connected Vehicles and Expo, ICCVE 2014 - Proceedings, 2015, pp. 391–395.

[24] B. Hassanabadi, C. Shea, L. Zhang, and S. Valaee, “Clustering in Vehicular Ad Hoc Networks using Affinity Propagation,” Ad Hoc Networks, vol. 13, Part B, pp. 535–548, Feb. 2014.

[25] S. Ucar, S. C. Ergen, and O. Ozkasap, “Multihop-Cluster-Based IEEE 802.11p and LTE Hybrid Architecture for VANET Safety Message Dissemination,” IEEE Trans. Veh. Technol., vol. 65, no. 4, pp. 2621–2636, 2016.

[26] L. Rivoirard, M. Berbineau, M. Wahl, D. Gruyer, and P. Sondi, “Using Real-World Car Traffic Dataset in Vehicular Ad Hoc Network Performance Evaluation,” IJACSA) Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 12, pp. 390–398, 2016.

[27] R. P. Surender Pal, “VANET BASED SIMULATION USING CLUSTER,” Int. J. Latest Res. Sci. Technol., vol. 6, no. 3, pp. 25–28, 2017.

[28] Z. Khan and P. Fan, “A Novel Triple Cluster Based Routing Protocol (TCRP) for VANETs,” IEEE Veh. Technol. Conf., vol. 2016–July, 2016.

Comparison between NFC/RFID and Bar Code Systems for Halal Tags Identification: Paired Sample T-Test Evaluation

Mohsen Khosravi, Najma Imtiaz Ali, Mostafa Karbasi*, Imtiaz Ali Brohi, Irfan Ahmed Shaikh**, Asadullah Shah
Department of Information Systems,
Kulliyah of Information Communication and Technology (KICT), International Islamic University Malaysia, Gombak,
Kuala Lumpur, Malaysia
*University of Ghiaseddin Jamshed Kashani, Iran
**University of Sindh, Jamshoro

***Abstract*—Malaysia is a modern Muslim country where the research on Halal product identification is at the peak. In this study, the authors have developed the mobile application which is based on Radio frequency identification near field communication RFID/NFC. The author first developed the database based on the data from Jabatan Kemajuan Islam Malaysia JAKIM, which is Malaysian Halal logo identification authority then the mobile application which uses the Near Field Communication to detect the Halal food using the Radio Frequency Identification. In this paper authors have performed the experimental analysis by comparing the Barcode system that is comprised of parallel line detected by the simple webcam for the Halal logo identification and the new developed RFID NFC mobile application. Paired sample T-Test was performed by using the SPSS 23.0 version. The results revealed that there is significantly difference between the usability, efficiency, affordability, security and satisfaction. The users are more satisfied with the newly developed mobile application as compared to old halal logo system in Malaysia.**

Keywords*—Halal products; RFID/NFC; JAKIM; paired sample T-test; Malaysia

I. Introduction

Malaysia is modern Muslim country heading towards being a global Halal hub; it has been well-known as a successful halal-hub in the world since 1997. Halal is often used in reference to food and drinks, i.e. food that is allowable for Muslims to eat or drink under Islamic Shariah (law). Halal is an Islamic religious manual for how Muslims ought to experience their lives from the way their nourishment is set up to how their own and social connections are directed.

In Halal industry, keeping up halal item trustworthiness is a fundamental variable so it is necessary to present a complete and fitting Tracking and Tracing Technology to keep up halal item uprightness and build up an innovative structure that can bolster the whole Halal Product Supply Chain [1]. Information technology (IT) is supposed to be the best formula to be used to make the world a more helpful and viable place for individuals from a wide range of culture and religion. Nowadays, smart spaces and smartphones are going to lead the world of business.

Barcode system is made with the help of Matlab and they are very to detect through the webcam [2] "Near Field Communication (NFC) is short-range, low bandwidth, high frequency, and wireless communication technology built on radio frequency identification (RFID) technology" [3]. RFID refers to Radio Frequency Identification and is a term that portrays an arrangement of ID [4]. RFID depends on putting away and remotely recovering data or information as it comprises of RFID tag, RFID peruse and back-end Database [5].

The mobile RFID empowers business to give new administrations to portable clients by securing administrations and exchanges from the end-client to an organization's current online business and IT frameworks [6]. RFID innovation has been acknowledged as an execution differentiator for an assortment of business applications, yet its capacity is yet to be completely used. In future brilliant spaces, the client ought to have the capacity to utilize his own particular versatile reader gadgets to recognize the things, look for the following thing of intrigue, restrict and explore. Fig. 1 demonstrates how RFID technology works.

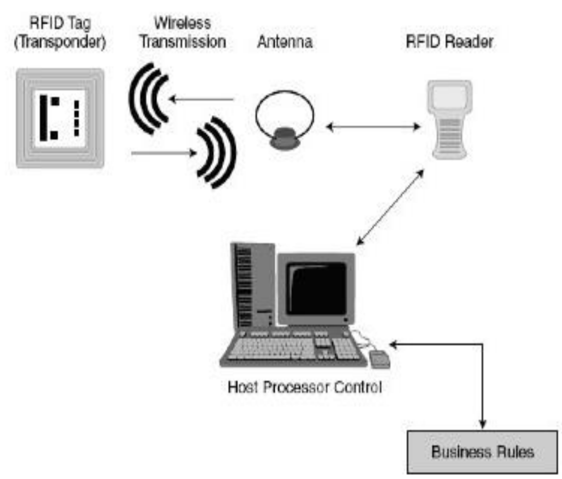


Fig. 1. RFID technology process.

II. Background

At the 2007 world halal forum, Malaysia's former Prime Minister, Tun Abdullah Haji Ahmad Badawi stated the government's aim of establishing Malaysia as a global halal hub. Subsequently, a large number of studies were conducted in various disciplines to help pursue the goal of "Malaysia as a global halal hub" [7]. One of the controversial issues in the halal food industry is detecting fake halal products from genuine ones. Malaysian Muslim specialist on the halal item, Jabatan Kemajuan Islam Malaysia (JAKIM) has created and actualized the halal logo as the halal approval due to the demand of market broker organizations. To check the originality of halal products, customers could either browse JAKIM's website or call JAKIM's office via phone.

On the one hand, these methods are time-consuming but on the other hand, cases misrepresentation and abuse of the halal logo have progressively been accounted by purchasers [8]. To handle this issue now some researchers stated to do research on this and have implemented RFID technology while others have suggested using mobile or Smartphone gadget for validating halal products. In the Malaysian Halal industry, RFID innovation is viewed as undeveloped since standardized tags are regularly utilized as programmed identifiers regardless of the presence of web-based interfaces and cell phone applications [9]. However, Muslim consumers still face problems in validating halal products.

My MobiHalal 2.0 is a mobile-based support application for Muslims to identify the Halal status [10] so that users can send and receive MMS as answers to their queries instead of entering 13-character barcodes in the SMS. This study thus discusses the barcode conception and its functions in customer product industry.

Another study has implemented 2-D barcode halal logo detector to identify halal products and UV hologram to spot fake halal logo [11]. The scanned image is decoded and used to match a database. Prototype hardware as halal detector device has also been designed so that the customer can hold the product in front of it, then the status of the product will appear on the device screen.

Another study proposes a system that would help the consumer to validate halal products through Smartphone barcode captured straightaway on time with the applied halal product alert database system [12]. The system is called MyHalal designed to focus on a new database structure which details the company's information, Halal certificate expiry alert and new applications technique using Smartphone without accessing the network. The only operating system that customers need to have is a smartphone with a minimum Android 2.1.

Other researchers have implemented RFID technology to identify and validate halal status [13]. One study states that barcodes, reader and ingredient information by far are not adequate to validate the information claimed by the manufacturer or food producer; instead, a real-time tool is needed to feed users with genuine and validated information to assist user-buying process that is RFID technology. The perception of Malaysian customers was measured and the result indicates users agreed that a real-time system is required for the information dissemination [14].

It is imperative for specialists and makers to give redress data since buyers depend basically on item bundling, including the Halal logo, fixings and producers so as to approve Halal status [15].

III. Evaluation

The important phase of any research is the evaluation process to determine the suitability or accuracy. The NFC enabled RFID mobile application was developed for the Halal logo recognition [16], [17]. The comparison between the Barcode system and the new RFID system was performed. To evaluate this research many qualifications attributed to this application were considered such as usability, efficiency, security, affordability and customer satisfaction. Usability referred to how easy the system interface is for the user and how easily a user can use the system [18]. The identification process might be quick for both using RFID and barcodes technology, but barcodes require special equipment and also more concentration on scanning process. The simplicity of use and inexpensive equipment in RFID identification technology increases the usability of this technology.

Efficiency can be defined as the lowest time by the system to perform the same job as compared to the old system [19]. Additionally, except for other technologies such as barcode reader which reads the codes from a printed reader, RFID does not need to read any pre-printed codes or signs. In order to identify products using barcode system, the reader must be clearly printed and without any damage to the reader. According to experience and studies, barcodes are easy to damage and sensitive to the environment. Moreover, the reader must be clean to obtain more effectiveness scan rates.

Furthermore, images for scanning the barcode, must be properly captured and need more concentration and focus on the capturing process which makes it less efficient. The RFID technology uses a wide range of radio frequency. Therefore, tags are easy to read by any customer and smartphone in any position. The third factor was very important for any system that is security [20], which referred to the process and methodologies to keep the personal data safe and secure [21]. The biggest issues in Halal product identification by consumers in Malaysia are misrepresentations of logos and product certificates. It is necessary to employ the technology which reduces fraudulent activities and increases the confidence of customers on shopping. The existing barcodes and logos are easy to reproduce and cheat the consumers. Therefore, Halal product identification methods, for example, online interface and SMS applications have been introduced to settle this issue. Using web portal has its own difficulties like consumers have to connect to the Internet and know how to search through the web portal. RFID technology offers a better security in product identification by using preregistered unique identification code for each product type. The tags are safe and protected against overwriting and altering the information which increases high security in the identification process.

The fourth factor which was evaluated was affordability it can be defined as "to bear the cost of without serious inconvenience" [22]. Despite barcode reader which is fixed to each product, the RFID tags are reusable for any other products or the same product for a long time. Moreover, new NFC tags are inexpensive and easily available which makes them more affordable to use; also the application is free of cost for consumers. The fifth factor which was considered as very imperative is customer satisfaction. Customer satisfaction is at the core of human experience, reflecting our liking of an organization's business activities. High levels of customer satisfaction (with pleasurable experiences are strong predictors of customer retention, customer loyalty, and product repurchase. Customer satisfaction is an important factor to remain in business in this modern world of competition [23]. Customer satisfaction is also a major goal of process improvement plans. Quality characteristics which are described in the previous sections are essential elements for customer satisfaction for this application.

In this section the evaluation is compared with traditional product identification (barcode) with developed RFID. Moreover, the other evaluation characteristics to evaluate the respondents' concern in the questionnaires. Section 1 was comprised of personal behavior including the questions regarding sex, age, householder status, marital status, educational level, job status, income per month, shopping times and system awareness. Section 2 was comprised of 30 questions related to the usability, efficiency, security, affordability and customer satisfaction. It is challenging to compare traditional and developed identification technologies particularly barcodes to newly developed ones like RFID. One had wide testing in uncommercial condition and alternate had restricted business reputation

IV. Methodology

In this paper, author used the Statistics Package for Social Sciences SPSS 23.0 to perform the experimental analysis. Paired Sample T-test was performed to compare the results of the old system and new system. The self-administrative questionnaire was distributed at one of the well-known and very big shopping Mall of Kuala Lumpur. The authors distribute the questionnaire for the old system, the respondent fills them and then the authors explain the use of the new system and do the practise among the many peoples. Although it took a long time more than 3 months to self-collect the data but the data collection phase was successful. Total 300 questionnaires were distributed. 280 were received which is a good number. Out of 280, 30 were having missing data so the pair of 30 questions were removed and finally we have 250 respondent. The respondent demographic details are given in Table I.

For sex there were (n=139,55.6%) were female whereas (n=111,44.4%) were male. For age (n=109,43.6%) were between the range 31-40, followed by range 21to30 were (n=48,19.2%) after that (n=40,16.0%) were between 41to50. (n=28,11.2%) were above 50 and the lowest range were (n=25,10%) were less than 20.In the reply of Are you house holder (n=177,70.8%) were answered Yes whereas (n=73,29.2%) were answered NO. In response to marital status (n=128,51.2%) respond against married with children, followed by (n=65, 26.0%) were Single and (n=57, 22.85) were married without children.

TABLE I. Demographic Details of the Respondent

Variable	Category	Frequency	Percentage (%)	Cumulative
Sex	Male	111	44.4	44.4
	Female	139	55.6	100.0
Age	Less than 20	25	10.0	10.0
	21 to 30	48	19.2	29.2
	31 to 40	109	43.6	72.8
	41 to 50	40	16.0	88.8
	Above 50	28	11.2	100.0
Are you a householder	yes	177	70.8	70.8
	No	73	29.2	100.0
Marital status	Single	65	26.0	26.0
	Married without children	57	22.8	48.8
	Married with children	128	51.2	100.0
Education	Less than High school	93	37.2	37.2
	Bachelor	132	52.8	90.0
	Master	20	8.0	98.0
	Doctoral	5	2.0	100.0
Job Status	Manager	20	8.0	8.0
	staff	30	12.0	20.0
	customer	200	80.0	100.0
Monthly Income (in Ringgit)	Under 2000 RM	74	29.6	29.6
	2000 t0 3000 RM	43	17.2	46.8
	3000 t0 4000 RM	38	15.2	62.0
	4000 t0 5000	37	14.8	76.8
	Above 5000 RM	58	23.2	100.0
How often you go for shopping	Once in a week	84	33.6	33.6
	Once in 2 weeks	67	26.8	60.4
	Once in 3 weeks	64	25.6	86.0
	Once in month	35	14.0	100.0
Did you aware of this System	No	250	100	100.0
	yes	0	.0	00

V. Instrument Reliability

The reliability of items used in the questionnaire was conducted through the consistency test that is an evaluation of Cronbach's Alpha. If the value of Cronbach alpha is o.7 or above then there is a strong relationship between the items used in the questionnaire [24].

1) Reliability Test for New System (RFID)

The overall Cronbach's alpha reliability for the items in the questionnaire was revealed .936 (Table II). There were total thirty items in the questionnaire. The summary item statistics as in Table III for this construct showed the mean value of the items means was 4.945 as shown in Table III.

TABLE II. Reliability Statistics (RFID System)

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	No. of Items
.936	.935	30

TABLE III. Summary Item Statistics (RFID System)

	Mean	Min.	Max.	Range	Max./ Min.	Variance	No. of Items
Item Means	4.945	4.610	5.154	.543	1.118	.021	30
Item Variances	2.737	1.991	3.484	1.493	1.750	.109	30
Inter-Item Covariances	.896	-.157	2.476	2.633	-15.727	.290	30

Table IV shows individual Cronbach's alpha value for the constructs used for the questionnaire of new RFID/NFC device. There were 6 items in each construct and all having reliability > 0.7 which is considered as the acceptable [25], [26].

TABLE IV. Overall Reliability for the RFID System

Construct	No. of Items	Cronbach's Alpha	Reliability Result
Usability	6	0.915	Excellent
Efficiency	6	0.946	Excellent
Affordability	6	0.931	Excellent
Security	6	0.884	Good
Satisfaction	6	0.931	Excellent

2) Reliability Test for Old System (Bar Code)

The overall reliability statistics of the constructs (Table V) revealed .820 Cronbach's alpha reliability for all construct, which comprised thirty items. The summary item statistics for all construct showed the mean value of the items means was 2.895 as shown in Table VI.

TABLE V. Reliability Statistics (Barcode System)

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	No. of Items
.820	.820	30

TABLE VI. Summary Item Statistics (Barcode System)

	Mean	Min.	Max.	Range	Max. / Min.	Variance	No. of Items
Item Means	2.895	2.536	3.524	.988	1.390	.057	30
Item Variances	2.767	1.776	3.576	1.800	2.014	.255	30
Inter-Item Covariances	.364	-.287	2.744	3.030	-9.573	.499	30

Table VII shows the overall reliability of the questionnaire for old Barcode system. There were 6 items in each construct and all having reliability > 0.7 which is considered as acceptable.

TABLE VII. Overall Reliability of Old System (Bar Code)

Construct	No. of Items	Cronbach's Alpha	Reliability Result
Usability	6	0.814	Good
Efficiency	6	0.950	Excellent
Affordability	6	0.930	Excellent
Security	6	0.843	Good
Satisfaction	6	0.733	Acceptable

A. Descriptive Analysis New System (RFID System)

In order to perform the parametric test, such paired sample T-test which was performed to evaluate the mean difference data should be normalized [27], [1], therefore authors have performed the descriptive analysis before performing the T-test. This section presents descriptive statistics of survey for a new system that is RFID based system, descriptive analysis is mandatory to conduct in order to perform the paired sample T-test to see whether our data is normalized or not. The descriptive analysis for the new system as follows.

To measure the usability, efficiency, security affordability and customer satisfaction of RFID (NFC) system, the survey respondents were first asked to indicate their level of agreement with the usability of the new system on a seven-point Likert scale ranging from 1 (Strongly disagree) to 7 (Strongly agree). Six items survey instrument were used to measure this construct. The results of the respondents' ratings for each item of this construct are reported as follows. The data is normalized if the value of Skewness and Kurtosis fall between (less than ± 1) and the mean values should be more than Neutral value [27] i.e. 4 in this research. The mean scores for usability as shown in Table VIII was ranged between 4.99 (1.688) and 5.12(1.825),the mean values are above the neutral value (i-e 4) , that shows that the new system is easy to be used by the respondent, while data normality values i-e Skewness and Kurtosis found between the acceptable range (less than ± 1) The mean scores for efficiency as shown in Table IX was ranged between 4.95 (1.555) and 5.12(1.408), which indicates that users are satisfied with the efficiency of new system, however data normality values i.e. Skewness and Kurtosis found between the acceptable range (less than + 1). Furthermore, the mean scores values for affordability shown in Table X was ranged between 4.68 (1.609) and 4.91(1.562), while data normality values i.e. Skewness and Kurtosis found between the acceptable range (less than + 1), which indicates that the new system can be afforded by the customers. The mean scores for security was ranged between 4.70 (1.669) and 5.11(1.594) as shown in Table XI, which indicates that users are satisfied with the security of the new system, furthermore, the data normality values i.e. Skewness and Kurtosis found between the acceptable range (less than + 1). The descriptive analysis for the customer satisfaction was evaluated as mean scores ranged between 4.61 (1.724) and 5.12(1.555) as shown in Table XII, While data normality values i.e. Skewness and Kurtosis found between the acceptable range (less than ± 1), which shows the satisfaction of the customer with the new system.

TABLE VIII. DESCRIPTIVE STATISTICS OF MEASURED ITEMS OF USABILITY FOR RFID NFC SYSTEM NFC SYSTEM

	Mean Statistics	Std. Deviation Statistics	Variance Statistics	Skewness Statistics	Kurtosis Statistics
U1	5.01	1.864	3.474	-.858	-.346
U2	5.08	1.786	3.190	-.842	-.357
U3	5.12	1.825	3.332	-.820	-.406
U4	4.95	1.760	3.098	-.813	-.213
U5	5.08	1.688	2.850	-.772	-.328
U6	4.99	1.668	2.783	-.787	-.193

TABLE IX. DESCRIPTIVE STATISTICS OF MEASURED ITEMS OF EFFICIENCY FOR RFID NFC SYSTEM NFC SYSTEM

	Mean Statistics	Std. Deviation Statistics	Variance Statistics	Skewness Statistics	Kurtosis Statistics
EF1	4.98	1.603	2.570	-.705	-.335
EF2	5.00	1.627	2.647	-.750	-.275
EF3	4.95	1.555	2.419	-.675	-.248
EF4	4.96	1.561	2.436	-.693	-.256
EF5	5.02	1.557	2.425	-.676	-.312
EF6	5.15	1.408	1.983	-.642	-.420

TABLE X. DESCRIPTIVE STATISTICS OF MEASURED ITEMS OF AFFORDABILITY FOR RFID NFC SYSTEM

	Mean Statistics	Std. Deviation Statistics	Variance Statistics	Skewness Statistics	Kurtosis Statistics
AF1	4.72	1.713	2.935	-.675	-.514
AF2	4.91	1.562	2.441	-.745	-.096
AF3	4.72	1.687	2.846	-.517	-.676
AF4	4.80	1.751	3.067	-.614	-.708
AF5	4.81	1.683	2.834	-.526	-.706
AF6	4.68	1.609	2.590	-.408	-.727

TABLE XI. DESCRIPTIVE STATISTICS OF MEASURED ITEMS OF AFFORDABILITY FOR RFID NFC SYSTEM

	Mean Statistics	Mean Statistics	Std. Deviation Statistics	Variance Statistics	Skewness Statistics
SE1	5.11	1.594	2.541	-.733	-.245
SE2	5.01	1.532	2.347	-.721	-.099
SE3	4.70	1.669	2.784	-.522	-.657
SE4	5.03	1.574	2.479	-.684	-.285
SE5	5.03	1.585	2.511	-.800	-.081
SE6	4.95	1.695	2.872	-.774	-.313

TABLE XII. DESCRIPTIVE STATISTICS OF MEASURED ITEMS OF SATISFACTION FOR RFID NFC SYSTEM

	Mean Statistics	Std. Deviation Statistics	Variance Statistics	Skewness Statistics	Kurtosis Statistics
SAT1	5.12	1.555	2.418	-.708	-.318
SAT2	4.88	1.598	2.555	-.596	-.443
SAT3	5.02	1.629	2.653	-.558	-.646
SAT4	4.61	1.724	2.973	-.357	-.913
SAT5	4.96	1.782	3.176	-.723	-.562
SAT6	5.03	1.682	2.830	-.763	-.320

B. Descriptive Analysis for Old System (Bar Code)

This section presents descriptive statistics of survey for the old system that is Bar Code based system.

To measure the customer satisfaction towards the old system, the survey respondents were also asked to indicate their level of agreement with the usability, efficiency, affordability, security and customer satisfaction on a seven-point Likert scale ranging from 1 (Strongly disagree) to 7 (Strongly agree). The same six items survey instrument was used to measure the construct that we used for the new system. The mean scores for the usability ranged between 2.85 (1.707) and 3.52(1.878) as shown in Table XIII, which indicates the means values lower than the neutral value (i.e. 4) it shows that the old system is not that much user friendly and not easy to use while data normality values i.e. Skewness and Kurtosis found between the acceptable range (less than + 1).

The mean scores for the efficiency was ranged between 2.54 (1.333) and 2.61(1.554) as shown in Table XIV, which indicates the means values lower than the neutral value (i.e. 4) therefore the old system is not that much efficient to use. While data normality values i-e Skewness and Kurtosis found between the acceptable range (less than + 1). The Mean values for affordability was lower than the neutral value (i-e 4), that indicates that the old system is not affordable by the respondent. The mean scores ranged between 2.54 (1.333) and 2.61(1.554) as shown in Table XV, which indicates the means values lower than the neutral value (i.e. 4) while data normality values i.e. Skewness and Kurtosis found between the acceptable range (less than + 1). Additionally, the mean scores ranged between 2.87 (1.693) and 3.02 (1.726) for the security construct, which indicates the means values lower than the neutral value (i.e. 4) it shows that the old system is not that much secure to use. While data normality values i.e. Skewness and Kurtosis found between the acceptable range (less than + 1) as shown in Table XVI.

The mean score for the customer satisfaction was ranged between 2.71 (1.483) and 3.41(1.891) as shown in Table XVII, which indicates the means values lower than the neutral value (i.e. 4) it shows that the old system is not that much secure to use while data normality values i.e. Skewness and Kurtosis found between the acceptable range that is (less than ± 1).

TABLE XIII. DESCRIPTIVE STATISTICS OF MEASURED ITEMS OF USABILITY FOR OLD SYSTEM

	Mean Statistics	**Std. Deviation Statistics**	**Variance Statistics**	**Skewness Statistics**	**Kurtosis Statistics**
U1	2.85	1.707	2.914	.990	.010
U2	2.89	1.831	3.353	1.017	-.063
U3	2.81	1.519	2.308	1.001	.452
U4	3.07	1.657	2.746	.862	-.120
U5	2.95	1.806	3.262	1.006	.002
U6	3.52	1.878	3.528	.417	-.991

TABLE XIV. DESCRIPTIVE STATISTICS OF MEASURED ITEMS OF USABILITY FOR OLD SYSTEM

	Mean Statistics	**Std. Deviation Statistics**	**Variance Statistics**	**Skewness Statistics**	**Kurtosis Statistics**
EF1	2.61	1.554	2.415	1.035	.270
EF2	2.57	1.469	2.158	.961	.188
EF3	2.54	1.333	1.776	1.062	.576
EF4	2.56	1.455	2.118	.995	.320
EF5	2.57	1.469	2.158	.961	.188
EF6	2.54	1.333	1.776	1.062	.576

TABLE XV. DESCRIPTIVE STATISTICS OF MEASURED ITEMS OF USABILITY FOR OLD SYSTEM

	Mean Statistics	**Std. Deviation Statistics**	**Variance Statistics**	**Skewness Statistics**	**Kurtosis Statistics**
AF1	3.03	1.733	3.003	.856	-.123
AF2	2.84	1.796	3.225	.941	-.104
AF3	2.93	1.588	2.521	.944	.001
AF4	2.90	1.603	2.569	.801	-.137
AF5	2.98	1.728	2.987	.912	.001
AF6	2.97	1.781	3.172	.823	-.319

TABLE XVI. DESCRIPTIVE STATISTICS OF MEASURED ITEMS OF SECURITY FOR OLD SYSTEM

	Mean Statistics	**Std,. Deviation Statistics**	**Variance Statistics**	**Skewness Statistics**	**Kurtosis Statistics**
SE1	2.92	1.633	2.668	.938	-.004
SE2	2.87	1.693	2.867	.883	-.039
SE3	3.08	1.768	3.126	.801	-.484
SE4	2.98	1.826	3.333	.914	-.240
SE5	3.02	1.726	2.980	.896	-.186
SE6	3.20	1.741	3.031	.705	-.534

TABLE XVII. DESCRIPTIVE STATISTICS OF MEASURED ITEMS OF USABILITY FOR OLD SYSTEM

	Mean Statistics	**Std. Deviation Statistics**	**Variance Statistics**	**Skewness Statistics**	**Kurtosis Statistics**
SAT1	3.01	1.850	3.422	.902	-.274
SAT2	3.41	1.891	3.576	.501	-.928
SAT3	3.00	1.655	2.739	.820	.108
SAT4	2.75	1.658	2.749	.997	.307
SAT5	2.71	1.483	2.198	1.00	.539
SAT6	2.78	1.528	2.335	.866	.303

C. Paired Sample T-Test

Paired Sample T-Test can be performed when there is only one group of people and we need to collect data from them on two different occasions, and if there is sig(2 tailed) value less than 0.5 than there is significant difference between the two scores [27]. "A paired-samples T-test compares the mean of two matched groups of people or cases or compares the mean of a single group, examined at two different points in time. If the same group is tested again, on the same measure, the t-test is called a repeated measures t-test" [28]. In this section, we will perform the paired sample T-test between the average means of constructs obtain from the survey of new system data and the old system data. For this purpose, we compute the average means for all the constructs separately for the old system and new system and then we performed the Pair sample T-test between means of new system and Means of the old system as follows.

Paired sample T-test was performed between the usability construct of RFID (NFC) system and the old system. Table XVIII shows that total mean for RFID (NFC) system is 5.0373 whereas total mean for usability of the old system is 3.0147. Table XIX shows the result of paired sample T-test. The mean difference is computed as 2.0226 with standard deviation of 1.953, whereas the t statistics is 16.360, degree of freedom df is 249 with p-value that is sig(2-tailed) is .000, it means there is significant difference in the usability of new system and old system and the new system is more usable as compared to the old one $t(249) = 16.369$, $p \leqslant .05$ as suggested by Julie Pallant (2011).

Paired sample T-test was also performed between the efficiency construct of RFID (NFC) system and the old system. Table XX shows that total mean for RFID (NFC) system is 5.010 whereas a total mean for usability of the old system is 2.564. Table XXI shows the result of paired sample T-test. The mean difference is computed as 2.446 with standard deviation of 1.834, whereas the t statistics is 21.080, degree of freedom df is 249 with p-value that is sig(2-tailed) is .000, it means there is significant difference in the efficiency of new system and old system and the new system is more efficient as compared to the old one $t(249) = 21.080$, $p \leqslant .05$.

Paired sample T-test between the efficiency construct of RFID (NFC) system and the old system was performed and Table XXII shows that total mean for RFID (NFC) system is 4.772 whereas total mean for affordability of old system is 2.942. Table XXIII shows the result of paired sample T-test. The difference between the mean of new system and old system was 1.8300 with standard deviation of 2.121, furthermore the t statistics is 13.637, degree of freedom df is 249 with p-value that is sig(2-tailed) is .000, it means there is significant difference in the affordability of new system and old system and the new system is more affordable as compared to the old one $t(249) = 13.637$, $p \leqslant .05$.

Paired sample T-test was performed between the security construct of RFID (NFC) system and the old system. Table XXIV shows that total mean for RFID (NFC) system is 4.968 whereas total mean for the security of the old system is 3.012. Table XXV shows the result of paired sample T-test. The difference between the mean of new system and old system was 1.956 with standard deviation of 1.771, furthermore the t statistics is 17.462, degree of freedom df is 249 with p-value that is sig(2-tailed) is .000, it means there is significant difference in the security of new system and old system and the new system is more secure as compared to the old one $t(249) = 17.462$, $p \leqslant .05$

Paired sample T-test was performed between the customer satisfaction construct of RFID (NFC) system and the old system. Table XXVI shows that total mean for RFID (NFC) system is 4.942 whereas total mean for the security of the old system is 2.942. Table XXVII shows the result of paired sample T-test. The difference between the mean of new system and old system was 2.000 with standard deviation of 1.633, furthermore the t statistics is 19.354, degree of freedom df is 249 with p-value that is sig(2-tailed) is .000, it means there is significant difference in the satisfaction level of new system and old system and the respondent are more satisfied with new system as compared to the old one $t(249) = 19.354$, $p \leqslant .05$

TABLE XVIII. PAIRED SAMPLES STATISTICS (USABILITY)

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Usability new system	5.0373	250	1.47905	.09354
	Usability old system	3.0147	250	1.25026	.07907

TABLE XIX. PAIRED SAMPLES TEST (USABILITY)

		Paired Differences						df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Usability new system – usability old system	2.02267	1.95382	.12357	1.77929	2.26604	16.369	249	.000

TABLE XX. PAIRED SAMPLES STATISTICS (EFFICIENCY)

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Efficiency new system	5.0102	250	1.37775	.08714
	Efficiency old system	2.5640	250	1.28611	.08134

TABLE XXI. PAIRED SAMPLES TEST (EFFICIENCY)

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Efficiency new system - Efficiency old system	2.44621	1.83484	.11605	2.21765	2.67476	21.080	249	.000

TABLE XXII. PAIRED SAMPLES STATISTICS (AFFORDABILITY)

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Affordability new system	4.7727	250	1.43530	.09078
	Affordability old system	2.9427	250	1.46833	.09287

TABLE XXIII. PAIRED SAMPLES TEST (AFFORDABILITY)

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Affordability new system - Affordability old system	1.83007	2.12191	.13420	1.56576	2.09439	13.637	249	.000

TABLE XXIV. PAIRED SAMPLES STATISTICS(SECURITY)

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Security new system	4.9689	250	1.27688	.08076
	Security old system	3.0127	250	1.29658	.08200

TABLE XXV. PAIRED SAMPLES TEST (SECURITY)

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Security new system - Security old system	1.95619	1.77126	.11202	1.73555	2.17683	17.462	249	.000

TABLE XXVI. PAIRED SAMPLES STATISTICS(SATISFACTION)

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Satisfaction new system	4.9427	250	1.18880	.07519
	Satisfaction old system	2.9427	250	1.10213	.06970

TABLE XXVII. PAIRED SAMPLES TEST (SATISFACTION)

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Satisfaction new system - Satisfaction old system	2.00000	1.63395	.10334	1.79647	2.20353	19.354	249	.000

VI. CONCLUSION

In this paper, the evaluation of the proposed system has been performed. Results were demonstrated to compare the proposed system with the old systems which are described in the state of art section in the literature review. According to the results, the proposed system using RFID (NFC) technology for Halal food identification gained better satisfaction among all customers in any quality characteristics as compared to old Bar Code system. The usability, efficiency, security, affordability and customer satisfaction characteristics have been qualified for evaluation purpose. Moreover, the other characteristics of respondents have been evaluated to demonstrate the better results in using the RFID system for Halal food. This system will be beneficial from the commercial point of view as well as the customer point of view, as it is easy to use, efficient, and affordable. Secure and gain the customer satisfaction.

REFERENCES

[1] Bahrudin, Siti Sarah Mohd, Mohd Iskandar Illyas, and Mohamad Ishak Desa. 2011. “Tracking and tracing technology for halal product integrity over the supply chain.” Electrical Engineering and Informatics (ICEEI), 2011 International Conference on. 1-7.

[2] Khairudin, N. E. Z. Bte. Halal Food Recognition System Using Barcode Nor Emilia Zetfy Bte Khairudin Report Submitted In Partial Fulfilment Of The Requirements For The Award Of the degree of Bachelor of Computer Sciences (Software Engineering) Faculty of Computer System & Softwa. (2010).

[3] Brohi, I. A. et al. Near field communication enabled payment system adoption: A proposed framework. *2017 IEEE 3rd Int. Conf. Eng. Technol. Soc. Sci.* 1–5 (2017). doi:10.1109/ICETSS.2017.8324199

[4] Bohn, Jurgen. 2006. “Prototypical implementation of location-aware services based on super-distributed RFID tags.” In Architecture of Computing Systems-ARCS 2006, 69-83. Springer.

[5] Schwieren, Joachim, and Gottfried Vossen. 2009. “A design and development methodology for mobile RFID applications based on the ID-Services middleware architecture.” Mobile Data Management: Systems, Services and Middleware, 2009. MDM'09. Tenth International Conference on. 260-266.

[6] Ahsan, Kamran, Hanifa Shah, and Paul Kingston. 2010. “RFID applications: An introductory and exploratory study.” arXiv preprint arXiv:1002.1179.

[7] Tan, Mohd Iskandar Illyas, Raziah Noor Razali, Mohammad Ishak Desa, and others. 2012. “Factors influencing ICT adoption in halal transportations: A case study of Malaysian halal logistics service providers.” International Journal of Computer Science Issues (International Journal of Computer Science Issues (IJCSI)) 9 (1): 62-71.

[8] Ali, N. I., Samsuri, S., Sadry, M., Brohi, I. A., & Shah, A. (2017). Online shopping satisfaction in Malaysia: A framework for security, trust and cybercrime. Proceedings - 6th International Conference on Information and Communication Technology for the Muslim World, ICT4M 2016, 194–198. http://doi.org/10.1109/ICT4M.2016.43

[9] Fortes, N., & Rita, P. (2016). Privacy concerns and online purchasing behaviour: Towards an integrated model. European Research on Management and Business Economics, 22(3), 167–176. http://doi.org/10.1016/j.iedeen.2016.04.002

[10] Julie, P. (2011). Survival Manual. Training.Malik, M. E., Ghafoor, M. M., & Iqbal, H. K. (2012). Impact of brand Image , service quality and price on customer satisfaction in Pakistan telecommunication sector. International Journal of Business and Social Science, 3(23), 123–130.

[11] Mohamed Syazwan Ab, T., & Mohd Remie Mohd, J. (2012). Issues in Halal Packaging: A Conceptual Paper. International Business and Management, 5(2), 94–98. http://doi.org/10.3968/j.ibm.1923842820120502.1080

[12] Neilson, J. (2012). Usability 101: Introduction to Usability. Retrieved from https://www.nngroup.com/articles/usability-101-introduction-to-usability/

[13] Ross, A., & Willson, V. L. (2017). Basic and advanced statistical tests.

[14] Techopedia. (2017). Information Systems Security (INFOSEC). Retrieved from https://www.techopedia.com/definition/24840/information-systems-security-infosec

[15] Nasir, Mohd, Azah Norman, Shukor Fauzi, and Masliyana Azmi. 2011. “An RFID-based validation system for halal food.” Int. Arab J. Inf. Technol. 8 (2): 204-211.

[16] Khosravi, M. *et al.* Halal Products Recognition Using Rfid/Nfc Technology. *Sci. Int.* **29,** 687–692 (2017).

[17] Khosravi, M., Karbasi, M., Shah, A., Brohi, I. A. & Ali, N. I. An adoption of halal food recognition system using mobile Radio Frequency Identification (RFID) and Near Field Communication (NFC). *Proc. - 6th Int. Conf. Inf. Commun. Technol. Muslim World, ICT4M 2016* 70–75 (2017). doi:10.1109/ICT4M.2016.74

[18] Neilson, J. Usability 101: Introduction to Usability. (2012). at https://www.nngroup.com/articles/usability-101-introduction-to-usability/

[19] Webster, M. Definition of efficiency. (2018). at <https://www.merriam-webster.com/dictionary/efficiency>

[20] Ali, N. I., Samsuri, S., Sadry, M., Brohi, I. A. & Shah, A. Online Shopping Satisfaction in Malaysia: A Framework for Security, Trust and Cybercrime. in 2016 6th International Conference on Information and Communication Technology for The Muslim World (ICT4M) 194–198 (IEEE, 2016). doi:10.1109/ICT4M.2016.048

[21] Techopedia. Information Systems Security (INFOSEC). (2017). at https://www.techopedia.com/definition/24840/information-systems-security-infosec

[22] Simon and Schuster, “Webster’s New World Dictionary”, Second College Edition, 1982.

[23] Malik, M. E., Ghafoor, M. M. & Iqbal, H. K. Impact of brand Image , service quality and price on customer satisfaction in Pakistan telecommunication sector. *Int. J. Bus. Soc. Sci.* 3, 123–130 (2012).

[24] Nunnally, J. (1978), Psychometric theory, New York: McGraw-Hill.

[25] Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika* **16,** 297–334 (1951).

[26] Tavakol, M. & Dennick, R. Making sense of Cronbach’s alpha. *Int. J. Med. Educ.* 2, 53–55 (2011).

[27] Julie, P. SPSS Survival Manual. Training (2011).

[28] Ross A., Willson V.L. (2017) Paired Samples T-Test. In: Basic and Advanced Statistical Tests. SensePublishers, RotterdamG. Eason, B. Noble, and I. N. Sneddon,“On certain integrals of Lipschitz-Hankel

Model of Interoperability of Multiple Different Information Systems using SOA Middleware Layer and Ontological Database on the Cloud

Meryem FAKHOURI AMR, Khalifa MANSOURI, Mohammed QBADOU, Bouchaib RIYAMI
Laboratory: Signals, Distributed Systems and Artificial Intelligence (SSDIA)
ENSET Mohammedia, Hassan II University of Casablanca, Morocco

Abstract*—The exponential evolution of technology and the environment surrounding the information systems (IS) forces companies to act quickly to follow the trend of business workflows with the use of high computer technologies and well adapted to the needs of the market. Currently, the performance of information system is considered a problem, we must intervene to make them more agile and responsive to better support the strategy of the company. The exchange and communication is a must between information systems to address these issues and new requirements, businesses are looking to integrate and interact even their information systems to interconnect applications. Interoperability is essential between information systems, it promotes alignment between the company's business strategy and IT strategy while respecting the existing technical heritage of the company. The interoperability solutions between information systems face major problems since the SI is independently developed and designed differently and that solutions must meet certain criteria, namely, autonomy, scalability and resolve trade problems of data. The MDA approach is the most suitable solution to our problem because it ensures a degree of independence between the logic the company's business and technology platform. Moreover, oriented architecture SOA service is used in this sense, it encapsulates the components of the information system into editable and reusable service. We want through this article to contribute to the development of a model for interoperability of several different IS, founded on a middleware layer compound of services according to the architecture SOA. The special feature of our model is that it uses an ontological database in the Cloud that will store concepts exchanged among interconnected IS and uses several transformation layers, integration, homogenization and adapting services.

Keywords*—Interoperability; middleware layer; SOA; information system (IS); cloud computing; ontological database (ODB)

I. Introduction

The information system is the nerve center of any organization because it allows it to share information with other IS and collaborates. Therefore, interoperability between information systems is crucial since the applications should be adapted in order to determine for each request, the relevant data sources, the required syntax for querying the terminology specific to the source, to combine fragments of the results from each source to construct the final result [1].

Today, the major concern of every business is to develop integrated applications to solve the interoperability problem. The increasing complexity of information systems has created a critical need for models and methods of reuse. Companies are obliged to deal with this problem by evolving their information systems and adapting their business strategies to the technical strategy [2]. In this paper, we provide a model of the interoperability which is going to interface several information systems which are different by founding itself on services oriented architecture SOA. This model affords an SOA middleware layer that eliminates the complexity of interconnected IS and makes them homogeneous for interoperability, this model also uses other conversion and adaptation layers. The new contribution of this paper is that it is based on the use of an ontological database will be hosted on the Cloud. The organization of this article will follow the following order. We start by introducing the main work on the study of interoperability between information systems and the role of architecture led by MDA models and the implementation of SOA at the IS to ensure better communication. We also present the different approaches used to design this model of interoperability as a state of the art, we explain the SOA middleware layer that presents our contribution. We go on to describe the layers that make up the model by describing it by a schema. We conclude by summarizing the main contributions and highlighting future work.

II. Related Work

Today the world is suffering from strong competition and from the uncertainty of markets. This turbulent change has led companies to rapidly evolve to keep this wave of transformation of enterprise information systems and ensure better interoperability between them. Several studies have focused on interoperability between information systems. According to Agostinho, [3] interoperability, is the best solution to the problem, forcing companies to adapt and organize themselves to an automated information exchange. He strongly believes in the dynamism of interoperability, this is the solution that counts for the future; however, it considers that the static interfaces that exist today are no longer valid for tomorrow, and the solution is to maintain interoperability networked information systems. He acknowledged that before, there was an unassailable gap between the requirements of companies and the implementers of information systems. But with the implementation of architectures led by MDA models and interoperability model-driven, the concept of integrated design attenuated this gap.

To overcome this problem of complexity of the software development process, integration of new application and interoperability, Ameedeen [4] proposes a method of interoperability models using model-driven development to bridge the difference between the design and analysis phases of software development. This method consists of a diagram of transition sequences to Petri nets.

Despite the rapid development of technology and the evolution of information systems, there are still few methodologies that have been adopted to facilitate the development of software and interoperability between them. Loukis [5] admitted that the implementation of the interoperability of the information systems of a company with those of its employees generates commercial value, though this point has been studied in a limited way.

Today, companies are moving towards service-oriented architectures seen its many benefits, although it is not easy to talk about the interoperability of business processes in an environment oriented service. To solve this problem faced by most businesses, Tebib [6] proposes the use of interaction protocols (IP) as an effective means to structure and organize the exchange of messages between partners. For this, she used the BPEL4WS language that is considered a language of Web services composition for the definition of business processes and description of interactions between services.

The interoperability of information systems is changing especially with the use of SOA services oriented architectures for IT systems involved. This combination of technologies related to ontologies, service models ensures better semantic interoperability of services provided and requested [7]. According to Jiucheng [7], the purpose of business is to bring their heterogeneous IS and ensure that critical applications work perfectly. An approach based on a model including semantic annotation service models with ontologies can support and enhance semantic interoperability in service-oriented systems. We cannot deny that in recent times, the development of models for the interoperability of information systems using SOA Service-oriented architecture has gained immense importance. Batra has a concise proposal in this area [8]. He proposed a system that uses the SOA-compliant web services concept, which involves exchanging provider data and transforming it into a standard data exchange format using two databases. Both databases can be in different system formats. So, it is necessary to convert them to standard format prior to storage in the data repository. The whole of the process is implemented using Web services and messages are exchanged using the SOAP protocol.

It is in this sense that our contribution fits, which involves the implementation of model of interfacing and adaptation based on MDA models and SOA middleware layer for interconnection of several different systems information. This model is founded on a number of business process transformation layer BPMN towards the language BPEL using an ATL execution language by using an ontology database hosted in the Cloud.

III. State of the Art: Approaches to Model Conception

A. *Approach MDA (Model Driven Architecture)*

The field of interoperability promotes the use of the MDA approach because it ensures the introduction of isolation between the business layer and the implementation layer. This approach is part of a field of engineering models driven by IDM (IDM or MDE Model Driven Engineering) models. This approach takes models into account and focuses development on them [6].

The importance of the MDA approach is the development of systems, it is oriented model because it provides the basis for the use of models to orient and guide the understanding, design, construction, deployment, maintenance and modification of the systems developed. MDA primary objectives are portability, interoperability and reusability [7].

The motivation behind the use of this approach is that it can meet the major needs of task reduction to re-design applications that have become a technological development, a requirement for companies. Since the design of the models is unchangeable, the codes, with the implementation of the MDA approach, will allow to keep the business requirements, to use the code, to ensure the integrity and coherence.

B. *Service Oriented Architecture SOA*

SOA is a functional architecture that provides the ability to structure functional services in relation to current technologies. SOA is used to ensure the integration of business applications [8]. The characteristic of SOA architecture is that they allow to separate the specification of the art treatments interfaces on the one hand, and to propose composition approaches for constructing process by assembling services on the other hand. Based on a set of de facto standards, the implementation of an SOA can meet, at least in technology, the problem of interoperability of IS components [8].

This approach allows the encapsulation of applications into services which is a reusable and detached software entity during its use [9]. This operation named the composition of application services or more technically, the orchestration of web services [8]. A web service is designed as a software program, its role is to guarantee the interconnection and the sharing of data between different distributed applications [8].

C. *Intermediate Approach*

Mediators are used as a generic model of mediation that ensures the interconnection between the IS [10]. Mediators represent themselves according to two categories that can be executed in order to fulfill their main mission of mediation [11], [12].

We intend to use the fundamentals of the MDA approach with great benefits, focusing on models as well as their designs compared to conventional programming [13].

The need for using the MDA approach appears in the conversion of BPMN to the language BPEL (Business Process Execution Language) of different IS interconnected. Later, we design a common executable BPEL language that includes the BPEL (Business Process Execution Language) languages of the interconnected IS. The differences will be dealt with using an ontological database. To create this language, we use the intermediary approach that will allow us from the resulting BPEL Global language, to transform it into a business process BPMN (Business Process Model and Notation) thanks to the retro-engineering.

IV. SOA Middleware Layer for Interoperability of Multiple IS

The upward approach is the method we used to design the model. This approach starts from a base to arrive at a complete model. Our added value through this article is the design of the SOA middleware layer that will make the IS interoperable and reduce the complexity of communication and sharing between them. The major principle on which is based is the encapsulation of each part in a service.

The model we have designed consists of a number of conversion and adaptation layers. The interconnection takes place between N information systems. The major rule of our model is that it can interface any information system whatever the number of interconnected IS. The principle of this model is based on the composition of business processes and encapsulation services to facilitate modeling and reuse through the SOA middleware layer. Through this model, we want to promote integrated communication via information technologies as a single entity that is based on the SOA architecture.

V. Layers Constituents the Model of Interfacing and Adaptation for the Interoperability of Several Sifferent IS

The proposed model as presented in Fig. 1 is based on MDA architectures and intermediates as well as SOA. It consists of several layers:

Layer 1: This layer transforms BPMN business processes into an executable BPEL language by using an ATL (Atlas Transformation Language) transformation language from MDA, and we want this conversion to make systems agile and consistent

Layer 2: From this layer, the common BPEL languages are gathered from the interconnected IS BPMNs. the ontological database is used to treat the differences between the BPELs obtained. The ontological database we use is hosted in a public cloud. The use of ontologies provides a formal description of concepts in a domain and guarantees reuse, knowledge sharing and interoperability.

Layer 3: It is the time of the correction and the adaptation of the common languages obtained towards the end of the transformation of BPMN business processes into the BPEL languages of each IS, an update is required of the source business processes of the interconnected IS to synchronize all the BPMNs. To do this, we use reverse engineering, which allows us, through transformation techniques and rules, to have raw models from their BPEL source code [14].

VI. Interoperability Model of Several IS using SOA Service-Oriented Architecture and Ontological Database on the Cloud

For the design of this model, we used the service-oriented SOA architecture to break business processes into services to facilitate their reuse or replacement by another service. The model makes it possible to communicate N System of information whatever the number of IS. In case the number of SI is odd, we gather the last SI with the last BPEL obtained. Our model will undergo some operations to arrive at generating the common global BPEL and transform it into a global BPMN that will help us, using an ontological database, to update the initial business processes BPMN of the interconnected IS.

We start by transforming the BPMN business processes from the first two ISs into the executable BPEL language using an ATL executable language. Subsequently, we bring the two BPEL together to a BPEL 1 Global. The differences that exist between BPELs are analyzed by a single ontological database. the same workflow is repeated for the other IS by bringing together the BPEL Global 1 obtained from the IS (1 and 2) with the BPEL Global 2 of the other IS until a cascading global model is obtained, which brings together the Global BPEL of all IS to have towards the end a single common GLOBAL BPEL. Thanks to the reverse engineering, we transform the BPEL GLOBAL language into a GLOBAL BPMN business process that will allow us to apply an update for all the source BPMN to ensure a better synchronization of the IT trades.

The model we proposed for the interoperability of multiple different IS characterized by a Cloud architecture that will allow it to store data exchanged between interconnected IS. The ontological database shared between the IS will be put into service on the Cloud is also the source of sharing and the communication link between the interconnected IS. The particularity of this architecture of an ontological database on the Cloud, is that it is more practical than the other architectures and presents many advantages for the implementation of our model, namely that all the concepts exchanged between the IS interconnected will be stored in this ontological database, the difference between all the obtained Global BPEL will be processed through this ontological database hosted on the Cloud. This architecture promotes the sharing and availability of data located at the level of the ontological database between all the interconnected IS as well as its rapid deployment and its simplicity of integration in addition to its high availability, since the ontological database will be on the Cloud which will guarantee the IS a better accessibility to resources.

The motivation behind the establishment of the ontological database on the Cloud is justified by the fact that this new architecture has many advantages in that the database will be maintained and usually stored on a secure server, it also allows ease of remote access to the ontological database or even a simple deployment of the database if a new information system will interconnect with other IS.

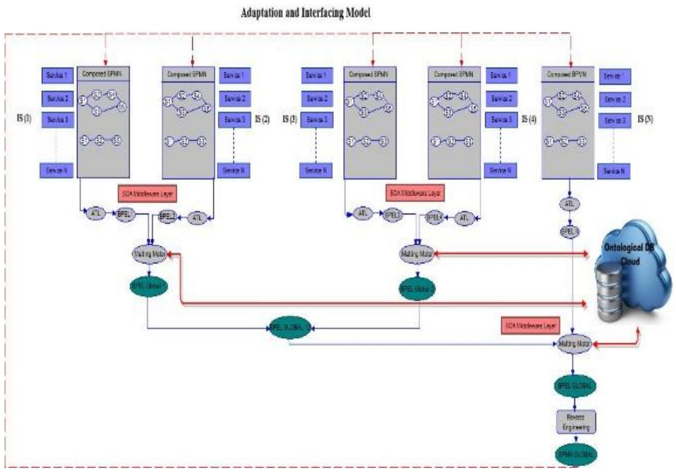


Fig. 1. Interworking and adaptation model for interoperability of multiple ISs using a SOA middleware layer and cloud database.

SOA is an approach that allows the management of information system components of a company as reusable services on this basis our interoperability model has been developed. Our solution is based on a middleware layer that presents the central and vital element for our SOA architecture, which will facilitate exchange and interoperability between information systems interconnected by integrating a database in a Cloud space. The establishment of a Cloud architecture facilitates easy access to computing resources by users of a company, the total availability of resources on the Internet and payment for the use of these resources.

Given the advantages offered by these two architectures, SOA and Cloud, it was thought to combine the two architectures to give information systems of interconnected companies more agility, ease of sharing and communication. It will also reduce the complexity of systems through the reuse of services and of course the optimization of resources. The model that we propose allows an easy communication between the information systems, it guarantees a high availability of the ontological database with a very high security.

VII. Cloud Architecture Adopted for the Proposed Model for Interoperability of IS

Today's markets need cloud computing which has had to situations in exponential growth. But its definition remains general especially that few people agree on a specific definition. Cloud means "applications delivered as Internet services, typically network, hardware and software systems in the data centers that provide these services" [15]. Cloud Computing allows any user to use multiple computer forums for a period of time to perform calculations, process data, or create Web applications [15].

Defined as a new technology, Cloud Computing allows the implementation of IT services through the management and use of information systems. Despite the difficulties through the use of Cloud expressed by the lack of standards and heterogeneous architectures, there is a compelling solution for migrating to new environments [16].

The choice of Cloud Computing for the implementation of our model was not random because the Cloud presents to the field of interoperability of information systems, several advantages [16], we quote:

- **Reducing infrastructure costs**: Information systems interconnected via this model must have a specific technical architecture to be able to interoperate with other IS which generates high costs. So, the proposed solution will guarantee IS the use of a shared Cloud infrastructure between them without the need for an internal architecture.
- **Reduced development costs**: The model deployed for interoperability will be centralized between IS interconnected which will generate a single centralized development task. All that comes under the maintenance of the technical infrastructure is supported by the provider of Cloud solution.
- **Reduce software costs:** Our model is based on an architecture based on a single ontological database so it will allow companies to save software costs used for its implementation and the costs associated with Cloud architecture.
- **Ease of access to shared resources and services:** This is the major advantage of the Cloud architecture, it facilitates an easy, secure and safe for even resources with the combination of SOA, there will be no difficulty in sharing services.
- **Increased processing power of shared data:** Cloud architecture with the data processing will be easier because this architecture has a set of powerful technical components in terms of capacity, strength and speed.

Cloud computing can be deployed according deployment models that have been treated and quoted by several studies. The four deployment models:

- **Public Cloud:** The peculiarity of the public Cloud is that it is open to the public, the infrastructure is managed by the provider who takes care to offer its services to users. All computing resources are shared among users who have no control or visibility into infrastructure [16].
- **Private Cloud:** The Private Cloud hosting infrastructure is controlled and in the ownership of the user of the resources. These are available and intended for private use.
- **Hybrid Cloud:** It's an environment that combines both models Public and Private. Resources are allocated from a private Cloud and a public Cloud.
- **Cloud Community:** The Community Cloud infrastructure is designed to use a very specific community of users. This infrastructure can be managed by one or more community organizations [16].

The model we propose for the interoperability of information systems involves a Cloud infrastructure to host the ontological database and computer technologies deployed for the implementation of this model between the information systems of interconnected companies. The choice of this infrastructure is justified by the fact that it has many advantages especially as information and shared data between the IS must be centralized and accessible by all IS interconnected.

Since the establishment of a specific architecture for each company wishing to interoperate with another, is too expensive and consuming IT resources, the solution we propose is to put this model on a Hybrid Cloud infrastructure that will allow users to share private data on the private Cloud and public data on the public Cloud, in this way all users will have equal access to shared resources.

VIII. Services Component the Model of Interoperability

A. BPMN-BPEL Transformation Service

The BPMN business process is used to model the company's business into a workflow that describes the system's functionalities and makes it possible to understand it in detail. This workflow can be used and shared with other systems. BPMN is designed by the Business Process Management Initiative (BPMNI) and managed by the Object Management Group (OMG) [17]. While the BPEL model is a programming language developed to handle orchestration processes [17].

The MDA approach ensures the process of transforming BPMN into BPEL through service transformation. You have to go from a conceptual model to an executable source code. To succeed this process, we must not go directly from one model to another but to go through their meta-models. The MDA approach ensures the process of transforming BPMN into BPEL through service transformation. You have to go from a conceptual model to an executable source code. To succeed in this process, we must not go directly from one model to another but to go through their meta-models. Typically, the processing service supports models (sources) and model outputs (target). Transformation allows us to gain a lot of development time by generating an automated code [17].

We start the model process by converting each BPMN business process into a BPEL executable code. We have presented the business process by its abstraction model which is a subset of information of this system. We speak in our case, meta-models that describe the general model, it defines the elements that constitute it, its overall structure and its semantics. We cannot proceed to the manipulation of the meta-model if the language that describes it is not ready [18]. In our case, we are talking about a meta-meta-model. Meta-meta-models are characterized by their self-defining capabilities, which is due to the principle of the Model Driven Engineering (MDE) approach [18]. There are some actions to put in place on the input models to have corresponding output models. It brings us to two categories of transformations, from models to models and models to text. As it is mentioned in Fig. 2, in our case, we are interested in the transformation of the model to the text, which allows us to transform the BPMN business process to a BPEL code and to gather all the BPEL code and generate a BPEL GLOBAL code. ATL (Atlas Transformation Language) is a language based on the definition of metamodel, its principle is the transition from one model to another via meta-models [17].

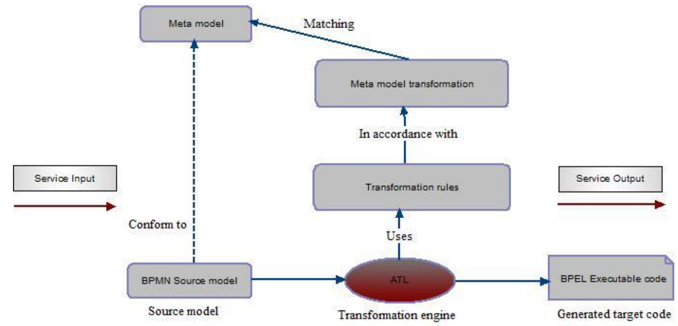


Fig. 2. BPMN to BPEL transformation service according to MDA [19].

B. Consolidation, Homogenization and Integration Service

The consolidation service provides a BPEL that contains different items and others in common. We do a BPEL analysis obtained by the homogenization and integration service, then bring the two BPELs into a common BPEL containing common elements and differences. These differences between BPEL will be analyzed by an ontological database hosted on the Cloud as it is described in Fig. 3. The consolidation, homogenization and integration service accepts the BPEL code as input and generates a BPEL executable code.

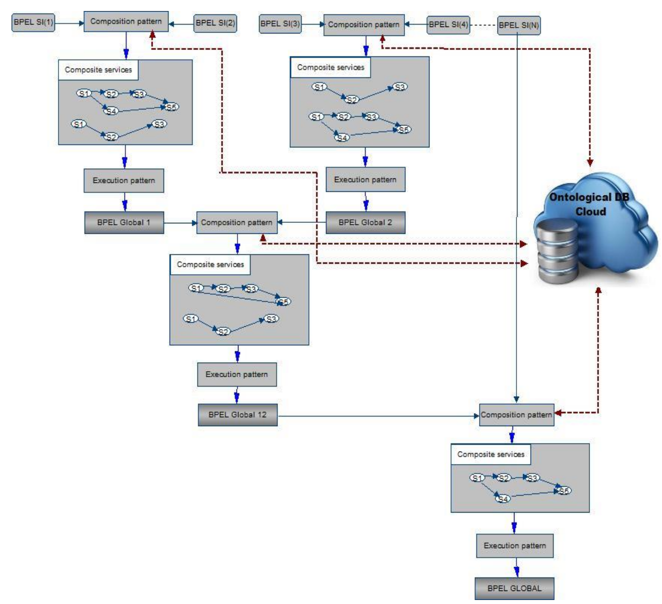


Fig. 3. Process for obtaining a GLOBAL BPEL using an ODB on Cloud.

Ontology is a conceptualization that makes it possible to represent the semantics of a domain. For this purpose, she uses models of consensual objects that associate each concept with an identifier [20]. The ontological database we use is hosted on the cloud. It makes it possible to treat the differences between the BPEL obtained [21]. Ontologies allow their reuse, so have other concepts of the domain which will ensure the interconnection and exchange of data sources [21].

C. *"Reverse Engineering" Service*

Thanks to the retro engineering and its techniques and rules as represented in Fig. 4, we will make an adaptation of the BPEL languages obtained. At the end of the BPMN business process transformation into a BPEL code, we assign an update to the source business workflows of the information systems to synchronize the Global BPMN. It is about managing the knowledge to answer the technical and functional specifications at the origin of the initial model, then to structure, formalize and update them [22].

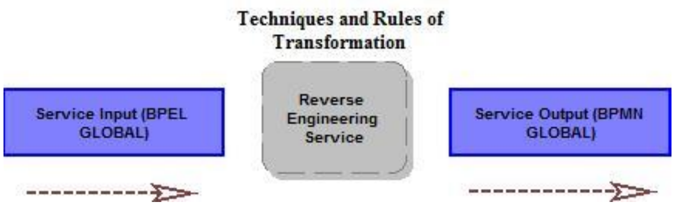


Fig. 4. "Reverse Engineering" Service [19].

IX. CONCLUSION

In this article, we developed a model for interoperability of multiple IS that remains valid regardless of the nature of the interconnected information system and its field of activity. This model can interoperate multiple information systems based on the SOA middleware layer. This model contains layers of transformation, consolidation, homogenization and integration of services. Throughout the development of this model, several approaches have been used in view of the advantages they present, including MDA, intermediate approach (ontologies) and the SOA approach for encapsulation of each IS component in a service. Towards the end reverse engineering was used, it was used for reconstitution and the updating of BPMN business process of each information system after the transformation and adaptation. The peculiarity of our contribution is that we used the ontological databases to treat the difference between BPEL obtained through the union of BPEL of interconnected SI by hosting this ontological database on the Cloud to benefit from its many advantages. In perspective, we consider the implementation of this model to interconnect different hospital information systems, also based on aspects of the SOA approach given its ability to optimize and improve business processes.

REFERENCES

[1] J. Moskolai, A.Ayissi R. Hoe-Ngouna, Integration and interoperability of heterogeneous information systems in distributed environments.. Towards a flexible approach based on the urbanization of information systems, International Research Conference computer, Dec 2013, Yaounde, Cameroon. 2013.

[2] S. Izza, Integration of Industrial Information Systems: a flexible approach is based semantic services. Modeling and simulation. Ecole des Mines de Saint-Etienne, 2006. French

[3] C. Agostinho, Y. Ducq, G. Zacharewicz, J. Sarraipa, F. Lampathaki, Towards a sustainable interoperability in networked enterprise information systems: Trends of knowledge and model-driven technology, Computers in Industry, Elsevier, 2016, Future Perspectives On Next Generation Enterprise Information Systems, 79, pp.64 – 76.

[4] M.A. Ameedeen, B. Bordbar, R. Anane, Model interoperability via Model Driven Development, Journal of Computer and System Sciences, Volume 77, Issue 2, March 2011, Pages 332-347.Elsevier, https://doi.org/10.1016/j.jcss.2010.01.011.

[5] J-M Jézéquel B. Combemale, D. Vojtisek Engineering Model Driven: concepts to practice, Ellipses. Ellipses, pp.144, 2012 References sciences 9782729871963

[6] M. RADGUI, decomposition and adaptive business processes BPMN for flexible information systems, in June 2015.

[7] A. Authosserre-Cavarero, F. Bertrand, F. Blay-Fornarino,O. Collet,D. Hu-bert, Interoperability of information systems: model-driven approaches, Inforsid, Montpellier, France. pp.11-30, 2012

[8] S. Baina, H. Panetto, K. Benali, The MDA approach Contribution to semantic interoperability: Interoperability of Enterprise Information Systems, Journal of Science and Information Technology -Series ISI. Systems Engineering Information Lavoisier, 2006, 11/13, pp.11-29

[9] K. Dahman, Governance and study the impact of changing business processes on service-oriented architectures. One approach model-driven, Hardware Architectures [cs.AR], University of Lorraine, 2012. French

[10] A. Zayati., Implementation of service-oriented architectures for industrial information systems. Business administration. INSA Lyon, 2012. French.

[11] M. Weske, Business Process Management: Concepts, Languages, Architectures. 2012 ,DOI 10.1007/978-3-642-28616-2 Springer

[12] P. Degoulet, L. Marin, L. Kleinebreil, B. Albiges, Present and future of information systems and hospital communication, Springer Paris, 2003.

[13] T. Erl, SOA: Principles of Service Design, USA: Prentice HALL PTR, 2007.

[14] I.Noe Garcia Garza, design and implementation models for mediation and integration services, Other [cs.OH] .University Grenoble, 2012. French

[15] M. Durut., PhD Thesis. Algorithms classfiication distributed cloud. Learning [cs.LG]. Telecom ParisTech, 2012. French

[16] H. Medhioub., Architectures and federation mechanisms in environments cloud computing and cloud networking. Other [cs.OH]. National Institute of Telecommunications, 2015, French.

[17] K. Mansouri, B. Riyami, M. Youssfi, O. Bouattane, MODEL OF AN ADAPTATION AND INTERFACING SOA MIDDLEWARE FOR THE INFORMATION SYSTEMS INTEROPERABILITY, JOURNAL OF THEORETICAL AND APPLIED INFORMATION TECHNOLOGY, FEBRUARY 2016.

[18] N. F. Noy, D.L. McGuinness, Ontology Development 101: A Guide to Creating Your First Ontology. Université Stanford

[19] M. FAKHOURI AMR, K. Mansouri, M. Qbadou, Towards a model of adaptation and interfacing based on a middleware layer SOA for interoperability of several different information systems., 2017 Morocco, IEEE Computer Society, The second International Conference on Intelligent Systems and Computer Vison (ISCV).

[20] A. kévin, A. Engel, Interoperability between models, Nancy, France, 2011.

[21] V. ARANEGA, Traceability to the development of models and correction of transformations, PhD Thesis, Graduate School: Engineering Sciences University lille1 November 2011, France.

[22] B.Mbaiossoum, S.Khouri, L.Bellatreche, S.Jean, M.Baron, Comparative Study based Database Systems Ontological, University of Poitiers.

Intelligent Educational Assistant based on Multiagent System and Context-Aware Computing

EducActiveCore: Experiment with Context-Aware and Multiagent Interaction

Fernão Reges dos Santos[1], Pollyana Notargiacomo[2]
Department of Post-Graduation in Electrical Engineering
Mackenzie Presbyterian University
São Paulo, Brazil

***Abstract*—This paper provides an overview of the current stage of EducActiveCore research, an orchestrated computational model, formed by different areas of artificial intelligent, combined to support personalized assistance to students in distance education process, mainly in interaction with Context-Aware environments. The Context-Aware environment applied in this research is observed in conjunction with IoT technologies. IoT is enabled by the latest developments in smart sensors, RFID, communication technologies, and Internet protocols. The basic premise is to have the resources availability and use arrangements, managed directly without human involvement to deliver a new class of smart environments for students. To support this central idea, a Multiagent model is proposed to assist students in interaction with context, determining autonomously the access to useful resources to students. This article introduces the overall research in progress and, the methods of an experiment tested with basic concepts of this scenario, implemented and used by a group of students in real locations. Results obtained during tests, indicates 93% of successful operation performed by this intelligent model on use prediction of resources and scheduling reservation.**

Keywords*—Context-aware computing; multiagent systems; artificial intelligence; internet of things

I. Introduction

This article resumes the current state of one domains covered within the research in progress to apply intelligent computational models to support personalization of student's educational activities when interacting with context-aware environments. Added to an exponential growth of different data and media content, new scenarios are emerging on the horizon in analysis of behavior to student's computing interaction. Some of most recent evolving related areas stand out, like Internet of Things, Artificial Intelligence, Mobile Computing and Context-Aware Computing. Understanding the combination of these computing areas as complementary researches, this work investigates the applicability of these domains in a combined intelligent computational model named EducActiveCore to support online and remote assistance on personalization within educational process.

The content of this paper is organized in the following structure: The overall resume of EducActiveCore research and the background related is described in Section II. An introduction to Context-Aware Computing and Internet of Things are presented in Section III. Multiagent Systems and Context-Aware Computing are introduced in a correlated domain in Section IV. The resume of computational models implemented to this experiment as its organization is presented in Section V. Method of experiment and preliminary results are presented in Section VI. Conclusions, issues of current stage and next phases of this research are introduced in Section VII.

II. Background and Related Work: Intelligent Computational Model EducActiveCore

In this section, we introduce the overall scope of the main research in progress and the focus of current work as an important implementation part within the proof of concept to applicability of proposed model.

A. Research Proposal and Objectives

The main objective in this research, considering its comprehensive scope, is established as a propose of an intelligent and adaptive computational model to support both educators and students through computer interaction within, initially, distance education process. Observing domains related to provide computational support in educational process, several challenges emerge regarding increase information available to educators during activities of task elaboration, collaborative process to collect and share information, interactivity between students and, hence, student's interaction with physical environments. The view of our main research proposing an intelligent computational model, named EducActiveCore, to cover these domains in a centric and self-orchestrated process, aims to act as intelligent assistant to help increase student's educational results since task elaboration (from educator's perspective) to activity execution by students.

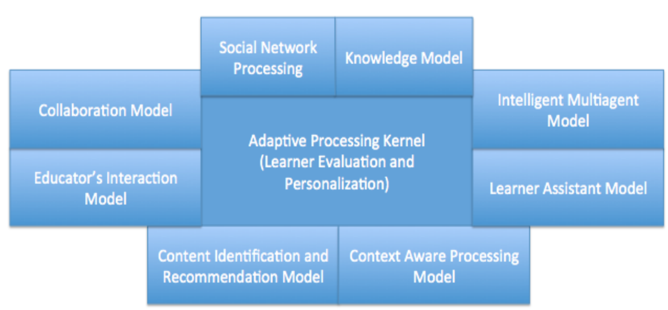


Fig. 1. Overview of model composition.

B. Adaptive Processing Kernel

The orchestration proposed by EducActiveCore, contains a centric computational set of algorithms, structured as Adaptive Processing Kernel, which commands and interprets the complementary engines of processing.

The organization diagram contained in Fig. 1, illustrates a macro view of entire scope from research, organized in different computational engines. Computational engines identified by Processing in diagram (except for Kernel case) indicates a component completely accessory to this research vision, commonly, with its services accessed by EducActiveCore without its internal structure changed by this research. The engines identified by Processing Model, contains changes of behavior or complementary elements proposed in research scope. Finally, the engines with Model identification, contains its central behavior defined by research.

The scope of domains was structured in conceptual modules, organized according to its processing objectives with aim to reduce the complexity of Adaptive Processing Kernel that is the central engine proposed on this research. One of the main operations expected to this kernel is to identify the proper combination of student's mobility behavior with computing interaction and, execute the orchestration of the coupled complementary models to increase student's results on educational tasks.

The set of algorithms in this Kernel is organized to perform precisely mining and tracking on data collected from the complementary models. This set of algorithms is divided into two groups of distinct operation; Student's Evaluation and Student's Personalization. Basically, within Student's Evaluation operation, the algorithms collects student's data results related to educational tasks and determine their acceptance according to specific parameters. The group of algorithms responsible for Student's Personalization performs interactivity identification and student's assistance. Students can perform educational tasks in different physical environments and, hence, in simultaneous contexts. The Personalization processing identify the most recurrent class of data during interaction and determine if this interaction occurs in social network or a context-aware environment, for example.

As indicated in previous works [1], beyond the actual digital consuming profile of students with multi-connected and decentralized computer interaction behavior, moving student's attention from screens like Smart TVs, console games, tablets, smartphones, wearable and mobile devices, all those devices interconnected with internet services providing convergent media content, researches related to Internet of Things and Context-Aware computing are evolving rapidly.

Observing student's mobility and multiconnection possibilities from the perspective of environment navigation flow and, considering this flow as a learning object to be collected and processed, reinforces the need of fine personalization and characterization like previous researches covered [2]. Hence, this article presents the continuity of previous work [3] to elaborate the purpose of Adaptive Processing Kernel. In this precedent work, initial results were obtained from experiments using machine learning technics (MLP Neural Networks) applied to classification tasks of environments and, an implementation of Multiagent protocol acting as assistant to route mapping and recommendation to physical locations. These results are data collected during the use of application implemented as proof of concept and feedback sent by students after experiment participation. Considering previous experiments and results obtained on research, this work evolves the proposed investigation of Adaptive Processing Kernel focusing on interoperability perspective correlating Context-Aware, Internet of Things and Multiagent system.

III. Context-Aware Computing and Internet of Things: Definition, Models and Learning Applications

A. Definition and Computing Models

Context-Aware computing can be described essentially by computational environments able to adapt itself and current context based on user presence and interactions, without explicit user supervision or intervention on this adaptive process [4]. There is no final and rigid boundary to specify what compose a context. Some authors refer to architecture key characteristics most appropriate to their research and field application to define resources and services of a context. Shared and common descriptions of these resources are presented as information describing and characterizing a person or object in interaction with computer application [5]. Beyond the informational aspects, physical elements and temporality of available resources [6] were published as features that are central to define a context.

Previous research (beginning of mobile research and development) resumed the context computing as a result on combination of three information domain regarding on where computational resources is, what kind of resources is provided and who is using it [7]. Modeling these characteristics in attributes, import and exchange it in form of relevant information to identify a context is a key process to support the context computing adaptive process [8]. The changing process of a context started by sensors, network events, mobile device location detection [9] or the combination of all these examples, provides several new different set of attributes to describe the context and related resources.

The Internet of Things (IoT) becomes an important study area since it denotes several impacts on local and global internet infrastructure and the services provided by computational environments [10]. The IoT bundles a variety of technologies (e.g. sensor hardware and firmware, communication technologies, semantic, cloud services, data modelling, storing, computational reasoning, high processing) and, eventually, combined together to establish its vision. The concept of IoT does not revolutionize or the existent fields of computing. It is evolving and complementing the Internet we already use.

The IoT holds the promise of establish a global network of useful and interoperable devices/resources supporting ubiquitous computing [11] and context-awareness among devices. Ubiquitous Computing and Context-awareness are key requirements of ambient intelligence, one of the main benefits considering the promises of the IoT [12]. IoT applied to ambient intelligence would allow objects of common use to understand their environments, interact with users and make decisions [13]. A world full of smart objects holds several possibilities to improve business processes and user's lives and, hence, it also comes with threats and technical challenges that must be overcome. From perspective of users, deal with a huge variety and quantity of smart objects like sensors/devices, with its raw data generated, presents a challenge to project useful application and its viability [14]. Considering this perspective and the current research focus, we observe the educational application of IoT as a resource within context-aware environments operation.

B. Context-Aware Learning Application

The variant set of IoT resources and its specific attributes within a context and, consequent different models of categorization and identification of actual resources and services content, requires specific observation and interaction regarding context computing and their applications. Previous works on context-aware computing applications within educational and mobile learning domains [15] helps to define the objective and scope of this computational model on this research. Investigations and development on context-aware applications to support language learning process can also be found on [16], using environment recognition to apply a compatible vocabulary according to user's location, supported by mobile devices and highlighting the importance of mobile devices to interact with context and evolving the application to detect the collaborative situational needs.

Another example based on context-aware computing implementing an instructional application [17] is built to identify tourist locations, relevant content and help tourist guides to take decisions considering the best matches for the group of visitors. Considerations of mobile device relevance within these tourist guide assistant and how embedding the solution on mobile devices to turn invisible and accessible to users were introduced.

On collaborative learning, previous research investigates learning environment built on a peer-to-peer architecture [18] and demonstrate the applicability of a common protocol (using meta-data structure) to determine context resources identification, information extraction and interoperability. The structure of meta-data, proposed in form of context-aware learning environment ontology, is processed with an adaptive model to evolve the ontology with aim to expand the interoperability. A range of supporting context-aware and IoT technologies to learning applications, including peripheral devices and sensors, image processing (recognition and augmented reality) can be found on [19].

C. MultiAgent Systems and Context-Aware Computing

Commonly, Context-Aware Computing considers the self-managed configuration. Previous investigations with Agent systems applied to Context-Aware Computing, covered e-commerce business domains applying intelligent negotiation through case base reasoning and context history analysis [20]. Related to user context personalization, frameworks based on Internet webservices, Internet address identification and segmentation proposes frameworks to act on behalf of users on pre-selection and definition of context content and services before turn it available to users [21]. Beyond the references related to relevant technical aspects to support multiconnected environments using agent-based model [22]. Previous researches covered Context-Aware Recommender Systems [23], investigating relevance of contextual information to recommend relevant content to users. The complementary research domain relevant to use on this current investigation is the Multiagent models applied to education. As precedent work, one specific approach proposed to Digital TV environment with adaptive content applied the combination of Multiagent model to educational content personalization [24]. In this work, as continuity of Adaptive Processing Kernel elaboration, we cover current step of research focusing the necessary behavior regarding to context-aware and Multiagent interaction aspects.

IV. Interaction with Context-Aware Environment Enabled by Multiagent System

A. Multiagent Model as Interaction Protocol

For the purpose of this research, as a solution using context aware computational model, we identify as strongly necessary a local processing with representation of last set of actions taken and with minimal data used to compute the actual action and recommendation reflected on student's perspective.

The concept *Local* in this scenario means independent processing algorithms (not connected with Kernel) able to execute basic tasks based on last instructions dispatched by Kernel. This *Local* processing is foreseen to synchronize frequently on line with Kernel but ability to process disconnected emerges as necessary due to common issues related to complementary resources like Data Repository based on cloud platforms and its availability [25] or while a network connection is not available, for example. The aim of applying Multiagent Model in this research is to simplify the different local computation models. From student's utility view, the Agent Model can be available as embedded application on personal devices searching for available connections and, service discovering with context-aware environment according to student's profile and needs.

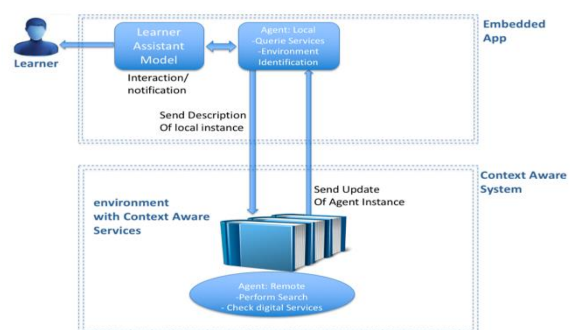


Fig. 2. Multiagent system as protocol interaction.

Fig. 2 illustrates the macro view of interaction through a Local instance on embedded application and its replica. The same Agent representation can be transported to an identified context-aware environment (in this case, acting as a protocol) and perform different computation using resources from environment instead of consuming local processing from student's device. This strategy reduces the problems related to device's energy efficiency and other questions that could affect devices with embedded application containing Agent Model.

Considering this approach to use the student's local Agent as a protocol to be transported to context-aware environments, allows device's embedded application to send data related to student's profile and educational tasks where this student is involved, previous activities performed and its results, next activities and all other necessary data used by local Agent to execute its processing.

This set of data that represent the student's local Agent, creates another instance of Agent on context-aware environment. This new instance is described in this research as remote Agent, due its additional behavior to process Agent according to environment model using initial state received as input parameters. Specialization applied to each Agent model attends the variability aspects of context-aware environments where some attributes and behavior related to student's profile are shared to perform remote Agent tasks but considering environment characteristics, specific behavior and attributes are inherent of processing. Remote Agent, beyond the data provided by students and environment, can receive data from other Agent instances running in same context platform or allowing Learning Multiagent System [26]. It is useful to permit a structure of data communication to help agent identify unknown information, like other similar environments or the ranking level of environment services, indicated by other students. Tasks to be performed by remote Agent on context-aware environment can return to student's device as a result to update and merge behavior with device's agent instance and reflect its results on Kernel synchronization and Student Assistant Model.

In this experiment, we extended the mobile application implemented on prior experiments presented in Santos et al. [3] to add a reduced set of Multiagent operations regarding to synchronize with Kernel, simulate the content search on context-aware environment, collect additional information about context and performs scheduling reservation on selected resources.

B. Context Aware Model - Service Discovery and Resource Identification

Considering the objectives of this research, the basic concept of context aware model must contain resources provided by a combination of physical location identification, descriptors and its digital services, observing the student's profile and needs. The model composition of services is variable and resources are intrinsically characteristic to environment accessed by student. A university laboratory can provide a 3D printer. A bookstore allows access on its book catalog. A commercial center indicates the store with specific scholar supplies. These distinct examples of environment differ on purpose and may vary on its characterization but they should present similarity with other environments under same objective classes.

Conceptually observing the local Agent behavior on Service Discovery and Resource Identification interaction, some common elements from context-aware platforms definition emerges as necessary to insure the abstraction level and reduced complexity model applied to local Agent.

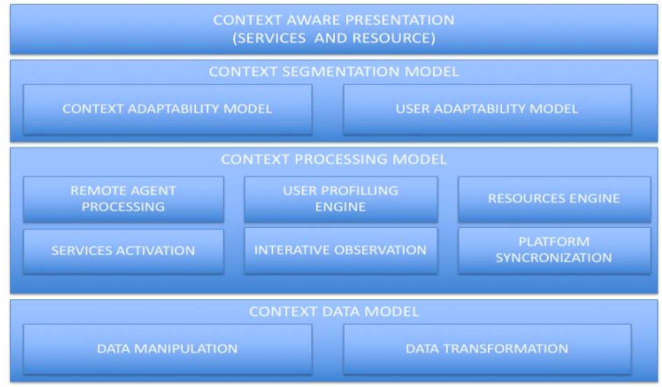


Fig. 3. Context aware platform organization.

Fig. 3 illustrates a macro model with minimal set of components and its layer organization proposed to establish a context-aware platform with based to this research vision.

The upper layer context aware presentation communicates with local Agent to present the Service Discovery and Resource Identification. The component Service Discovery provides the catalog of available services in environment. Resource Identification component permits the compatibility verification with a local Agent action intent related to a determined service previously discovered. Segmentation Model layer contains the management of user adaptability that control the level of interaction with student. The context adaptability uses this level of interaction as parameter to calibration and setup with student without interaction history. In example, if a student does not intend to share personal data with context, this combined engines process according to this definition. The layer Context Processing Model contains the engines responsible to process de dynamic of environment operation. In this layer we define the services orchestration and provisioning. Context Data Model layer contains the operations to access (query, transform and organize) data related to context and its resources.

In an overall and integrated example, if a local Agent queries and discovers an available 3D print service, resources engine verifies if the necessary drivers and supplies are compatible to attend the print request for student's project. If an incompatibility between service and requisition is detected, the resources engine could be update the available drivers to attend the request, if applicable. Eventually, a request for supplies reposition can be started if it turns insufficient. In case of impossibility with supplies reposition or driver's updates, the context segmentation configures the unavailability of this service specifically to user or group of users, and reflects it on presentation model to avoid new requests to same service.

V. Enabling Computational Resources, Middleware Implementation and Context-Aware Environment

In this section, first we present the computational engines and its implemented operations involved in the scope of proof of concept implemented. To local Agent operation, the following sequence of processing flow was implemented as initial behavior with the mobile application: 1) local Agent instance identify the context-environment through provided network WiFi name containing the defined pattern “CTX-ENV-'NAME|ID'”, where 'NAME|ID' pattern represents the the combination of environment and its unique identifier. 2) With local Agent connected to environment; firstly, application provides a copy of local Agent containing the current instance and all related attributes to desired services, and initiate the service request processing with context.

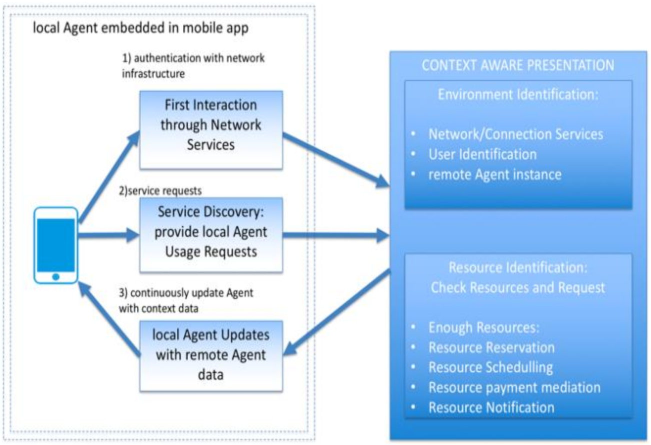


Fig. 4. Agent interactivity processing.

Fig. 4 contains the illustration of mobile application processing, embedding the Agent operations. Service Discovery and Requests are the main processing acting in environment resources. This block of operations executes the navigation on services catalog provided by environment and performs requests of resources reservation and content search according to desired needs. In a continuously flow, the mobile application receives an updated version of remote Agent and merges the local instance with new attributes and data received.

Fig. 5 contains an example of data received from context, representing the message describing the protocol of one remote Agent instance. This message contains group of data related to user’s profile, resources available on environment, resources reserved, beyond environment information and the attributes collection representing data shared by remote Agent instances of other students (agentShared indicative). It helps Multiagent learning process and increase the data information about environment during synchronization with Kernel.

With experimental objectives in simulate the basic operations of our proposed Context-Aware architecture, the middleware was implemented using base components from platform AWARE (http://www.awareframework.com). This open-source framework permits plugin development, extensions and complements, attending objectives of this initial experiment to provide an engine readily modifiable while investigation evolves and, consequently, new refinements and reviews on proposed architecture turns necessary adjust the model and the implemented platform.


```
{
   "environment":{
      "enviromentID":"CTX-ENV-EletroLAB|Duz+3q6u/t6tyr7",
      "contextURI":"https://10.25.120.6/Duz+3q6u/t6tyr7",
      "contextClass":"GU"
   },
   "profile":{
      "agentProfileID":"35416018D6C8938749DB2CC0BC528478",
      "studentName":"student 01",
      "studentEmail":"usr001@gmail.com",
      "lastKernelSync":"20171203134635"
   },
   "resourcesAvailable":{
      "printer3D":"Impressora 3d Creality Cr-10s 300x300x400",
      "contentCatalogCategories":[
         "eletronica profissional",
         "eletrica",
         "eletronica hobista"
      ]
   },
   "resourcesBooked":{
      "printer3D":true,
      "reservation":true,
      "availableSupplies":true,
      "scheduledtime":"17:00"
   },
   "agentShared":[
      {
         "agentSharedID":"C73BE44A6CF9C3DAD21F488E074E6762",
         "agentSharedUserTags":[
            "facilidadedeacesso",
            "pertodometro",
            "materialparavenda",
            "protoboardmaissimples",
            "amploespacodetrabalho"
         ]
      }
   ],
   "contentCatalog":[
      "Arduino For Dummies",
      "Arduino: 101 Beginners Guide",
      "Arduino Programming in 24 Hours",
      "Arduino Workshop: A Hands-On Introduction with 65 Projects",
      "Getting Started with Arduino: The Open Source Electronics Prototyping Platform",
      "Exploring Arduino: Tools and Techniques for Engineering Wizardry",
      "Arduino Robotics",
      "30 Arduino Projects for the Evil Genius",
      "Arduino: A Technical Reference"
   ]
}
```


Fig. 5. Segment of remote Agent data.

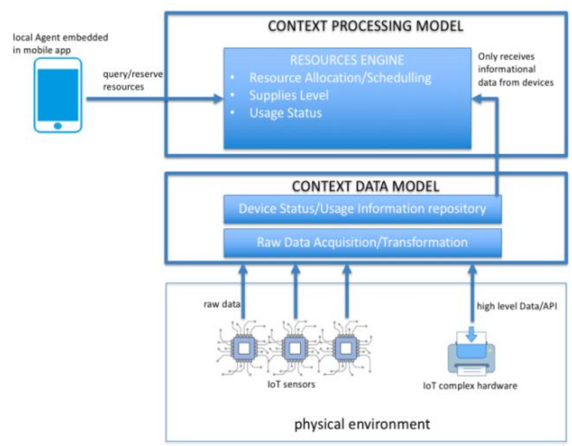


Fig. 6. Interaction Flow of Context and IoT.

Fig. 6 contains the illustration of layers and components organized to respond Agent’s requests. The lowest layer is responsible to integrate physical devices to platform. In case of IoT hardware sending only binary/raw data, without embedded application or driver/software providing interfaces to translate its state to consumer application, in this experiment we call it as IoT Sensor. This group of devices can contain temperature, presence or light sensors, providing their signals to component that performs data acquisition/transformation and turns this data available to component responsible of data managing, both components in Context Data Model layer. To categories of hardware which embed any level of interfaces, drivers or operational software, we call it as IoT Complex Hardware. Within layer Context Processing Model, the implemented component Resources Engine contains the logic to handle local Agent requests to query services, status, supplies level and performs scheduling of usage reservations. This engine also manages remote Agent instance and performs the synchronization between agent instances.

VI. Experiment with Activities Supported by Multiagent Interaction and Context-Aware Environment

Since the intent of current stage of this research, requires identify the useful behavior of this computational engine proposed, a field investigation and simulated tests in real environments become important to map difficulties of real world application. This experiment was organized in the following steps:

- Find and select pre-defined physical locations with available devices prepared to act like IoT/environment resources as its services.
- To each physical location, the same base copy of middleware was installed, adjusted and parameterized according to its environment restrictions, number of devices and type of resources.
- A mobile application from a previous work was complemented to simulate the proposed Multiagent operations to perform interactions with Context-Aware environment, executing resources reservation and content search. This application also contains the operations of Multiagent synchronization with Adaptive Processing Kernel.
- A group of volunteer students tested the mobile application during real use of physical location services on educational tasks.
- Results and feedback collected to analysis.

A. Experiment Method and Test Flow

The first list of physical locations invited to participate on this experiment, was prepared with co-work offices (https://coworkingbrasil.org), public libraries (http://www.prefeitura.sp.gov.br/cidade/secretarias/cultura/bibliotecas/) and private learning entities.

TABLE I. Distribution of Physical Locations

Locations	Invited	Interested	Participating	Volunteers
Co-Work Offices	60	3	0	0
Libraries	20	18	0	0
Learning Entities	20	17	6	21

Table I contains the categories distribution of location's list prepared. From a total of 100 locations contacted and invited to participate, 38 demonstrated interest on experiment. From this group of 38 interested, 6 locations attended the criteria of this experiment on sharing data containing history of resources usage and having technological infrastructure/support to participate. Each physical location involved in this experiment, invited some users to test the mobile application during their educational activities, and shared a dataset with history of resources used by invited volunteer users. In this phase of research and, in this proof of concept, data regarding volunteer users are not profile identifiable to preserve participant's privacy.

The history of user's access on resources, provides data that was prepared (pre-processed) to train and test a dedicated layer of Multilayer Perceptron Neural Network within Adaptive Processing Kernel, to predict the next resource to be scheduled/provided to user. This layer is dedicated to this processing since for each user's operation of accept/reject a scheduling and content recommended by Kernel, a new cycle of this MLP Neural Network training is performed.

For each access on mobile application, hence, identifying an activity start by user, the local Agent starts to find to identify a new available context or connect to a new context. Once integrated with context, the local Agent performs its action function receiving from Kernel the resources reservation recommendation. Initially, an indicative with resource and agenda to reserve is received from Kernel.

A rule based on proximity of next available agenda is applied on local Agent as conflict resolution strategy [27], without Kernel synchronization to this task needed. In this operation, adapted to this experiment, a pre-reservation of resource occurs after Kernel recommendation and remains valid waiting to user's confirmation on the proposed agenda during a defined time by context.

Within Agent operation, a level of data sharing between Agent instances running on same environment occurs through remote Agent protocol, with a group of data indicating a collection of tags provided by users during the use of application. These tags are collected in context by local Agent and synchronized with Kernel to indicate attributes/properties related to environment that should be seen by every user in same environment. This strategy helps simulate a communication level with multiple instances of Agents that could take local action based on shared specific information. The last received data from Kernel, data collected from environment (device status for example) remains on local Agent and are synchronized with remote Agent on context. In parallel with remote Agent updates, the context manages devices status information (supplies availability and activity status) and performs scheduling operation.

Every operation executed by users on application are tracked to understand the navigation flow, including other previously implemented activities performed by application like routing recommendation that remains in this extended version.

B. Preliminary Results

A total of 21 volunteers participated, using the mobile application during four weeks in support of their educational tasks interacting with the selected locations. Weekly, feedback of each user and the data resulted from tests were collected, keeping same implementation of Agent model and middleware of context provided.

TABLE II. DATA FROM LOCATION WITH HIGHEST ACCES

Location A	Week 1	Week 2	Week 3	Week 4
Scheduling Recommended	34	46	71	74
Scheduling Accepted	12	37	62	69
Content Validated	3	21	26	43
Conflicts Solved by Agent	29	13	9	7

Table II contains the data collected from Location A, with highest number of application access by its users selected to this experiment. The evolution of model operating with this location can be observed with growing of recommendation acceptance.

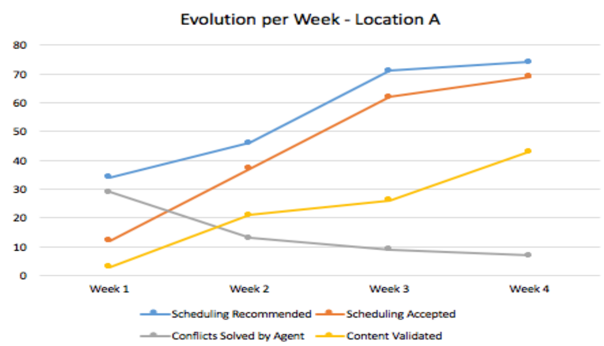


Fig. 7. Graphic representation of growing recommendation acceptance on Location A.

With this Location A can be observed an evolution from 35% of acceptance of Agent's recommendation by users during activities on first week, to around 93% on fourth week.

TABLE III. DATA FROM LOCATION WITH LOWEST ACCESS

Location B	Week 1	Week 2	Week 3	Week 4
Scheduling Recommended	12	19	41	59
Scheduling Accepted	5	9	34	52
Content Validated	7	11	17	31
Conflicts Solved by Agent	4	7	5	4

Table III contains the data from Location B, with lowest access observed. With this Location B, at first week, evolution observed started from 41% of acceptance of Agent's recommendation on first week, to around 88% on fourth week.

The graphic illustration indicating recommendation and acceptance in convergent tendency to Location A and Location B are contained in Fig. 7 and 8, respectively. Conflicts resolution can be observed as a decreasing event for both Locations presented.

After finished the period of four weeks of tests, the locations and the group of volunteers sent final feedbacks about their participation on this experiment.

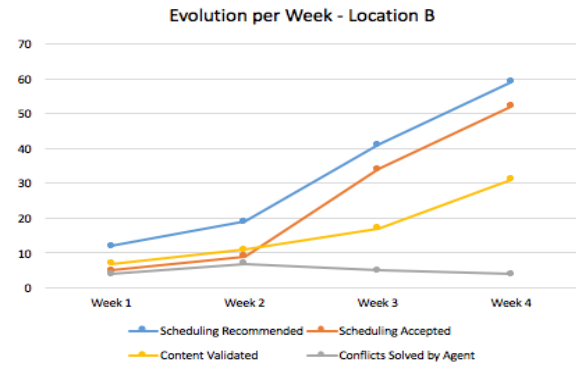


Fig. 8. Graphic representation of growing recommendation acceptance on Location B.

VII. CONCLUSION AND FURTHER WORK

Lessons learned selecting locations, indicates the need of a study to investigate privacy and technology aspects to each category of location. During investigation on details to map the reasons of lower interest from Co-Work Offices on experiment, a questionnaire answered reveals business concerns on avoid sharing data of services usage and, eventually, negative ratings by users.

To libraries category, the public institutions invited to participate presented a high number of interested locations, however, none of institutions contacted owns the necessary infrastructure to participate on tests. Some of these libraries, sharing a public platform to manage the book collection does not have an automated control over other resources provided like room reservation and digital/media resources. Given this scenario, an open space emerges to an intelligent model as proposed in this research to complement an automated platform applied to this environment category.

The group of Learning Institutions participating in this experiment, essentially centers with Arduino/IoT specialization courses, presented the necessary requirements and improvements in its environment to participate, hence, totality of volunteers to test. These spaces with intrinsic technology profile can be useful to understand the necessary complement to scale the research model to other environment categories.

Considering the main premise of this research about a modular intelligent computational engine, this experiment preparation with setup and customizing a framework of context middleware, facilitates the process and indicates a fluid strategy to evolve the model proposed and future tests.

The modeling of the experiment indicates the context architecture proposed, acting as a computational tool to manage and mediate the access to the IoT devices of the environment, as favorable strategy. This architecture enables formation of data repositories that can be accessed when needed by Kernel and reduce the complexity of managing multiple devices.

Owing to the space limit and the focus of this work, many complementary aspects covered during research could not be discussed in details here. This includes Multiagent self-sustainability that may be applied to maintaining stability of the system (with respect to an appropriate size of the population of agents), the capability to respond and attend to the emergency situations occurring in the environment (breakdown hardware/software of infrastructure), as well as in the environment and user requirements (e.g. change of computation parameters or devices/resources coupled). Not detailed in this resume, the Neural Networks layer added to Kernel, its topology and data normalization method used on prediction process process of resources reservation, followed the basic method presented by Dos Santos et al. 2017 [3].

Feedback collected with volunteers during the tests, pointed desirable features on future version of model. About 80% of volunteers indicated automated activation of resources and communication of execution status as important improvement. Around 68% of volunteers indicated as necessary, the feature of previous software/drivers compatibility check to access electronic resources. As desirable resource, about 90% of volunteers indicated active recommendation of new location, added to resources scheduled reservation in case of unavailability/error on current location resource.

The next stage of research will cover the computational engine model responsible to process and predict the influence of resources on activities results.

Aspects of educator's perspective, regarding to support intelligent task composition is one of next research phase to cover.

Conclusions initially obtained of this current state of research provides basement to evolve the investigation and future experiments.

References

[1] Ally, Mohamed. Mobile learning: Transforming the delivery of education and training. Athabasca University Press,

[2] Silva, Luciano et al. Using conceptual lattices to represent fine granular learning objects through scorm meta-objects. The Electronic Journal of e-Learning, v. 4, n. 2, p. 141-148, 2006.

[3] Dos Santos, Fernao Reges; Notargiacomo, Pollyana. EducActiveCore: Computational Model to Educational Personalization Based on Multiagent and Context-Aware Computing. International Journal of Learning, Teaching and Educational Research, v. 16, n. 11, 2017.

[4] Baldauf, Matthias; Dustdar, Schahram; Rosenberg, Florian. A survey on context-aware systems. International Journal of Ad Hoc and Ubiquitous Computing, v. 2, n. 4, p. 263-277, 2007.

[5] Abowd, Gregory D. et al. Towards a better understanding of context and context-awareness. In: Handheld and ubiquitous computing. Springer Berlin Heidelberg, 1999. p. 304-307.

[6] Bellavista, Paolo et al. A survey of context data distribution for mobile ubiquitous systems. ACM Computing Surveys (CSUR), v. 44, n. 4, p. 24, 2012.

[7] Schilit, Bill; Adams, Norman; Want, Roy. Context-aware computing applications. In: Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on. IEEE, 1994. p. 85-90.

[8] Makris, Prodromos; Skoutas, Dimitrios N.; Skianis, Charalabos. A survey on context-aware mobile and wireless networking: On networking and computing environments' integration. Communications Surveys & Tutorials, IEEE, v. 15, n. 1, p. 362-386, 2013.

[9] Ogata, Hiroaki; Yano, Yoneo. Context-aware support for computer-supported ubiquitous learning. In: Wireless and Mobile Technologies in Education, 2004. Proceedings. the 2nd IEEE International Workshop on. IEEE, 2004. p. 27-34.

[10] Al-Fuqaha, Ala et al. Internet of things: A survey on enabling technologies, protocols, and applications. IEEE Communications Surveys & Tutorials, v. 17, n. 4, p. 2347-2376, 2015.

[11] Bandyopadhyay, Soma et al. Role of middleware for internet of things: A study. International Journal of Computer Science and Engineering Survey, v. 2, n. 3, p. 94-105, 2011.

[12] Dohr, Angelika et al. The internet of things for ambient assisted living. In: Information Technology: New Generations (ITNG), 2010 Seventh International Conference on. Ieee, 2010. p. 804-809.

[13] Braun, Andreas; Wichert, Reiner; Maña, Antonio. Ambient Intelligence. Springer International Publishing AG, 2017.

[14] Lee, In; Lee, Kyoochun. The Internet of Things (IoT): Applications, investments, and challenges for enterprises. Business Horizons, v. 58, n. 4, p. 431-440, 2015.

[15] Huang, Yueh-Min; Chiu, Po-Sheng. The effectiveness of a meaningful learning-based evaluation model for context-aware mobile learning. British Journal of Educational Technology, v. 46, n. 2, p. 437-447, 2015.

[16] Cheverst, Keith et al. Developing a context-aware electronic tourist guide: some issues and experiences. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems. ACM, 2000. p. 17-24.

[17] Yang, Stephen JH. Context aware ubiquitous learning environments for peer-to-peer collaborative learning. Journal of Educational Technology & Society, v. 9, n. 1, p. 188-201, 2006.

[18] Kwon, Oh Byung; Sadeh, Norman. Applying case-based reasoning and multiagent intelligent system to context-aware comparative shopping. Decision Support Systems, v. 37, n. 2, p. 199-213, 2004.

[19] Traxler, John; Kukulska-Hulme, Agnes (Ed.). Mobile learning: The next generation. Routledge, 2015.

[20] Kwon, Ohbyung; Choi, Sungchul; Park, Gyuro. Nama: A context-aware multiagent based web service approach to proactive need identification for personalized reminder systems. Expert Systems with Applications, v. 29, n. 1, p. 17-32, 2005.

[21] Soldatos, John et al. Agent based middleware infrastructure for autonomous context-aware ubiquitous computing services. Computer Communications, v. 30, n. 3, p. 577-591, 2007.

[22] Aouatef, Chaib; Iman, Boussebough; Allaoua, Chaoui. Adaptive composition of services in context-aware ambient intelligent systems. In: Proceedings of the 2014 International C* Conference on Computer Science & Software Engineering. ACM, 2014. p. 26.

[23] Adomavicius, Gediminas; Tuzhilin, Alexander. Context-aware recommender systems. In: Recommender systems handbook. Springer, Boston, MA, 2015. p. 191-226.

[24] Santos, F. R. ; Mustaro, P. N. . Educatutor: A Tutoring System Based On Multiagent Approach And Social Network Analysis For Development A Digital Tv Environment with Adaptive And Collaborative Proposal. In: 23rd ICDE World Conference on Open Learning and Distance Education including the 2009 EADTU Annual Conference, 2009, Maastricht. Book of Abstracts, 2009.

[25] Alshwaier, Abdullah, Ahmed Youssef, and Ahmed Emam. "A new Trend for E-learning in KSA using educational clouds." Advanced Computing 3.1 (2012): 81.

[26] Weiss, Gerhard (Ed.). Multiagent systems: a modern approach to distributed artificial intelligence. MIT press, 1999 - p241-278.

[27] Cicirelli, Franco; Giordano, Andrea; Nigro, Libero. Efficient environment management for distributed simulation of large-scale situated multi-agent systems. Concurrency and Computation: Practice and Experience, v. 27, n. 3, p. 610-632, 2015.

Helitron's Periodicities Identification in C.Elegans based on the Smoothed Spectral Analysis and the Frequency Chaos Game Signal Coding

Rabeb Touati[1], Imen Messaoudi[2], Afef Elloumi Ouesleti[3], Zied Lachiri[4]
SITI Laboratory, National School of Engineers of Tunis (ENIT), University of Tunis El Manar, Tunisia
Tunis, Tunisia

***Abstract*—Helitrons are typical rolling circle transposons which make up the bulk of eukaryotic genomes. Unlike of other DNA transposons, these transposable elements (TEs), don't create target site duplications or end in inverted repeats, which make them particular challenge to identify and more difficult to annotate. To date, these elements are not well studied; they only attracted the interest of researchers in biology. The focus of this paper is oriented towards identifying the helitrons in C.elegans genome in the perspective of signal processing. Aiming at the helitron's identification, a novel methodology including two steps is proposed: the coding and the spectral analysis. The first step consists in converting DNA into a 1-D signal based on the statistical features of the sequence (FCGS coding). As for the second step, it aims to identify the global periodicities in helitrons using the Smoothed Fourier Transform. The resulting spectrum and spectrogram are shown to present a specific signature of each helitron's type.**

***Keywords*—Helitrons; C.elegans; Frequency Chaos Game Signal (FCGS) coding; spectral analysis; tandem periodicities**

I. INTRODUCTION

The helitrons are a distinguished type of the transposable elements (TEs) DNA which transposes by a rolling circle replication mechanism. Due to their ability to move rapidly and replicate within genomes, helitrons play a major role in genomes evolution. In fact, by transferring a DNA segment from one genomic site to another, these elements are responsible for intragenomic multiplication [1]. Many types of genetic variation caused by TEs in animals and plants are described in [2].

The helitrons are part of the transposable elements (TEs) class 2 [3]. They were discovered by in silico genome-sequence analysis. Newly discovered in all eukaryotic genomes [4]-[7], helitrons have shown a remarkable ability to capture gene sequences [8], [9]. In fact, on a large scale, these elements are suggested to duplicate and shuffle exon domains [10]-[12].

Helitrons are present in various organisms: like plants (such as corn, maize, rice, Arabidopsis Thaliana) [13], nematodes (such as the worm Cænorhabditis elegans) [14], fungus [15], [16] and animals [17] (such as lucifugus [8] and specifically in the vertebrates genomes like the fishes Danio rerio and Sphoeroides nephelus [16]). Approximately, the helitron DNA constitutes at most 2% of the A. thaliana, the C.elegans and the maize genomes [12], [14], [18] and 4.23% of the silkworm genome [19]. These elements are mostly represented by non-autonomous elements.

Since their discovery, helitrons have attracted widespread attention. Many computational tools were developed to identify and analyze the helitron in genomes from which are: HelitronFinder [20], HelSearch [21], a combination of BLAST search and hidden Markov models [22] and HelitronScanner [23].

HelitronFinder and HelSearch are similar [24], [25]; they are based on the conserved sequences at the termini 5′-TC and CTAG-3′(R = A or G) of most Helitrons. Both programs look for the hairpin structure and the CTRR 3′ terminus.

The users of HelSearch have to manually search for the 5′end of Helitrons whereas users of HelitronFinder can identify the 5′end automatically. The combination of BLAST search and hidden Markov models is a limited method to identify helitrons with more diverse termini. The HelitronScanner, a two-layered local combinational variable (LCV), is a tool that identifies helitrons missed by both HelSearch and HelitronFinder [23]. This tool aims to extract helitron features like the hairpin structure, CTRR at the 3′ end and the TC dinucleotide at the 5′ end, and the A and T residues flanking the 5′ and 3′ ends, respectively. In an automated way, this tool uncovers many new helitrons which were missed by other tools. But this method presents also limitations in finding helitrons which don't have hairpins. Given that these transposable elements lack typical transposon features, the helitron's investigation is limited. The automated identification and localization of helitrons remain purely based on previously known sequences.

Helitrons elements are widespread and highly heterogeneous which makes their identification a difficult task. Here, the novelty of this work consists in the use of signal processing techniques to search helitrons in a large genomic database without prior knowledge about the content of the DNA sequences. The idea consists in finding a way to identify each helitron's type based on a signature that characterizes it.

To apply the signal processing methods on biological sequences, it is required to convert DNA into a digital signal known as DNA coding. The numerical representation of DNA using coding techniques is important since it plays a great role in visualizing, characterizing and highlighting the information contained in it. Different coding techniques exist: the binary coding [26], the structural bending trinucleotide coding (PNUC) [27], the electron-ion interaction pseudo-potential (EIIP) mapping [28], the Frequency Chaos Game Signal [29]-[31], etc.

In the signal processing field, several techniques were applied with success to detect some biological sequences. For example, genes were segmented into coding and non-coding regions using the windowed Fourier Transform and based on the 3-bp periodicity that characterizes exon [32], [33]. In addition, the Fast Fourier Transform (FFT) was used to reflect the correlation properties of the coding and non-coding DNA sequences [32]. In [34], another technique of analysis is used to detect the short exons: the Fourier transform infrared spectroscopy. The latent periodicities in different genome elements including exons and microsatellite DNA sequences were detected and used the Fourier transform [35]. In addition, the modified Gabor-wavelet transform was used to identify the protein coding regions [36]. The Auto-Regressive technique was also used to predict genes and exon location using allpass-based filters [37].

Further, the Wavelet Transform (WT) allowed one to balance resolution at any time and frequency, which gave the ability to automatically capture different periodicities (frequencies inverse): periodicity 3 in exons [38] and periodicity 10 in nucleosomes [39]. The wavelet transform was also shown to reflect the characteristic signature associated to the tandem repeats. Indeed, some biological sequences (such as tandem repeats) were characterized by periodicities; the scalograms served to visualize the way these features appear as well as their locations [40]. Given the efficacy of these signal processing tools, analyzing the helitron DNA category (which is governed by complex latent periodicities) is particularly challenging.

Our main goal is to characterize helitrons within the framework of signal processing. For this, the Frequency Chaos Game Signal (FCGS) is selected to preserve the statistical proprieties of DNA sequences. Secondly, signal processing analysis techniques are applied to identify each helitron type by a genomic signature. The key component of this system is the combination of the DNA coding with the Frequency Chaos Game Signal (FCGS) and the windowed Smoothed Fourier technique which enhances the spectral signature of DNA (the overall periodicities). This paper distinguishes four main sections: the First section introduces the work. The second section describes the methodology adopted for the helitron identification. The third section explains the establishment of the helitron's signal database. The fourth section explains the spectral analysis used to characterize helitrons. Then, the experimental results are provided and discussed. Finally, a summary is put forth describing the effectiveness to capture all the periodicities in helitrons sequences based on this new approach. The rest of this paper end with a perspective that open questions concerning the biology of this type of TEs.

II. METHODOLOGY OF HELITRON'S IDENTIFICATION

With the wealth of genomic sequences now available, the identification of a specific DNA element has to be an automatic task. The following methodology consists of three steps. The flowchart describing this work is given in Fig. 1.

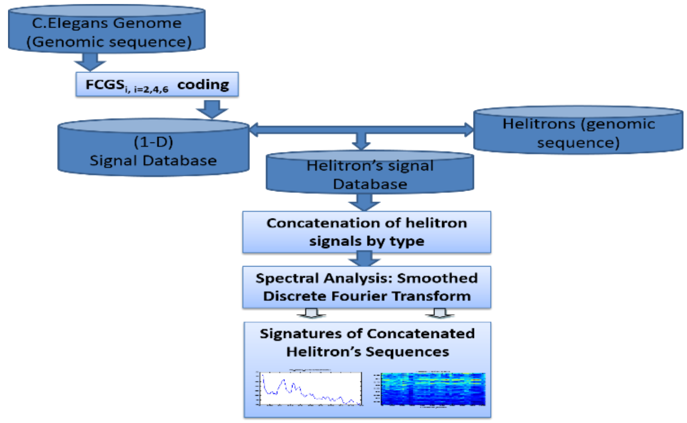


Fig. 1. The Helitron identification establishment flowchart in C.elegans organism.

The first step: A DNA sequence must be converted into a numerical sequence before processing. This step consists in generating different signals for each C.elegans chromosome. For this, chromosomes sequences from the NCBI data base for the C.elegans model (http://www.ncbi.nlm.nih.gov/Genbank) are extracted. Then, 1-D signals are generated by applying the Frequency Chaos Game Signal of order i: $FCGS_{i,\ (i=2,\ 4,\ 6)}$ for the whole chromosomes. This coding technique is based on the apparition's probability of N successive nucleotides groups in an entry DNA sequence [29]-[31]. Here, three values of the length: N=2, 4, 6 are selected.

The probability (P_{N_nuc}) of given N nucleotides in the chromosome is as follows:

$$P_{N\_nuc}= N_{N\_nuc}/N_{ch} \tag{1}$$

N_{N_nuc} represents the number of apparition of the N nucleotides in the whole sequence. N_{ch} represents the length in base pairs of the DNA sequence. In other words, the number of occurrence of each of these elements (N nucleotides) in the genome are counted. It is important to note that the FCGS signal depends on the DNA sequence to be encoded since it is based on counting the words contained in it. It reflects, therefore, the statistical-features of the DNA sequence itself. In fact, changing the DNA sequence would affect the probability of apparition of words and the FCGS values. The interesting point here is that this coding technique is useful in terms of enhancing the repetitive DNA (such is the case of helitronic DNA) at any desired scale (i.e. for any word's size). Specifically, the highest level of repetitions in DNA are detected by high order FCGS's coding techniques. It must be noted that hidden information at a certain scale can be highlighted at another scale; that's make the FCGS very suitable to investigate helitrons.

The second step: It consists in the establishment of a 1-D signal database of the helitrons. This is done through the association of the $FCGS_i$ values to each group of letters in the helitron sequence.

In position (k), the oligomer (i), which consists of N nucleotides, is replaced by the correspondent occurrence probability:

$$S_{N\_nuc}(k)=\sum_i P_{N\_nuc}(i,k) \quad (2)$$

The sum of the N nucleotide indicators (S_{N_nucl}) can be computed as following:

$$H_w[n,k] = H[n]W[n-k\Delta n] \quad (3)$$

Then, a database of helitrons signals regarding different oligomers is prepared. As a result, helitrons are represented by three levels of FCGS ($FCGS_{i,i=2,4,6}$) .

As example, Fig. 2 outlines the resulted signals: $FCGS_2$, $FCGS_3$, $FCGS_4$, $FCGS_5$ and $FCGS_6$ of an helitron type NDNAX2 with a size of 341 base pairs and positioned at: [274811bp: 275151bp] in the chromosome II of the C.elegans genome.

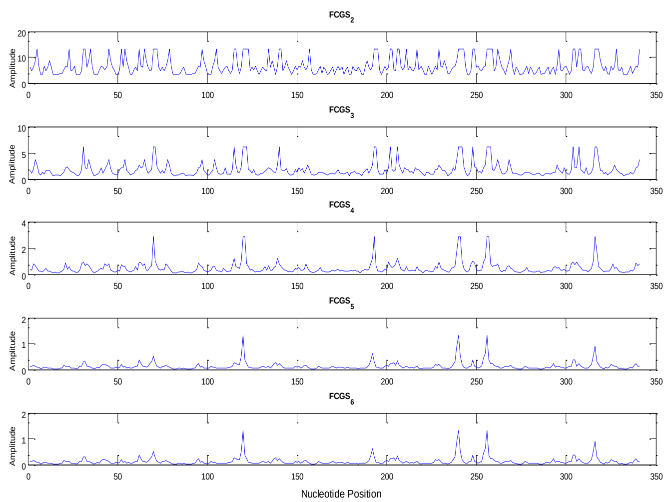


Fig. 2. NDNAX2_CE Helitron presentation by FCGS2 to FCGS6.

After coding process, several numerical signals, denoted by H[n], are obtained for each helitron. A signal database of each helitron's type and for each chromosome in the considered genome is then established. For a desired level i of the FCGS (FCGSi) and for a specific class of the helitron, signals of all existing helitron elements are concatinated in such a way to obtain one global sequence. The resulting database comprises three $FCGS_i$ signals (i=2, 4, 6) for each class of helitron contained in the chromosomes of the C.elegans genome.

The third step: The spectral analysis method is adopted to detect the periodicity of each helitron's type. Based on the classic discrete Fourier transform related to numerical sequence (4), a windowed local analysis is used to give more precise location in time and frequency.

The reason for selecting the spectral analysis is to characterize the helitron elements by a global spectral signature.

$$H[k] = \sum_{n=0}^{N-1} H(n)e^{-j\frac{2\pi}{N}nk} \quad (4)$$

At this stage, two types of signatures for each helitron class are revealed: the1-D spectrum and the 2-D spectrogram. The spectral and the time-frequency signatures are established for different orders of our coding technique (FCGSi, i=2,4,6).

III. Helitron's Signals Database

For experimentation, the C.elegans genome whose chromosomal DNA sequences in the NCBI database [40] are considered. This genome contains ten helitron families diffused through six chromosomes (5 autosomes: ChI , ChII ,..., ChV and 1 gonosome: ChX). These helitron classes are: Helitron1 (H1), Helitron2 (H2), HelitronY1 (Y1), HelitronY1A (Y1A), HelitronY2 (Y2), HelitronY3 (Y3), HelitronY4 (Y4), NDNAX1 (N1), NDNAX2 (N2) and NDNAX3 (N3). The helitron sequences are very heterogeneous. In fact, the apparition number of these elements and the size varies from one chromosome to another as described in Table I. In addition, for a particular type of helitron this size varies from one sequence to another; which makes their identification very difficult.

From Table I, it is obvious that HelitronY1A is the most frequent (occurrence number of 1093) and the longest (size of 425487bp) in the genome.

The least frequent helitron in the genome is Helitron NDNAX1 (with an occurrence number of 77). At the chromosomal level, Helitron NDNAX1 and HelitronY3 are the least frequent elements; they are present in chromosomeII (with an occurrence number of 8). As for the smallest class of helitron is HelitronY3 with a size of 16404 bp; in other words it presents the shortest helitron in the genome.

Because of the variability of helitron's number and size in the six chromosomes, regrouping all the helitron signals in a way to obtain a global signature of each class comes in handy. The idea consists in concatenating all the signals in one vector for a well-defined order of the FCGS, while keeping the apparition order of helitrons in chromosome. This reunification allows linking all the local signatures of helitrons. Our main goal is to find the structure that could be repeated in each helitron.

IV. Smoothed Discrete Fourier Transform

The smoothed spectral analysis is a convenient tool to search for base periodicities within DNA sequences. In fact, the periodically placed motifs may indicate the presence of genes, regulatory elements or other significant hotspots like helitrons; hence the need to locate them and exhibit the correspondent frequency (or periodicity). This technique is applied to investigate the global signature of helitrons within the genome by considering the concatenated FCGS signals of all helitrons which constitutes the novelty of this work. Therefore, two tasks are carried out here:

- Revealing periodicities by enhancing the global periodicities in the helitron sequences by using the Smoothed Fourier Transform.
- Locating the frequencies points using the Smoothed Discrete Fourier Transform applied on sliding window along the DNA sequence [26], [27], [41], [42].

The technique consists in:

a). After converting the DNA sequence into a numerical one, the signal $H[n]$ must be divided into frames of L length with an overlap Δl.

b). Using a sliding analysis window $w[n]$, each L portion is also divided into N overlapped segments with an overlap length Δn:

$$H_w[n,k]=H[n]w[n-k\Delta n] \qquad (5)$$

Where, the index of the frequency ([0, N-1]) is noted by k.

The choice of the windowing function determines the time-frequency resolution.

c). Using the Discrete Fourier Transform (DFT), each weighted block of the frame $H_w[n]$ is transformed in the spectral domain. The DFT of each segment is expressed as follows:

$$H_w^i[k] = \sum_{n=0}^{N-1} H_w\,[n,i]e^{\frac{-j\,2\,\pi nk}{N}} \qquad (6)$$

d). Calculating the DFT mean value for each portion (1: L); then, performing the same operation to the N segments. The mean smoothed spectrum is expressed as:

$$Hmean_w^j[k] = \frac{1}{L*N}\sum_{j=o}^{L-1}\sum_{i=0}^{N-1} H_w^i[k] \qquad (7)$$

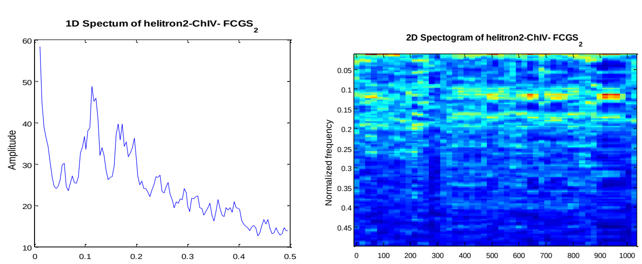
Noted that i corresponds to the index frame of N frames ([1...N]), k is the index of the frequency and j corresponds to the index frame of L frames ([1: L]).

e). With the obtained values, a matrix containing the join time frequency information is constructed:

$$\mathrm{Matrix(j,k)} = Hmean_w^j[k] \qquad (8)$$

This representation consists of the spectrogram amplitude for a specific index periodicity in a specific nucleotide position in the DNA sequence.

Fig. 3 provides an example of the spectra and the spectrograms of the concatenated helitrons of type Helitrons2_CE that exist in chromosome IV of C.elegans. The frequency and the time-frequency representations are generated by the mean valued technique based on the smoothed Discrete Fourier Transform. As for the sliding window, we suggest using a blackman window with the parameters: L=1024, Δl=512, N=256 and Δn=64. In this example, two levels of the coding technique are used: FCGS2 and FCGS4.



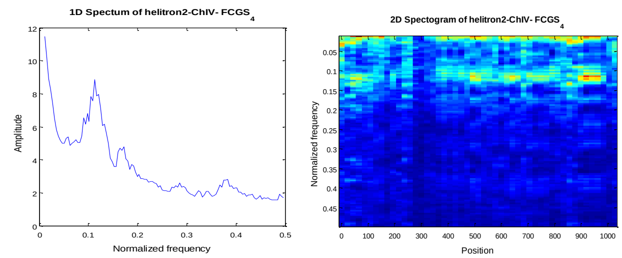
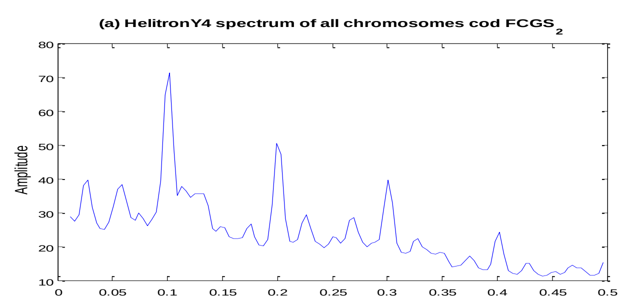


Fig. 3. 1D spectra and 2D spectrograms of helitron2_CE [chromosome IV with size of 25932bp] when coded with $FCGS_2$ and $FCGS_4$.

From these figures, it is noticeable that increasing the FCGS order induces a general smoothing of the 1D spectrum and the 2D spectrogram; which allows to enhance the spectral and the time-frequency behaviours of the considered helitron by highlighting its periodicities.

V. Helitron's Characterization in C.Elegans: Tests and Results

A characterization of helitron DNA at the aim of its identification is the main focus of this paper. Therefore, such a goal can be reached by studying the global resident periodicities of each helitron's class. Thus, considering the helitrons representations in the frequency and the time-frequency plans, a specific signature (pattern or form) characterizing the repetitions of the sequence is allowed by the spectral attributes for each type. The windowed smoothed Fourier analysis allows following the path of existing periodicities (engendered by the repetitions) in each helitron's class. For experimentation, all helitrons existing in the C.elegans genome are encoded by the $FCGS_{i,(i=2,4,6)}$ coding technique. Secondly, the $FCGS_i$ fragments which correspond to the helitron sequences in the same chromosome are concatenated for each helitron's type. For comparison, the concatenation of all helitrons contained in the genome (all chromosomes) is considered. Finally, the smoothed spectral analysis is applied to these sequences. In this step, many types of windows with different values of length and overlap are tested. The optimal parameters are fixed to L=1024, Δl=512, N=256, Δn=64 and the most accurate smoothed spectrum is given by a Blackman window. Consequently, two types of spectral representations are obtained: the 1-D spectrum and the 2-D spectrograms. For illustration, the example of HelitronY4 is selected and encoded with $FCGS_2$. The resulting 1-D spectra are presented in Fig. 4.



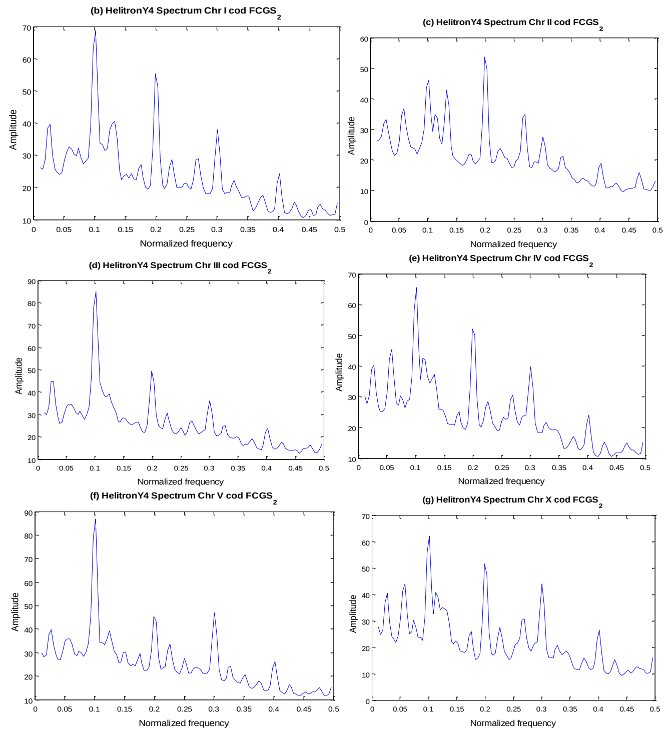


Fig. 4. HelitronY4 spectra- (a) spectrum of the concatenated chromosomes- (b:g) spectrum of each chromosome apart.

The sub-figures (b, c, d, e, f, g) provide the spectrum shape for each chromosome apart; the sub-figure (a) gives the shape spectrum of the overall genome. The horizontal axis of each sub-figure indicates the frequency (which is equivalent to the inverse of periodicity) measured by the Smoothed Fourier Transform and the vertical axis indicates the spectrum's amplitude.

By examining each subfigure closely, it is crystal clear that HelitronY4 is identified by five peaks whose power value differs from one chromosome to another. These peaks are located around remarkable frequencies which are: 0.02 (periodicity 50), 0.1 (periodicity 10), 0.2 (periodicity 5), 0.3 (periodicity 3) and 0.4 (periodicity 2). Since the C.elegans genome kept the same behavior as for each chromosome apart, these periodicities characterize the HelitronY4 type. Here, the idea is to identify each helitron class with a specific spectrum since the behavior of all spectra remains almost the same with a small variation in amplitude.

A. *Helitron's Signature and Effect of the FCGS Order on the Spectrum*

In this section, the role played by the FCGS coding in enhancing the helitron signature is thoroughly examined. For this purpose, the mean values of the Smoothed Discrete Fourier Transform for each type of helitron sequence are computed and thus for different orders of FCGS (order=2,4,6).

The spectrum representations of all helitron classes in C.elegans (Helitron1, HelitronY1, HelitronY1A, Helitron2, HelitronY2, HelitronY3, HelitronY4, NDNAX1, NDNAX2 and NDNAX3) are given in Fig. 5.

These representations reflect the specific periodicities (frequencies) that characterize each helitron's type which forms a pertinent tool for their identification. In fact, each helitron class is shown to possess a specific spectral signature.

Further, for each helitron type, the effect of increasing the FCGS order in the spectrum shape evolution is remarkable. But the overall shape remains when increasing the FCGS level. Based on that, the fact that the helitron can be identified by its spectral signature is confirmed. And that is emphasized even in the spectrum shape evolution.

On other hand, a high similarity between these three helitrons is striking:

- Helitron1_CE
- HelitronY1_CE
- HelitronY1A_CE

These helitrons have in common the frequencies; 0.02734 (Periodicity 37), 0.05859 (Periodicity 17) and 0.1 (Periodicity 10); which have the highest amplitudes. These periodicities reflect the presence of hidden minisatellite into the helitrons sequences.

In addition, Helitron2_CE have important amplitudes within the frequencies bands [0.09 0.125] and [0.15: 0.2]. They share several frequencies with the HelitronY2_CE class. The latter helitron (HelitronY2_CE) is shown to be easily recognized through its distinctive spectrum. In fact, during tests, from a chromosome to another the HelitronY2_CE is the only subclass which possesses an invariant behavior along the whole genome.

The HelitronY3 have the main highest amplitudes around the frequencies: 0.015, 0.15 (periodicity 7), 0.09 (periodicity 11) and 0.05 (periodicity 20). Regarding the FCGS signal patterns, this helitron class is shown to have a regular behavior.

The HelitronY4_CE have four remarkable amplitude around the frequencies which are: 0.02734 (periodicity 36), 0.05859 (periodicity 17), 0.1 (periodicity 10), 0.2 (periodicity 5), 0.3 (periodicity 3) and 0.4 (periodicity 2).

The Helitron NDNAX1 have two principal frequency bands which are centered around the frequency 0.1 (periodicity 10) and around the frequency 0.03 (periodicity 30).

The Helitron NDNAX2_CE is characterized by the frequencies: 0.01563 (periodicity 54), 0.04 (periodicity 24), 0.1(periodicity 10), and 0.168 (periodicity 6). Also, their spectrogram contains repetitive patterns around the frequencies band 0.1.

Finally, the Helitron NDNAX3_CE has pronounced amplitude around the frequency 0.1(periodicity 10). Other important frequencies are within reach, such as 0.01953 which corresponds to periodicity 51 and 0.02344 which corresponds to periodicity 42. In addition, the periodicities 10 and 9 are present in this class of helitrons families.

B. *Helitron's Spectrograms with different FCGS Orders*

The time-frequency presentation allows an energetic interpretation of signals. This representation has shown high-ability in detecting different hotspots in DNA sequences. In this part, the specific regions that can exist in the helitron sequences are examined.

Fig. 6 presents the spectrogram of each helitron type considering three levels of FCGS: $FCGS_2$, $FCGS_4$ and $FCGS_6$. The vertical axis of each sub-figure indicates the frequency measured by the Smoothed Fourier Transform. As for the horizontal axis, it indicates the position in base-pairs.

Based on these sub-figures, each helitron class seems to possess a unique time-frequency behavior. In other words, it forms a time-frequency signature which will be taken as a basis for the helitron identification.

In this case, the $FCGS_2$ coding seems to allow the best way of characterizing helitrons since it shows more details about the energy distribution in spectrograms. On the other hand, increasing the coding order has no major impact on the overall signature for all helitron types. Nevertheless, smoothed spectrograms with high levels of the FCGS signals are obtainable.

VI. Conclusions

Exploring the latent periodicities of helitrons (eukaryotic rolling-circle transposons) can play a key role in the identification of this subclass of DNA. The way these periodicities appear can mark specific regions of helitrons in a unique manner; which constitutes a signature allowing us to distinguish these elements.

This study was done based on the spectral analysis. In fact, the recognition of each class's periodicities can be very useful for helitron classification.

To be able to apply the spectral method, the DNA sequence is converted into a numerical 1-D signal using the FCGS coding. This technique offers the possibility to encode DNA into several signals according to a well-defined order. Three levels of the representation are taken into account: $FCGS_2$, $FCGS_4$ and $FCGS_6$.

To know more about how similar are these helitrons along the genome, all the helitron sequences that exist in the C.elegans chromosomes are associated in order to obtain a global genomic signature.

After that, the Smoothed Fourier Transform Analysis is applied and the 1-D spectra and the 2-D spectrograms are collected as characteristic signatures of the studied elements. In fact, for each helitron type, the periodicities shape was involved in the spectral and the time-frequency representations in a unique manner; which forms a pertinent tool for the helitron characterization and identification.

Comparing the signature of helitrons in a chromosome with the signature of the overall genome, the great similarity between them is plain to see. In addition, increasing the FCGS order has not affected the global behavior in the spectra and the spectrograms: the main periodicities of each helitron class have remained. A smoothing effect is spotted in the representations. This confirms that the helitron adopts a distinctive behavior which characterizes it and thus permits its identification.

In this work, the major advantage of these frequency representations (spectrum) and time-frequency representation (Spectrograms) is that it captures all the periodicities (repetitive sequence) for the heletronic sequences. It is worth mentioning that this approach can be used in detecting periodicities of other transposable elements. The limitation of the approach depends on the limitation of the Fourier transform. Therefore, this approach remains limited as it does not allow the temporal localization using a fixed window size.

VII. Future Work

In this work, the importance of helitron characterization has been highlighted and a method to perform their identification based on the Smoothed Fourier Analysis has been suggested.

Classification methods based on the feature analysis technique might be investigated to improve the classification accuracy of helitrons as a future work.

TABLE I. The Occurrence Number and the Length (in Base Pairs) of Helitrons in C.Elegans.

Type	Characteristics	ChI	ChII	ChIII	ChIV	ChV	ChX	Genome
H1	N helitrons	42	32	37	28	42	14	197
	Size(bp)	12911	59232	11836	9456	11012	3486	107933
H2	N helitrons	106	74	72	77	100	40	469
	Size(bp)	42643	28078	31715	25932	38158	15788	182314
Y1	N helitrons	60	44	82	83	188	26	483
	Size(bp)	21567	21193	35148	32612	89573	12171	212264
Y1A	N helitrons	176	132	211	155	366	53	**1093**
	Size(bp)	60860	43949	75437	75410	145489	24342	**425487**
Y2	N helitrons	54	44	39	79	70	51	337
	Size(bp)	11026	8184	6851	15253	12337	10091	63742
Y3	N helitrons	11	**8**	17	24	25	32	117
	Size(bp)	1578	1264	2521	3496	3167	4378	**16404**
Y4	N helitrons	68	68	81	114	148	44	523
	Size(bp)	29522	36026	35311	55409	72431	24371	253070
N1	N helitrons	9	**8**	14	15	19	12	77
	Size(bp)	6197	6701	6896	10422	7761	5730	43707
N2	N helitrons	30	40	22	43	37	16	188
	Size(bp)	7762	19810	12476	16698	21105	7256	85107
N3	N helitrons	20	20	24	33	23	14	134
	Size(bp)	18229	14357	21517	29163	16545	17049	116860

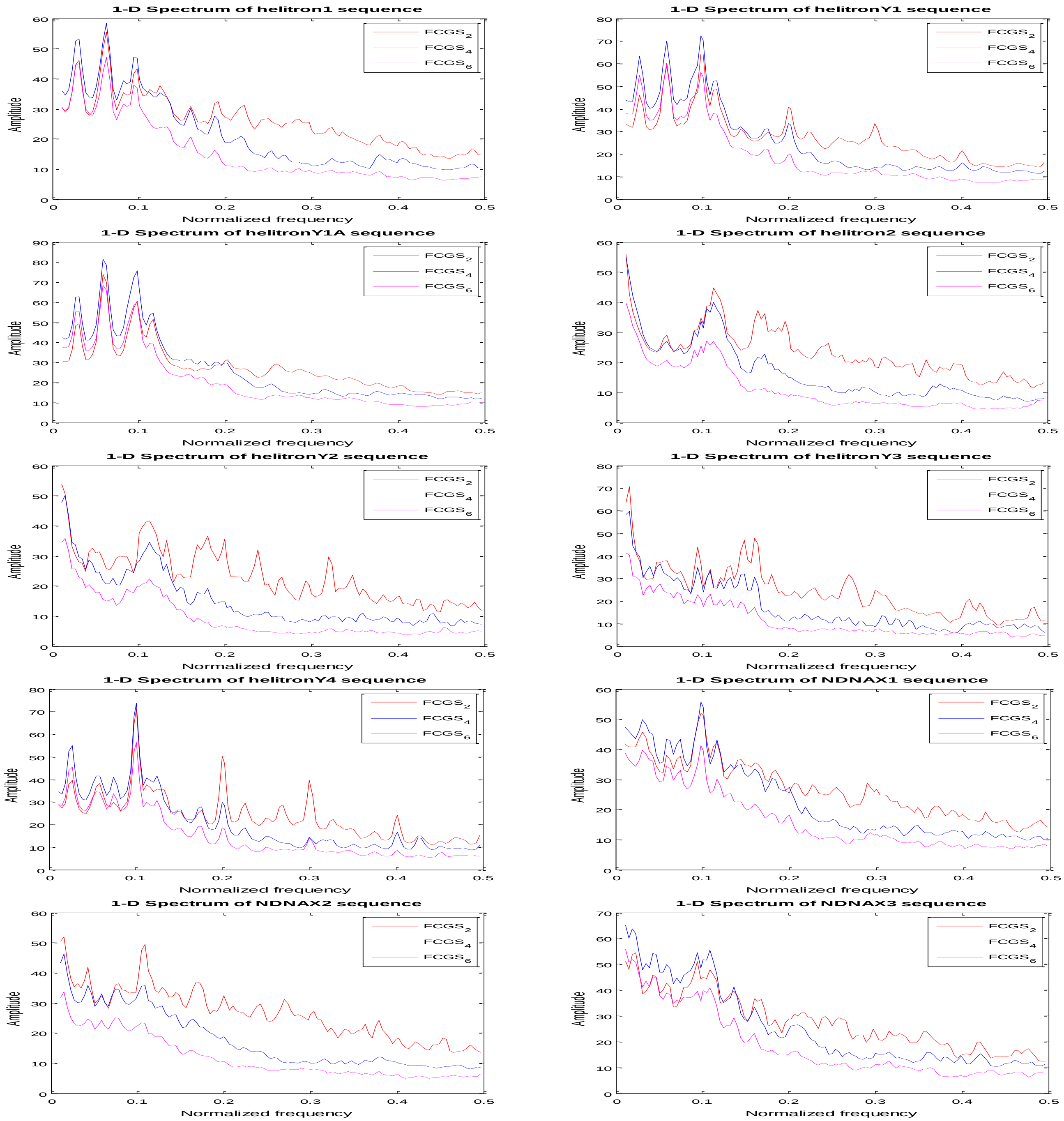
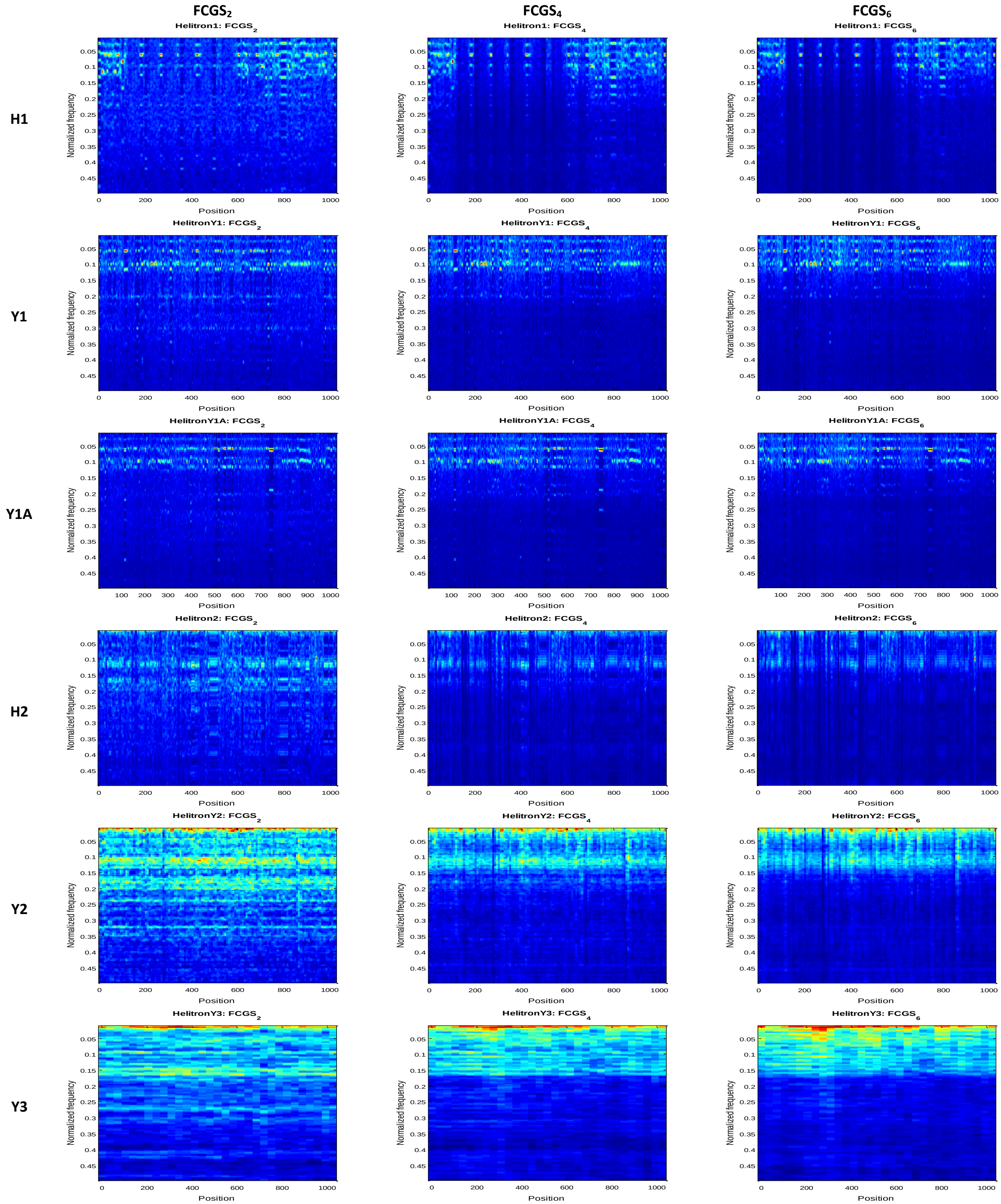


Fig. 5. 1-D Spectrum of each helitron's type in the C.elegans genome when coded by $FCGS_2$, $FCGS_4$ and $FCGS_6$.


FCGS2
FCGS4
FCGS6
H1
Y1
Y1A
H2
Y2
Y3
Helitron1: FCGS2
Helitron1: FCGS4
Helitron1: FCGS6
HelitronY1: FCGS2
HelitronY1: FCGS4
HelitronY1: FCGS6
HelitronY1A: FCGS2
HelitronY1A: FCGS4
HelitronY1A: FCGS6
Helitron2: FCGS2
Helitron2: FCGS4
Helitron2: FCGS6
HelitronY2: FCGS2
HelitronY2: FCGS4
HelitronY2: FCGS6
HelitronY3: FCGS2
HelitronY3: FCGS4
HelitronY3: FCGS6
Normalized frequency
Position

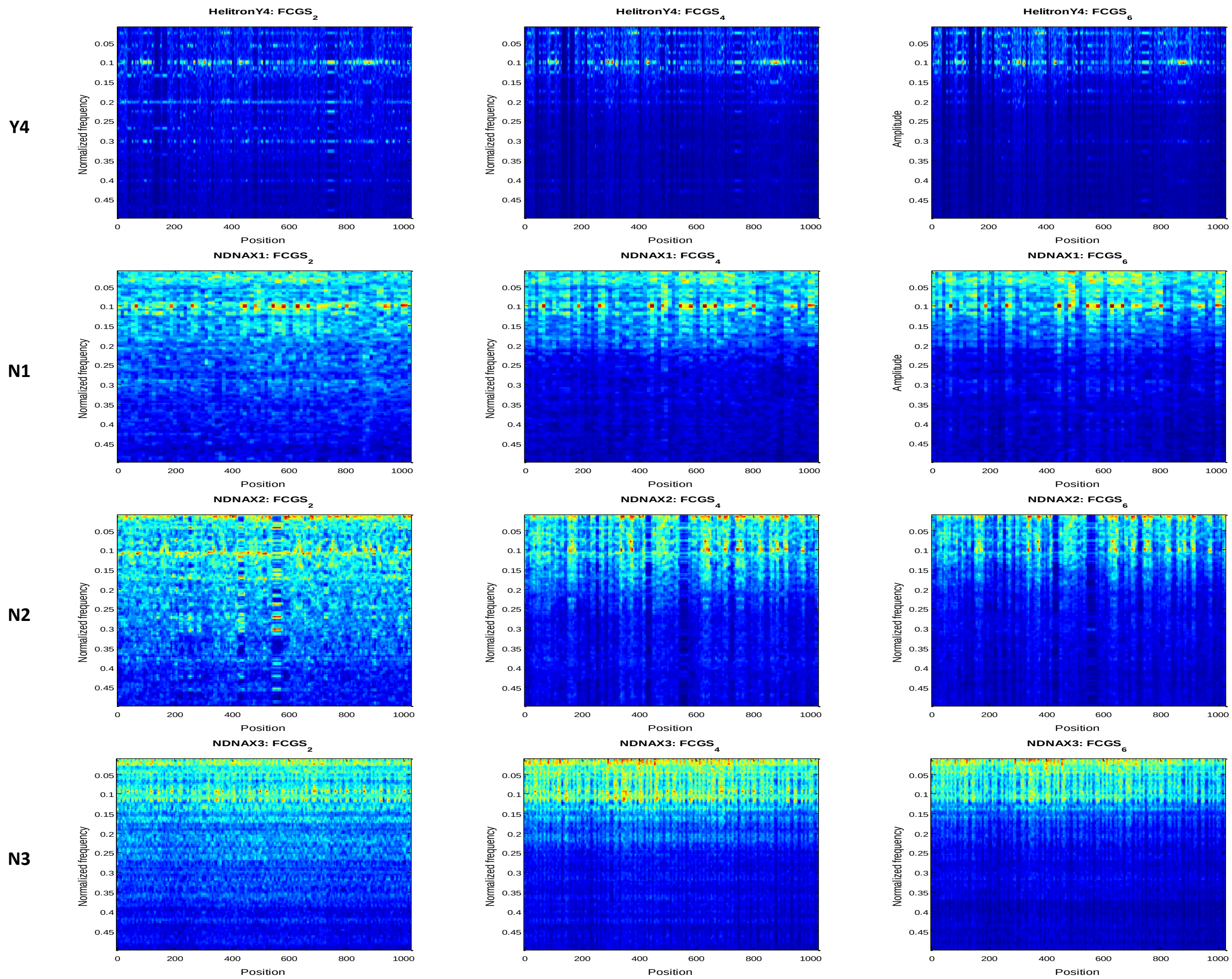


Fig. 6. 2-D Spectrogram of each helitron type in C.elegans when coded with $FCGS_{2,4,6}$.

REFERENCES

[1] N. V. Fedoroff, "Transposable elements as a molecular evolutionary force," Annals. New York Acad. Scien, vol.870, no. 1, pp. 251-264, 1999.

[2] M. G. Kidwell and Lisch, D, "Transposable elements as sources of variation in animals and plants", Proce. National. Acade. Scien, vol. 94, no. 15, pp. 7704-7711, 1997.

[3] D. J. Finnegan, "Transposable elements," Current opinion in genetics & development, vol. 2, no. 6,pp. 861-867, 1992.

[4] C. Feschotte and S. R. Wessler, "Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes," Proceedings of the National Academy of Sciences,vol. 98, no.16, pp.8923-8924, 2001.

[5] V. V. Kapitonov and J. Jurka, "Helitrons on a roll: eukaryotic rolling-circle transposons," TRENDS in Genetics, vol.23, no.10, pp. 521-529, 2007.

[6] E. Cocca, S. De Iorio and T. Capriglione, "Identification of a novel helitron transposon in the genome of Antarctic fish," Molecular phylogenetics and evolution, vol. 58, no. 3, pp. 439-446, 2011.

[7] J. Thomas, E. J. Pritham, "Helitrons, the eukaryotic rolling-circle transposable elements," Microbiology spectrum, vol. 3, pp. 4, 2015.

[8] A. Borgognone, R. Castanera, E. Muguerza, A. G. Pisabarro, and L. Ramírez, 'Somatic transposition and meiotically driven elimination of an active helitron family in Pleurotus ostreatus," Dna Research, vol. 24, no.2, pp. 103-115, 2017.

[9] J. Thomas, C. D. Phillips, R. J. Baker, E. J. Pritham, "Rolling-circle transposons catalyze genomic innovation in a mammalian lineage," Genome biology and evolution, vol. 6, no. 10, pp.2595-2610, 2014

[10] M. Morgante, S. Brunner, G. Pea, K. Fengler, Zuccolo, A., Rafalski A, "Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize". Nature genetics, vol. 37, no.9, pp. 997-1002, 2005.

[11] J. Lai, Y. Li, J.Messing, H.K. Dooner, "Gene movement by Helitron transposons contributes to the haplotype variability of maize," Proceedings of the National Academy of Sciences of the United States of America vol. 102, no.25, pp.9068-9073, 2005.

[12] P. S. Schnable, D. Ware, Fulton, J. C. Stein, F. Wei, S. Pasternak, and Minx, P , "The B73 maize genome: complexity, diversity, and dynamics," science vol. 326, no.5956, pp. 1112-1115, 2009.

[13] S. Tempel, J. Nicolas, A. El Amrani, I. Couée, "Model-based identification of Helitrons results in a new classification of their families in Arabidopsis thaliana," Gene, vol. 403 no. 1, pp.18-28, 2007

[14] V. V. Kapitonov, J. Jurka, "Rolling-circle transposons in eukaryotes," Proceedings of the National Academy of Sciences vol. 98, no. 15, pp.8714-8719 , 2001

[15] M. E Hood, "Repetitive DNA in the automictic fungus Microbotryum violaceum," Genetica, vol. 124, no. 1, pp. 1-10, 2005.

[16] R. T. Poulter, T. J.Goodwin, M. I. Butler, "Vertebrate helentrons and other novel Helitrons," Gene, vol. 313, pp.201-212, 2003.

[17] Zhou Q, et al. Helitron transposons on the sex chromosomes of the Platyfish Xiphophorus maculatus and their evolution in animal genomes. Zebrafish 3, 39–52 (2006)

[18] L. Yang, JL. Bennetzen, “Distribution, diversity, evolution, and survival of Helitrons in the maize genome,” Proc Natl Acad Sci USA vol. 106, no. 47, pp. 19922–19927, 2009.

[19] M. J. Han, Y. H.Shen , M. S. Xu and H. Y. Liang, “Identification and evolution of the silkworm helitrons and their contribution to transcripts,” DNA research, dst024 , 2013.

[20] C. Du, J. Caronna, L. He and H. K.Dooner, “Computational prediction and molecular confirmation of Helitron transposons in the maize genome” Bmc Genomics, vol. 9, no. 1, pp. 51, 2008

[21] L. Yang and J. L.Bennetzen, “Structure-based discovery and description of plant and animal Helitrons,” Proceedings of the National Academy of Sciences vol. 106, no.31, pp. 12832-12837, 2009.

[22] M.Sweredoski, L. DeRose-Wilson and B. S. A. Gaut, “comparative computational analysis of nonautonomous helitron elements between maize and rice,” BMC genomics vol. 9, no.1, pp. 467, 2008.

[23] W.Xiong, L. He, J. Lai, H. K. Dooner and C. Du, “HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes,” Proceedings of the National Academy of Sciences, vol. 111, no.28, pp.10263-10268, 2014.

[24] L.Yang and J. L. Bennetzen, “Distribution, diversity, evolution, and survival of Helitrons in the maize genome” Proceedings of the National Academy of Sciences, vol. 106, no.47, pp.19922-19927, 2009.

[25] C. Du, N. Fefelova, J. Caronna, L. He and H. K. Dooner, “The polychromatic Helitron landscape of the maize genome,” Proceedings of the National Academy of Sciences vol. 106, no.47, pp.19916-19921, 2009.

[26] A. Elloumi-Oueslati, N. Ellouze, and Z. Lachiri, “3D spectrum analysis of DNA sequence: application to Caenorhabditis elegans genome” In Bioinformatics and Bioengineering (BIBE 2007) pp. 864-871, 2007.

[27] A. Elloumi-Oueslati, I. Messaoudi, N. Ellouze and Z. Lachiri, “Spectral analysis of global behaviour of C. elegans chromosomes,” INTECH Open Access Publisher, 2012

[28] E.Pirogova, Q.Fang, M. Akay, and I. R. E. N. A Cosic, “Investigation of the structural and functional relationships of oncogene proteins,” Proceedings of the IEEE, vol. 90, no. 12, pp. 1859-1867, 2002.

[29] I. Messaoudi, A. Elloumi-Oueslati, and Z. Lachiri, “2D DNA Representations Generated Using a New Coding and the Time-Frequency Analysis,” Journal of Medical Imaging and Health Informatics, vol. 5, no.5, pp.1035-1044, 2015.

[30] R. Touati, I. Massaoudi, A, Elloumi-Oueslati, and Z. Lachiri, “Classification of Helitron’s Types in the C.elegans Genome based on Features Extracted from Wavelet Transform and SVM,” In Bioinformatics, pp. 127-134, 2018.

[31] I. Messaoudi, A. Elloumi-Oueslati, and Z. Lachiri, “Building specific signals from frequency chaos game and revealing periodicities using a smoothed Fourier analysis,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 11, no.5, pp. 863-877, 2014.

[32] A. A.Tsonis, J. B. Elsner, and P. A. Tsonis, “Periodicity in DNA coding sequences: implications in gene evolution,” Journal of Theoretical Biology, vol. 151, no. 3, pp. 323-331, 1991.

[33] L.Wang, and L. D. Stein, “Localizing triplet periodicity in DNA and Cdna sequences,” BMC Bioinformatics vol. 11, pp. 550, 2010.

[34] M. M. Hlaing, B. R.Wood, D McNaughton, and D. Ying, “Effect of drying methods on protein and DNA conformation changes in Lactobacillus rhamnosus GG by Fourier transform infrared spectroscopy,” Journal of Agricultural and Food Chemistry, 2017

[35] C.Yin, and J.Wang, “Periodic power spectrum with applications in detection of latent periodicities in DNA sequences,” Journal of mathematical biology, vol. 73, no. 5, pp. 1053-1079, 2016.

[36] J.Mena-Chalco, H.Carrer, Y.Zana, and Jr. R. M. Cesar, “Identification of protein coding regions using the modified Gabor-wavelet transform,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 5, no. 2, pp.198-207, 2008.

[37] P. P. Vaidyanathan, and B. J.Yoon, “Gene and exon prediction using allpass-based filters,” In Proc. IEEE Workshop on Gen. Sig. Proc and Stat, 2002.

[38] X. Zhang, Z.Shen, G. Zhang, Y. Shen, and M. Chen, “Short Exon Detection via Wavelet Transform Modulus Maxima,” PloS one, vol. 11, no. 9, pp. e0163088 , 2016.

[39] R.Touati, I. Massaoudi, A, Elloumi-Oueslati, N. ellouze and Z. Lachiri, Nucleosome location method based on morlet wavelet analysis scalograms investigation. In Advanced Technologies for Signal and Image Processing (ATSIP), pp. 307-312, 2016.

[40] A. Elloumi-Oueslati, I. Messaoudi, Z. Lachiri and N. Ellouze, “A new way to visualize DNA’s base succession: the Caenorhabditis elegans chromosome landscapes,” Medical & biological engineering & computing, vol. 53, no. 11, pp. 1165-1176, 2015.

[41] D. Anastassiou,“Genomic signal processing,” IEEE signal processing magazine, vol. 8, no. 4, pp.8-20, 2001.

[42] A. Fukushima, T. Ikemura, M. Kinouchi, and T. Oshima, “Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis,” Gene, 300(1), 203-211 (2002).

Rainfall Prediction in Lahore City using Data Mining Techniques

Shabib Aftab, Munir Ahmad, Noureen Hameed, Muhammad Salman Bashir, Iftikhar Ali, Zahid Nawaz
Department of Computer Science
Virtual University of Pakistan
Lahore, Pakistan

***Abstract*—Rainfall prediction has extreme significance in countless aspects and scopes. It can be very helpful to reduce the effects of sudden and extreme rainfall by taking effective security measures in advance. Due to climate variations, an accurate rainfall prediction has become more complex than before. Data mining techniques can predict the rainfall through extracting the hidden patterns among weather attributes of past data. This research contributes by exploring the use of various data mining techniques for rainfall prediction in Lahore city. Techniques include: Support Vector Machine (SVM), Naïve Bayes (NB), k Nearest Neighbor (kNN), Decision Tree (J48) and Multilayer Perceptron (MLP). The dataset is obtained from a weather forecasting website and consists of several atmospheric attributes. For effective prediction, pre-processing technique is used which consists of cleaning and normalization processes. Performance of used data mining techniques is analyzed in terms of precision, recall and f-measure with various ratios of training and test data.**

Keywords—Rainfall prediction; data mining; classification techniques

I. Introduction

Time series data mining is one of the hot research topics in the domain of knowledge discovery [19]. The data with time series approach is collected over a specific period of time such as daily, weekly, monthly, quarterly or yearly [13]. This data can be used for predictions in different domains such as finance, stock market and climate change etc. Data mining techniques are used to extract the hidden knowledge from time series data for future use [13], [17], [25], [28]. Weather prediction with time series data is beneficial but quite challenging task [16], [27], [29]. It comes with an array of complexities which needs to be tackled for optimal results [18]. The statistical weather data has a wide variety of fields which are called features such as humidity, pressure, wind speed, pollutants, concentrations etc. Data mining techniques can predict the weather on the basis of hidden patterns among these features [27], [29]. Rainfall prediction is an important aspect of climate forecasting. Accurate and timely rainfall prediction is crucial for the planning and management of water resources, flood warnings, construction activities and flight operations etc. [14], [15]. This study used 5 data mining techniques for rainfall prediction in Lahore, capital of Punjab province, Pakistan. In Lahore, development and construction activities are increasing exponently, so timely rainfall prediction is crucial for better assessment of future requirements and planning. The used data mining techniques include: Support Vector Machine, Naïve Bayes, k Nearest Neighbor, Decision Tree and Multilayer Perceptron. These algorithms belong to supervised data mining class where pre-classified data is required first for training purpose. During training, these algorithms make rules of classification for input dataset (test data) [20]-[25], [30]. In this research, dataset is obtained from weather forecasting website [10] from December 1, 2005 to November 31, 2017 (12 years), which contains several weather related attributes such as Temperature, Atmospheric pressure, Relative humidity etc. For rainfall prediction, a classification framework is used in which the dataset gone through cleaning and normalization process before classification. Cleaning is performed to deal with the missing values and the purpose of normalization is to keep the attribute values in a certain limits. These pre-processing activities are crucial for the smooth classification process as well as for good results [9], [12]. Prediction performance of used data mining techniques is evaluated in terms of precision, recall and f measure, which are the important metrics of information retrieval. Finally the results are shown in tables and graphs.

Further organization of this paper is as follows. Section II describes the related work. Section III discusses the materials and methods used in this research. Section IV presents results and discussion. Section V finally concludes the paper.

II. Related Work

Many researchers have been working to achieve high accuracy in rainfall prediction using data mining techniques; some of the selected studies are discussed here. Researchers in [1] performed a comparative analysis of multiple classifiers for rainfall prediction in Malaysia. Classifiers include Naïve Bayes, Support Vector Machine, Decision Tree, Neural Network and Random Forest. Dataset was obtained from multiple stations of Selangor, Malaysia. Pre-processing tasks were applied before classification to deal with the noise and missing values. According to results, Random Forest performed better as with small training data it correctly classified large amount of instances. In [2], researchers presented Clusterwise Linear Regression (CLR) method, which is the combination of clustering and regression techniques. The proposed technique is used to predict monthly rainfall in Victoria, Australia, by using input data of 8 geographically diverse weather stations. To analyze the performance of proposed CLR, results were compared with other techniques such as: CLR using the maximum likelihood framework by the expectation-maximization algorithm, multiple linear regression, artificial neural networks and the support vector machines. It was observed that proposed algorithm performed better than other methods in most of the locations. Researchers in [3] performed a comparative analysis of a modern technique named Markov Chain (extended with rainfall prediction) and six other well-known machine learning techniques: Genetic Programming, Support Vector Regression, Radial Basis Neural Networks, M5 Rules, M5 Model trees, and k-Nearest Neighbors. For prediction, rainfall time series data of 42 cities with different climatic features is used. The results reflected that machine learning techniques have the capacity to perform better than Markov Chain technique moreover this study has also pointed out the correlations between different climatic attributes and predictive accuracy. In [4], Artificial Neural Network (ANN) was used to develop one-month and two-month ahead forecasting models for rainfall prediction using monthly rainfall data of 141 years from various weather stations in the North India. In these models, Feed Forward Neural Network (FFNN) using Back Propagation algorithm and Levenberg-Marquardt training function was used. The performance of both the models was analyzed on the basis of Regression Analysis, Mean Square Error (MSE) and Magnitude of Relative Error (MRE). ANN showed optimistic results for both the models and found that one month ahead forecasting model performed better than two months model. Researchers in [5] proposed an algorithm which combined data mining and statistical techniques. The likely predictors with highest confidence level, based on association rules were selected. Those predictors were derived from local and global conditions. From local conditions: sea level pressure, wind speed, and maximum &minimum temperatures were recorded. On the other hand from global condition, southern oscillation and Indian Ocean dipole conditions were taken. The algorithm predicted the rainfall in five categories: Flood, Excess, Normal, Deficit and Drought. Researchers in [6] presented Wavelet Neural Network model (WNN) for rainfall prediction which is the combination of wavelet technique and Artificial Neural Network (ANN). Proposed WNN and ANN, both models were applied on monthly rainfall data of Darjeelin grain gauge station, west Bengal, India. Statistical methods were used to analyze the performance of both techniques and it was observed that WNN performed much better than ANN model. In [7], researchers implemented a rainfall forecasting model using Focused Time-Delay Neural Networks (FTDNN). The parameters for neural networks were taken from several experiments to perform prediction with one step ahead. For prediction, the daily rainfall data was obtained from Malaysia Meteorological Department (MMD) and then converted to monthly, biannually, quarterly and yearly basis. Models were trained and tested on each dataset and corresponding accuracies were evaluated using Mean Absolute Percent Error. According to results, most accurate forecasts were made with yearly rainfall dataset. The authors have pointed out that more parameters such as temperature, humidity and sunshine data should be incorporated into the neural networks to make the performance more accurate. Researchers in [8] presented a methodology to predict maximum temperature in the day, which followed the Support Vector Regression approach. Proposed technique performed prediction on the basis of several features, obtained from different measuring stations in Europe. Weather related features included temperature, precipitation, relative humidity, air pressure, specifically synoptic situation of the day and monthly cycle. The proposed technique performed well when compared with other neural networks, multi-layer perceptron and an extreme learning machine.

III. Materials and Methods

This research aims to analyze the performance of data mining techniques on rainfall prediction in Lahore city using a classification framework (Fig. 1). Dataset used in this research consists of several attributes along with the known output class. Output class is one which is going to be predicted on the basis of other available attributes. The reason of including the output class in dataset among other features is to analyze the performance and accuracy of data mining techniques [20], [24]. The output result after processing is compared with the known class and performance is measured in terms of precision, recall and f measure [1], [20], [21], [24], [26].

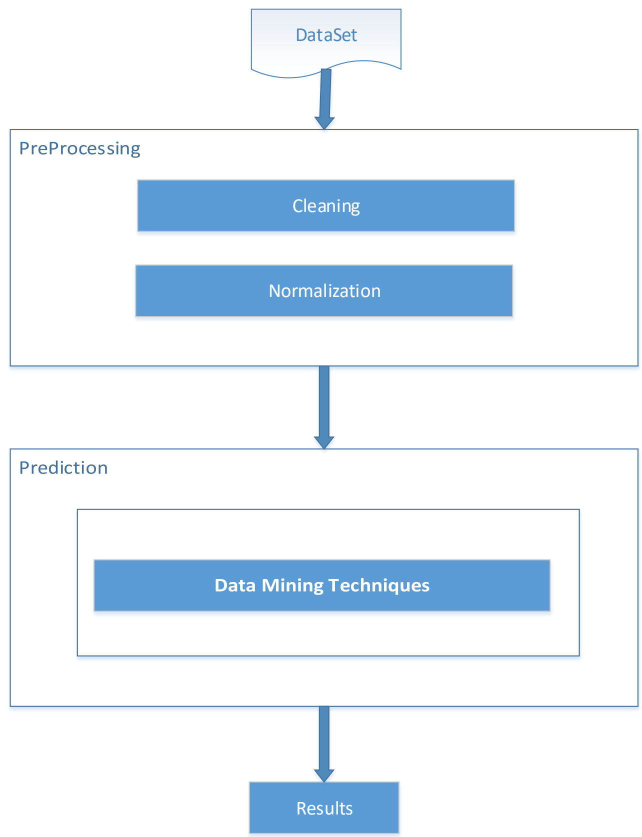


Fig. 1. Classification framework.

Weka [22], [23] is used in this study for classification and performance analysis. It is one of the extensively used data mining softwares. Weka is developed in Java language at the University of Waikato, New Zealand. It is famous and widely accepted tool among students and researchers due to its easy to use GUI interface, portability and General Public License.

The classification framework used in this research consists of four stages: Selection of appropriate dataset, Preprocessing, Prediction and Simulation of results. The input dataset for rainfall prediction is obtained from weather forecasting website [10] and consists of several atmospheric attributes. Name, type and measurement unit of selected attributes are given in Table I.

TABLE I. DATASET ATTRIBUTES

Attribute Name	Attribute Type	Measurement
Temperature	Continuous	Degrees Celsius
Atmospheric Pressure (weather station)	Continuous	Millimeters of Mercury
Atmospheric Pressure (sea level)	Continuous	Millimeters of Mercury
Pressure Tendency	Continuous	Millimeters of Mercury
Relative Humidity	Continuous	%
Mean Wind Speed	Continuous	Meters per Second
Minimum Temperature	Continuous	Degrees Celsius
Maximum Temperature	Continuous	Degrees Celsius
Visibility	Continuous	Km
Dew Point Temperature	Continuous	Degrees Celsius

Dataset contained missing values as shown in Table II. The incomplete data can affect the accuracy of results as the attribute which has the missing value cannot fully participate in prediction process. Beside the missing values, dataset also contained noise where value resides below or exceeds from a certain limits. For effective data mining results it is recommended to keep the values in a certain limits [1], [11]. Pre-processing of input data is a crucial stage in classification framework which ensures the high accuracy of mining results. This stage consists of two activities: cleaning and normalization. Cleaning process deals with the missing values by using average mechanism. In this mechanism sum of all the instances of selected attribute is divided by the number of samples. On the other hand normalization process deals with the noise by limiting the values within a specific interval. Such interval can effectively facilitate the prediction process where the values will be mapped onto a particular range. In this research the normalization process is performed in Weka. Prediction is the final stage of classification framework where data mining algorithms perform classification by exploring the hidden patterns.

TABLE II. VALID RECORDS AND MISSING VALUES

Attribute Name	Valid Record	Missing Values
Temperature	25846.0	73
Atmospheric Pressure (weather station)	23689.0	2230
Atmospheric Pressure (sea level)	23714.0	2205
Pressure Tendency	11320.0	14599
Relative Humidity	25790.0	129
Mean Wind Speed	25890.0	29
Minimum Temperature	2415.0	23504
Maximum Temperature	4174.0	21745
Visibility	25829.0	90
Dew Point Temperature	25865.0	54

Performance of any supervised machine learning technique can be analyzed by comparing the output result with known class (pre-classified data). Performance evaluation of used data mining techniques is performed with 10 proportions (10:90-90:10) of training data and test data. For comparative analysis, three evaluation parameters of information retrieval are used: Precision, Recall and F Measure.

The aim of Precision is to evaluate the True Positive (TP) entities with respect to False Positive (FP) entities. It can be calculated as follows:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{1}$$

TP is used for the entities, which are correctly classified, and FP is for those entities, which are wrongly classified.

The aim of recall is to evaluate the True Positive entities with respect to the (FN) False Negative entities, which are not classified at all. It can be calculated as follows:

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{2}$$

There may be a point where performance evaluation will not be possible with precision and recall, for example if one mining algorithm has higher precision but lower recall than another algorithm so the question arises that which algorithm is better. Solution to this issue is to use F-measure, which provides the average of precision and recall. F-measure can be computed as bellow:

$$\text{F-measure} = \frac{\text{Precision} * \text{Recall} * 2}{(\text{Precision} + \text{Recall})} \tag{3}$$

IV. RESULTS AND DISCUSSION

With SVM the results are almost same (Table III, Fig. 2) in all three accuracy parameters (Precision, recall and f-measure). Results for no-rain class with first seven proportions from 10:90 to 70:30 in precision, recall and f-measure are 0.941, 1, and 0.955 respectively however minor improvement were seen when proportions 80:20 and 90:10 were used. The notable point is that the result for rain class with all proportions in all accuracy parameters is 0, which means that this technique could not classify a single instance correctly for rain class even with 90:20 ratios.

TABLE III. SVM RESULTS

Proportion	Class	Precision	Recall	F-Measure
10:90	No Rain	0.914	1	0.955
	Rain	0	0	0
20:80	No Rain	0.914	1	0.955
	Rain	0	0	0
30:70	No Rain	0.914	1	0.955
	Rain	0	0	0
40:60	No Rain	0.914	1	0.955
	Rain	0	0	0
50:50	No Rain	0.914	1	0.955
	Rain	0	0	0
60:40	No Rain	0.914	1	0.955
	Rain	0	0	0
70:30	No Rain	0.914	1	0.955
	Rain	0	0	0
80:20	No Rain	0.919	1	0.958
	Rain	0	0	0
90:10	No Rain	0.919	1	0.958
	Rain	0	0	0

The results with Naive Bayes are shown in Table IV and Fig. 3. It can be seen that with no-rain class the 10:90

performed better in precision, 30:70 and 40:60 in recall and 90:10 in f-measure. With rain class 40:60 performed better in precision, 10:90 in recall and 50:50 in f-measure.

TABLE IV. NB RESULTS

Proportion	Class	Precision	Recall	F-Measure
10:90	No Rain	0.941	0.85	0.893
	Rain	0.215	0.435	0.287
20:80	No Rain	0.932	0.955	0.943
	Rain	0.353	0.258	0.298
30:70	No Rain	0.931	0.957	0.944
	Rain	0.356	0.254	0.296
40:60	No Rain	0.933	0.957	0.945
	Rain	0.366	0.264	0.307
50:50	No Rain	0.934	0.953	0.943
	Rain	0.362	0.283	0.318
60:40	No Rain	0.935	0.943	0.939
	Rain	0.331	0.299	0.314
70:30	No Rain	0.933	0.942	0.938
	Rain	0.316	0.282	0.298
80:20	No Rain	0.936	0.954	0.945
	Rain	0.34	0.268	0.3
90:10	No Rain	0.938	0.954	0.946
	Rain	0.355	0.286	0.317

The results with KNN are shown in Table V and Fig. 4. With no-rain class the 80:20 and 90:10 performed better in precision, 10:90 in recall and 90:10 in f-measure. With rain class 10:90 performed better in precision, 80:20 in recall and 90:10 in recall.

TABLE V. KNN RESULTS

Proportion	Class	Precision	Recall	F-Measure
10:90	No Rain	0.935	0.95	0.942
	Rain	0.358	0.295	0.324
20:80	No Rain	0.935	0.946	0.941
	Rain	0.35	0.306	0.326
30:70	No Rain	0.936	0.948	0.942
	Rain	0.365	0.317	0.339
40:60	No Rain	0.937	0.948	0.943
	Rain	0.368	0.321	0.343
50:50	No Rain	0.936	0.948	0.942
	Rain	0.363	0.314	0.337
60:40	No Rain	0.937	0.948	0.942
	Rain	0.364	0.317	0.339
70:30	No Rain	0.936	0.948	0.942
	Rain	0.361	0.313	0.335
80:20	No Rain	0.942	0.943	0.943
	Rain	0.35	0.346	0.348
90:10	No Rain	0.942	0.949	0.946
	Rain	0.373	0.343	0.357

The results with Decision Tree (J48) are shown in Table VI and Fig. 5. With no-rain class the 80:20 performed better in precision, 20:80 in recall and 90:10 in f-measure. With rain class 20:80 performed better in precision, 80:20 in recall and 80-20 in recall.

TABLE VI. DECISION TREE RESULTS

Proportion	Class	Precision	Recall	F-Measure
10:90	No Rain	0.936	0.976	0.955
	Rain	0.527	0.286	0.371
20:80	No Rain	0.926	0.993	0.958
	Rain	0.673	0.154	0.25
30:70	No Rain	0.931	0.988	0.959
	Rain	0.636	0.224	0.331
40:60	No Rain	0.933	0.98	0.956
	Rain	0.549	0.255	0.348
50:50	No Rain	0.932	0.985	0.958
	Rain	0.599	0.244	0.347
60:40	No Rain	0.934	0.985	0.959
	Rain	0.626	0.262	0.37
70:30	No Rain	0.934	0.984	0.959
	Rain	0.614	0.266	0.371
80:20	No Rain	0.939	0.981	0.96
	Rain	0.568	0.287	0.381
90:10	No Rain	0.938	0.985	0.961
	Rain	0.607	0.257	0.361

The results with MLP are shown in Table VII and Fig. 6. It can be seen that with no-rain class 80:20 performed better in precision, 90:10 in recall and 90:10 in f-measure. With rain class 90:10 performed better in precision, 60:40 in recall and 80:20 in f- measure.

TABLE VII. MLP RESULTS

Proportion	Class	Precision	Recall	F-Measure
10:90	No Rain	0.927	0.993	0.959
	Rain	0.709	0.173	0.278
20:80	No Rain	0.935	0.983	0.959
	Rain	0.61	0.28	0.384
30:70	No Rain	0.931	0.99	0.96
	Rain	0.681	0.221	0.334
40:60	No Rain	0.937	0.979	0.958
	Rain	0.579	0.303	0.397
50:50	No Rain	0.931	0.992	0.96
	Rain	0.724	0.22	0.337
60:40	No Rain	0.937	0.98	0.958
	Rain	0.587	0.304	0.4
70:30	No Rain	0.93	0.99	0.959
	Rain	0.67	0.212	0.322
80:20	No Rain	0.94	0.984	0.962
	Rain	0.619	0.296	0.401
90:10	No Rain	0.935	0.997	0.965
	Rain	0.846	0.21	0.336

A. *Critical Analysis*

Data mining techniques used in this study showed good results for no-rain class in all accuracy measures (Precision, recall and f-measure) however for rain class these techniques did not perform well and results are not accurate enough. F-measure is a high-quality accuracy measure as it provides the average of precision and recall. Table VIII is arranged according to highest f-measure score in each mining technique along with its class and proportion. There could be several reasons for the lower results with rain class such as, missing values as mean value cannot reflect the actual one, absence of one or more important climatic attributes and the most important is the lower rainfall rate in the city. Due to climate variations, rainfall rate in most of the locations is not, what it used to be. Moreover the dataset does not include the rainfall quantity/measure, instead it only includes the rainfall polarity (yes/no). So the data is reflecting the number of times it rained but not how much. There might be only one rainy day in a week but that might have been catastrophic with extreme rainfall. With overall lower rainfall rate (number of times it rained), less patterns were provided to classification algorithms which resulted in poor performance with rain class whereas on the other hand in no-rain class, more patterns were available for training of classification techniques, resulted in high accuracy.

TABLE VIII. DM TECHNIQUES WITH HIGHEST F-MEASURE

DM Algorithm	Class	Proportion	F-Measure
SVM	No Rain	80:20	0.958
	Rain	90:10	0
NB	No Rain	90:10	0.946
	Rain	50:50	0.318
KNN	No Rain	90:10	0.946
	Rain	90:10	0.357
Decision Tree	No Rain	90:10	0.961
	Rain	80:20	0.381
MLP	No Rain	90:10	0.965
	Rain	80:20	0.401

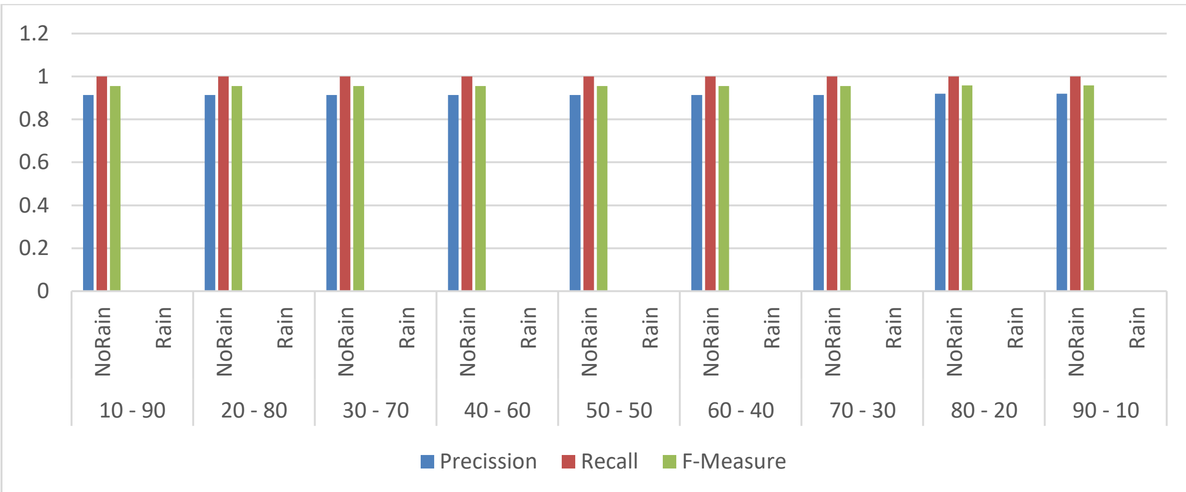


Fig. 2. SVM results.

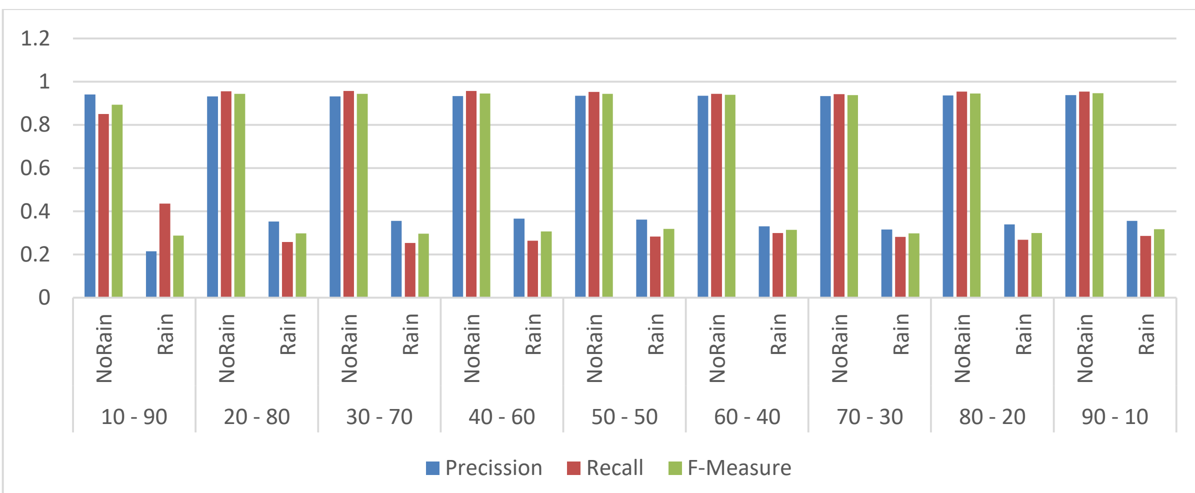


Fig. 3. NB results.

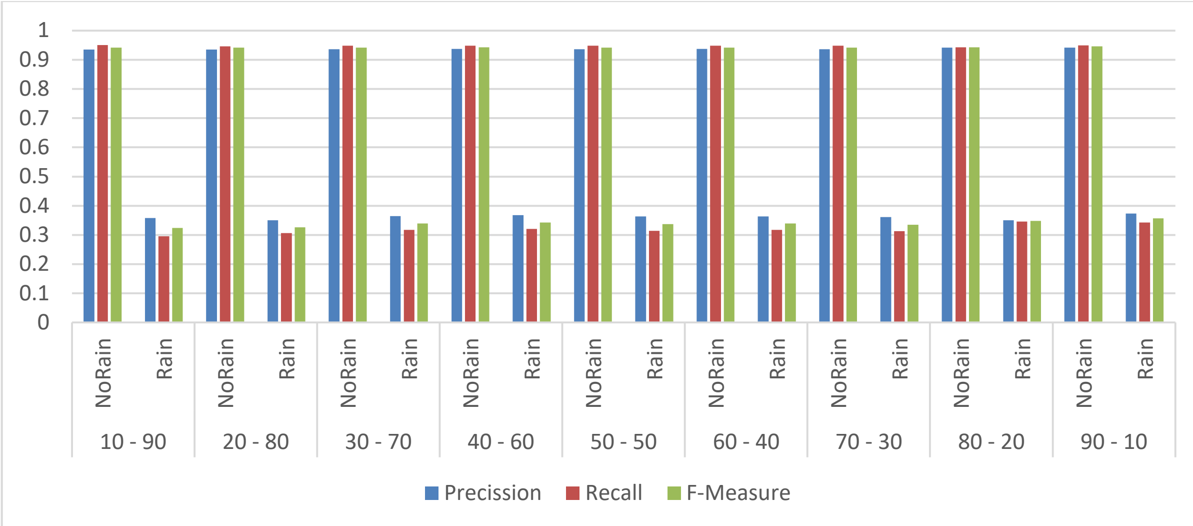


Fig. 4. KNN results.

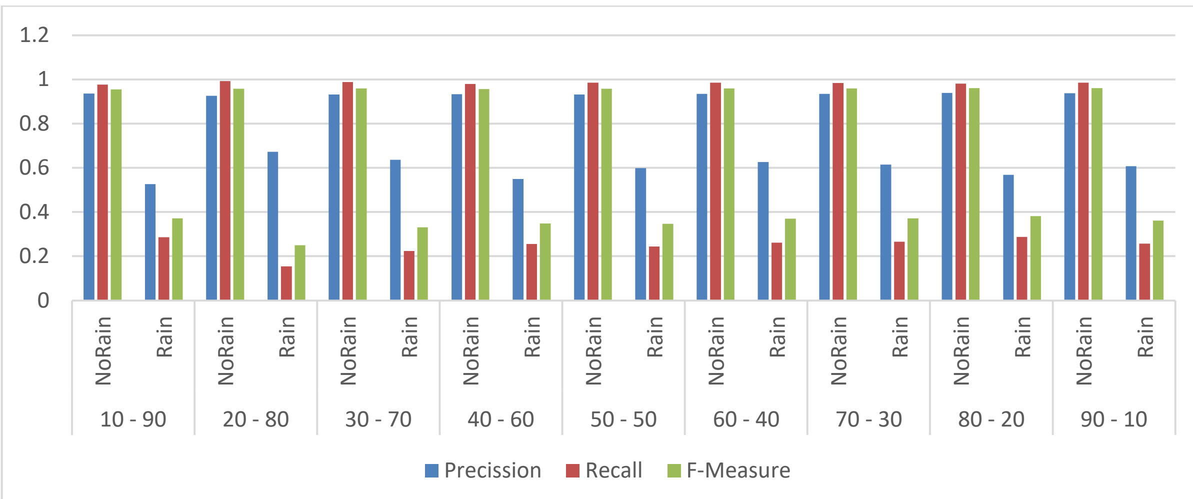


Fig. 5. Decision-tree results.

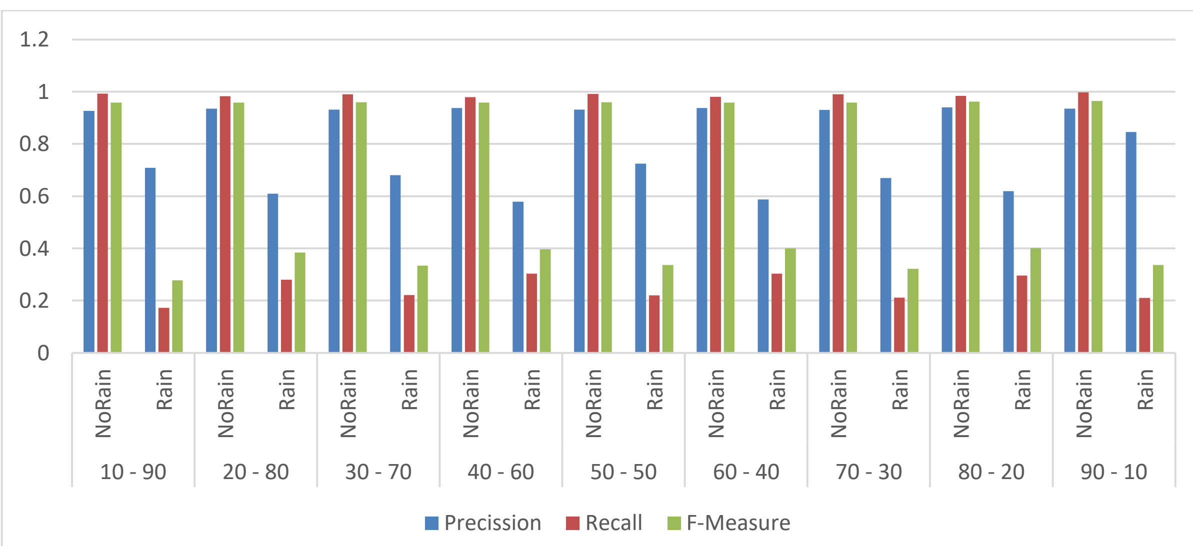


Fig. 6. MLP results.

V. Conclusion and Future Work

This research performed rainfall prediction in Lahore city using five data mining techniques: Support Vector Machine, Naïve Bayes, k Nearest Neighbor, Decision Tree and Multilayer Perceptron. 12 years of past weather data from December 1, 2005 to November 31, 2017, is used for prediction in this research. Performance analysis of used data mining techniques is performed using three accuracy measures: precision, recall and f-measure and results are presented in tables and graphs. For effective prediction, a classification framework is used in which the input data went through a pre-processing stage and got cleaned and normalized before classification process. To analyze the performance dependency of classification techniques on training data, ten ratios of training and test data (training data: test data) are used from 10:90 to 90:10. According to results, used classification techniques performed well for no-rain class however for rain class, the techniques did not perform well. The reasons behind the lower accuracy in rain class may include: missing values, absence of important climatic attributes in dataset and overall lower rate of rainfall in the city. It is suggested for future work that further predictions should be performed by exploring more classification techniques and climatic attributes on different weather data.

References

[1] S. Zainudin, D. S. Jasim, and A. A. Bakar, “Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction,” Int. J. Adv. Sci. Eng. Inf. Technol., vol. 6, no. 6, pp. 1148–1153, 2016.

[2] A. M. Bagirov, A. Mahmood, and A. Barton, “Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach,” Atmos. Res., vol. 188, pp. 20–29, 2017.

[3] S. Cramer, M. Kampouridis, A. A. Freitas, and A. K. Alexandridis, “An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives,” Expert Syst. Appl., vol. 85, pp. 169–181, 2017.

[4] N. Mishra, H. K. Soni, S. Sharma, and A. K. Upadhyay, “Development and Analysis of Artificial Neural Network Models for Rainfall Prediction by Using Time-Series Data,” Int. J. Intell. Syst. Appl., vol. 10, no. 1, pp. 16–23, 2018.

[5] H. Vathsala and S. G. Koolagudi, “Prediction model for peninsular Indian summer monsoon rainfall using data mining and statistical approaches,” Comput. Geosci., vol. 98, pp. 55–63, 2017.

[6] R. Venkata Ramana, B. Krishna, S. R. Kumar, and N. G. Pandey, “Monthly Rainfall Prediction Using Wavelet Neural Network Analysis,” Water Resour. Manag., vol. 27, no. 10, pp. 3697–3711, 2013.

[7] K. K. Htike and O. O. Khalifa, “Delay Neural Networks,” Int. Conf. Comput. Commun. Eng., no. May, pp. 11–13, 2010.

[8] A. Paniagua-Tineo, S. Salcedo-Sanz, C. Casanova-Mateo, E. G. Ortiz-García, M. A. Cony, and E. Hernández-Martín, “Prediction of daily maximum temperature using a support vector regression algorithm,” Renew. Energy, vol. 36, no. 11, pp. 3054–3060, 2011.

[9] C. Sivapragasam, S. Liong, and M. Pasha, “Rainfall and runoff forecasting with SSA-SVM approach,” J. Hydroinformatics, no. April 2016, pp. 141–152, 2001.

[10] “http://ru8.rp5.ru/Weather_archive_in_Lahore.” [Online]. Available: http://ru8.rp5.ru/Weather_archive_in_Lahore.

[11] S. S. Monira, Z. M. Faisal, and H. Hirose, “Comparison of artificially intelligent methods in short term rainfall forecast,” Proc. 2010 13th Int. Conf. Comput. Inf. Technol. ICCIT 2010, no. Iccit, pp. 39–44, 2010.

[12] D. Isa, L. H. Lee, V. P. Kallimani, and R. RajKumar, “Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine,” IEEE Trans. Knowl. Data Eng., vol. 20, no. 9, pp. 1264–1272, 2008.

[13] N. Mishra, H. K. Soni, S. Sharma, and A. K. Upadhyay, “A Comprehensive Survey of Data Mining Techniques on Time Series Data for Rainfall Prediction,” vol. 11, no. 2, pp. 168–184, 2017.

[14] K. W. Chau and C. L. Wu, “A hybrid model coupled with singular spectrum analysis for daily rainfall prediction,” J. Hydroinformatics, vol. 12, no. 4, p. 458, 2010.

[15] J. Wu, J. Long, and M. Liu, “Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm,” Neurocomputing, vol. 148, pp. 136–142, 2015.

[16] W. C.L. and K.-W. Chau, “Prediction of Rainfall Time Series Using Modular Soft Computing Methods,” Eng. Appl. Artif. Intell., vol. 26, no. 852, pp. 1–37, 2012.

[17] D. Gupta and U. Ghose, “A Comparative Study of Classification Algorithms for Risk Prediction in Pregnancy,” pp. 0–5, 2015.

[18] K. Abhishek, A. Kumar, R. Ranjan, and S. Kumar, “A rainfall prediction model using artificial neural network,” 2012 IEEE Control Syst. Grad. Res. Colloq., no. Icsgrc, pp. 82–87, 2012.

[19] M. A. Nayak and S. Ghosh, “Prediction of extreme rainfall event using weather pattern recognition and support vector machine classifier,” Theor. Appl. Climatol., vol. 114, no. 3–4, pp. 583–603, 2013.

[20] M. Ahmad and S. Aftab, “Analyzing the Performance of SVM for Polarity Detection with Different Datasets,” Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 10, pp. 29–36, 2017.

[21] A. Sharma and S. Dey, “A boosted SVM based sentiment analysis approach for online opinionated text,” Proc. 2013 Res. Adapt. Converg. Syst. - RACS ’13, pp. 28–34, 2013.

[22] G. Holmes, A. Donkin, and I. H. Witten, “WEKA: A Machine Learning Workbench.” 1994.

[23] “http://www.cs.waikato.ac.nz~mlweka.” [Online]. Available: httpwww.cs.waikato.ac.nz/mlweka.

[24] M. Ahmad, S. Aftab, and I. Ali, “Sentiment Analysis of Tweets using SVM,” Int. J. Comput. Appl., vol. 177, no. 5, pp. 25–29, 2017.

[25] M. Ahmad, S. Aftab, and S. S. Muhammad, “Machine Learning Techniques for Sentiment Analysis: A Review,” Int. J. Multidiscip. Sci. Eng., vol. 8, no. 3, p. 27, 2017.

[26] M. Ahmad, S. Aftab, M. S. Bashir, and N. Hameed, “Sentiment Analysis using SVM : A Systematic Literature Review,” vol. 9, no. 2, pp. 182–188, 2018.

[27] Y. Radhika and M. Shashi, "Atmospheric Temperature Prediction using Support Vector Machines," Int. J. Comput. Theory Eng., vol. 1, no. 1, pp. 55-58, 2009

[28] K. Lu and L. Wang, "A Novel Nonlinear Combination Model Based on Support Vector Machine for Rainfall Prediction," 2011 Fourth Int. Jt. Conf. Comput. Sci. Optim., pp. 1343-1346, 2011.

[29] V. B. Nikam and B. B. Meshram, "Modeling Rainfall Prediction Using Data Mining Method: A Bayesian Approach," 2013 Fifth Int. Conf. Comput. Intell. Model. Simul., pp. 132-136, 2013.

[30] M. Ahmad, S. Aftab, I. Ali, and N. Hameed, “Hybrid Tools and Techniques for Sentiment Analysis: A Review,” Int. J. Multidiscip. Sci. Eng., vol. 8, no. 3, 2017.

Concentrated Solar Power Site Suitability using GIS-MCDM Technique taken UAE as a Case Study

Mohammad Basheer Alqaderi[1]
Dept. of Engineering projects management, Isra University
Amman, Jordan

Walid Emar[2]
Dept. of Electrical Engineering
Isra University
Amman, Jordan

Omar A. Saraereh[3]
Dept. of Electrical Engineering
Hashemite University
Zarqa, Jordan

***Abstract*—In recent years, countries have begun to reduce the consumption of fossil fuels and replace them with renewable energy resources in order to mitigate the effects of fossil fuels on the environment and save money besides increasing energy security. This paper has investigated the suitability map for the large-scale projects in concentrated solar power (CSP) in United Arab Emirates (UAE) using GIS data and multi-criteria decision making technique (MCDM). The suitability map is composed of multi-maps (layers) of solar irradiation [Direct Normal irradiance (DNI) component], land slope, protected areas, land use, proximity to water bodies, power grid and the roads. Then, Analytic Hierarchy Process (AHP) method is applied to identify the weights of ranking criteria. The paper has highlighted the most-suitable location as well as non-suitable location among UAE to install CSP projects. The study's results proved that UAE have multi-hotspot locations can be exploited for CSP projects.**

Keywords—Analytic hierarchy process (AHP); concentrated solar power (CSP); multi-criteria decision making (MCDM); direct normal irradiance (DNI); United Arab Emirates (UAE); hot-spot locations

I. Introduction

Middle East has great potential for Renewable Energy (RE), but right now RE projects form only 1% of energy mix within the region [1]. Abu Dhabi's renewable target, one of UAE emirates, is to produce 7% of its electricity from renewable sources by 2020. Despite the huge ongoing projects in Abu Dhabi, a glance at the statistics shows that the emirate's installed generation capacity from renewable sources in 2013 was 111 megawatts which is 1% of total energy mix [3].

United Arab Emirates set in its main energy plan to generate electricity from CSP and PV systems, however selecting the suitable location and areas to install large-scale CSP and PV projects is not easy but represents a primary task that must be done early in the stage of planning. The selecting of suitable locations that will host CSP and PV plants depend not only on the amount of solar radiation received but also depends on several conflicting factors which shall be taken into decision maker's consideration like technological factors (grid connectivity), environmental factors (protected areas and water bodies), economic, social factors (population density), topographic factors (water availability, road proximity and grid proximity) and elevation [2].

To solve such a conflicting problem's factors, this research will use the geographical information system (GIS maps) and multi-criteria decision making (MCDM) technique to produce suitability map of the entire UAE state to have large scale CSP installation. Through MCDM approach, the analytic hierarchy process (AHP) technique has been applied to determine the multi-stage hierarchical structure.

Multiple-criteria decision-making (MCDM) or multiple-criteria decision analysis (MCDA), popularized by Stanley Zionts (1979) is MCDM or MCDA are well-known acronyms for multiple-criteria decision-making and multiple-criteria decision analysis. Conflicting criteria are typical in evaluating options: cost or price is usually one of the main criteria, and some measure of quality is typically another criterion, easily in conflict with the cost. MCDM is concerned with structuring and solving decision and planning problems involving multiple criteria. The purpose is to support decision-makers facing such problems. Typically, there does not exist a unique optimal solution for such problems and it is necessary to use decision-maker's preferences to differentiate between solutions.

The Analytic Hierarchy Process (AHP), introduced by Thomas Saaty (1980), is an effective tool for dealing with complex decision making, and may aid the decision maker to set priorities and make the best decision. The AHP generates a weight for each evaluation criterion according to the decision maker's pairwise comparisons of the criteria. The AHP generates a weight for each evaluation criterion according to the decision maker's pairwise comparisons of the criteria. Finally, the AHP combines the criteria weights and the options scores, thus determining a global score for each option, and a consequent ranking. The global score for a given option is a weighted sum of the scores it obtained with respect to all the criteria.

The data were collected using various methods such as Geographic Information System (GIS) analysis combined with Multi-criteria Decision Making (MCDM) technique. The pairwise matrix data, used to develop MCDM-AHP technique, were conducted from literature studied based on experts opinions, the data used to build model then determine proposed location and usage for RE projects and resources from multi Hotspot locations in UAE using commercial software ArcMap (version 10.3.1). Data were analyzed to investigate the site suitability.

The GIS data were downloaded from different online sources based on satellite images. The GIS combined with MCDM will be used to identify CSP and PV hotspot area in United Arab Emirates. The process of identifying hotspot areas will be discussed in two parts, in the first part, the unsuitable areas within UAE will be discussed just to mask the areas that are not suitable to build CSP and PV projects. Then in the second part MCDM represented by AHP technique will be used in order to weight the decision areas and ranking criteria as shown in Fig. 1.

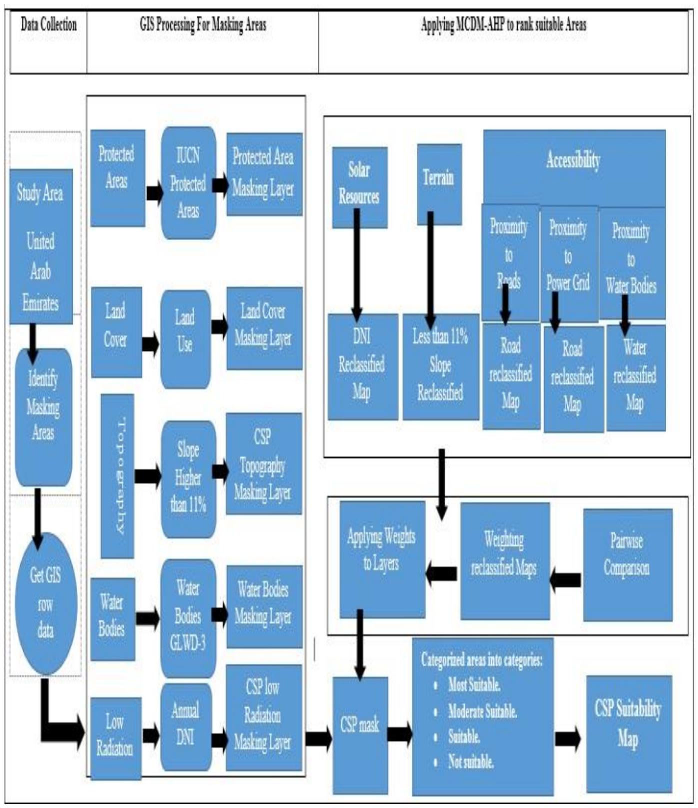


Fig. 1. Methodology process.

II. GIS Tool

GIS is a helpful tool for the analysis of maps and geographical information. Furthermore GIS represents powerful tool to identify renewable energy potential sites that help policymaker to decide optimal locations for installing renewable energy projects.

The idea of the solution is that the optimal location, where to install solar plant, must comply some conditions, for example, the location must be unused, it should have a suitable solar radiation, doesn't represent water body, has an elevation suitable for adapting the proposed technology and far from future urban location as well as near to power Grid. The five *steps* in the analysis process are:

1) Exclusion of Unsuitable Areas

Firstly, the study will exclude areas that are not suitable for building large-scale CSP projects. Identifying those areas requires spatial data analysis. The data includes geographical factors, infrastructure factors, terrain factors, land cover factor, solar potential factor, land use factor, elevation and slope factor.

A. Protected Areas

International union for conservation of nature IUCN defined the protected areas as: "a clearly defined geographical space, recognized, dedicated and managed, through legal or other effective means, to achieve the long term conservation of nature with associated ecosystem services and cultural values "[3]. IUCN defined six categories of protected areas which are: Strict Nature Reserve, Wilderness Area, National Park, Natural Monument or Feature, Habitat/Species Management Area and Protected Landscape/Seascape [3].

United Arab Emirates has 44 protected areas, the area of land area protected is 12296 Km^2 while the area of marine area protected is 2336 Km^2. Protected areas include parks, wild areas and nature reserves, those areas represent biodiversity conservation [4].

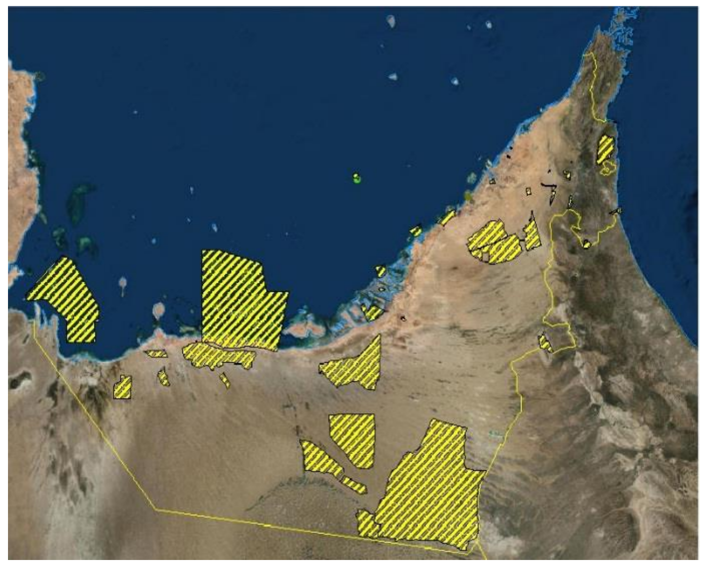


Fig. 2. UAE protected areas map.

Fig. 2 shows the UAE protected areas which shaded in yellow color. The information regarding protected areas was obtained from project of world database on protected area (WDPA). The map shown in Fig. 2 represents first mask layer to be excluded when performing GIS analysis.

B. Land Cover Area

A large scale CSP projects require large-scale areas compared with traditional power generating plants. Therefore, it is necessary to assign areas that not occupied by other land uses. The information regards land cover was obtained from European space Agency Globcover database map (ESA) 2009.

International Globcover map contains 23 categories for land use, among those categories only two categories mainly existing in UAE that will be considered in exclusion areas. Fig. 3 shows in colors such areas:

- Ginger Pink color: refers to value "190 " in GlobCover Land Cover Classes, which means Artificial surfaces and associated areas (Urban areas >50%). This area represents 1% of total country's area [5].

- Yellow color: refers to value "210 " in glob cover land cover classes, which means small size water bodies, similarly this area covers less than 1% of total area [5].

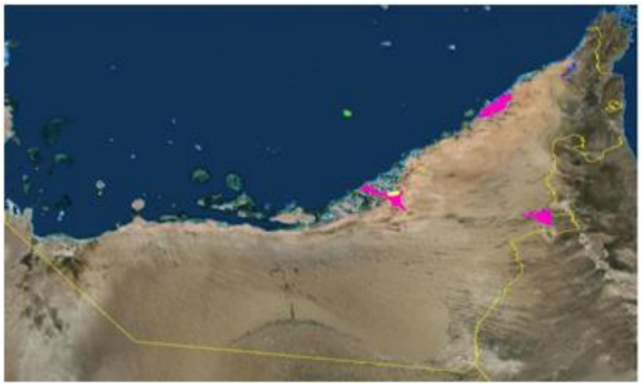


Fig. 3. UAE land cover map.

C. Topography

It is recommended for large scale solar projects to be installed on flat area especially CSP plants. The terrain slope value is the main factor that determines whether this area is suitable to construct solar power plant or not, so the study will use the slope as ranking criteria in our study. Many more studies were done to determine the proper value for suitable land slope. The studies concluded that the semi-flat areas is suitable being have high solar exposure, conversely the steep areas increase the project installation cost as well as maintenance cost [2]. In this paper, the lands with slope less than 4% are considered, while the slopes higher than 4% are not considered. Fig. 4 shows the digital elevation model obtained from U.S. geological survey (USGS) with one arc second resolution (30 M) with 95% confident level [8]-[10].

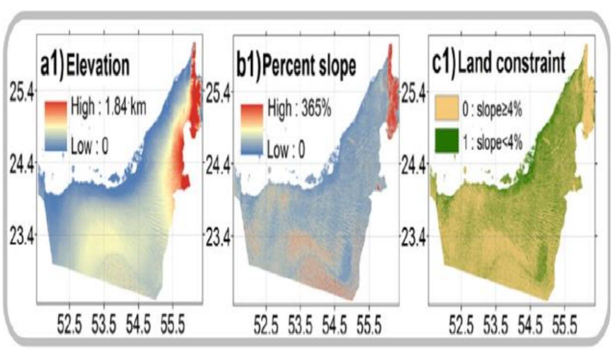


Fig. 4. a1) UAE map elevation; b1) percent slope; c3) land constraint.

D. Water Bodies

The data regards water bodies including lakes, reservoirs and rivers, water bodies that will be used as excluded mask from suitable areas map [6]. Fig. 5 below shows the water bodies' map, where the water body location is shaded by Blue color. The data regarding water bodies is provided by the Global Lakes and Wetlands Database (GLWD). GLWD consists of three data levels: GLWD-1, GLWD-2, and GLWD-3. In this paper GLWD-3 is used which has high resolution (30 arc-second) as well as contains all data exist in the lower two levels [6].

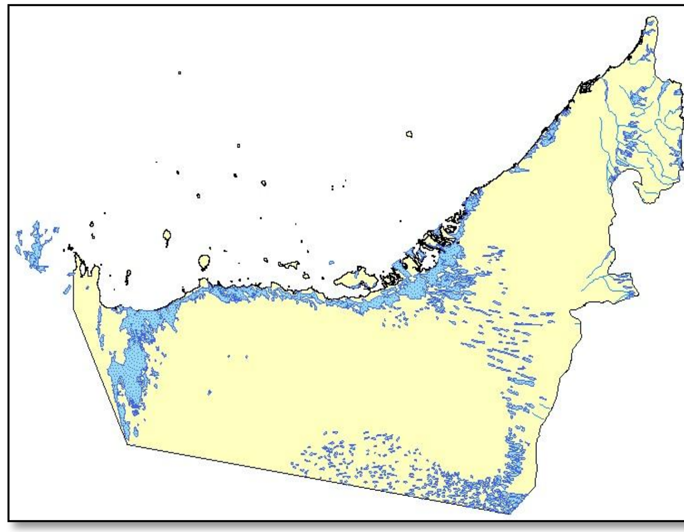


Fig. 5. UAE water bodies map.

E. Low Solar Radiation

Sun radiation goes through multiple changes until it reaches the surface of earth. These changes are related to the atmosphere as well as the topography of the earth. Radiation reaches the earth surface either direct or reflected or diffused, all those three types represent global solar radiation. Solar radiation is the main source of solar power CSP and PV technologies. The direct normal *irradiance* (DNI) represents the source of specific cooling *power* (SCP) technology. DNI is the amount of solar radiation at the flat earth's surface perpendicular to sun's beam while the sky radiation is blocked. In order to install feasible large-scale solar project, the area intended to be used shall meet the minimum radiation levels required by each technology. Experimentally the commercial CSP projects have been installed on areas with annual DNI greater than 1800 KWh/m^2. Hence throughout this study the areas with annual DNI less than 1800 KWh/m^2 will be excluded in mask map (assigned low weight) [7]. The source data regarding Global Horizontal Irradiance (GHI) and DNI are obtained from MASDAR Solar atlas [8].

The UAE monthly maps for the DNI and GHI are shown in Fig. 6. Note that for DNI component, it presents higher values approximately 170 W/m^2, during March, April, May (spring season), September and October (autumn). Furthermore, the lower values were registered in summer and winter in which the minimum value was approximately 130W/m^2 [7]. The average annual sum of DNI varies between 1600 KW/m^2 and 2300 KW/m^2 as shown in Fig. 7. The best areas to install CSP projects are the places where DNI is greater than 1800 KW/m^2 whereas the locations that are greater than 1800 MW/m^2 represent more than 70% of country area at the southern part of UAE.

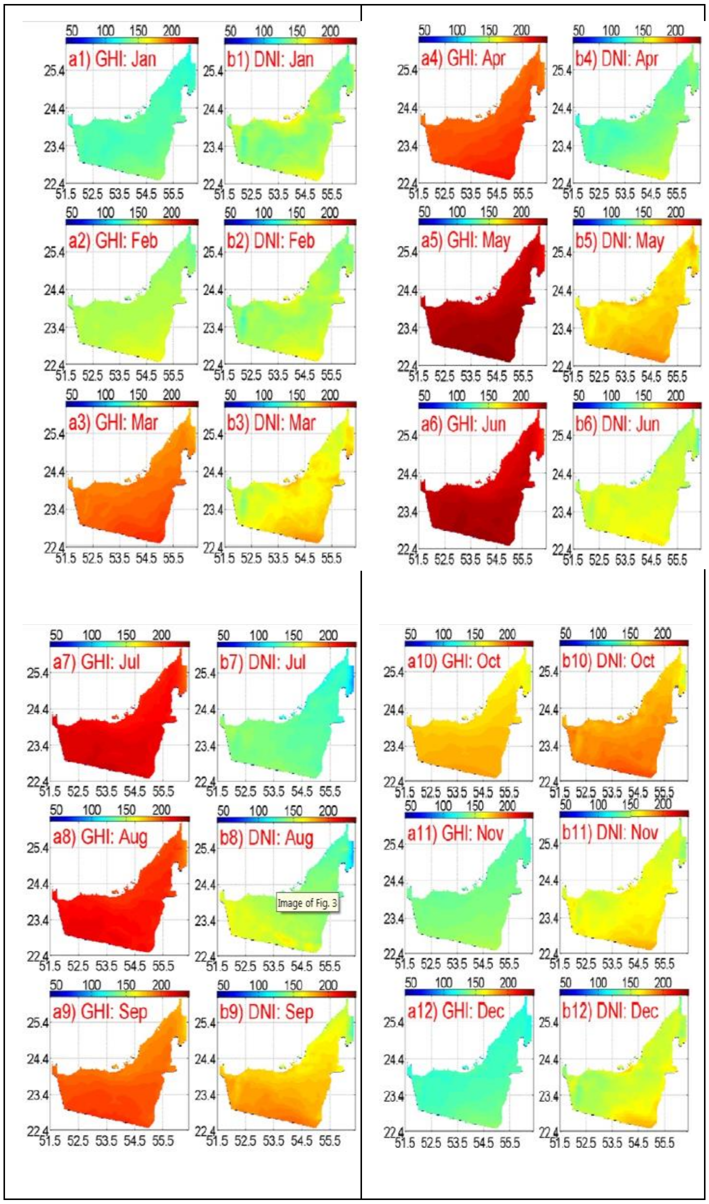


Fig. 6. Monthly Avg. of GHI , DNI in KWH/m^2.

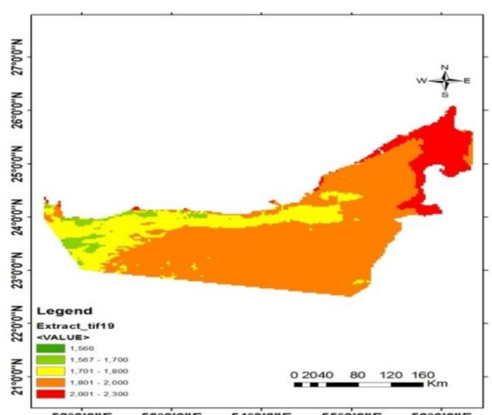


Fig. 7. UAE annual sum of DNI

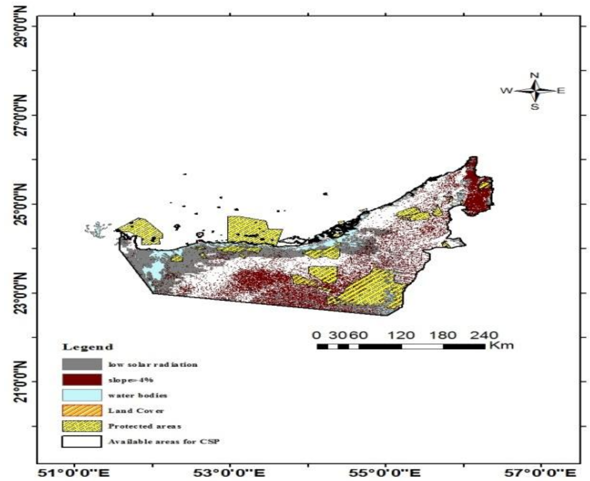


Fig. 8. UAE exclusion map.

The above exclusion maps shown in Fig. 8 are accumulated in one map that includes all inclusion criteria's. The areas that don't represent any of the above exclusion criteria will be considered as available areas to install CSP projects. Fig. 8 shows the exclusion areas, where the white colored pixels represent suitable places to install CSP projects which will be the main target for ranking criteria in the next section.

2) Ranking Suitable Area

In this section, the research will define ranking criteria to be used within MCDM. The MCDM will identify the hot-spots within suitable areas shown in Fig. 8. The ranking criteria determine the weights for each criterion, and then determine main hot-spots within that area. The main idea of ranking criteria is to give a weight to each ranking factor; this weight is calculated depending on how much the factor is important as compared to other factors. The factor may have conflicted objective with other factors, therefore, it is recommended to install solar projects at high solar irradiation areas, but these areas may have a high elevation that makes the installation process difficult and costly. Through-out the research it will be considered the following factors: irradiation level, water availability, proximity to roads and power grid as well as slope.

A. Solar Irradiation

As mentioned in the previous section the solar radiation is the main source for solar power technology. The sites that have higher levels of DNI will have great impact in the feasibility of solar projects. The map is classified with interval of 50 KWh/m^2 as annual radiation. We have got 10 radiation interval, each interval had been assigned ~~a~~ value from 0-10 according to its importance as shown in Table I.

TABLE I. ANNUAL DNI RANKING CRITERIA

Irradiation	<1800	1800-1850	1850-1900	1900-1950	1950-2000	2000-2050	2050-2100	2100-2150	2150-2200	>2200
Score	0	2	3	4	5	6	7	8	9	10

B. Water Availability and Proximity to Water Sources

In CSP technology, the water is necessary for cooling operation of CSP plants. The CSP uses the thermal energy absorbed from the solar field that is used to drive the turbines. Therefore, it is recommended for CSP projects to be built near water sources in order to make the project techno-economic feasible. In Table II, the water sources proximity were assigned scores according to the closeness to the water bodies ranging from 0 -10. Remembering that the water body itself is excluded from the ranking process, it has been considered as an exclusion area. Furthermore, a buffer zone of 20Km was applied to the processed map, where the pixels inside buffer zone has score from 2-10, and the pixels outside buffer zone are assigned a score of 0 value.

TABLE II. PROXIMITY TO WATER BODIES RANKING CRITERIA

Distance (Km)	0 - 2	2 – 4	4 - 6	6 - 8	8 - 10	10 - 12	12 - 14	14 - 16	16 - 20	>20Km
Score	10	9	8	7	6	5	4	3	2	0

C. Distance from Roads

The roads play significant role in any project or service intended to be supplied. The proximity to the roads presented in Table III determines the installation cost of the solar projects; therefore, when starting to build a site, the vehicles need roads for their easy access to the site. Furthermore, when the site is close to the roads, then the maintenance cost will be less as compared to the remote sites. Remote sites need additional cost represented by constructing new roads to access the sites. Fig. 9 shows the main roads within UAE, this map will be used to construct the overall suitability map [9].

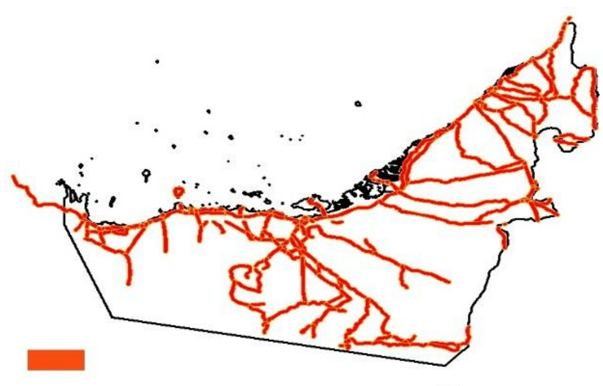


Fig. 9. UAE roads map.

TABLE III. PROXIMITY TO ROADS RANKING CRITERIA

Distance (Km)	0 - 5	5 – 10	10 - 15	15 - 20	>20Km
Score	10	9	8	7	1

In this paper, a buffer zone of 20Km was considered for a main target of this study. So the distances greater than 20Km were given a low score equal to 1, while the distances less than 20Km were given a high score (2-10) reciprocal to the closeness.

D. Distance from Transmission Lines

The proximity to the power grid is an important feasibility factor in the project planning phase. Solar projects need to be connected to the high voltage power grid. There, where the power grid is not exist, the project managers need to install transformers as well as running high quality cables for the national power grid which represent additional cost to the project. Fig. 10 illustrates the main power transmission lines within UAE, whereas this map will be used to construct the overall suitability map.

Throughout the research a surrounding area with 40 Km was considered as a buffer zone around the transmission lines. A high scores from (3-10) were given in Table IV to pixels inside buffer zone while a low score (equal 0) were given to the pixels out of the buffer zone. Accordingly, the manager at the project plan phase should look for sites that are close to the power grid in order to reduce cost and increase the project feasibility.

TABLE IV. PROXIMITY TO TRANSMISSION LINE RANKING CRITERIA

Distance (Km)	0 - 5	5 - 10	0 - 15	5 - 20	0 - 25	5 - 30	0 - 35	5 - 40
Score	10	9	8	7	6	5	4	3

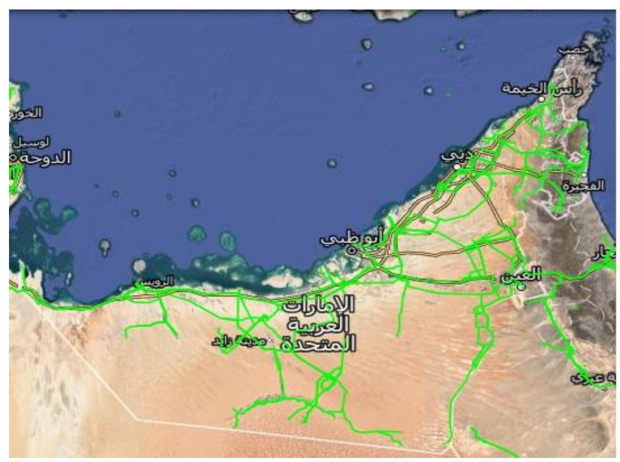


Fig. 10. UAE power lines map.

All the above mentioned factors can be represented as GIS map layer, where the GIS technique enables the combination of above conditions. Hence, the decision criteria are defined based on relative weight as a result of using AHP method. Map's layers are digitized then overlaid on the top of each other, thus the final map will show the occupied areas as well as free areas (suitable regions).

Fig. 11. AHP hierarchy.

AHP is multi-criteria decision making divided the problem into small parts then construct hierarchy in which the goal is in the top and criteria in the sub-level of hierarchy, where each level has adequate weight and the alternatives exist in the bottom of hierarchy, Fig. 11. The best alternative is chosen by performing series of activities in order to select best alternative from multi-alternatives [10].

TABLE V. RATIO SCALE FOR PAIR WISE COMPARISON OF IMPORTANCE

Significance	Definitions
1	Same importance
3	Moderate importance
5	Strong importance
7	Very strong importance
9	Extreme importance
2,4,6,8	Intermediate value between the above adjacent values

The first step in AHP techniques is to construct the decision hierarchy in which the decision consists of the objective for criteria then sub-criteria and finally the goal. The goal is represented at the top, whereas the criteria and sub-criteria are represented at the middle and the alternatives are at the lowest level.

The next step is to set the priorities at each level, multi-comparison matrices of all levels with respect to higher level that are made [11]. The pairwise comparisons are set according to how much "X " element is more important than "Y ", the preferences are given numbers using nine-point scale based on Saaty's pairwise comparison shown in Table V.

The next step is to generate the matrix of ranking for each level (based on Saaty's pairwise comparison as shown in Table V. After that the relative weights called Eigen vectors are obtained.

The matrix is constructed as comparison between objectives denoted by (A_1, A_2, …, A_n) and their weights denoted by (W_1,W_2,….,W_n). Table VI represents the pairwise comparisons matrix [12].

TABLE VI. MATRIX CONTAINING WEIGHTS

	A_1	A_2	…….	A_n
A_1	W_1/W_1	W_1/W_2		W_1/W_n
A_2	W_2/W_1	W_2/W_2		W_2/W_n
. . .				
A_n	W_n/W_2	W_n/W_2		W_n/W_n

The above matrix has a positive entries, if we multiply the above matrix by matrix W^T (W transpose), the result is the victor nW, where AW = nW.

Table VI presents the A matrix that is a unity rank, where each row is considered to a constant value multiplied by first row. $Aw = \lambda w$, the difference, if any, between λ_{max} and n is an indication of the inconsistency of the judgments. If λ_{max} = n then the judgments have turned out to be consistent.

The solution of W is any column of A, this solution is normalized so the total sum is equal to zero as shown in (1).

$$(A - \lambda\max)W = 0, \text{if } \sum Wi = 1 \quad (1)$$

The largest Eigen-Value of the above Matrix denoted by (λ_{max}) is calculated as shown in (2).

$$\lambda\max = \frac{1}{n}\sum_{i=1}^{n}\frac{(AW)i}{Wi} \quad (2)$$

Furthermore, matrix A satisfied consistency property represented by (3).

$$\text{Consistency ratio CR} = \text{CI} / \text{RI} \quad (3)$$

Where CI is consistency index (CI = (λ_{max} -n) / (n -1)), and RI is a random consistency index having fixed values as listed in Table VII. The value of CR ≤ 0.10 indicates that consistent exists.

TABLE VII. VALUES OF RI INDEX

n		2	3	4	5	6	7	8
RI		0	0.52	0.89	1.11	1.25	1.35	1.4

III. EVALUATION AND RESULTS OF APPLYING MCDM TO RANK SUITABLE AREAS

The GIS layers are now improved for selecting the suitable site for installing solar projects. In order to apply GIS-processing to the GIS layers, a weights are required to scale each layer to get suitability (Hot-Spots) Map. MCDM technique could be applied to the factors in order to get the required weights MCDM follow pre and post steps in order to get the required weights as follows:

Step 1: Set the goal / defined the problem

The goal should be specific, measurable, and relevant and time bounded. The main goal here is to attain overlay measurable weights which are relevant to be applied to the GIS-layers in order to get the hot-spots map specific.

Step 2: Determination of the criteria and constrains.

Throughout the previous sections the study discussed the factors as well as constrains, where the factors were the solar irradiation (DNI), terrain elevation represented by slope, as well as proximity to the roads, water bodies and power grids. The constraints were the protected areas, water bodies, land cover areas and low solar radiation areas as well as steep areas [13].

Step3: Standardize the factors-criterion scores

Throughout the research the study has combined multi-factors in different measurement scale, for instance the slope is percentage value while elevation is measured by meters and so on. To tackle this problem each factor was reclassified to standard range in which each interval is represented by standard value according to its importance.

Step 4: Determine the weight of each factor

AHP is represented by importance matrix that contains a comparison between criterion (factors) shown in Table V. The decision matrix values had been determined according to literature studies in this field based on international experts' opinions. The weights are calculated based on the related formula, then the result used to calculate a consistency ratio

(CR) of the pairwise comparisons by equation CR=CI/RI mentioned above. The CR must be less than 10% to ensure consistency. Otherwise the process is not consistent and the summation of overall weights is equal to 1. Tables VIII and IX show the calculations of AHP decision required to determine factors weights in addition to main criteria and sub-criteria. The consistency of the factors is less than 10%, which means that the estimation is consistent such that [14]:

- CR= CI/RI, CI = (λ_{max} - n)/ (n-1), so CI = (3.065 - 3) / (3 – 1) = 0.0325.
- CR = 0.0325 / 0.52 = 0.0625
- Weights w can be found by solving the matrix (A-λ) w=0 (Table VIII).
- CR= ((3.018 - 3)/ (3-1))/0.52 = 1.7%.
- Weights w can be found by solving the matrix (A-λ) w=0, Table IX.

Step 5: Aggregate the criteria

The aggregation is made by linear combination of all weights multiplied by constraints Boolean value. Equation (4) represents the decision role:

$$S = \sum \mathrm{wi.xi} \times \prod \mathrm{cj} \qquad (4)$$

Where, S – aggregate suitability score, xi – pixel's factor value (radiation, slope, proximity), wi – weights of factor, cj – constraints (exclusion areas), product of constraints (1-suitable area, 0-unsuitable area).

TABLE VIII. AHP METOD TO DECISION MAIN CRITIRION (FACTORS): THE PAIRWISE COMPARISON MATRIX.

Factor (criteria)	Decision matrix			Eigen value	Eigen vector solution	Consistency Ratio CR	Weights
Solar irradiation	1	3	7	3.065	4 iterations, delta = 2.6E-8	6.25%	64.9%
Slope	0.33	1	5				27.9%
Proximity	0.14	0.2	1				7.2%
SUM=							100%

The results we have got from Step 4 were applied into (5) of Step 5 using ArcGis maps software (Raster calculator command). The final result is the maps shown below in Fig. 12, the red areas represent hot-spots areas within Dubai Emirate, while the orange represent the suitable areas, the green color represent moderate and less suitable areas (not-suitable).

Step 6: Verify the results

Site survey is a method to validate the results, fortunately UAE have implemented two large-scale successful projects and on coming project in swihan, where the implemented projects exist in: Masdar City and Shams1 solar projects [14]-[16].

It is evident from Fig. 12 that the two projects in orange areas and macaw green areas are suitable and moderate suitable areas for successful implementing large-scale CSP projects in UAE. It is good to mention here that one of the most successful CSP projects known as "Swihat project" has been already installed in this orange area. This improve that the results we gained are valid [17].

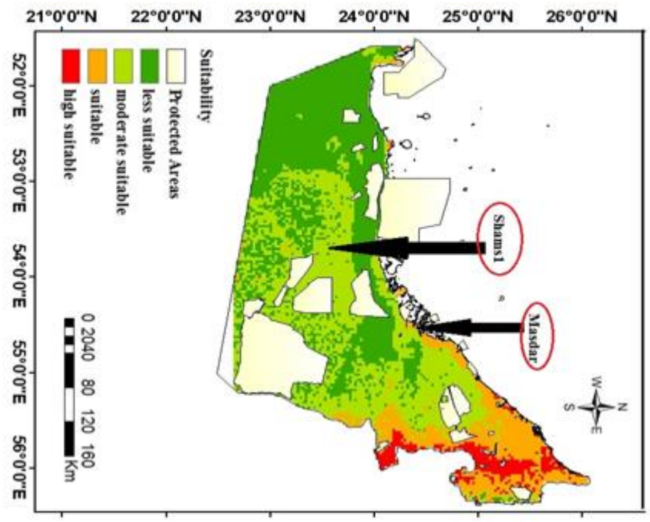


Fig. 12. UAE solar suitability map.

IV. CONCLUSION

Identifying the optimal location to install CSP solar project in UAE is the main concern of this paper, in terms of low cost, right location, low risk as well as appropriate delivering time.

A combination between GIS-maps and MCDM-AHP approach has been adapted in order to determine the optimal location to house large-scale solar CSP projects in UAE. This is the first study to identify suitability map for large-scale CSP solar locations using GIS-MCDM technique on a country level of UAE.

The research answered the main question of this study "where and how to determine the suitable locations to install large scale solar projects in UAE?" The decision factors, including exclusion areas plus ranking suitable areas, are the key elements to generate the suitability map.

To mask unsuitable areas, the study has identified five exclusion criteria (i.e. protected areas, land cover areas, topography, water bodies, and low solar radiation). Only three main ranking criteria have been included (i.e. sola irradiation, slope of the area and accessibility). Concerning the accessibility, three sub-criteria have been identified (i.e. proximity to roads, proximity power transmission lines and proximity to water sources) [18].

The AHP approach has been applied to determine the relative weights of the decision-criteria and the final weights of the ranking criteria. Then the combination between GIS data and AHP approach was used to determine the potential areas to install large-scale solar projects. The suitability maps show that Dubai Emirates has more suitability for CSP projects, whereas the hotspot areas to install CSP projects are shown in red color on the map [19].

The study prepared in this research can help policy and decision makers to adapt suitable policies as well as suitable locations for renewable investments.

TABLE IX. AHP METHOD TO DECISION SUB-CRITIRON (FACTORS)-PROXIMITY

Sub-Factor (criteria)	Decision matrix			Eigen value	Eigen vector solution	Consistency Ratio CR	Main weight	Sub-weight	Sub-criteria weight
Roads	1	0.5	3	3.018	4-iterations, delta = 2.3E-10	1.7%	7.2%	32%	2.23%
Power Grid	2	1	4					55.8%	4%
Water sources	0.33	0.25	1					12.2%	0.8%
Sum=								100%	7.2%

REFERENCES

[1] D. R. JALILVAND, "Renewable Energy forthe Middle East and North Africa", Friedrich-Ebert-Stiftung, Amman, 2012.

[2] S. S. A. B. Ahmed Aly, "Solar Power Potential in Tanzania", elsevier, Amestrdam , Holand, 2017.

[3] Marei, I., "The law and policy for electricity generated by rnewable energy: Grenning the power in three middle eastern Jurisdictions", Ms.C. thesis. Brisbane, Australia: Queensland University of Technology, 2015.

[4] IUCN, "Protected areas", International union for conservation of nature IUCN, 2008.

[5] WDPA, "protected planet", 14 october 2017. https://www.protectedplanet.net/country/AE.

[6] European space Agency ESA, "ESA", 2009. http://due.esrin.esa.int/page_globcover.php.

[7] L. a. Doll, "Global lakes and wetlands database", 17 Oct. 2017. https://www.worldwildlife.org/publications/global-lakes-and-wetlands-database-lakes-and-wetlands-grid-level-3.

[8] Gherboudj I, Ghedira H, "assessment of solar energy potential over UAE", ELSEVIER, 2015.

[9] IRENA, "RECREMA", 2017. https://atlas.masdar.ac.ae/.

[10] o. s. m. f. OSMF, "open street map", 2017. https://goo.gl/1EtZRh.

[11] M. Stojanovic, "multi-criteria decision-making for selection of renewable energy system", saftey engineering, pp. 115-120, 2013.

[12] M. STOJANOVIĆ1, "MULTI-CRITERIA DECISION-MAKING FOR SELECTION OF RENEWABLE ENERGY SYSTEM", www.safety.ni.ac.rs, 2013.

[13] P. Kousalya, Analytical Hierarchy Process approach – An application of engineering education.

[14] DECC, "Statistical Data Set - Industrial energy price statistics", 2014.

[15] N. W. Station, "Solar PV Geneartion - Live Performance Data", 2014.

[16] A. Haque, M. Rahman, and Q. Ahsan, "Building Integrated Photovoltaic system: Cost effectiveness", in Electrical & Computer Engineering (ICECE), 2012 7th International Conference on, 2012, pp. 904-907.

[17] M. Koller, T. Borsche, A. Ulbig, and G. Andersson, "Defining a degradation cost function for optimal control of a battery energy storage system", in PowerTech (POWERTECH), 2013 IEEE Grenoble, 2013, pp. 1-6.

[18] F. O'Flaherty, J. Pinder, and C. Jackson, "Determination of payback periods for photovoltaic systems in domestic properties", 2012.

[19] Z. Wu, X.-P. Zhang, J. Brandt, and S. Zhou, "Three Control Approaches for Optimized Energy Flow with Home Energy Management System", IEEE Power and Energy Technology Systems Journal 2015.

Formal Specification and Analysis of Termination Detection by Weight-throwing Protocol

Imran Riaz Hasrat, Muhammad Atif
Department of Computer Science
and Information Technology
The University of Lahore
Lahore, Pakistan

Muhammad Naeem
Department of Electronics
and Electrical Systems
The University of Lahore
Lahore, Pakistan

***Abstract*—Termination detection is a critical problem in distributed systems. A distributed computation is called terminated if all of its processes become idle and there are no in-transit messages in communication channels. A distributed termination detection protocol is used to detect the state of a process at any time, i.e., terminated, idle or active. A termination detection protocol on the basis of weight-throwing scheme is described in Yu-Chee Tseng, "Detecting Termination by Weight-throwing in a Faulty Distributed System", JPDC, 15 February 1995. We apply model checking techniques to verify the protocol and for formal specification and verification the tool-set UPPAAL is used. Our results show that the protocol fails to fulfil some of its functional requirements.**

***Keywords*—Termination detection; weight-throwing protocol; formal specification and verification; model checking**

I. Introduction

Termination detection is an important problem for distributed systems. For a distributed system, termination detection is based on the concept of a process state. During a distributed computation, a process can either be in alive or dead state. An alive state means that a process is still performing its task whereas dead state represents that the process becomes idle simultaneously. Dead and alive states are referred as *passive* and *active* states as shown in Fig. 1. At the start of the computation, all the processes are supposed to be in *active* state. Processes can take several actions discussed below:

- Only the processes in *active* state can send *basic* messages to other processes.
- Any *active* process can reach a *passive* state at any time.
- A *passive* process becomes *active* again by receiving a *basic* message.

A distributed computation is called terminated if all of its processes become *passive* and there are no in-transit messages in the communication channels.

Many applications of distributed systems depend on termination detection of a computation to guarantee a proper operation. In multiphase algorithms [2], one phase depends on proper completion of other phase. So, initiation of new phase needs termination detection of previous phase. In distributed databases, deadlock detection is a critical problem and this problem is purely related to termination detection [3]. Garbage collection [4] and token loss detection in a token ring are other examples of termination detection problems. Termination detection solution allows a system to guarantee that all tasks in the system are obviously complete and this permits the dependent systems to start their computations. About more than three decades ago, termination detection problem was separately suggested by Dijkstra and Scholten [5] and Francez [6]. Many researchers started to tackle this problem by developing different termination detection algorithms described in [7]–[15].

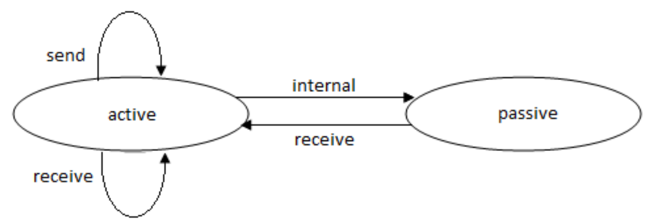


Fig. 1. Termination detection [1].

Formal methods are based on mathematical tools and techniques used for design, specification and verification of different hardware and software systems. Formal methods provide correctness for all requirements and inputs of a given system. In the past, formal methods were used for only safety critical and defense related systems [16]–[18]. Now a days, high demand of error free and secure systems is giving much importance to formal methods.

We present a formal modeling and analysis of the termination detection by weight-throwing in a faulty distributed system presented in [7]. It is a comprehensive analysis of all possible versions of protocol along with verification of detailed functional requirements. Basic concept in weight-throwing protocol is that each process sends some weight with every *basic* message. On reception of this message, recipient process adds this weight to its current weight. We present formal verification of the weight-throwing protocol using a model checker known as UPPAAL.

UPPAAL has a simulator that is used to develop the model [19]. The verifier in UPPAAL has capability to create the traces that can lead the action sequences where system's required property fails. To investigate this situation, action sequences are replayed in simulator.

We formally model and analyze both parts of the protocol. We present a formal specification of the protocol in the timed-automata language of UPPAAL. Then, we specify functional requirements for safety and liveness of the protocol. We also present specification and verification of its invariants. We also analyze the protocol in the situations when some processes may tend to fail. We discover some situations in which protocol fails to satisfy its functional requirements, termination is not detected in these situations. The protocol also fails to satisfy some of its invariants. We present counterexamples for the requirement violations in the form of message sequence charts.

II. Weight-Throwing Protocol

Weight-throwing protocol works in a distributed system. In this protocol a process sends half of its weight with its message when it communicates to some other process. Let S is the set of processes i.e. $S = \{P_1, P_2, \cdots, P_n\}$. The S is supposed to be fault-free. In fault-free system, a process never fails. A process from S with minimum id is said to be the *leader*. Total weight in the system is 1. Every process is *active* at the start of the computation and weight of each process P_i is $w_i = 1/n$, where n is the cardinality of the set S. The *leader* collects all the weights in the system and announce termination. Upon every message from a process P_i to P_j following actions are performed:

1) P_i divides its weight w_i into two equal positive real parts x, y so that $w_i = x + y$.
2) P_i sends a *basic* message $B(x)$ along with a weight x.
3) P_i updates its current weight as $w_i = y$.
4) On reception of the *basic* message, P_j's weight increases as $w_j = w_j + x$.

Any process in the system can become *passive* at any time. When a process other than the *leader* becomes *passive*, it sends a *control* message to the *leader* for submitting its weight. After sending the *control* message this process sets its weight to 0. Any *passive* process can become *active* again by receiving a *basic* message from some other process. The following are the invariants in the protocol:

- Each process and in-transit messages in the system have a non-zero weight at any time.
- If we take the sum of the weights of all processes and in-transit messages at any time, it is always 1.

When the *leader* becomes *passive*, it accumulates all the weights in the system. If accumulated weight becomes equal to 1, the *leader* calls termination. Weights should be handled precisely. Fractional values of all weights make it nearly impossible to make the sum to 1 again due to rounding errors of float values. This problem can be solved by representing the weight using two integer values as $[1, n]$ instead of $1/n$.

A. Flow-detecting Scheme for Flushing/Freezing of Channels

In case of faulty distributed system, some processes may fail during a computation. Overall weight of all processes can be less than 1 due to holding weights of failed processes and weights carried by undelivered messages in the channels. This problem is solved by introducing a flow detecting scheme. Let H be a subsystem which contains all healthy processes and all their communication channels. During a computation, the weight change in H at time interval I is equal to the difference of weight flowing into H and the weight flowing out of H. With the help of this scheme, the weight information of failed processes can be obtained from the outgoing weight records of healthy processes in the system because each process keeps the record of incoming and outgoing weights.

Assume that the intended system provides the facility of flushing or freezing of channels connected to faulty processes. Flushing or freezing mean preventing and ceasing further communication between a healthy and a faulty process. There is no global clock in the system that makes it very difficult to get global views of the system weight. A *snapshot* is taken to get the global views. The *leader* sends the *snapshot request* to all the healthy processes. On reception of this request, each healthy process flushes or freezes all of its communication channels connected to faulty processes and submits its incoming and outgoing weights to the *leader*. The *leader* uses these incoming and outgoing weight values to calculate the overall weight of the system.

B. Data Structure of Weight-throwing Protocol

Before the formal modeling of the protocol we need to know the specific keywords and data structure used in the protocol. Let P_i be any process in the system where $i = 1 \ldots n$, n is any arbitrary positive number. The data structure for P_i is given in Table I.

III. Modeling in UPPAAL

Our formal specification in UPPAAL has three participants, i.e. the *Termination*, the *MessageBuffer* and the *SnapBuffer*. The main process is the *Termination* process. This process receives and sends messages to other processes to communicate. Each message holds some non-zero weight. A process adds the incoming weight to its current weight when it receives a message. The communication is asynchronous. The *MessageBuffer* process holds the *basic* and *control* messages when receiver is not ready to receive them. These messages are moved to their receivers when they become ready. The *SnapBuffer* process temporarily stores the *snapshot request* messages sent by the current *leader*. It then sends the *snapshot request* messages to all the healthy processes to inform them the faulty set of processes known to the *leader*. This process also temporarily keeps the *snapshot reply* messages sent by the processes to the *leader*. These messages are delivered to the *leader* when it becomes ready.

The protocol has two parts. In first part, the termination detection is done using a fault-free distributed system. In the second part, a faulty distributed system is used to detect proper termination of processes. In faulty distributed system any number of processes may tend to fail. We present the formal specification of both parts separately to check the correctness of the protocol in both cases.

TABLE I. INFORMAL SPECIFICATION OF WEIGHT-THROWING PROTOCOL

$leader_i$	This variable stores id of current *leader*. At the beginning, $leader_i = 1$.
w_i	This field keeps the value of weight for each process P_i. Initially, the value of w_i is $1/n$, where n is the number of processes.
sum_i	This variable records total weight in the system assumed by the *leader*. It is a real number. At the beginning, $sum_i = 1$.
$IN_i[1 \ldots n]$	This is an array of real numbers. The $IN_i[j]$ keeps the record of all weights thrown out from process P_j to P_i. At the beginning, $IN_i[j] = 0$ for all $j = 1 \ldots n$.
$OUT_i[1 \ldots n]$	This is an array of real numbers. The $OUT_i[j]$ stores all the weights thrown to process P_j from P_i. At the beginning, $OUT_i[j] = 0$ for all $j = 1 \ldots n$.
F_i	This array represents a set of faulty processes. If a process P_i knows P_j to be faulty and P_i has flushed or frozen all the channels to P_i then it belongs to F_i . At the start $F_i = \phi$.
SN_i	This array contains set of all processes to which *snapshot request* is to be sent. Let's suppose P_i initiates the *snapshot request*. If a process P_j replies to the request or it is found faulty by P_i then it is removed from the SN_i. Second *snapshot* can be initiated only when SN_i becomes empty. At the start $SN_i = \phi$.
$temp\_sum_i$	This field stores a real number. During the *snapshot*, it temporarily calculates the total remaining weight.
$consistent_i$	This field indicates a boolean value which keeps the record of a *snapshot's* consistency.
$B(x)$	This represents the *basic* message with a weight x.
$C(x)$	This indicates a *control* message. The *control* message is used for reporting the weight x to the *leader* of the system.
$Request(F_i)$	This represents the message for snapshot request that is sent by the current *leader* of the system P_i. With the help of F_i, message receiving process is informed about the set of faulty processes already known to the *leader*.
$Reply(F_i, IN_i, OUT_i)$	This indicates the reply to the *leader's* request message. This reports the state of the replying process.

IV. MODEL1: TERMINATION DETECTION IN A FAULT-FREE DISTRIBUTED SYSTEM

We specify all the processes of fault-free part of the protocol. The *Termination* and the *MessageBuffer* are the participants in Model1. We present the functionality and formal specification of these participants in this section.

A. Channels

This protocol uses four channels which are described below. To model the functionality of termination detection in a fault-free distributed system, we use hand shaking channels. The working of these channels is described below:

1) *basicMessageS:* This channel is very important because the system uses this channel for the *basic* message communications. It sends *basic* messages to the *MessageBuffer* to hold them until their receivers become ready.
2) *basicMessageR:* For moving stored *basic* messages from the *MessageBuffer* to the receiver process, system uses *basicMessageR* channel.
3) *controlMessageS:* This is a channel for *control* message communications. System uses this channel to send *control* messages to the *MessageBuffer* to hold them until the *leader* becomes ready.
4) *controlMessageR:* This channel moves stored *control* messages from *MessageBuffer* to the *leader*.

B. Global Declarations

Some variables and arrays are declared globally so that each participant can access them and use them according to their needs. Table II represents the global declarations and data types for Model1.

TABLE II. GLOBAL DECLARATIONS FOR MODEL1

numberofweights	This is a constant that represents the total parts of weight.
max_weight_limit	This describes the maximum value of weight that can be sent through communication channels.
maxproc	It is a variable that tells the number of concurrent instances of *Termination* process.
max_proc_id	This variable stores the highest id of concurrent instances of the *Termination* process.
leader	This variable keeps id of the current *leader* of the system.
W1 and *W2*	These variables record the weights when a *basic* or a *control* message is received.
Out_arr[] and *In_arr[]*	Two customized data structures *Weight_out* and *Weight_in* are introduced to define the *Out_arr[]* and *In_arr[]* arrays respectively. The *Out_arr[]* keeps the records of out going weights and *In_arr[]* stores the records of incoming weights of each process of the *Termination* process.

C. The Automaton for Termination Process in Model1

The automaton for the *Termination* process is depicted in Fig. 2. The protocol model has a number of parallel processes, each of which is triggered by a certain communication among each other. The specification of the *Termination* process comprises four communicative choices i.e. sending *basic* message, receiving *basic* message, sending *control* message and receiving *control* message.

For overall working of the *Termination* process, we discuss the functionalities which take place between *active* and *A1* states. The *basicMessageS!* sends two weight values of *basic* message to the *MessageBuffer* for any other process. Two weights actually represent the single weight because we are using two values (numerator and denominator) just to avoid floating point errors. The guards for weight values, limit the number of *basic* messages that a process can send to other processes to reduce transition state space. Going from *A1* to *active* state, the *updateOut()* function updates the *Out_arr[]* to record the outgoing weights. Also the *weight[1]* value is doubled because doubling the denominator value, overall weight of a process becomes half. For a process taking a transition from the *active* to *A2* state, the channel *basicMessageR?* receives two weight values and sender id from its *MessageBuffer* to record incoming weight against a specific sender. In next transition, the *updateIn()* function updates the *In_arr[]* to record incoming weight from that specific process. The *updateWeight()* function records the overall current weight of receiver process. Same procedure is followed when going from *passive* to *active* state because both the transitions are identical and perform exactly same functionality. While taking transitions from *active* to *A3* and *A3* to *passive* state, the channel *controlMessageS!* sends weight values of this process alongwith its id to the *leader*. The function *updateIn()*

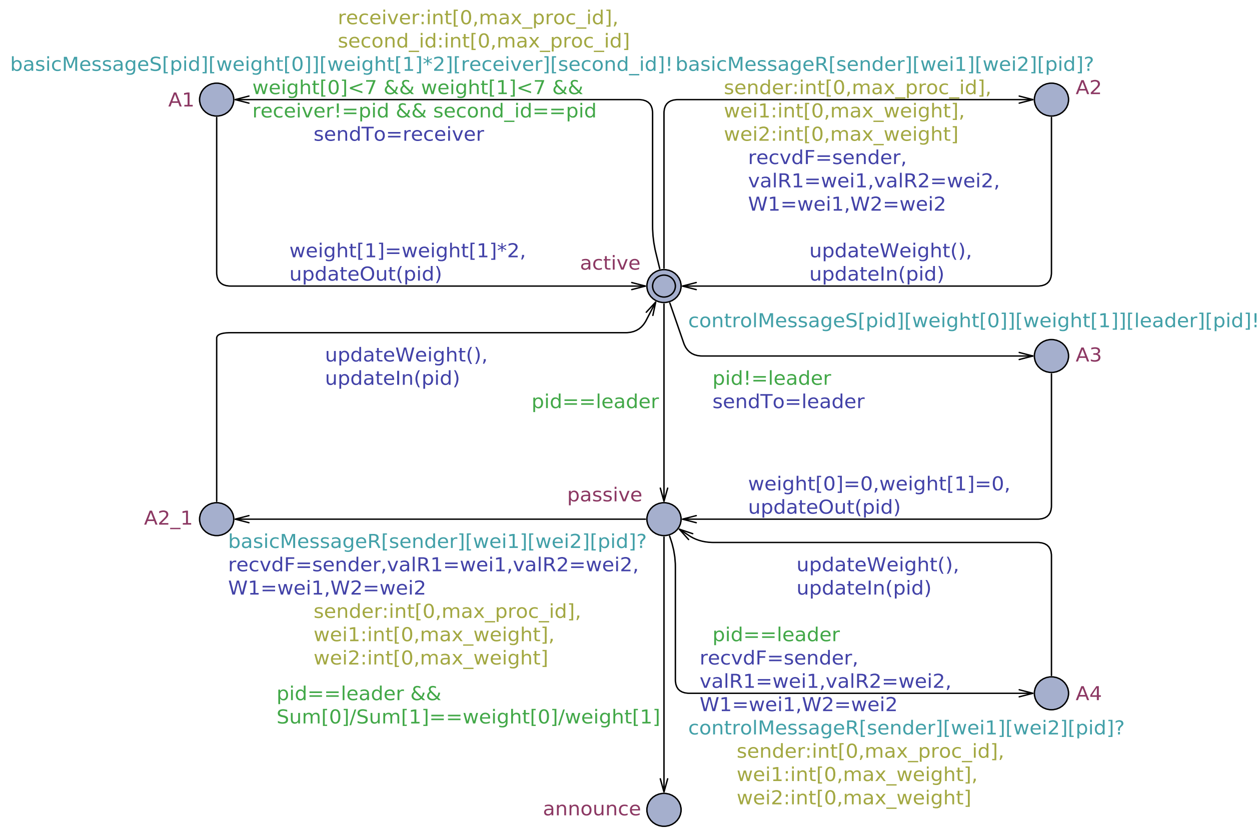


Fig. 2. Formal model for *Termination* process in Model1.

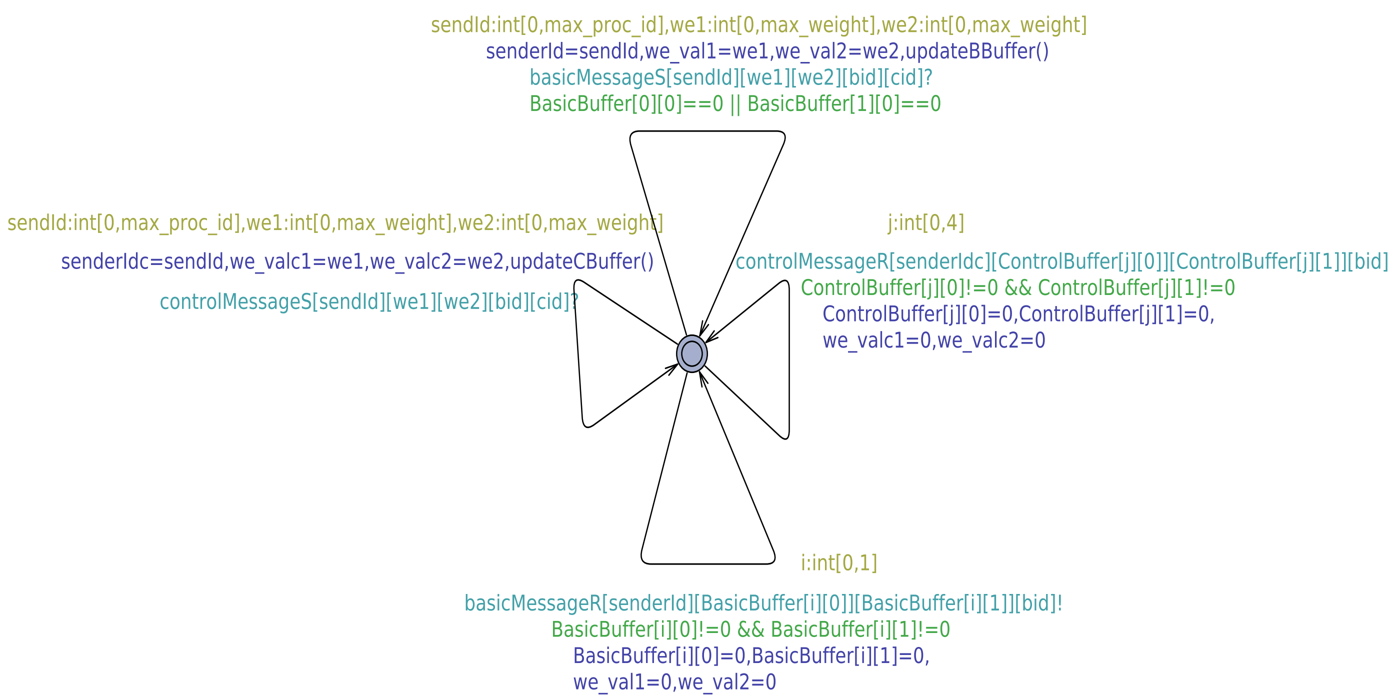


Fig. 3. Formal model for *MessageBuffer* process.

updates *Out_arr[]* of this process. The value of *weight[0]* and *weight[1]* is set to zero to make sure that all the weight is transferred to the *leader*. The working of transitions from *passive* to *A4* is similar to moving from *active* to *A2* state. The only difference is, in *controlMessageR?* channel, the *leader* receives the *control* message from *MessageBuffer*. All other variable definitions are exactly same. Taking transition from *A4* to *passive* state uses exactly same functions as described previously for *A2* to *active* state. Only the *leader* can take the transition from *passive* to *announce* state. The *leader* takes this transition only when it has collected all the weight from all the processes through *control* messages. If this weight becomes equal to the weight defined at the beginning of the computation, the *leader* announces termination.

The local declaration and data types of *Termination* process are given in Table III.

TABLE III. LOCAL DECLARATIONS FOR *Termination* PROCESS IN MODEL1

valR1 and *valR2*	These variables temporarily store first and second value of weight respectively at different transitions.
recvdF and *sendTo*	System uses *recvdF* variable to store process id of the message sender and *sendTo* records process id of the message receiver.
weight[]	*weight* array permanently stores the two weight values for every process at any time.
Sum[]	The *Sum[]* a value container that represents the total weight of the system that is actually 1.
tempIn[]	This array temporarily stores the sum of all incoming weights to a process.
tempOut[]	*tempOut[]* keeps the sum of all the outgoing weights from a process.

In UPPAAL system models, we can declare functions with in the process or alongwith global declarations. The functions can have return types and passing parameters. The *Termination* process also uses some functions to perform its functionality.

1. *updateWeight()*: This function is very important because it is called at different transitions when a *basic* or a *control* message is received. If the receiver has zero weight, the received weight is directly moved to *weight[]* of the process. If the process already has some non-zero weight then function adds the incoming and current weight to calculate the new weight.

2. *updateIn()*: A process uses this function when it receives a *control* or a *basic* message. This function updates the *In_arr[]* to record the incoming weight from a particular process. The incoming weight is directly moved to the *In_arr[]* if receiver is receiving the first message from the sender. The current weight and incoming weight is added to update the incoming weight in *In_arr[]* if the process already has received some messages from the same sender.

3. *updateOut()*: A process calls this function when it sends a *control* or *basic* message. This function updates the *Out_arr[]* to record the outgoing weight to a particular process. The outgoing weight is directly moved to the *Out_arr*array if receiver is receiving the first message from the sender. The current weight and outgoing weight is added to update the *Out_arr[]* if the process already has sent some messages to the same receiver.

D. The Automaton for MessageBuffer Process

The automaton for the *MessageBuffer* process is depicted in Fig. 3. The process has three instances. The specification of the *MessageBuffer* process in the protocol provides four communication choices i.e. receiving a *basic* message from the *Termination* process, sending a stored *basic* message to the *Termination* process, receiving a *control* message from the *Termination* process and sending a stored *control* message to the *Termination* process.

Now we discuss the functionalities for the *MessageBuffer* process when a *basic* message is received through *basicMessageS?* channel. The function *updateBBuffer()* updates the *basic* message buffer to keep the record of this incoming *basic* message until it is not delivered to the respective recipient. The *basicMessageR!* channel sends the stored *basic* message when the receiver *Termination* process becomes ready. The two guards prevent the communication between *MessageBuffer* and *Termination* process when their is no stored message. The function *updateCBuffer()* is called to update the *control* message buffer to keep the record of this incoming *control* message until its recipient is not ready. This happens when a *control* message is received through *controlMessageS?* channel. The *controlMessageR!* sends the stored *control* message when the *leader* becomes ready. The two guards prevent the communication between the *MessageBuffer* and the *leader* when their is no stored *control* message for the leader.

The local declaration and data types of the *MessageBuffer* process are given in Table IV.

TABLE IV. LOCAL DECLARATIONS FOR *MessageBuffer* PROCESS

we_val1 and *we_val2*	The variables store first and second value of *weight[]* for receiving and sending *basic* messages at different transitions of *MessageBuffer* process.
senderId and *senderIdc*	*senderId* variable stores the process id of the *basic* message sender process and *senderIdc* keeps the process id of the *control* message sender process.
we_valc1 and *we_valc2*	These variables contain the first and second value of *weight[]* for receiving and sending *control* messages at different transitions of *MessageBuffer*.
BasicBuffer[] and *ControlBuffer[]*	The *BasicBuffer* array list is to store the two *basic* messages for each process at any time. The limit to store only two *basic* messages is for achieving reduced transition state space. The *ControlBuffer[]* can store five *control* messages for the *leader* process at any time.

The *MessageBuffer* process uses some functions to perform different tasks on different transitions. We discuss these functions below.

1. *updateBBuffer()*: It is called on a transition when a *basic* message is received. It locates the available free space in *BasicBuffer[]* and then stores the incoming *basic* message at that space.

2. *updateCBuffer()*: A transition calls this function when it receives a *control* message at *MessageBuffer* process. The function checks the free space in *ControlBuffer[]* and then stores the incoming *control* message at that position.

V. MODEL2: TERMINATION DETECTION IN A FAULTY DISTRIBUTED SYSTEM

We specify all the concurrent processes of faulty part of the protocol. The three participants are the *Termination*, the *MessageBuffer* and the *SnapBuffer*. The functionality and formal specification of these participants is presented in this section. The functionality of the *MessageBuffer* process is already described in Model1. Here we discuss the functionality of remaining participants.

A. Channels

This protocol uses nine channels for communication among processes. Four of them are same as described in Model1. The other five channels are described here in detail.

1) *failReport:* The *Termination* process uses this channel when a process fails. It tells the failed status of a process to other processes. On the other side, a process receives the status of a failed process using this channel.
2) *failRequest:* This channel sends *snapshot request* message to the *SnapBuffer* process.
3) *failRequestR:* The channel sends the stored *snapshot request* message from *SnapBuffer* process to the recipient *Termination* process.
4) *failReply:* This channel is used to send the *snapshot reply* message to the *SnapBuffer* process.
5) *failReplyR:* The channel delivers the stored *snapshot reply* message from *SnapBuffer* process to the recipient *Termination* process.

B. Global Declarations

This system is modeled for termination detection in faulty distributed environment. This is the enhancement of the Model1 (fault-free model). Therefore some global variables are common in both the systems. We present here the description of variables that are not present in Model1 but are present in Model2. The global declarations for Model2 are described in Table V.

TABLE V. GLOBAL DECLARATIONS FOR MODEL2

FIn[] and *FOut[]*	The *FIn* array stores all the incoming weights and *FOut[]* saves the outgoing weights of failed processes which are known to the *leader* during the *snapshot*.
FI_FO_Diff[]	The *FI_FO_Diff* array stores the difference of all the incoming and outgoing weights of failed processes known to the *leader* during the *snapshot*.
S[]	System uses *S* array to store ids of all instances of *Termination* process participating in the system.
F[]	The *F* is a global array. It records the failed processes known to each process. The value may be different for each process.
Ftemp[]	*Ftemp[]* keeps the actual record of failed processes in the system.

C. The Automaton for Termination Process in Model2

We have discussed the *Termination* process for Model1 in previous section. The automaton for *Termination* process in Model2 is depicted in Fig. 4. The specification of *Termination* process has ten communicative choices four of which are same as in Model1. The other six choices are sending *snapshot request* message, receiving *snapshot request* message, sending *snapshot reply* message, receiving *snapshot reply* message, sending *fail report* message and receiving *fail report* message.

Now *Termination* has nine actions from *A1* to *A5* and from *F1* to *F4*. The actions *A1* to *A5* are already discussed in Model1. So we discuss here only the actions *F1* to *F4*. Fig. 5 represents the formal model for *F1*. The *failReport?* channel detects the failed process when no snapshot is in progress. In next transition, the process adds the failed process in its *F[]* and *Flush[]*. The function *Leader()* is called to determine the *leader*. If current process is not the *leader* then it reaches to *active* state. If this process is the *leader* then it reaches to *Snap* state and starts calculating the healthy processes to send them *snapshot request* message. The *SnapBuffer* stores this message until the receiver of this message is not ready. After sending these messages the process is allowed to reach at *active* state.

Fig. 8 represents the formal model for *F2*. A process receives the stored *snapshot request* message from *SnapBuffer* process through *failRequestR?* channel. Then it sends the *snapshot reply* message to the *SnapBuffer* for the *leader* through *failReply!* channel. It records the new *leader*. It matches and updates its *F[]* to record the failed processes known to the *leader*.

Fig. 6 shows the formal model for *F3*. The *leader* receives the stored *snapshot reply* message from *SnapBuffer* process through *failReplyR?* channel. It checks the difference of *F[]* of sender and its own *F[]*. The snapshot is marked as inconsistent if this difference is greater than zero or the snapshot is not consistent. The process which has sent this *snapshot reply* message is removed from *SN[]*. The process reaches to *active* state if a snapshot is in progress otherwise reaches to *Snap* state.

Fig. 7 represents the formal model for *F4*. The *failReport?* channel detects a failed process. The action makes sure that a snapshot is already in progress. In next transition this process adds the failed process in *F[]* and *Flush[]*. It also sets the snapshot as inconsistent. Then after removing the failed process from *SN[]*, if still *SN[]* is non empty then process calls the snapshot.

Fig. 9 represents the formal model for the process when it fails. At failure, the process updates the *FTemp[]* to record its entry in that array and moves from *active* to *fail* state. The channel *failReport!* at *fail* state continuously tells other processes that its status is failed.

We have discussed some local declarations of *Termination* process in Model1. Now we discuss the local declarations of remaining variables in Table VI.

calSN(), *calcDiff()*, *isAvailable()*: The three functions perform the combined functionality of identifying the healthy processes for sending *snapshot request* messages. The function *calSN()* calls the *calcDiff()* function. The *calcDiff()* function calls *isAvailable()* function for every process id. If a true value is returned it means that the process is present in the *F[]* of calling process and there is no need to send *snapshot request* message to that process.

AddIn(): The function *AddIn()* calculates the total incoming weights of all the failed processes known to the current *leader* during the *snapshot*. It adds the incoming weights of every failed process to make a sum of incoming weights in *FIn[]*.

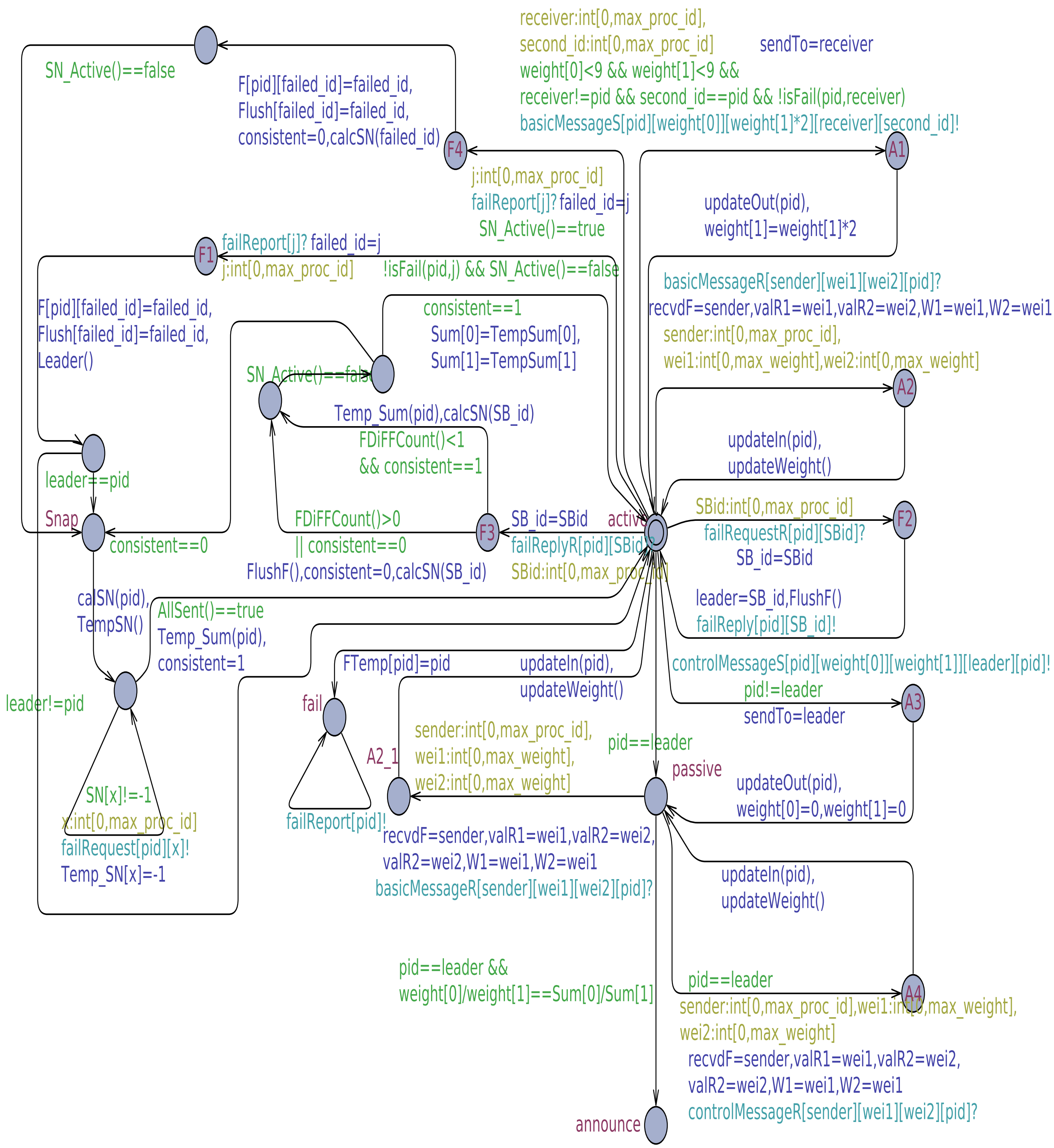


Fig. 4. Formal model for *Termination* process in Model2.

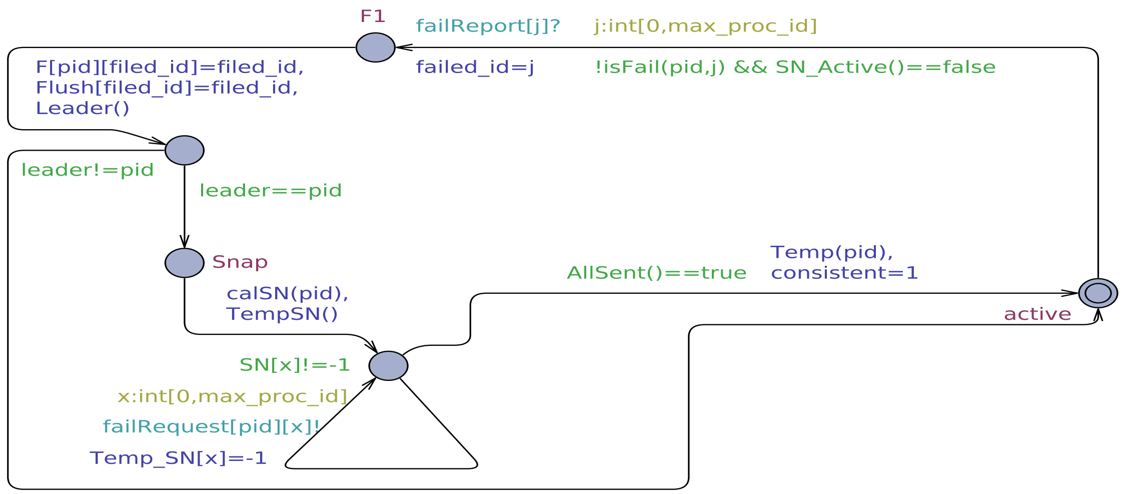


Fig. 5. Formal model for Action F1.

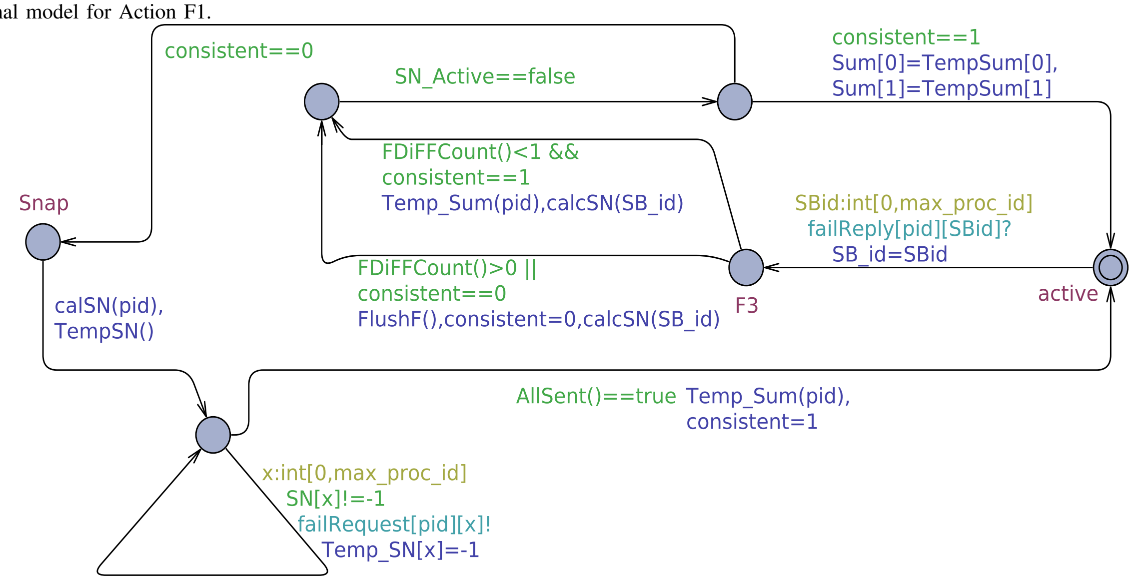


Fig. 6. Formal model for Action F3.

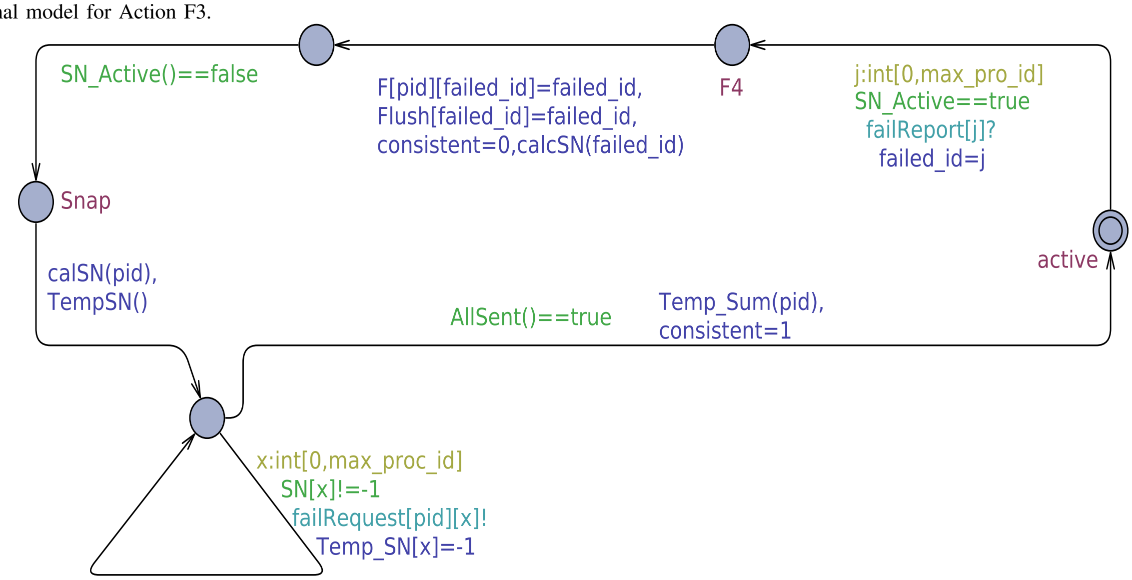


Fig. 7. Formal model for Action F4.

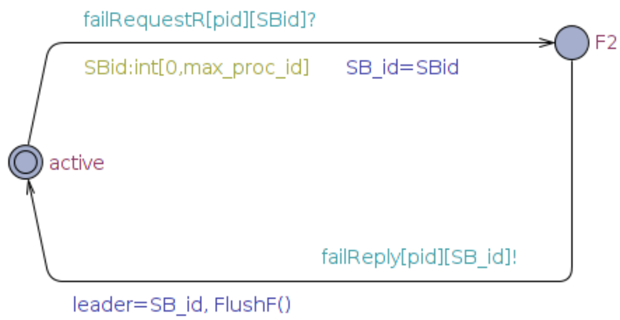


Fig. 8. Formal model for Action F2.

Fig. 9. Formal model for failed process.

TABLE VI. LOCAL DECLARATIONS FOR TERMINATION PROCESS IN MODEL2

leader	This variable stores the id of the current *leader* of the system known to a process. It may be different for each process.
SB_id and *SB_id2*	These variables record the ids of the message sender process at different transitions.
consistent	The boolean variable that shows the consistency of the *snapshot request* sent by the *leader*.
TempSum[]	The *TempSum* array stores the sum of all the weights during the snapshot.
SN[]	The *SN[]* keeps the set of processes to which the *snapshot request* is to be sent by the *leader*.
Temp_SN[maxproc]	The *Temp_SN* array temporarily records the set of processes to which the *snapshot request* message is to be sent by the *leader*. After sending the *snapshot request* to a process, the sender removes this process from *Temp_SN[]*. This array becomes empty after sending *snapshot request* message to all the healthy processes.
failed_id	A local variable that stores the id of the failed process at different transitions.
Flush[]	This array keeps the record of failed processes for which all further communications are flushed.
Stemp[]	The *Stemp* array contains the list of all the instances of *Termination* process taking part in the system.

AddOut(): It calculates the total outgoing weights of all the failed processes known to the current *leader* during the *snapshot*. It then adds the outgoing weights of every failed process to make a sum of outgoing weights in *FOut[]*.

FIin_FOut_Diff(): This function checks the difference of incoming weights and outgoing weights. It uses the *FIn[]* for incoming weights and *FOut[]* for outgoing weights. It subtracts the outgoing weights from incoming weights. Then it moves the difference to *FI_FO_Diff[]*.

Temp_Sum(): This function adds the *FI_FO_Diff* values with *[1/n]*. The sum is stored in *TempSum[]*. The *Temp_Sum()* function calls *FIn_FOut_Diff()* function, the *FIn_FOut_Diff()* function calls *AddOut()* function, the *AddOut()* function calls *AddIn()* function. In this way all the calculations are done properly. The benefit of calling functions inside other functions is that we just call the *Temp_Sum()* function on a transition and all the calculations for *TempSum[]* are done properly.

AllSent(): It checks if the *snapshot requests* have been sent to all the healthy processes.

isFail(): A process uses this function to check the entry of a failed process in *F[]*. If record found then current process can not detect the failure of this process again.

Leader(): The *Leader()* function makes the new *leader* when a process detects failure of some other process. This function is very important because if the *leader* fails then the system needs a new *leader* to collect the weights and send *snapshot request* messages. This function makes the *leader* to a healthy process with minimum id. If the failing process is not the *leader* then this function again selects the previous *leader*.

SN_Active(): System uses this function at different transitions to check if a *snapshot* is already in progress. The function returns a true value if *snapshot* is already in progress otherwise returns a false value.

FDiFFCount(): It calculates the difference of *F[]* of *snapshot reply* sending process and the *F[]* of the *leader* when the *leader* receives the *snapshot reply* message.

D. The Automaton for SnapBuffer Process

The automaton for the *SnapBuffer* process is depicted in Fig. 10. The process is initiated four times to make four instances, each of which is triggered by a certain communication with the *Termination* process. The specification of the *SnapBuffer* process in the protocol has following communicative choices:

1) Receiving a *snapshot request* message from *leader* to store it.
2) Sending a stored *snapshot request* message to a *Termination* process.
3) Receiving a *snapshot reply* message from a *Termination* process to store it.
4) Sending a stored *snapshot reply* message to *leader*.

The process receives a *snapshot request* message through *failRequest?* channel. The guard makes sure that buffer is empty and can receive this message. After receiving this message the *IsEmpty* variable is assigned a false value to show that now buffer is non-empty. The *SnapBuffer* process delivers the stored *snapshot request* message to the recipient through *failRequestR!* channel. The guard makes sure that buffer contains a message for sending. After sending this message the variable *IsEmpty* is set true to show that now buffer is empty again. The *SnapBuffer* process receives a *snapshot reply* message through *failReply?* channel. The guard ensures that buffer is empty and can receive this message. After receiving this message the *IsEmpty2* variable is assigned a false value to show that now buffer is non-empty. The *SnapBuffer* process sends the stored *snapshot reply* message to the *leader*

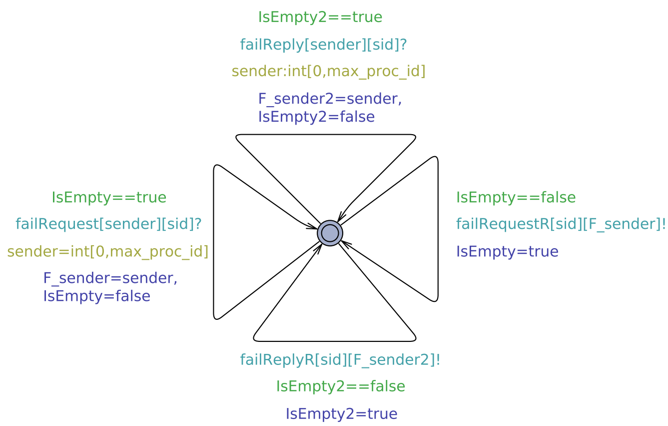


Fig. 10. Formal model for *SnapBuffer* process.

through *failReplyR!* channel. After sending the message the variable *IsEmpty2* is set true to show that now buffer is empty again.

The local declarations of *SnapBuffer* process are described in Table VII.

TABLE VII. LOCAL DECLARATIONS FOR *SnapBuffer* PROCESS

F_sender and *F_sender2*	These variables store the process id of the *snapshot request* sender and process id of the *snapshot reply* sender, respectively.
IsEmpty and *IsEmpty2*	Boolean variables which indicate free space for incoming *snapshot request* message and incoming *snapshot reply* message, respectively.

VI. FUNCTIONAL REQUIREMENTS

The functional requirements illustrate the behaviour of the system and explain what an intended system should do. In other words, they describe the functionality of the system. Every protocol has some functional requirements. We discuss and verify these requirements for both models separately.

A. Functional Requirements for Model1

We extract three functional requirements from the protocol for Model1. These are given as:

R1: No deadlock is supposed to be there except when the *leader* process is at *announce* state and all other processes are at *passive* state. This indicates that all the processes are idle and the *leader* has collected all the weights successfully resulting in proper termination of the system.

R2: This requirement states that after doing certain communications and collecting the weights of other processes, the *leader* process eventually reaches at *announce* state.

R3: According to this requirement, after doing certain message communications all the processes must be idle at *passive* state. All the processes eventually reach at *passive* state.

There are three invariants given in the protocol at page 12 of [7]. These invariants are expected to be preserved by the system.

INV1: In Model1, no process fails. It means all the processes are healthy. The process with minimum id is the *leader* of the system. This invariant is for all the healthy processes other than the *leader*. This invariant states that at any time, if a process is at *passive* state then it must have a zero weight. Similarly, if a process has zero weight then it must be at *passive* state.

INV2: This invariant is related to message sending. All processes can pass *basic* messages to each other. Every process can also send *control* messages to the *leader*. A process sends some weight with *basic* and *control* messages. This invariant describes that the weight sent with any message must be greater than zero.

INV3: At the start of the computation, every process is given an equal initial weight. A process sends some of its weight when it sends a *basic* or *control* message. Each process receives some weight when it receives that message. It updates its weight after sending or receiving a message. All processes also record their incoming and outgoing weights. This invariant states that for all healthy processes at any time, the sum of current weight and all outgoing weights of a process must be equal to the sum of its initial weight and all incoming weights.

B. Functional Requirements for Model2

For Model2, we extract three functional requirements.

R1: This requirement describes that some process reaches at the *announce* state to make sure that all the weights are collected and the system is terminated properly.

R2: A faulty process cannot be the *leader* of the system. This requirement is not satisfied. We discuss a scenario in which this requirement is violated. We have 4 instances of the *Termination* process. These are *p0*, *p1*, *p2* and *p3*. The *p0* is the *leader* of the system. The *leader* sends the *snapshot request* message to all the processes. These messages are yet stored in buffers and not delivered to the recipients. Meanwhile, the *leader* fails. The *p1* detects the *leader* to be faulty and calculates the new *leader* with minimum id from healthy processes. It becomes the *leader* itself. But in future, when it receives the stored *snapshot request* message sent by *p0*, it makes the *p0* as the *leader* of the system. The *p0* is faulty and is supposed to be the *leader* of the system again. That is why this is the clear violation to this functional requirement.

R3: This requirement states that every time the *leader* fails, the *snapshot* is called. The healthy process with minimum id calls the *snapshot*. But this requirement is trivially violated when the *p0* fails and the *p1* becomes passive without detecting the fault. The *p2* and *p3* also become *passive*. Now the *snapshot* is never called by any process.

The three invariants discussed for Model1 must be preserved for Model2 also.

VII. Formal Specification of Requirements

In this section, we describe formal specification of the requirements and invariants. We also present the formalism for these requirements.

A. Requirement Formal Specification for Model1

According to requirement R1, there should not be a deadlock if the *leader* is not at *announce* state or any of the other process is not at *passive* state. The formula for this requirement is given as.

```
A[] deadlock imply(Termination(0).announce and
        Termination(1).passive and
          Termination(2).passive)
```

The requirement R2 says that the system is terminated properly if the *leader* reaches at *announce* state for every path of execution. The formula for the requirement is given as:

```
A<> Termination(0).announce
```

All the processes must reach at *passive* state for proper termination of the system according to the requirement R3. Given below is the formula for R3:

```
E<> forall(i:id_t) Termination(i).passive
```

The formula for INV1 is presented below. A process moves from *passive* to *A2_1* state after receiving a *basic* message. Then this process updates its weight in next transition. It means, like *passive* state this process has a zero weight at *A2_1* state. So, we are including this state in the formula for INV1.

```
A[] ((Termination(1).passive imply
 Termination(1).weight[0]==0) and
(Termination(1).weight[0]==0 imply
Termination(1).passive or Termination(1).
A2_1)) and ((Termination(2).passive imply
 Termination(2).weight[0]==0) and
(Termination(2).weight[0]==0 imply
Termination(2).passive or Termination(2).
A2_1))
```

Now we discuss the formula for INV2. Each process updates two global variables W1 and W2 after receiving a *basic* or a *control* message. The W1 records the first value and W2 records the second value of received weight. If these variables always keep some non-zero value of weight then it means every message sent by any process has a non zero weight. The formula for this invariant is given as:

```
A[] W1!=0 and W2!=0
```

The invariant INV3 is for all healthy processes. The formula for this invariant is:

```
A[] forall(i:id_t)
Termination.In_Out_Equal(i)==true
```

This invariant INV3 uses five functions for its calculations. These functions are *AddInWeights()*, *InSum()*, *AddOutWeights()*, *OutSum()*, and *In_Out_Equal()*. All these functions perform combined functionality for verification of INV3. The function *AddInWeights()* adds all the incoming weights recorded in *In_arr[]* of calling process. The function *InSum()* adds the current weight and the some of incoming weights and stores the result in *tempIn[]*. The *AddOutWeights()* function adds all the outgoing weights recorded in *Out_arr[]* of a process. The function *OutSum()* calculates the sum of initial weight and all outgoing weights of a process and stores the result in *tempOut[]*. At the end *In_Out_Equal()* checks the equality of *tempIn[]* and *tempOut[]*. If both arrays are equal then this function returns a true value otherwise a false value.

B. Requirement Formal Specification for Model2

According to requirement R1, some process reaches at the *announce* at some time. The formula for this requirement is given as:

```
A<> exists(i:id_t) Termination(i).announce
```

The requirements R2 and R3 are clearly discussed with examples in Model2 requirements part. Now we discuss the formula for invariant INV1. Formalism for this invariant is given as:

```
A[] forall(i:id_t((Termination(i).notMin(i)==
  true and Termination(i).passive imply
    Termination(i).weight[0]==0) and
   (Termination(i).notMin(i)==true and
    Termination(i).weight[0]==0 imply
Termination(i).passive or Termination(i).A2_1))
```

The function *notMin()* checks if the calling process belongs to faulty set of processes or it is a healthy process with minimum id. It returns false if the process is faulty or it is healthy with minimum id. It returns true otherwise allowing other processes to check their weight at the *passive* state.

The formalism for INV3 for Model2 is similar to formalism of INV3 for Model1. All the functions and their definitions are same. But Model2 uses an extra function *Equal()* that checks if the calling process is faulty. It means we are just concerned with the calculations for healthy processes. If calling process is healthy then it returns true if the invariant is preserved and returns false if the invariant is violated. The formula for this invariant is given as:

```
A[] forall(i:id_t) Termination(i).Equal(i)==
true
```

VIII. Verification Results for Model1

This section shows the simulation results of formalism for functional requirements and invariants for Model1. These results are collected by executing the formulas in verifier of the UPPAAL toolset. For simplicity we use the *Buffer* instead of the *MessageBuffer* in all counterexamples. Results for Model1 are given below in Table VIII. We verify our system model for,

Total Number of *Termination* Process Instances = 3

Total Weight of the System = 1

TABLE VIII. REQUIREMENT RESULTS FOR MODEL1

Requirement	Status	Computational Time
R1*	Satisfied, 225504590 states	57600m23.345s
R2*	Satisfied, 32045480 states	7200m53.167s
R3	Satisfied, 18986 states	3.564s
INV1*	Satisfied, 28006031 states	7210m34.453s
INV2*	Satisfied, 25933265 states	7206m12.560s
INV3*	Satisfied, 20568945 states	7002m19.350s

Weight of Each Instance = 1/3

Requirement R3 is satisfied. For R1, R2, INV1, INV2 and INV3, state space is not fully explored. We are not sure about their final results. We explored these requirements for breadth first searching technique. Our claims are limited to the number of states and time given in Table VIII.

IX. VERIFICATION RESULTS FOR MODEL2

Results for Model2 are given below in Table IX. We verify our system model for,

Total Number of *Termination* Process Instances = 4

Total Weight of the System = 1

Weight of Each Instance = 1/4

TABLE IX. REQUIREMENT RESULTS FOR MODEL2

Requirement	Status	Computational Time
R1	Not satisfied	06.307s
INV1*	Satisfied, 25166435 states	7215m33.873s
INV2*	Satisfied, 42542339 states	11520m21.212s
INV3	Not satisfied	20.353s

In Model2, for INV1 and INV2, state space is not fully explored. We are not sure about the final results of these invariants. We explored these requirements using breadth first searching technique and we claim these results to the number of states given in Table IX. The requirement R1 is not satisfied. The counterexample for this violation is shown in Fig. 11. The *p3* sends a *basic* message to *p0*. The *p1* and *p2* send *control* messages to *p0*. Then *p0* sends a *basic* message to *p3*. After this, *p3* receives the *basic* message sent by *p0* and sends the *control* message to *p0*. After doing all these communications, *p0* fails and other three processes become *passive*. In this case, there is no process that is *active* and detect *p0* as faulty. So, it is not possible to collect the weights carried by *p0*. Therefore, the *announce* state is not reachable in this scenario.

The invariant INV3 violates in the given scenario. The *p0* sends a *basic* message to *p1* and becomes *passive*. The *p1* sends a *basic* message to *p0* and reaches at *A1* state. The *p2* sends a *control* message to *p0* and visits *A3* state. After this, *p0* receives the *control* message of *p0* and becomes *passive*. At *passive* state, it receives the *basic* message from *p1* and reaches at *active* state. In the whole activity *p3* stays at *active* state. The invariant is not satisfied in this situation. The counter example for this violation is shown in Fig. 12.

X. LIMITATIONS AND CHALLENGES

There are some limitations for verification of intended termination detection protocol. We limit the concurrent *termination* processes to three in case of fault-free distributed system and four for faulty distributed system. The *basic* message sending limit for any process is two. The models generate a huge state space with millions of states. The purpose of these limitations is to reduce the state space of our computations. The machines and servers used in our verification have limited resources for memory and speed. We performed some computations on the machine with 16GB RAM, core i7(4th Gen) CPU and 3.4 GHz clock speed. We also performed some computations on mammoth server (tue.nl) which has 56 machines with 2 GHz speed and a total of 935GB RAM.

We faced some challenges during the modeling and verification of the protocol. The total weight of the system is 1 and initial weight of each process is *1/n* where *n* is the total number of processes. The expression *1/n* returns fractional values. At the end of the computation it becomes hard to make the sum of all weights to 1 due to possible rounding errors. The first challenge was to manage a single weight in the form of a pair of integers. We presented the weight *1/n* in the form of two values as *[1,n]*. This method created many difficulties for accurately calculating the incoming, outgoing and current weights of the processes. This was also a big challenge to correctly model the protocol and its invariants in UPPAAL. We used model abstraction for reducing the state space that was also a challenge.

XI. CONCLUSION

We formalised both parts of termination detection protocol as specified in [7] in the timed-automata-theoretic formalism of UPPAAL. We formally specified and performed verification analysis of the protocol with respect to its functional requirements and invariants. During our formal analysis, we found some scenarios in which the protocol does not meet its functional requirements. Counter example are there to witness the claimed results.

REFERENCES

[1] W. Fokkink, *Distributed algorithms: an intuitive approach*. MIT Press, 2013.

[2] K. M. Chandy and J. Misra, "Asynchronous distributed simulation via a sequence of parallel computations," *Commun. ACM*, vol. 24, no. 4, pp. 198–206, Apr. 1981. [Online]. Available: http://doi.acm.org/10.1145/358598.358613

[3] K. M. Chandy, J. Misra, and L. M. Haas, "Distributed deadlock detection," *ACM Trans. Comput. Syst.*, vol. 1, no. 2, pp. 144–156, May 1983. [Online]. Available: http://doi.acm.org/10.1145/357360.357365

[4] G. Tel and F. Mattern, "The derivation of distributed termination detection algorithms from garbage collection schemes," *ACM Trans. Program. Lang. Syst.*, vol. 15, no. 1, pp. 1–35, Jan. 1993. [Online]. Available: http://doi.acm.org/10.1145/151646.151647

[5] E. W. Dijkstra and C. Scholten, "Termination detection for diffusing computations," *Information Processing Letters*, vol. 11, no. 1, pp. 1–4, 1980. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0020019080900216

[6] N. Francez, "Distributed termination," *ACM Trans. Program. Lang. Syst.*, vol. 2, no. 1, pp. 42–55, Jan. 1980. [Online]. Available: http://doi.acm.org/10.1145/357084.357087

[7] Y. Tseng, "Detecting termination by weight-throwing in a faulty distributed system," *J. Parallel Distrib. Comput.*, vol. 25, no. 1, pp. 7–15, 1995. [Online]. Available: https://doi.org/10.1006/jpdc.1995.1025

[8] X. Wang and J. Mayo, "A general model for detecting distributed termination in dynamic systems," in *18th International Parallel and Distributed Processing Symposium, 2004. Proceedings.*, April 2004, p. 84.

[9] G. Tel and F. Mattern, "The derivation of distributed termination detection algorithms from garbage collection schemes," *ACM Trans. Program. Lang. Syst.*, vol. 15, no. 1, pp. 1–35, Jan. 1993. [Online]. Available: http://doi.acm.org/10.1145/151646.151647

[10] F. Mattern, H. Mehl, A. A. Schoone, and G. Tel, "Global virtual time approximation with distributed termination detection algorithms," Tech. Rep., 1991.

[11] S. Chandrasekaran and S. Venkatesan, "A message-optimal algorithm for distributed termination detection," *J. Parallel Distrib. Comput.*, vol. 8, no. 3, pp. 245–252, mar 1990. [Online]. Available: http://dx.doi.org/10.1016/0743-7315(90)90099-B

[12] J. Pang, *Analysis of a Security Protocol in μCRL*, C. George and H. Miao, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. [Online]. Available: http://dx.doi.org/10.1007/3-540-36103-$0_4$0

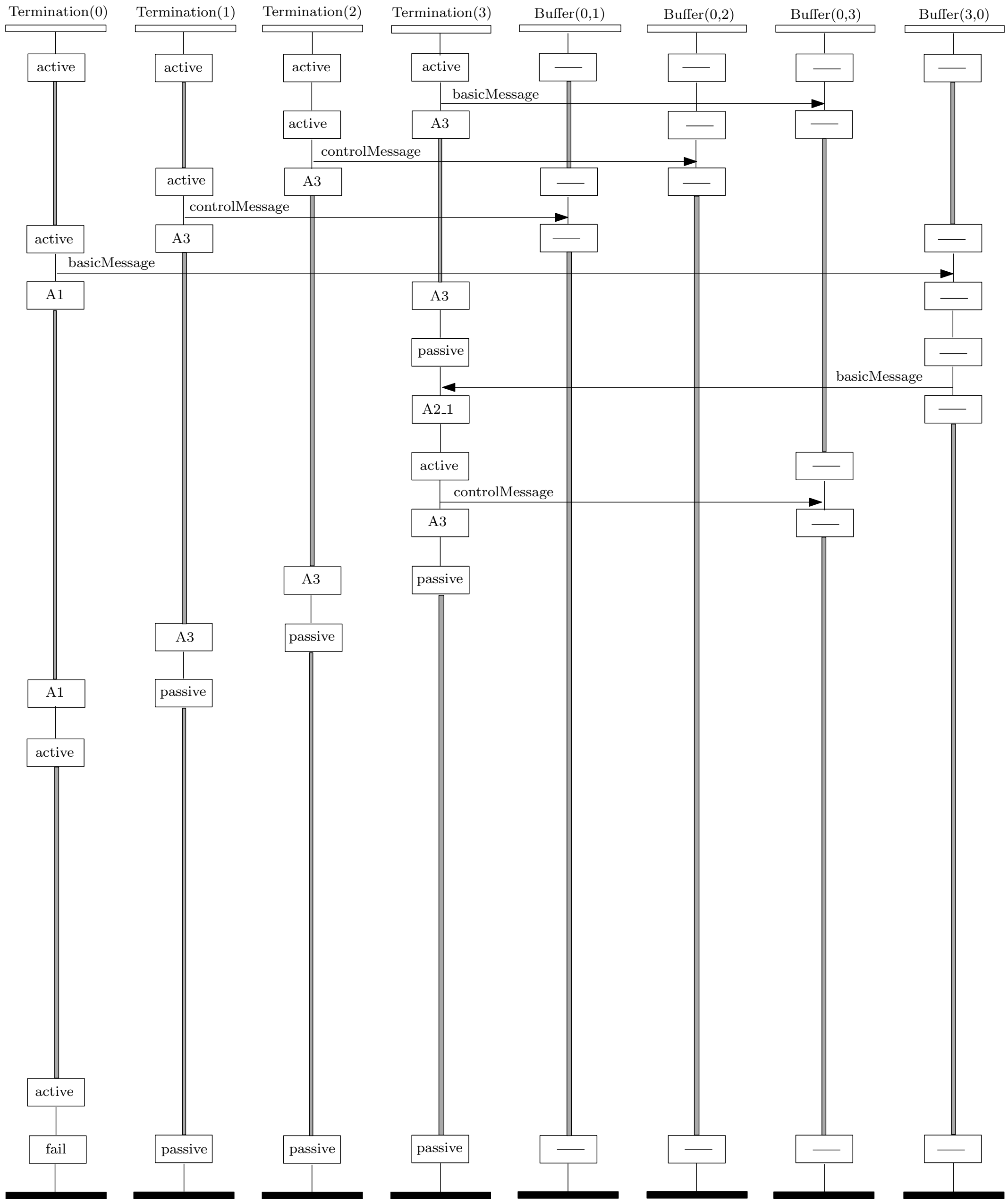


Fig. 11. Trace for R1 in Model2.

[13] W. H. J. Feijen and A. J. M. van Gasteren, *Shmuel Safra's Termination Detection Algorithm*. New York, NY: Springer New York, 1999. [Online]. Available: http://dx.doi.org/10.1007/978-1-4757-3126-2$_2$9

[14] N. Mittal, S. Venkatesan, and S. Peri, *Message-Optimal and Latency-Optimal Termination Detection Algorithms for Arbitrary Topologies*, R. Guerraoui, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-30186-8$_2$1

[15] N. R. Mahapatra and S. Dutt, "An efficient delay-optimal distributed termination detection algorithm," *Journal of Parallel and Distributed Computing*, vol. 67, no. 10, pp. 1047–1066, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0743731507000998

[16] J. S. Ostroff, "Formal methods for the specification and design of real-time safety critical systems," *J. Syst. Softw.*, vol. 18, no. 1, pp. 33–60, Apr. 1992. [Online]. Available: http://dx.doi.org/10.1016/0164-1212(92)90045-L

[17] A. O. Gomes and M. V. M. Oliveira, *Formal Specification of a Cardiac Pacing System*, A. Cavalcanti and D. R. Dams, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-05089-3$_4$4

[18] Y. K. H. Lau, "The design of distributed safety critical software using csp," in *IEE Colloquium on Safety Critical Software in Vehicle and Traffic Control*, Feb 1990, pp. 8/1–8/5.

[19] K. G. Larsen, P. Pettersson, and W. Yi, "Uppaal in a nutshell," 1997.

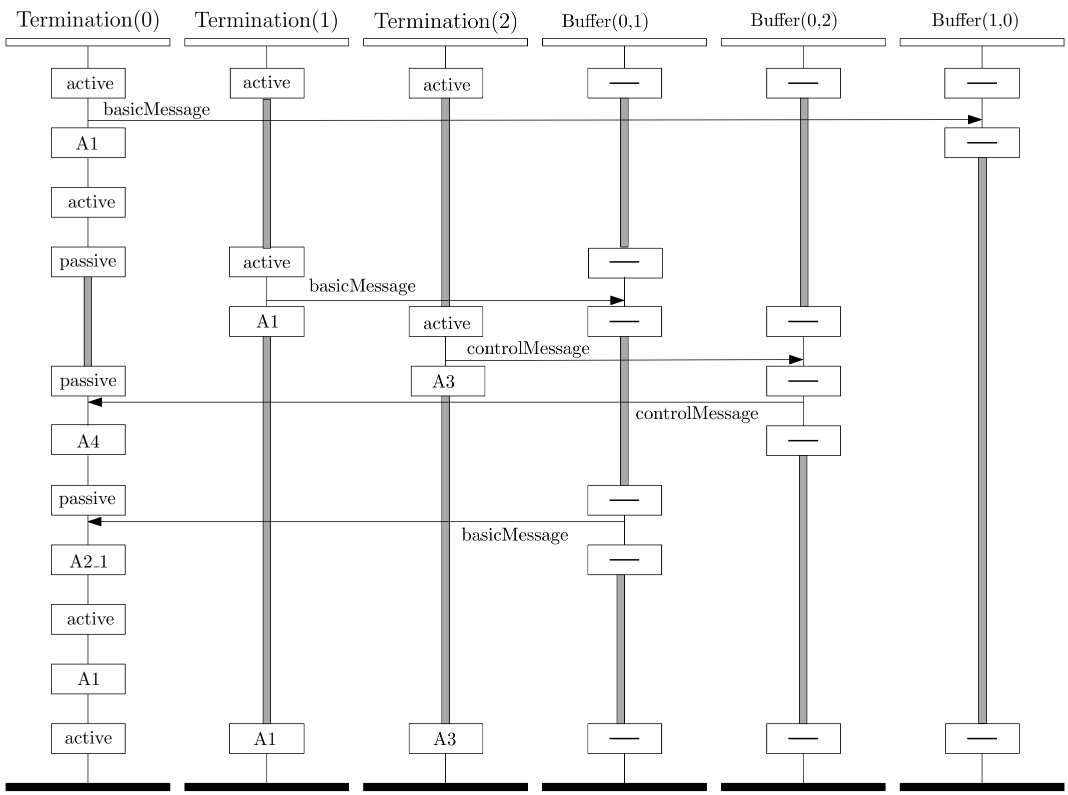


Fig. 12. Trace for INV3 in Model2.

Trajectory based Arabic Sign Language Recognition

Ala addin I. Sidig, Sabri A. Mahmoud
Collage of Computer Science and Engineering
King Fahad University of Petroleum & Minerals
Dhahran 31261, Saudi Arabia

***Abstract*—Deaf and hearing impaired people use their hand as a tongue to convey their thoughts by performing descriptive gestures that form the sign language. A sign language recognition system is a system that translates these gestures into a form of spoken language. Such systems are faced by several challenges, like the high similarities of the different signs, difficulty in determining the start and end of signs, lack of comprehensive and bench marking databases. This paper proposes a system for recognition of Arabic sign language using the 3D trajectory of hands. The proposed system models the trajectory as a polygon and finds features that describes this polygon and feed them to a classifier to recognize the signed word. The system is tested on a database of 100 words collected using Kinect. The work is compared with other published works using publicly available dataset which reflects the superiority of the proposed technique. The system is tested for both signer-dependent and signer-independent recognition.**

***Keywords*—Trajectory processing; sign language recognition; ensemble classifier; polygon description; parameters tuning; signer independent**

I. INTRODUCTION

Communicating thoughts and feelings is an essential need for human beings. Hearing disabilities hinder the natural speech based communication. To communicate with each other and with speaking people, deaf has invented nonverbal languages that use descriptive gestures to convey their thoughts. These languages are developed by the deaf communities in different regions of the world. Sign languages are full featured languages with their own vocabularies and grammar. They make use of hands-motion, fingers-configurations, facial-expressions, and body lane in parallel to express different terms. Unfortunately, speaking people find it hard to learn these languages which increases the barrier between them and the deaf community. To communicate with deaf, speaking people need skilled professional translators that knows the spoken and signed languages. These skilled translators are few and can‘t be available all the time. Sign language recognition systems tries to fill this gap by exploiting the advanced technologies to automatically translate signed language to a form of spoken language such as text or speech. To effectively translate a signed language all its components need to be considered. Of these components, the hands-motion is one of the most important modalities of signed language.

This work proposes to use the 3D trajectory of hands to recognize signs. The 3D trajectory, in contrast to 2D, provides information about the front-back hand motion. We record both of 2D and 3D trajectory using Kinect device. The proposed system is composed of three stages: Preprocessing, Features representation, and Classification. The preprocessing stage removes the noise and compresses the trajectory to form a polygon. The compression is done by finding N key points that represent the polygon corners. The feature representation stage builds a features vector that describes this polygon. These features are used to train and test different classifiers to recognize the signs in the third stage. Fig. 1 shows the pipeline of the proposed system. The main contributions of this work are:

- Propose a trajectory based sign language recognition system.
- Propose a trajectory compression algorithm.
- Propose a two features representations for 2D and 3D trajectories applied to signer dependent and independent recognition.

The rest of this article is organized as follows: Section II describes some of the related works. Followed by Section III on the trajectory preprocessing. Features representation is described in Section IV. Then Section V on classification. Experimental evaluation is shown in Section VI. Finally, we conclude this article in Section VII.

II. RELATED WORKS

Arabic sign language recognition is addressed by many researchers using different scales and strategies. The work on Arabic sign language recognition in the literature can be classified into three levels. Arabic sign alphabets and numbers recognition level [1]–[4], isolated words recognition level [5]–[9], and sentences recognition level [10]–[12]. This work proposes a system for isolated words recognition based on hands‘‘ trajectories. Trajectory processing exists in a wide range of applications. Therefore, a lot of work is done on trajectory processing in on-line character recognition [13], [14], action recognition [15], [16], gesture recognition [17] and more.

Lin and Hsieh in [18] proposed a kernel based trajectory representation using Kernel Principal Component Analysis (KPCA) and Nonparametric Discriminant Analysis (NDA). In their method a 2D/3D trajectory is first min-max normalized then projected to higher dimensional space using KPCA. The dimensionality is reduced using NDA with the hope of maximizing the interclass variability and minimizing the within-class variability. The resulting representation is hoped to be more discriminative. The classification is done using the nearest neighbor rule. The approach is tested on a limited set of 38 words from the Australian sign language and reported accuracy of 69% for 2D trajectory and 78% for 3D.

Naftel and Khalid in [19] encoded the 2D trajectory along x and y dimensions using Discrete Fourier Transform (DFT)

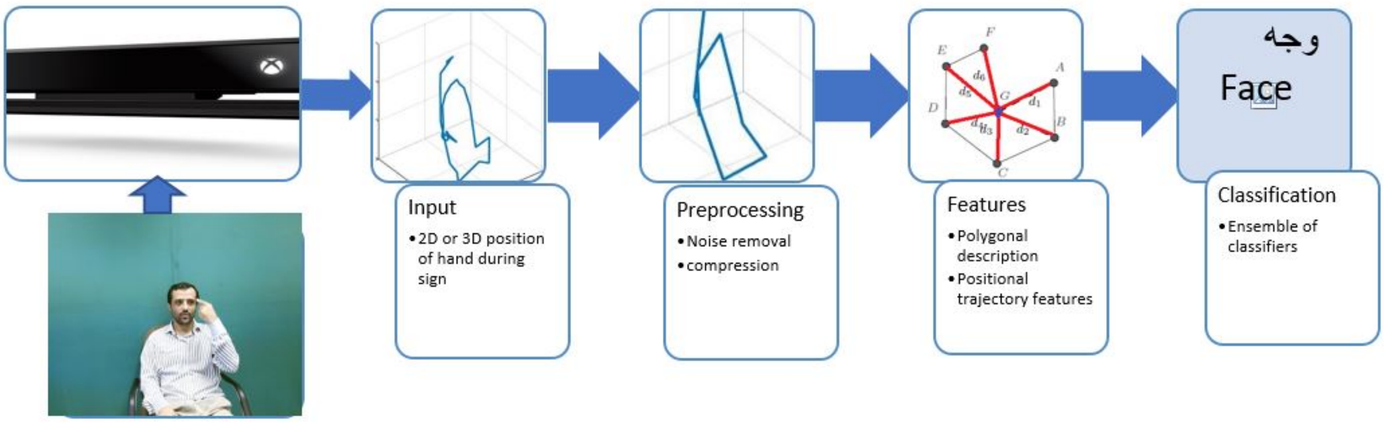


Fig. 1. The stages of the proposed system. The signer performs the sign in front of the Kinect which tracks the hands. The resulting trajectory is then fed to the system.

separately. Then the first four coefficients are used as feature vector that represent the trajectory. The coefficients are then clustered using Self Organized Map (SOM). They tested the approach on 24 words from Australian sign language and reported an accuracy of 70.1%.

Pu et al. [20] modeled the trajectory as a sequence of M sub-motions and used HMM to model the transition between these sub-motions. For each point on the sub-motion trajectory, they find the *shape context* as a histogram of relative coordinates of other points on the sub-motion trajectory. Then a codebook is generated from these shape contexts. The features vector of each sub motion curve is composed as a weighted histogram of the code book centers. The wights are found by soft clustering the shape context of each point. Finally the sign curve feature is a sequence of M sub-motion features. They tested the system on a database of 100 signs from the Chinese sign language and reported an accuracy of 67.3% for signer dependent and 54.4% for signer independent.

Boulares in [21] extracted signatures from 3D hands trajectories and used SVM to classify different signs. To extract trajectory signature, they used non linear regression to fit the trajectory points to a conic section. The trajectory signature along with hand shape and other features is used to train and test SVM classifier. Curve fitting does not accurately represent complex trajectories that include cycles.

Geng et al. in [22] used a combination of trajectory modeling and hand shape representation as a feature to train an Extreme Learning Machine (ELM) classifier. A combination of 3D trajectories of hand, wrist, and elbow are used. They normalized the values of trajectory points to [0 , 1] range and smoothed the trajectory by average convolution. To form a feature vector from the smoothed trajectory, they subtract the starting point of the trajectory from all following points. The difference between the hand trajectory and wrist trajectory is represented by spherical coordinates system and similarly for the hand-elbow trajectory difference. The final features vector is concatenation of hand trajectory, hand-wrist spherical difference, hand-elbow spherical difference, and hand shape features from depth image. These features are used to train ELM and 82.8% accuracy is reported on a limited database of 8 words from the Chinese sign language. Normalization of trajectory points to the range of [0 , 1] results in loss of information about where was the hand motion with respect to body when signing the word.

Wang et al. in [23] formed the trajectory of hands as a combination of hands location and orientation. The hand location is defined as the hand location with respect to the face centroid and with respect to the non dominant hand location. Similarly, the orientation is defined as the direction between successive hand locations. For single handed signs the trajectory of non dominant hand is set to zeros. All trajectories are normalized to have the same length. Similarities between trajectories are measured by dynamic time wrapping (DTW). Based on the trajectory matching the top 10 accuracy of the sign search results is about 74% and was improved to 78% when incorporating additional hand shape feature. They slightly modified the trajectory feature in [24] by including the hand velocity and defining separate feature for single handed signs doesn't include the hand location with respect to non dominant hand location. However the information of single or two handed sign need to be given by the user.

Bhuyan et al. in [25] modeled the trajectory as a combination of shape and motion features. The shape features include, the trajectory length, and the number of curves in the trajectory. The motion features include, the average speed, standard deviation of the speed, and the number of minima in the velocity. The classification of gestures is done in two stages. First candidate signs are included based on the trajectory shape similarities using maximum boundary deviation as similarity measure. In the second stage trajectories are aligned using DTW then the trajectory features are classified based on the nearest candidate template.

Mohandes and Deriche proposed a system for Arabic sign language recognition [26]. The trajectory is composed of 3D position and orientation with 12 dimensional vector for both hands. For each dimension the acquired readings are partitioned into 5 equal partitions. From each partition the mean and standard deviation is calculated. That results in 120 dimensional features vector. LDA is used to reduce the dimensionality to 20. The nearest neighbor classifier is used to find the class of a sign. They reported an accuracy of 84.7% on a dataset of 100 words.

III. Preprocessing

In this work, Kinect is use to record signs. A synchronized color image, depth image, and 25 body joints locations are recorded. For each joint the 3D locations of joints and the 2D mapping to both color and depth images are recorded. For this work the sequence of hands locations in 3D is used to recognize signs.

Trajectory preprocessing includes: Noise removal and Compression. The joints‘ locations obtained by Kinect are noisy and include some outliers. The noise removal stage smooths out these outliers by using median filter. Since the frame-rate for recording is at 30 frames per second, fine details of part of second trajectory is not very useful and results in redundant information. Trajectory compression stage compresses the trajectory into few key points. To find such key points the trajectory is treated as a polygon formed by connecting the locations of the hand while signing. The key points are obtained by reducing the number of vertices of this polygon to a specific number. The reduction is done by recursively calculating the importance of each vertex based on angle and segment length and then removing the least important. The process is repeated until the desired number of vertices is reached. Fig. 2 shows the calculation of vertex importance. The algorithm for trajectory compression is shown in Algorithm 1. Fig. 3 shows the effect of 3D trajectory preprocessing. The preprocessing of a 2D trajectory is shown in Fig. 4.

Algorithm 1 TrajectoryCompression

procedure Compress(Traj,NumVers)
  $TrajLength \leftarrow \text{Length}(Traj)$
  for all points v in Traj **do**
    $IMP(v) \leftarrow \text{CalcImpotrance}(Traj, v, NumVers)$
  end for
  while $TrajLength > NumVers$ **do**
    $I \leftarrow IndexOfMin(IMP)$
    $Traj \leftarrow Traj - Traj(I)$ ▷ The - sign is set difference
    $IMP \leftarrow IMP - IMP(I)$
    $TrajLength \leftarrow TrajLength - 1$
    update IMP by recomputing the importance of the removed vertex’s neighbors.
  end while
  function CalcImpotrance(Traj,v,NumVertices)
    $imp \leftarrow D_{vp} \times D_{va} \times \Theta$
    return imp
  end function
end procedure

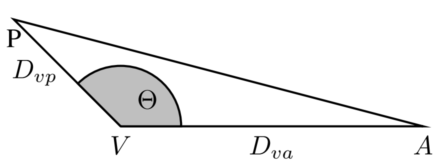


Fig. 2. The importance calculation for vertex V is found by multiplying the distances from v to adjacent vertices P (previous), A (after) and the angle Θ as $IMP_v = D_{vp} \times D_{va} \times \Theta$.

Some of the previous works as stated in Section II include another stage in preprocessing called min-max normalization. In this stage the trajectory is nominalized to be in [0-1] range. In this work, such stage is exclude arguing that it leads to loss of discriminative features. Signs can have similar trajectory pattern but at different locations. Min-max normalization leads to loss of the localization feature of the trajectory.

IV. Features Representation

After noise removal and compression, features are extracted from each sign trajectory. Here we describe two types of features.

A. Polygon Description

In this method the 3D hand trajectory is represented as a polygon. The description of this polygon is represented by: it is center of gravity and the distances from the perimetric points to the center of gravity point. The center of gravity point is approximated by the mean of perimetric points calculated as $G = (\bar{x}, \bar{y}, \bar{z})$ where $\bar{r} = \frac{1}{N}\sum_{i=1}^{N} r_i$ and N is the number of perimetric points. The distance from G to permetric points is calculated using the Euclidean distance formula $d_i = ||G - P_i||, i = 1, 2, 3, ..., N$. Fig. 5 illustrates the polygon description procedure.

Then the polygonal description feature is formed by concatenating G and d_i as

$$F = [\bar{x}, \bar{y}, \bar{z}, d_1, d_2, d_3, ..., d_n]$$

This feature representation captures both of the trajectory shape and more importantly the position of hand motion. The position of hand motion is important as it distinguishes between signs with similar trajectories but different body positions.

B. Positional Trajectory Feature

In this feature representation only perimetric points of the trajectory polygon are included. The feature vector is a concatenation of perimetric points formed as

$$F2 = [x_1, y_1, z_1, x_2, y_2, z_2, ..., x_N, y_N, z_N]$$

This feature representation although is simple, but have shown very good discrimination and generalization as will be shown in the experimental results section.

V. Classification Techniques

After preprocessing and features representation of all trajectories at hand, features are used to train and test classifiers. In this work several classifiers are tested and the best accuracy is obtained when using ensemble of classifiers. Specifically, the best performing classifier is Ensemble Subspace KNN.The tested classification algorithms are listed in Table I. We use five folds cross validation.

In subspace ensemble algorithm, a set of N weak learners each is trained on a randomly chosen partition of the features vector of M dimensions less than the D dimensions of the original feature vector. On prediction, the average score from weak learners is calculated and the class with the highest average score is chosen as the true class [27]. This work used KNN as a weak learner to build the ensemble subspace

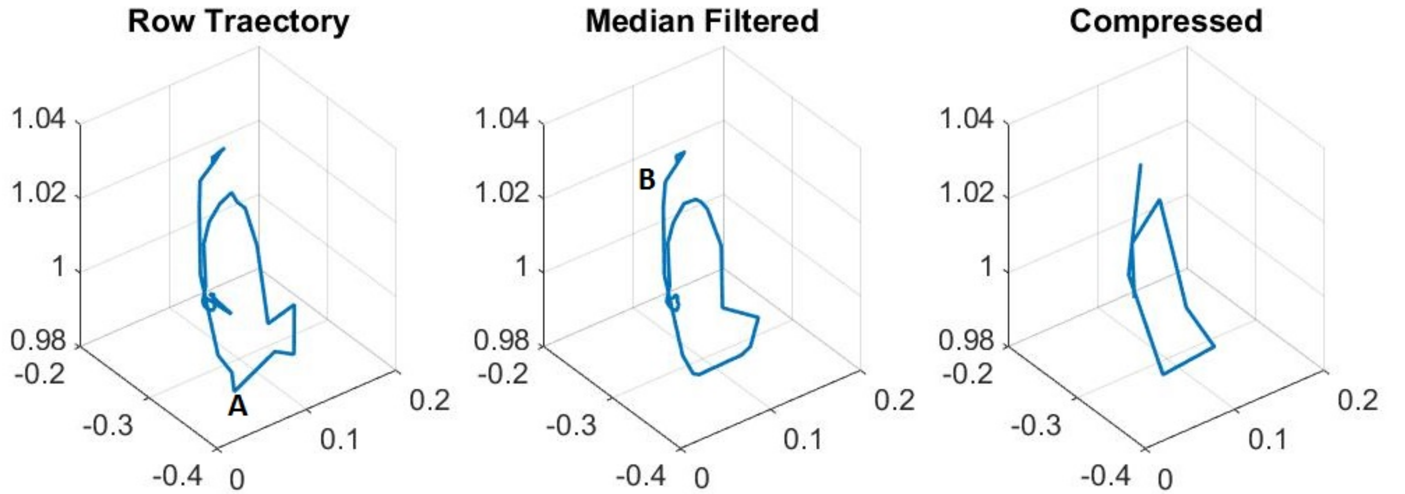


Fig. 3. The preprocessing stage of the trajectory. 'A' is a noisy point smoothed out by the median filter. 'B' is a less important point removed by compression stage.

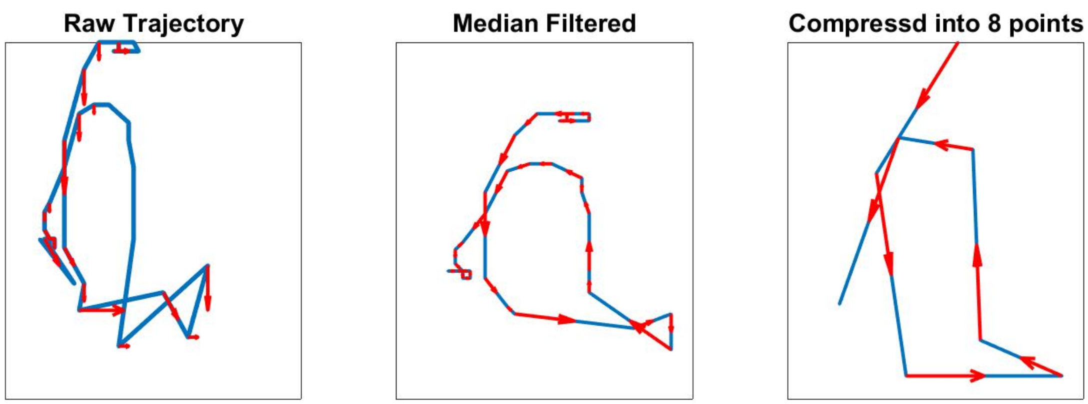


Fig. 4. The preprocessing stage of a 2D trajectory. The median filter reduce the noise of the trajectory resulting in smoother one. The compression stage finds the most important 8 points in the trajectory. The arrows indicate the direction of motion.

TABLE I. LIST OF CLASSIFIERS USED IN THE EXPERIMENTS

Tree	Linear Discriminant	Quadratic Discriminant
SVM Linear	SVM Quadratic	SVM Cubic
SVM Gaussian	KNN Euclidean	KNN Cosine
KNN Cubic	Ensemble Boosted Trees	Ensemble Bagged Trees
Ensemble Subspace Discriminant	Ensemble Subspace KNN	Ensemble RUSBoosted Trees

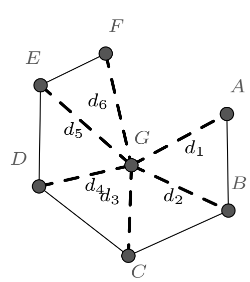


Fig. 5. The polygon description feature is found by the center of gravity G and distances $[d_1, d_2, d_3, d_4, d_5, d_6]$ form G to perimetric points [A,B,C,D,E,F] respectively.

classifier. It is clear that N, M and K (of the KNN) are hyper parameters that need to be chosen for best performance of the classifier. To find the best values for these parameters cross validation is used as shown by Algorithm 2.

The algorithm first runs KNN with different values of K to find the best performing one (BestK). Then it fixes the number of weak classifiers to 100 and K to BestK and searches for the best number of partitions, BestM. With BestK and BestM the algorithm then searches for best number of weak learners BestN.

VI. EXPERIMENTAL RESULTS

A set of experiments are carried out to evaluate each stage of the proposed system. Starting by the preprocessing stage to

Algorithm 2 Fine Tune parameters of Ensemble Subspace KNN

function FINETUNE(SampleFeatures)
  $S \leftarrow NumberOfSamples$
  $D \leftarrow NumberOfDimentions$
  $KCanidates \leftarrow$ set of 10 values between 1 and $\log S$
  for all k in KCanidates **do**
    $Loss(k) \leftarrow$ CROSSVALIDATEKNN$(SampleFeatures, KCanidates[k])$
  end for
  $BestK \leftarrow KCanidates[MinimalLoss]$ ▷ Find the best K
  $MCanidates \leftarrow$ set of 10 values between 1 and D
  $N \leftarrow 100$ ▷ Fixed Number of weak classifiers
  for all m in MCanidates **do**
    $Loss(m) \leftarrow$ CROSSVALIDATEENSEMKNN$(SampleFeatures, MCanidates[m], BestK, N)$
  end for
  $BestM \leftarrow MCanidates[MinimalLoss]$ ▷ Find the best M
  $NCanidates \leftarrow$ set of 100 values between 1 and 100
  for all n in NCanidates **do**
    $Loss(m) \leftarrow$ CROSSVALIDATEENSEMKNN$(SampleFeatures, BestM, BestK, NCanidates[n])$
  end for
  $BestN \leftarrow NCanidates[MinimalLoss]$ ▷ Find the best N
  return $BestK, BestM, BestN$
end function

the classification stage to fine tune the hyper parameters and then test the sign language recognition.

A. Arabic Sign Language Dataset

To our knowledge, there is no public dataset for Arabic sign language, so we collected a dataset of 100 words from the health chapter of Arabic sign language dictionary [28]. The dataset is recorded using Kinect to record synchronized color video, depth data, and 25 skeletal joints of body. The dataset was recorded by 3 signers repeated each sign 50 times on different sessions. For this work, only the hands joints‘ trajectories are employed to recognize signs. A list of the words in this database are shown in Table VI.

B. Effect of Trajectory Compression

This section investigates the effect of the number of vertices used to represent the trajectory as a polygon on the accuracy. This experiment used the trajectories of all signs performed by one signer and apply the preprocessing stage by varying the number of vertices from 4 to 18. Fig. 6 shows the classification error rates for different representations of the trajectory features. In this figure, F1 represent the *polygon description* feature representation of trajectory (see Section IV-A) while F2 stands for the *positional trajectory* feature representation. The 1H and 2H encodes the usage of only one hand trajectory or both hands respectively in building the feature vector. In 1H the features encode only the trajectory of the dominant hand while in 2H a concatenation of features that encode both hand trajectories is used. The 2D and 3D for which trajectory points representation being used, X-Y or X-Y-Z respectively. From this figure, many properties can be inferred. First, the best average accuracy can be obtained when using a polygon with 12 vertices. Using small number of vertices does not capture the complex trajectories well, and using very high number of vertices includes noisy details that mix up distinct classes. Second, the usage of 3D trajectory always performs better than the 2D one. This can be attributed to the fact that the Z dimension captures front-back motion of hands, and there are some signs in the database with only front-back motion pattern. Third, the inclusion of non-dominant hand in the feature representation increases the discrimination power. The state of non-dominant hand in sign language can either be static, mirrors the motion of dominant hand, or moving in different way than the dominant hand. In all cases of non-dominant state, its motion pattern helps in distinguishing similar signs that are of similar dominant hand trajectory. Forth, as a comparison between the two features representation the *positional trajectory* feature representation outperforms the *polygon description* feature representation of the trajectory.

C. Fine Tuning EnsembleSupspaceKNN Classifier

This experiment applied Algorithm 2 on the same set used in Section VI-B to find the best parameters for each feature representation. Table II lists the best parameters' settings for each feature representation. In this table the best value for K is 1 for all features, the best value for M for feature F1 is roughly half D which is similar to the findings in [27]. The values in BestN column are for the value of N after which no significant drop in loss is seen. Based on this table, the parameters settings for following experiments will be: K=1, N=40, M= BestM from the table.

TABLE II. BEST PARAMETERS FOR ENSEMBLESUPSPACEKNN CLASSIFIER

Feature	D	BestK	BestM	BestN
F1-1H-2D	14	1	8	40
F1-1H-3D	15	1	9	25
F1-2H-2D	28	1	13	40
F1-2H-3D	30	1	14	25
F2-1H-2D	24	1	6	40
F2-1H-3D	36	1	9	40
F2-2H-2D	48	1	6	40
F2-2H-3D	72	1	9	40

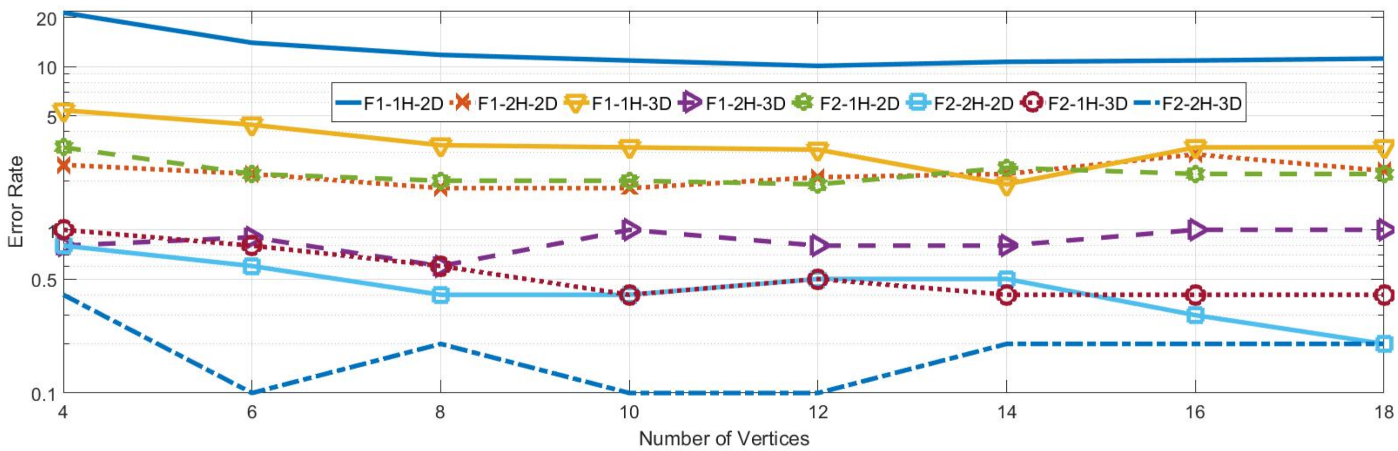


Fig. 6. The compression effect on the classification error rate for different versions on the proposed features representation. The Y axis is log scaled for better visualization.

D. Evaluation of the Proposed Features

After choosing the best trajectory compression ratio and the best parameters settings for the classifier, the system is tested on the collected database. Table III lists the recognition rates obtained when using each feature representation for each signer in the database. The results reflect that the 3D trajectory is more informative and discriminative than the 2D one, and the inclusion of non dominant hand status improves the accuracy for both types of trajectories. The third signer shows better accuracies than the other two which can be attributed to the less variability in his performance of signs, and the samples used for fine tuning the hyper parameters are performed by him. The fifth column lists the accuracies when using mixed samples from all signers for both training and testing. This shows the scalability of the system to larger number of samples and different signers.

TABLE III. SIGNER DEPENDENT CLASSIFICATION RECOGNITION RATE

Feature	Signer1	Signer2	Signer3	All Signers
F1-1H-2D	89.8	88.8	91.5	84.4
F1-1H-3D	96.2	95.0	97.6	94.7
F1-2H-2D	97.6	96.8	97.9	96.4
F1-2H-3D	99.3	98.8	99.4	99.5
F2-1H-2D	97.7	96.0	98.2	95.6
F2-1H-3D	99.2	99.0	99.8	99.2
F2-2H-2D	99.5	98.9	99.7	99.1
F2-2H-3D	99.7	99.6	100	99.7

Although the number of signers is not big enough to evaluate the system for signer independent recognition, experiments are done to get initial intuition about the generalization of the system to unseen signer. Table IV lists the accuracies of the different types of features in signer independent mode. Each column is named by the test signer when the training is done by samples performed by the other two signers. The lower results of the second signer are due to the different signing style, some signs are repeated more than once in the same sample. Overall average performance is around 53% for all features 48%, and 57% for F1, and F2 features, respectively.

TABLE IV. SIGNER INDEPENDENT CLASSIFICATION RECOGNITION RATE

Feature	Signer1	Signer2	Signer3
F1-1H-2D	40.1	27.2	43.1
F1-1H-3D	44.6	30.7	50.9
F1-2H-2D	57.5	48.6	60.1
F1-2H-3D	60.0	51.0	64.8
F2-1H-2D	56.8	41.9	58.7
F2-1H-3D	60.2	43.2	65.7
F2-2H-2D	58.9	47.8	63.9
F2-2H-3D	61.3	49.7	64.4

E. Comparison with Published Work

This experiment tests the proposed features representation and classification algorithm on a publicly available dataset and compares the results of the proposed method with published work on the same dataset. The dataset is composed of 95 Australian sign language words. Each word is performed by 1 signer 27 times. For each sample a vector of 22 measures is recorded per frame. These measures include the 3D position of hands (X,Y,Z), the orientation of hands (Roll, Pitch, Yaw), and the status of fingers. Some previous work used only the (x,y) points to form 2D trajectory while others used 3D. This work, uses the 2D/3D trajectory as well as the hand orientation. The same steps of trajectory preprocessing, features representation, and classification are applied on this database. In this dataset, the signer starts with his hands on the rest position and return them back to the rest position after signing. This makes the center of gravity of some signs to be the same. To avoid that, the compression stage is applied twice. First with 14 vertices which will include the starting and ending rest position. Then it finds the 12 vertices after excluding the first and last points which results in removing the rest position from the calculation of the center of gravity. Table V shows the accuracy reported by different previous works along with our work (the last 4 lines). The first row shows the number of classes out of 95 used. In this table, F1 stands for the polygonal description feature representation and F2 for the positional feature. 3D stands for the only use of 3D hand position to form the feature while 3DO for inclusion of the hand orientation too.

Note that the work in [33] uses the 22 features while ours use three - in case of 3D feature - or six - in case of 3DO -

TABLE V. COMPARISON WITH PUBLISHED WORK ON AUSLAN

Reference	2 Words	4 Words	8 Words	16 Words	29 Words	38 Words	All Words
Khalid et al. [29] 2D	98%	92%	88%	83%	-	-	-
Khalid et al. [30] 2D	99%	95%	92%	88%	-	-	-
Bashir et al. [31] 2D	96%	92%	86%	78%	69%	66%	-
Wu et al. [32] 3D	93%	89%	83%	-	-	-	-
Naftel et al. [19] 2D	96%	90%	82%	76%	-	-	-
Wei et al. [18] 2D	98%	93%	86%	78%	72%	69%	-
Wei et al. [18] 3D	99%	96%	92%	89%	82%	78%	-
Simao et al. [33]	-	-	-	-	-	-	86.7%
F1-2D	100%	100%	95.4%	76.7%	63.7%	58.8%	46.8%
F1-3D	100%	100%	98.1%	90.4%	76.4%	70%	58.3%
F1-3DO	100%	100%	99.1%	95.4%	89.5%	86.5%	82.8%
F2-2D	100%	100%	96.3%	85.2%	74.3%	68.8%	61.7%
F2-3D	100%	100%	99.1%	94.8%	86.2%	79.7%	74.5%
F2-3DO	100%	100%	98.1%	95.9%	92.8%	88.7%	88.4%

of them. It is included to compare with a work that examined the whole database. Although it shows better performance than some of the proposed features, yet it uses more measures that are not related to the hand trajectory. The proposed system features lower dimensionality and simplicity.

VII. CONCLUSIONS

This work proposes a system for Arabic sign language recognition based on the trajectories of hands. It models the trajectory as a polygon and proposes two polygonal description features. The system shown good performance for both signer dependent and signer independent recognition. Th accuracy of the system reaches 99% for signer dependent and 64% for signer independent recognition. The proposed system is tested on two different datasets and is compared with published works that use the same dataset and shown better performance than most of them. The proposed system features simplicity, scalability, and generalization to unseen signer. The work in database collection is still in progress to extend the vocabulary size and number of signers.

ACKNOWLEDGMENT

The authors like to acknowledge the support provided by King Fahd University of Petroleum and Minerals (KFUPM) for funding this work through project number IN151008.

REFERENCES

[1] A. Tharwat, T. Gaber, A. E. Hassanien, M. Shahin, and B. Refaat, "Sift-based arabic sign language recognition system," in *Afro-european conference for industrial advancement*. Springer, 2015, pp. 359–370.

[2] N. El-Bendary, H. M. Zawbaa, M. S. Daoud, A. E. Hassanien, and K. Nakamatsu, "Arslat: Arabic sign language alphabets translator," pp. 590–595, 2010.

[3] K. Assaleh and M. Al-Rousan, "Recognition of arabic sign language alphabet using polynomial classifiers," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 2136–2145, 2005.

[4] M. F. Tolba, M. Abdellwahab, M. Aboul-Ela, and A. Samir, "Image signature improving by pcnn for arabic sign language recognition," *Can. J. Artif. Intell. Mach. Learn. Pattern Recognit*, vol. 1, no. 1, pp. 1–6, 2010.

[5] A. A. Ahmed and S. Aly, "Appearance-based arabic sign language recognition using hidden markov models," in *Engineering and Technology (ICET), 2014 International Conference on*. IEEE, 2014, pp. 1–6.

[6] A. S. Elons, M. Abull-Ela, and M. F. Tolba, "A proposed pcnn features quality optimization technique for pose-invariant 3d arabic sign language recognition," *Applied Soft Computing*, vol. 13, no. 4, pp. 1646–1660, 2013.

[7] M. Mohandes, M. Deriche, U. Johar, and S. Ilyas, "A signer-independent arabic sign language recognition system using face detection, geometric features, and a hidden markov model," *Computers & Electrical Engineering*, vol. 38, no. 2, pp. 422–433, 2012.

[8] T. Shanableh and K. Assaleh, "Arabic sign language recognition in user-independent mode," pp. 597–600, 2007.

[9] M. Al-Rousan, K. Assaleh, and A. Talaa, "Video-based signer-independent arabic sign language recognition using hidden markov models," *Applied Soft Computing*, vol. 9, no. 3, pp. 990–999, 2009.

[10] M. F. Tolba, A. Samir, and M. Aboul-Ela, "Arabic sign language continuous sentences recognition using pcnn and graph matching," *Neural Computing and Applications*, vol. 23, no. 3-4, pp. 999–1010, 2013.

[11] K. Assaleh, T. Shanableh, M. Fanaswala, H. Bajaj, and F. Amin, "Vision-based system for continuous arabic sign language recognition in user dependent mode," in *Mechatronics and Its Applications, 2008. ISMA 2008. 5th International Symposium on*. IEEE, 2008, pp. 1–5.

[12] K. Assaleh, T. Shanableh, M. Fanaswala, F. Amin, H. Bajaj *et al.*, "Continuous arabic sign language recognition in user dependent mode," *Journal of Intelligent learning systems and applications*, vol. 2, no. 01, p. 19, 2010.

[13] C.-L. Liu, I.-J. Kim, and J. H. Kim, "Model-based stroke extraction and matching for handwritten chinese character recognition," *Pattern Recognition*, vol. 34, no. 12, pp. 2339–2352, 2001.

[14] M. Kherallah, L. Haddad, A. M. Alimi, and A. Mitiche, "On-line handwritten digit recognition based on trajectory and velocity modeling," *Pattern Recognition Letters*, vol. 29, no. 5, pp. 580–594, 2008.

[15] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3551–3558.

[16] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4305–4314.

[17] M. Elmezain, A. Al-Hamadi, J. Appenrodt, and B. Michaelis, "A hidden markov model-based continuous gesture recognition system for hand motion trajectory," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[18] W.-Y. Lin and C.-Y. Hsieh, "Kernel-based representation for 2d/3d motion trajectory retrieval and classification," *Pattern Recognition*, vol. 46, no. 3, pp. 662 – 670, 2013.

[19] A. Naftel and S. Khalid, "Motion trajectory learning in the dft-coefficient feature space," in *Fourth IEEE International Conference on Computer Vision Systems (ICVS'06)*, Jan 2006, pp. 47–47.

[20] J. Pu, W. Zhou, J. Zhang, and H. Li, "Sign language recognition based on trajectory modeling with hmms," in *International Conference on Multimedia Modeling*. Springer, 2016, pp. 686–697.

[21] M. Boulares and M. Jemni, "3d motion trajectory analysis approach to improve sign language 3d-based content recognition," *Procedia Computer Science*, vol. 13, pp. 133–143, 2012.

TABLE VI. LIST OF THE WORDS IN THE DATABASE

Sign	ID	Sign	ID	Sign	ID	Sign	ID
mongoloid منغولي	146	diabetes مرض السكر	121	anesthetist مخدر	96	Skeleton هيكل عظمي	71
bacterium بكتريا	147	heart attack أزمة قلبية	122	surgery عملية جراحية	97	Skull جمجمة	72
microbe جرثومة	148	cancer سرطان	123	dressing شاش / ضمادة	98	Back bone عمود فقري	73
virus فيروس	149	aids الإيدز	124	tape شريط لاصق adhesive	99	Chest قفص صدري	74
spread إنتشار	150	hair loss تساقط الشعر	125	pharmacy صيدلية	100	system جهاز تنفسي Respiratory	75
hindrance إعاقة	151	heart failure سكتة قلبية	126	blood analysis تحليل دم	101	Trachea قصبة هوائية	76
hindrance إعاقة ذهنية mentality	152	hemiplegia شلل نصفي	127	فحص سريري physical examination	102	Lungs رئتان	77
اعاقة جسدية	153	paralysis شلل دماغي cerebral	128	sight test فحص النظر	103	Inspiration شهيق - زفير exhalation -	78
اعاقة بصرية	154	pressure ضغط الدم blood	129	ميزان حرارة thermometer	104	system جهاز هضمي digestive	79
hindrance إعاقة سمعية hearing	155	allergy حساسية	130	stethoscope سماعة أذن	105	Face وجه	80
epidemic وباء	156	itch حكة	131	جهاز قياس الضغط sphygmometroscope	106	pharynx بلعوم	81
immunity مناعة	157	medicine دواء	132	heartbeat نبض القلب	107	liver كبد	82
nerve عصب	158	دورة شهرية menstruation	133	analysis تحليل طبي medical	108	Pancreas البنكرياس	83
healthy معافى	159	sick مريض / مرض	134	معمل التحاليل analysis laboratory	109	intestine الأمعاء الدقيقة Small	84
eat يأكل	160	capsule كبسولة	135	X-ray صورة اشعة	110	intestine الأمعاء الغليظة Large	85
drink يشرب	161	medicine دواء شراب liquid	136	inflammation التهاب	111	appendix الزائدة الدودية Vermfiorm	86
sleep ينام	162	ointment مرهم	137	swelling تورم	112	system جهاز عصبي Nervous	87
up wake يستيقظ	163	dropper قطارة	138	cold زكام	113	heart قلب	88
hear يسمع	164	acupuncture أخذ إبرة	139	infection عدوى	114	senses حواس خمس Five	89
silence يسكت	165	pollination تلقيح	140	headache صداع	115	muscle عضلة	90
inhale يشم	166	inoculation تطعيم	141	pain ألم	116	tissue أنسجة	91
rise يصعد	167	laser ray أشعة ليزر	142	fever حمى	117	hospital مستشفى	92
descend ينزل	168	drug's مخدرات	143	diarrhea إسهال	118	first aid إسعافات أولية	93
open يفتح	169	addiction إدمان	144	constipation إمساك	119	wound جرح نازف	94
close يقفل (يغلق)	170	oneness توحد / أوتيزم	145	colic مغص	120	burning حروق	95

[22] L. Geng, X. Ma, H. Wang, J. Gu, and Y. Li, "Chinese sign language recognition with 3d hand motion trajectories and depth images," in *Intelligent Control and Automation (WCICA), 2014 11th World Congress on*. IEEE, 2014, pp. 1457–1461.

[23] H. Wang, A. Stefan, S. Moradi, V. Athitsos, C. Neidle, and F. Kamangar, "A system for large vocabulary sign search," in *European Conference on Computer Vision*. Springer, 2010, pp. 342–353.

[24] P. Jangyodsuk, C. Conly, and V. Athitsos, "Sign language recognition using dynamic time warping and hand shape distance based on histogram of oriented gradient features," in *Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, 2014, p. 50.

[25] M. Bhuyan, P. Bora, and D. Ghosh, "Trajectory guided recognition of hand gestures having only global motions," *World Academy of Science, Engineering and Technology*, vol. 21, pp. 753–764, 2008.

[26] M. Mohandes and M. Deriche, "Arabic sign language recognition by decisions fusion using dempster-shafer theory of evidence," in *Computing, Communications and IT Applications Conference (ComComAp), 2013*. IEEE, 2013, pp. 90–94.

[27] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[28] *LAS: Second part of the Unified Arabic Sign Dictionary*, 2006.

[29] S. Khalid, "Motion-based behaviour learning, profiling and classification in the presence of anomalies," *Pattern Recognition*, vol. 43, no. 1, pp. 173–186, 2010.

[30] S. Khalid and S. Razzaq, "Frameworks for multivariate m-mediods based modeling and classification in euclidean and general feature spaces," *Pattern Recognition*, vol. 45, no. 3, pp. 1092–1103, 2012.

[31] F. I. Bashir, A. A. Khokhar, and D. Schonfeld, "Object trajectory-based activity classification and recognition using hidden markov models," *IEEE transactions on Image Processing*, vol. 16, no. 7, pp. 1912–1919, 2007.

[32] S. Wu and Y. F. Li, "Flexible signature descriptions for adaptive motion trajectory representation, perception and recognition," *Pattern Recognition*, vol. 42, no. 1, pp. 194–214, 2009.

[33] M. Simão, P. Neto, and O. Gibaru, "Using data dimensionality reduction for recognition of incomplete dynamic gestures," *Pattern Recognition Letters*, 2017.

A Traffic Congestion Framework for Smart Riyadh City based on IoT Services

[1]Hailah Ghanem Al-Majhad, [2]Arif Bramantoro, [2,3]Irfan Syamsuddin, [2,4]Arda Yunianta, [2]Ahmad Hoirul Basori, [2,5]Anton Satria Prabuwono, [2]Omar M. Barukab
[1]College of Computer and Information Sciences, Al Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia
[2]Faculty of Computing and Information Technology Rabigh, King Abdulaziz University, Rabigh 21911, Saudi Arabia
[3]Department of Computer and Networking Engineering, School of Electrical Engineering, Politeknik Negeri Ujung Pandang, Makassar, Indonesia
[4]Faculty of Computer Science and Information Technology, Mulawarman University, Indonesia
[5]Master in Computer Science Program, Budi Luhur University, Jakarta 12260, Indonesia

***Abstract*—Internet of Things has become one of the most challenging issues in many researches to connect physical things through the internet by creating a virtual identity for everything. Traffic congestion in Riyadh city is chosen due to the proliferation in the number of vehicles on Riyadh roads that is resulting in grumbling by residents. Currently, there are few reliable services offered to residents from the traffic department enabling them to access traffic information. A new traffic congestion framework for Riyadh is proposed to help the development of traffic congestion services. This framework aims to benefit from the current Riyadh road infrastructure and apply the Internet of Things paradigm for detecting traffic congestion with Everything as a Service approach. Sensing devices are used to identify the congestion of the traffic flow through providing multiple proposed services such as a vehicle counting, live streaming video and rerouting services. Users are able to access the services by using proposed mobile application connected to the internet, as those services are integrated with public map service. By using the services, the users are able to identify the exact location where congestion occurs and an alternate solution can be provided easily. To achieve this, Business Process Execution Language is embedded as a supporting framework layer. Due to the effectiveness in this layer, executable workflows are designed to combine the proposed services with the legacy Riyadh services as individual model. This approach clearly defines how the services are executed through the proposed models. A quantitative evaluation is provided to support the usability of this research.**

Keywords*—Traffic congestion framework; internet of things; smart city; business process execution language; everything as a service

I. Introduction

Nowadays, the rapid increase in traffic vehicles appears to be a major problem in urban and sub-urban regions. Traffic congestion in Riyadh city is an ever growing problem as the number of vehicles is growing exponentially and the road infrastructure cannot be increased proportionally. Moreover, with the fast development of Riyadh city, the scale of the problem is expanding day by day as the population of the city is also surging. According to a survey prepared by the High Commission for the Development of Riyadh city, the population of Riyadh is approximately 4.9 million that is flooding the roads with about 985,000 cars daily and an average of 1.6 cars per family which 90% of them are private cars [1]. It leads to an increase in congestion of the roads which is rapidly becoming a phenomenon that is seen every day. Traffic congestion is one of the main reasons for increasing transportation costs due to the extra fuel and wasted time. Studies in Riyadh have proved that the costs of congestion create a heavy loss of SR 28 billion [2].

The growing number of vehicles and increased population have triggered to the requirement of effective traffic management systems. An Intelligent Transportation System (ITS) is an effective approach to solve traffic problems without building any extra physical infrastructure such as tunnels and bridges [3]. It applies Information and Communication Technologies (ICT) to the online transportation systems to improve performance and help to alleviate traffic congestion and optimize fuel consumption [4]. The global ITS market is expected to grow and reach 38.68 USD billion by 2020 [5].

Most researchers are engaged in exploring different technologies to monitor road traffic and detect traffic congestion to make congestion management more efficient. There are several technologies that are being used to detect traffic congestion, including loop-coils and intelligent video cameras known as CCTV systems. As stated in the Ministry of Transportation's report in Saudi Arabia [6], Riyadh City has some ITS systems which disseminate traffic information via the internet to a control center. Therefore, it offers pre-trip traffic information and on-trip traffic information services to their residents through radio and Dynamic Messages Signs (DMS). Traffic information is gathered from a variety of sources such as CCTV cameras, policeman maintenance, contractors and road travelers. However, CCTV cameras have lots of drawbacks such as being affected by weather conditions and high cost of installation [7], [8]. Therefore, Riyadh city must benefit from the internet and communication technologies, especially concerning traffic, to be a smart city in traffic areas and provide residents more confidence and reliable services.

Recent fast advancements in various technological fields including hardware miniaturization, wireless communications, sensing devices and embedded computing allow for increasing physical world with a unique identification; and also the capabilities to analyze and process information to sense and respond to the environment, thus making them smart. By connecting smart objects to the internet, an Internet of Things (IoT) is formed. The IoT is such a key in ICT that is rapidly becoming one of the most influential development and research topics, through enabling people and devices to communicate with each other in real time. The IoT describes everyday physical objects connected with the internet and able to identify themselves to other devices. It provides more advanced services to people through connecting several devices, systems and applications beyond traditional machine-to-machine.

The concept of IoT is closely related to Radio Frequency Identification (RFID) as the way of communication and also it may include other wireless or sensor technologies. RFID is shaping up to be an important building block for the IoT. The affordable cost with improved benefits has made RFID a reliable technology with a competitive advantage [3]. RFID is highlighted as one of the converging technologies and main catalyst playing a significant role in this project. Another reason is that RFID technology is considered as the most applicable for the country, such as in Saudi Arabia, where privacy is a big concern. The crowdsensing from user's smartphone [9] is getting more popular these days, however, it is argued that RFID is better in term of privacy as well as its support to the convergence of IoT [10].

This research proposes an intelligent traffic congestion framework for smart Riyadh city based on using the same infrastructure and the IoT services through identification of vehicles causing congestion in specific regions. The physical basis of the framework is IoT, and realized computation and storage of services through a cloud computing platform. The objective of this project is to offer residents access to road information and detecting traffic congestion by providing them automated services such as counting vehicle service, video streaming service, dynamic message signs service and alternative routes in real-time. Business Process Execution Language approach (BPEL) is proposed to integrate more than one service in one model publicly available for residents. BPEL is a XML-based language used to define business processes' workflow to unify the format of business process flow definition. By using BPEL technology, the proposed services are integrated with the existing Riyadh services such as public map and CCTV services to offer the users a new service which is a traffic congestion service.

II. Related Works

A particular IoT service is useful to make application development and implementation in agile manner. Authors in [11] categorize IoT services into four categories, which are identity-based services, data collection services, collaboration-supported services, and ubiquitous-oriented services. Generally, IoT development starts from data collection to collaboration-supported and ubiquitous-oriented. However, not all services of IoT necessarily develop to the stage of ubiquitous convergence because some applications and services only require data collection and are not intended for ubiquitous-oriented due to privacy required by minority group of people.

Riyadh city is the pioneer of smart city in Saudi Arabia since 2008, when Riyadh Municipality began to implement a smart street on Prince Muhammad Ibn Abdul-Aziz road by providing a free wireless internet service for all residents on this road through WiFi and WiMAX technologies [12]. In 2013, Riyadh city is continuing its efforts by launching smart roads that are developed to combine extensions of Oruba and Abu Bakr Al-Siddiq Street, and are centrally operated by a control room. It aims to guarantee the premium safety standards for residents by offering dynamic message signs, ventilation fans, and 58 centers for emergency calling, with surveillance cameras along the tunnels, bells and alarms and escape doors between the tunnels with lighting [13].

Authors in [14] develop a traffic information interface based on historic traffic data. It aims to show the drivers an output in public map as image processing of traffic speed details of speed limit on the available routes with different colors by giving day and time limit information to help drivers make their choice in selecting a route or change their current route for any kind of incident such as traffic congestion. The traffic speed data are collected on the basis of video analysis from several segment routes. Several videos are collected from each road segment, then each recorded video is analyzed to measure the speed data. By using the data, an interface is developed to show the traffic conditions on public map with speed limits denoted in different colors, which can be a guideline for developing countries where navigation is still unavailable. On the other hand, this method has many limitations. The main problem is that collecting many videos is difficult and time consuming. Moreover, it does not provide a real time traffic information, only historical data are available.

Authors in [15] propose a vehicle cloud computing architecture based on three IoT layers: device layer, communication layer and service layer. By using a cloud computing technique, three layers enable devices, network and services to exchange information with each other in a real-time manner. Combining devices with cloud computing technologies allows essential services to be offered to residents.

It is observed that almost all the previous researches and studies that talked about IoT frameworks focus only on three or more essential layers, but each researcher names the layer differently. To date, the integration layer that can be used for integrating the services of IoT is not available yet. Therefore, BPEL tool is used to effectively collaborate and integrate several existing and new services with each other by defining a business workflow for each service in one BPEL model. Moreover, by using the BPEL tool the technical team can then easily develop several applications due to the visibility of any possible workflows amongst available service.

III. Traffic Congestion Framework

Riyadh city should benefit from internet and communication technologies to provide smart roads in order to have more confidence and reliable services. The ability to know the traffic conditions on the road ahead via effective services on the mobile devices allows residents to find out the number of vehicles on a specific road, display live video streams captured from CCTV cameras, view warning messages on DMS and find alternative routes. This eventually leads to the reduce of time consuming and vehicle's fuel consumption. Moreover, the DMS boards on the roads display warning messages automatically to residents. The more residents utilize these services, the more traffic congestion decreases and the more residents are confident and satisfied.

The proposed traffic congestion framework contains a fixed and mobile platform including the support of sensors mounted on vehicles. The sensor devices collects data and transfers it to cloud via the internet to offer new services to residents. More specifically, Fig. 1 shows the layers of the proposed traffic congestion framework for Riyadh city. This proposed framework consists of four different layers that have different purposes: a physical layer, a communication layer, an integration layer and an application layer.

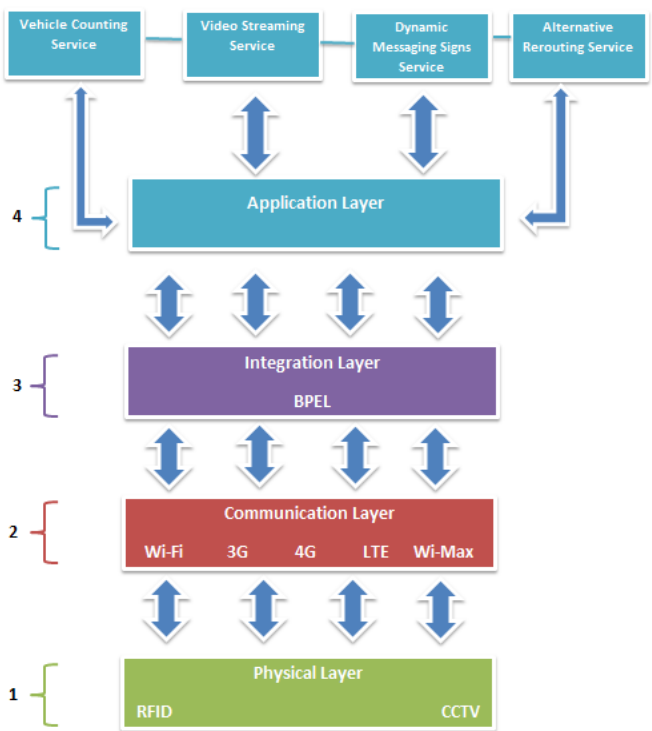


Fig. 1. Conceptual traffic congestion framework for Riyadh City.

An architectural view of the framework is shown in Fig. 2. The bottom layer is called the physical layer, whose function is recognizing and collecting data of vehicles through RFID technology and CCTV cameras systems. When data are pooled synchronously, transmission is required on second layer, which is called a communication layer. The formation of network in the communication layer consist of wireless technologies such as 3G, 4G, LTE or WiFi. This layer is responsible for transmitting data with high reliability and security to cloud storage. The third layer is an integration layer. BPEL is used in this layer to design the workflow process of each service on the traffic congestion framework. BPEL is proposed to integrate several services with each other to build composite service; when more than one service is combined together in one BPEL model, then residents have a new service. The top of the framework is an application layer which consists of different services to be presented to residents, which are vehicle counting service, live video streaming service, dynamic messages signs service and alternative rerouting service.

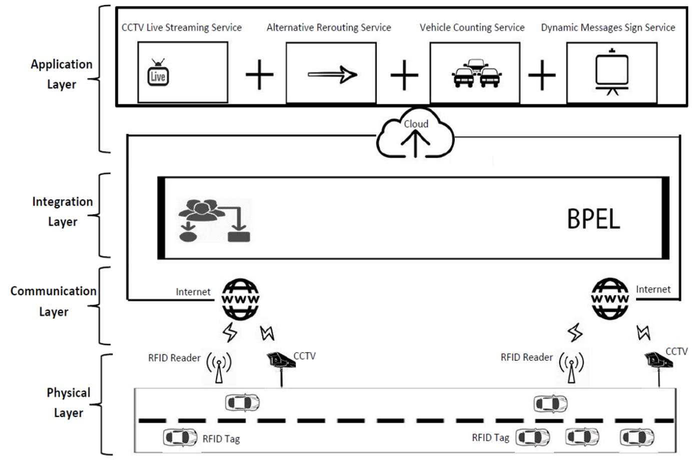


Fig. 2. Traffic congestion architecture for smart riyadh city based on IoT services.

To achieve an effective traffic congestion control system in Riyadh, both RIFD technology and CCTV devices are used. RFID technology provides automatic recognition of vehicles with the tags, RFID tags transfer the signal containing the tag number captured by RFID reader. For this approach, vehicles should be installed with RFID tags; whenever the vehicle reaches the specified range of RFID readers, the RFID reader is able to read the information of each vehicle's tag. The stored information in the RFID tag for each vehicle can only be a unique tag number while the location coordinates and time-stamp is in the RFID reader. The location coordinates should be integrated into the RFID reader through the built-in GPS sensor to provide information regarding where a vehicle tag has been interrogated and at what time. The microcontroller in the RFID reader is programmed to calculate the exact time at which the tag ceases to be read and then link the time with the location coordinate of the RFID reader. On the other hand, the CCTV video surveillance cameras that are installed on Riyadh roads is able to take images or recordings for surveillance to monitor the traffic congestion through the monitoring screen in the traffic department center. This research uses the already installed CCTV cameras through Everything as a Service (XaaS) technology to get the live video of each road to also be displayed to residents.

The information trunk for the IoT is a communication layer. The communication layer provides the connection between the physical devices of RFID readers and CCTV and the cloud through the internet. The layer makes use of 3G, 4G, LTE, fiber optic or WiFi networks to transfer the RFID data and video data to the cloud storage in application layer. These communication technologies generally characterized by high reliability and high transfer rates.

In integration layer, BPEL tool is used to display a work-flow for integrating more than one service with each other. Because more than one service is combined with another service in one BPEL model, the users have a new service at the end of the model. Each service should be able to achieve automation of process by defining a workflow of activities which is called a business process. A business process is defined as a set of all possible sequences of activities for delivering a specific business objective, and business process of services is often modeled and performed through BPEL. The BPEL model is a language used to define how the business processes involving web services are combined and executed in smooth manner. Each notation represents a specific activity in the process. Fig. 3 presents the notations legend used in Eclipse BPEL designer and executed with Active BPEL Engine for traffic congestion framework.

Notation	Definition
Invoke	Invoke the operation
Receive	Get an incoming request data
Reply	Reply request to external source
Assign	Copying data from one place to another
If	Checks to see if a statement is true or false and then does one of the two depending on the result
Sequence	Set of activities in ordered sequence
Flow	Set of activities that will be processed in parallel

Fig. 3. BPEL Notations for the Framework.

The goals of using BPEL models in the framework are:

- To define the business processes by using XML language.
- The BPEL layer clearly represents the flow of each service such as vehicle counting service, CCTV live streaming service, rerouting service and DMS service in a graphical representation.
- BPEL has multiple graphical notations to show how the business process works and allow their use to be blended as seamlessly as possible, such as combining multiple services to one, therefore residents can choose the desired services from cloud without having prior knowledge about the application development.
- BPEL is able to combine multiple services in one service such as combining RFID service with the internet service and public map service to get vehicle counting service.
- The developers can easily develop applications and services by using the BPEL model XML code to create database modeling from where non developer end users can benefit.

The application layer in the top of the framework provides several services to residents to complete data processing, data exchange, data calculation and data storage. The cloud storage in this layer stores RFID tags, RFID readers' data and video data. The cloud is considered as XaaS and used to provide services in the application layer, hence, residents can access the cloud to choose their required services from their mobile devices. There are four proposed services: vehicle counting service, CCTV live streaming service, dynamic messages signs service and alternative rerouting service. These services can be executed from a mobile application to benefit residents in detecting traffic congestion on Riyadh roads and then reducing the congestion costs as general. Moreover, the proposed DMS boards service in the roads that relied on IoT also benefits residents to get reliable and real-time warning messages.

A. Vehicle Counting Service

The objective of Vehicle Counting Service is to display the exact number of vehicles in a specified street for one lane. It also displays the average number of vehicles for a specific period of time to detect the exact traffic congestion. There are two ways to achieve this objective. One is to use the BPEL model to identify the workflow of this service. The other is to develop a proposed vehicle counting web-based application.

BPEL model is used to combine street sensor services RFID internet service and user defined map service to become one service which is vehicle counting service by defining a set of activities for the service. The RFID readers in the roads are supposed to receive RFID tag information from each vehicle equipped with the RFID tag. The received information from the RFID tags and reader is then transferred through the internet to the cloud storage. The stored information is the tag number of each vehicle, location coordinates and time-stamp. Based on this information, it is possible to identify the location of each vehicle and the time when it crossed the RFID readers in the roads. The application in cloud is able to retrieve this information from the cloud storage to calculate and identify how many vehicles per lane are congested in a specific RFID reader range installed on the roads and the average number of the traffic congestion for a specific period of time. Once the user connects to internet, the result of the congested vehicles in a specific RFID reader is received through the proposed mobile application and integrating with the public map service for displaying to residents the average and exact number of vehicles per specific time in the congested area on public map.

A web-based application is developed which is used by Riyadh residents to assess the actual number of vehicles at each RFID reader for one lane in addition to the average number of vehicles for a specific period of time to detect the traffic congestion. The heavily congested roads are almost the result of bottlenecks; these bottlenecks are frequently due to the junctions, accidents, reduced number of lane or reduced carriage way width. Based on the previous researches, it is noticed that most of researchers are stated several metrics to measure the traffic congestion such as average traffic speed [16], intensity and density [17]. This research proposes vehicle counting web-based application to calculate and measure the congestion of roads based on three important measures: the number of vehicles within RFID tags read by one RFID reader, the exact number of vehicles at each RFID reader based on the number of lanes, and the average number of vehicles per lane during a certain period of time by each RFID reader.

B. Vehicle Counting Web-Based Application

The presented web-based application is developed by using Microsoft SQL server and ASP.net programming language, which is a server side web application development structure that was generated for the web creation to develop dynamic web pages with Hyper-Text Markup Language (HTML), JavaScript and Cascading Style Sheet (CSS) to design the user interface. The vehicle counting service application is simulated by using ASP.net in visual basic (VB.net) and Microsoft SQL server database. Hence, the data are stored in the database. Moreover, by using public map API, the vehicle counting web-based application data are connected with the public map. The API can be retrieved from the provider by using an API key and then integrate it with the web-based application. The application interface is shown in Fig. 4.

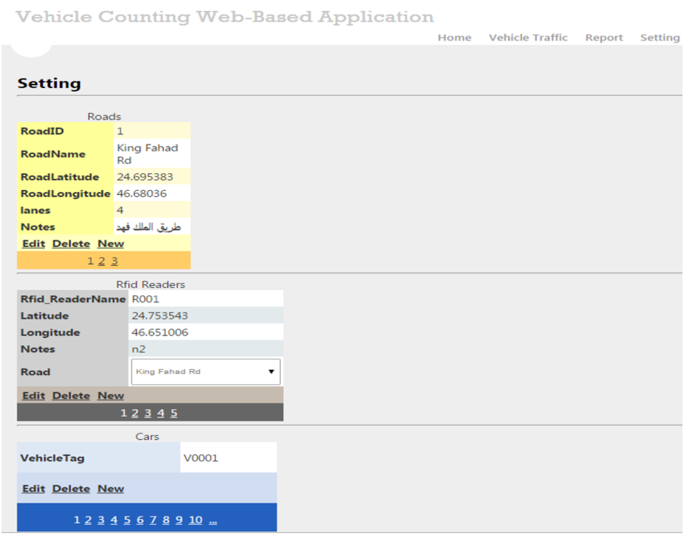


Fig. 4. Setting page in vehicle counting web-based application.

There are three tables in the setting page. The information in the tables is stored in database as inputs. The first table is for adding Riyadh roads on the public map. A user can insert new road by filling the following information: Road ID, Road Name, Road Latitude, Road Longitude, Number of lanes in each road and Notes. The second table is for adding the RFID Readers in Riyadh Roads by filling the RFID Reader Name, Latitude and Longitude of this reader and the Road that want to add this RFID reader to it. Finally, the last table is for adding vehicles which represent by the vehicle tag ID for each one. The inputs are stored in the database, and then this information is retrieved back through ASP.net with the integration of public map API services to give the output to Riyadh residents.

In Fig. 5, the vehicle traffic page aims to insert vehicle tag to particular RFID reader. In other words, the RFID reader reads the tag information for each vehicle that passes along the road. Each vehicle tag is inserted to one of the defined RFID reader in setting page. Once the vehicle tag is added to particular RFID reader, the detailed table is shown in the vehicle traffic page. This table contains the information of RFID reader and roads. It displays the exact date and time when the RFID reader and vehicle tag are inserted.

C. CCTV Live Streaming Service

The current CCTV system in Riyadh city is used for the video surveillance of roads that is accessed only by the traffic department. A traffic department employee should see the status of the roads and then assign a traffic policeman to a specific site to manage the traffic congestion or any other services but this method is not of much use to the Riyadh residents. This research project improves the performance of the current CCTV monitoring system by allowing access a live streaming video service to residents. The BPEL model

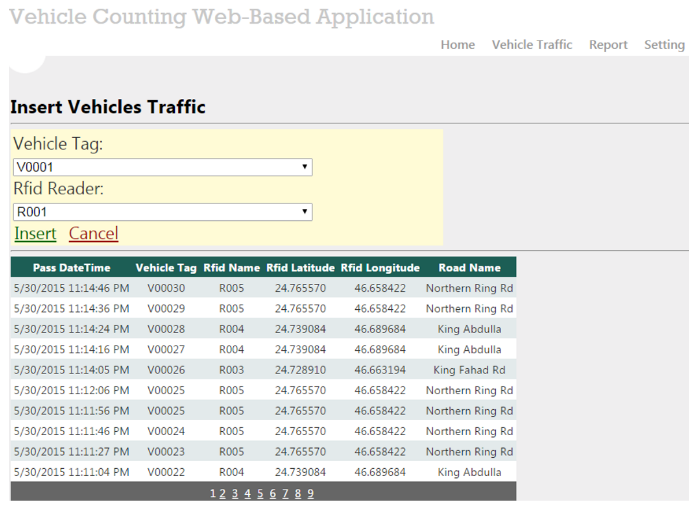


Fig. 5. Vehicle traffic page in vehicle counting web-based application.

combines the CCTV service, the internet service and user defined map service (in this research, public map service is utilized) to provide residents map-based video live stream service. The BPEL model also displays the workflow process of the CCTV live streaming service.

The CCTV cameras that are installed along Riyadh roads are connected to the internet to store the received video data in the cloud storage, then the same video data are streamed from the cloud to residents by using an application to view the live streaming video of the desired road. Once video data are received, public map is invoked to display the CCTV locations on the roads. Then, residents choose the desired location that has CCTV cameras and finally residents are able to view and display the live streaming of the chosen road.

D. Dynamic Messages Signs (DMS) Service

Some roads in Riyadh city have DMS boards displaying warning messages regarding the traffic status to residents. It is the responsibility of the traffic department to update the messages based on CCTV system. Residents are unable to get real-time information because the DMS is not updated automatically. BPEL is able to combine street sensor services (RFID), internet services and public map service in one model.

IoT is proposed to connect the DMS to the internet and then the dynamic message board is able to provide real-time messages to residents who are travelling on the roads. The DMS is completely controlled by an application itself and the information is retrieved from cloud storage based on the RFID readers' information, not from updating the DMS application manually through the monitoring of the traffic department to CCTV cameras. The RFID readers installed on the roads are able to receive the installed RFID tags information from the vehicles. A DMS application in the cloud retrieves only the RFID readers' information that is placed after the DMS from the cloud storage to compare which RFID reader shows the most congestion and then an automated warning message is displayed on the DMS board on the roads to residents. Since the availability of the internet for DMS on roads and on users' mobile applications, residents are now able to see warning messages about any congestion via a DMS board on the road and also via public map on a mobile application, in addition to receiving any warning messages from the traffic department regarding to weather effects, accidents and others.

E. Alternative Rerouting Service

In alternative rerouting service, if there is traffic congestion on some roads, the service gives residents an alternative road based on the real-time information of the RFID data. BPEL combines the sensor street service, the internet service and public map service in one model to act as one service, which is a rerouting service.

The RFID readers read the RFID tag information from each vehicle equipped with the RFID tag. The received information is transferred through the internet to cloud storage. The alternative rerouting application in cloud compares the congested vehicle data of the nearest RFID readers to the congested road and then the application stores the least congested value of the road. Thus, users can open the application and connect to the internet service to receive the value of the least congested road. Public map is also required to be invoked, so that the application provides the users with the least congested road as a rerouted road for residents, visually available on the map.

F. Traffic Congestion Service

The traffic congestion service combines the vehicle counting service, CCTV live streaming service, dynamic messages signs service and rerouting service together in one BPEL model. These services are available to Riyadh residents from the mobile applications on a public map at any time and from anywhere in Riyadh by XaaS, in addition to the DMS boards that are installed on Riyadh roads that give real-time warning messages. These valuable services benefit residents by detecting real-time information on traffic congestion, thus saving valuable time, and warning residents about traffic congestion in a specific location, thus enabling resident to easily choose a recommended alternative route to avoid the traffic congestion. BPEL is used to combine the proposed and existing services in one model to act as one service.

IV. Traffic Congestion Service Models

The BPEL model shows a list of the activities of each service. Residents can benefit from choosing their desired services from the cloud; the application can be developed by the developers with reference to the BPEL models for the benefit of both the traffic department and residents of Riyadh. Although BPEL is a modelling language, the user can design their models by using an interactive any graphical BPEL designer tool and then execute the model by instantiating abstract services in the model with concrete available services. Hence, there is no programming task required to execute the BPEL models, especially if concrete services are already available.

A. Vehicle Counting Service BPEL Model

A vehicle counting service BPEL model is shown in Fig. 6. The model shows the list of activities that should be followed in each service to assess the number of vehicles congested in a specific road per every lane as well the average number of vehicles for a specific period of time.

The service starts with receiving a request, which then invokes the vehicle counting service to begin the processes. By assuming there is a flow of readers on the roads, and taking one RFID reader in the Northern Ring Road and another in the King Fahad Road, both RFID readers read the vehicle's information from the RFID tags that the vehicles on the roads are equipped with and assigned a RFID tag's information to each RFID reader; the RFID readers on the two roads then assign the information from the RFID readers as tag numbers, location coordinates and time-stamps to the cloud storage by invoking the internet service. The application in the cloud retrieves the RFID readers' information to calculate the number of tags for each RFID reader. Then, resident invokes the internet service to connect with the mobile application to receive the RFID reader's values from the cloud. Once the values are received, the public map API service is invoked to retrieve the maps from the cloud and then the retrieved RFID reader's values are assigned to the map. Finally, the total number of vehicles are indicated to residents on the public map for each reader.

The developed web-based application for vehicle counting was built to display on public map the actual number of vehicles per specific minutes in one hour and in one lane for each RFID reader. Overall, the application operated well and as intended. The Vehicle Counting Web-Based Application is able to provide the result from the inputs of RFID readers, Roads and the Vehicle tags. Hence, the inputs data are combined to give the output. Fig. 7 shows the number of vehicles per lane in addition to the average number of vehicles per specific time of period based on resident's preferences. The user preference is the key concept of XaaS as coined in [18].

From report page, resident can choose specific time in hour to display the average number of vehicles per lane either in five minutes or above. As each vehicle is mapped to RFID reader, the equations are executed to get the traffic congestion results. Once resident chooses show reports menu, the public map is displayed to resident. The location of RFID readers are marked in red icon. Then, the user can zoom in the map to choose the desired reader of specific road and click on it. The pop-up screen appears with the name of reader, road, road lanes, actual number of vehicles in any lanes, number of vehicles per one lane and the average number of vehicles per specific period of time. Through these information, resident can view the congested road for each RFID reader and detect the traffic congestion.

B. CCTV Live Streaming Service BPEL Model

A CCTV live streaming service BPEL model is shown in Fig. 8. This model shows a list of activities that should be followed to apply this service to achieve the goal of a live streaming service to residents. Hence, residents are able to see the live video status of the chosen roads in Riyadh roads.

The CCTV live streaming service is invoked to get the result of the CCTV video footage on public map. By assuming there are two CCTV cameras processed in parallel, one in Northern Ring Road and another one in King Fahad Road, both cameras take live videos of the roads and the video data are assigned to the CCTV in the same road. After the CCTV video data have been received, it is assigned to the cloud

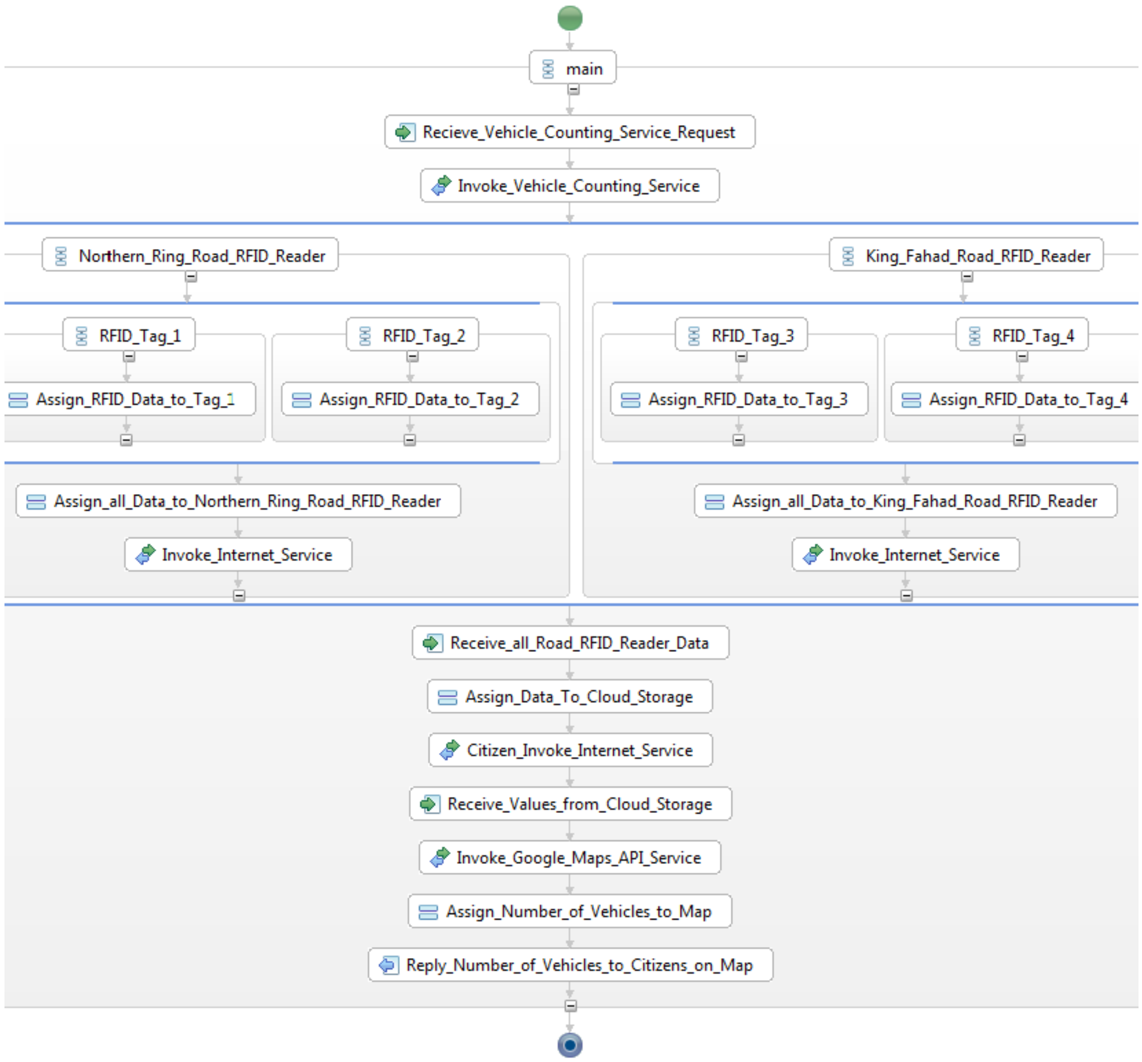


Fig. 6. Vehicle counting service BPEL Model.

storage by invoking the internet service. The video data from the cloud storage are retrieved by invoking the internet service from resident to connect with the mobile application. Once the video data are received, the public map API service is invoked to retrieve the maps from the public cloud, then the video data and map data are combined together to show the location of the CCTV cameras installed on the roads of Riyadh. Accordingly, residents can choose the CCTV cameras of the needed location from public map. The desired location of CCTV is assigned to public map and finally, the mobile application is able to give a pop-up screen to show the live stream of the video data for the chosen road. The output of the CCTV live streaming service model is shown in Fig. 9.

C. Dynamic Messages Signs Service BPEL Model

A dynamic messages signs service BPEL model for the Dynamic Messages Signs (DMS) in roads and mobile application are shown in Fig. 10. The BPEL model shows a list of activities that should be followed to apply these services to achieve the goal of displaying real-time warning messages about the traffic congestion status on the roads.

The model starts with invoking the Dynamic Messages Signs Service. By assuming there are three parallel RFID readers after the DMS board in the Northern Ring Road, these readers read the tag information for each vehicle that passes along the northern ring road and assign each RFID tag information to the RFID reader. The three RFID readers then assign the RFID information to cloud after invoking the internet service. A DMS application in cloud is programmed to receive the information of the RFID readers that is installed after the DMS from the cloud storage and then identify the congestion of vehicles by comparing the results of each RFID reader's information. For example, Northern Ring Road RFID Readers (2) and (3) are placed after the DMS board. Northern Ring Road RFID Reader (2) has 30 vehicles and Northern Ring Road RFID Reader (3) has 100 vehicles. The highest number of vehicles in the two RFID readers are identified by the application and considered as more congested, which is Northern Ring Road RFID Reader (3).

The DMS board on the road and the mobile application

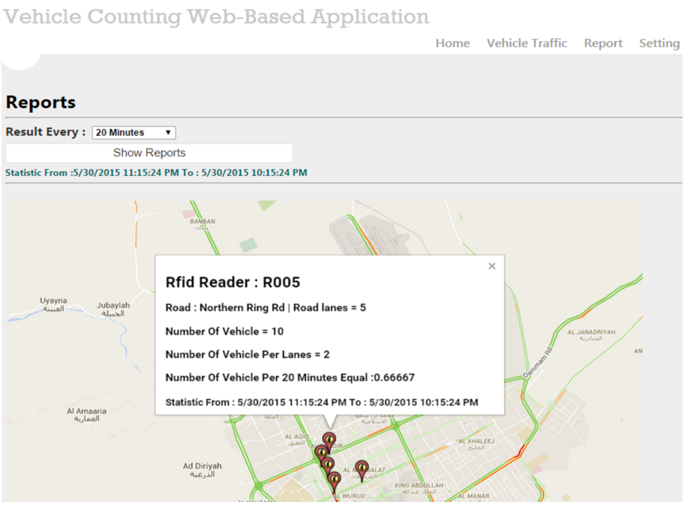


Fig. 7. Vehicle counting service output.

Fig. 8. CCTV Live Streaming Service BPEL Model.

must connect to the internet to receive the Northern Ring Road RFID Reader (3) information as the highest congested reader from the cloud. An automated message is displayed on the DMS board on the roads after assigning the Northern Ring Road RFID reader (3) data to the DMS, while the warning message in the mobile application appears to the users after invoking the public map service and assigning the Northern Ring Road RFID reader (3) data to the map. Thus, Riyadh residents can get warning messages about congestion or any updated status regarding traffic problems from the roads and from the mobile application. Unlike the existing DMS in Riyadh, the DMS in the framework can also be accessed by any user's device, since XaaS technology is utilized. The output of the dynamic messages signs service on roads is shown in Fig. 11, while the output of the dynamic messages signs service in mobile application is in Fig. 12.

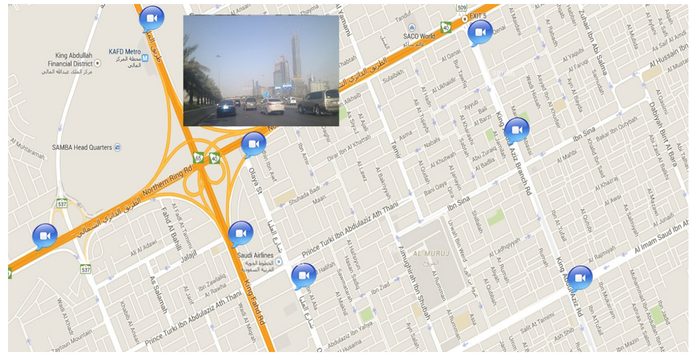


Fig. 9. CCTV live streaming service output.

D. Alternative Rerouting Service BPEL Model

An alternative rerouting service BPEL model is shown in Fig. 13. This model shows a list of activities that should be followed to apply the service to achieve the goal of providing residents an alternative route that is not congested.

The alternative rerouting service is invoked to identify the least congested route. Assuming there are three RFID readers: in Northern Ring Road, King Fahad Road and King Abdullah Road. These readers read the tag information for each vehicle that passes them and then each RFID tag's information is assigned to the RFID reader. After that, the three RFID readers assign the RFID tag's information to cloud storage after invoking the internet service to store this information in cloud storage. The alternative rerouting application in the cloud receives the RFID reader's information, then the congestion comparison is taken to identify the lowest number of RFID tags, for example the RFID reader in Northern Ring road receives 300 RFID tags while the RFID reader in King Abdullah road receives 58 RFID tags. The lowest one, King Abdullah road, is chosen through this application and the rerouting path is identified as King Abdullah road. When resident invokes the internet service, then the mobile application receives King Abdullah road information from the cloud, then public map is invoked to retrieve the maps from public cloud and King Abdullah road information is assigned to the map. Hence, residents are notified by the mobile application to reroute to the least congested road to reach their needed location without any congestion. The output of the Alternative Rerouting service model is illustrated in Fig. 14.

E. Traffic Congestion Service BPEL Model

A Traffic Congestion Service BPEL model shows the list of activities that should be followed to combine the proposed services with each other in one BPEL model (it is not shown due to the limitation of the page). As the proposed services

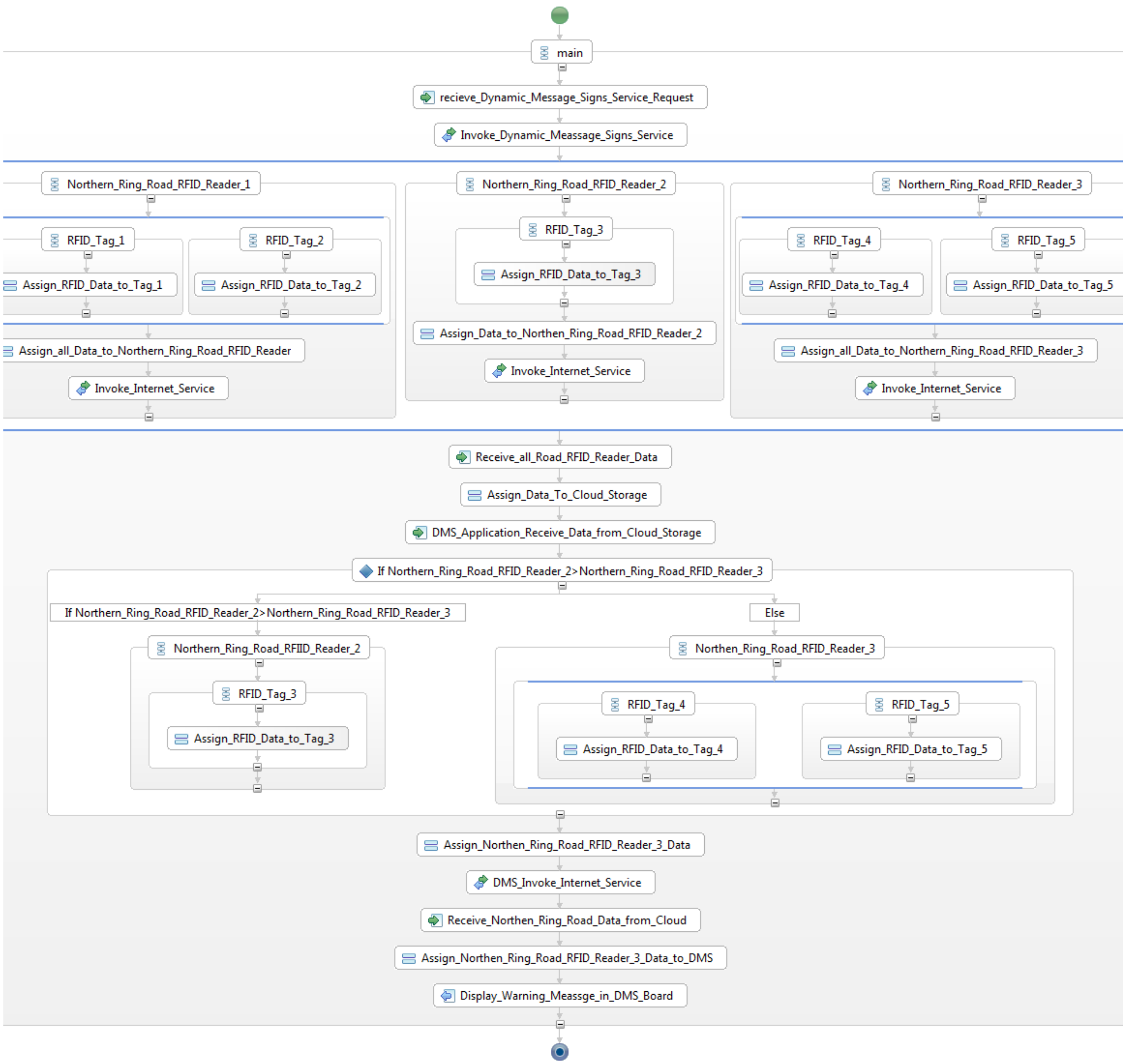


Fig. 10. Dynamic messages signs service in roads BPEL Model.

are in cloud as a service, residents can choose any service they need from public map in the proposed mobile application or from the DMS in the roads. The appropriate output of each service is given through the application layer.

The physical devices in traffic congestion service model comprises RFID and CCTV cameras. The RFID sensors services can be used to access the output of the received RFID tags, and is used by the following three services: Vehicle Counting, Dynamic Messages Signs and Alternative Rerouting services. The output of RFID services is stored in cloud storage. Similarly, the CCTV cameras that are installed on Riyadh roads are able to retrieve the video data of each road and then store it in cloud storage. Each application on the cloud is designed depending on each service and is able to retrieve the needed data from the cloud storage and then process it. The returned values from the cloud are combined with the public map API to get the output of services in one application. Hence, residents are able to choose their desired services from the proposed mobile application. Furthermore, the returned RFID reader's values from the cloud are also used for the DMS on the roads. Residents have the option to use the mobile application or the DMS boards on Riyadh roads.

In Fig. 15, the traffic congestion service starts with receiving a request, and then invokes the traffic congestion service to initiate the processes to work. In this traffic congestion service model, there are two physical devices: RFID and CCTV devices. Both of them are accessed in parallel by using the flow activity. RFID services start by assuming there are two RFID readers, one in Northern Ring Road and the other in King Abdullah Road. The readers read the vehicle's

Fig. 11. Dynamic messages signs service output.

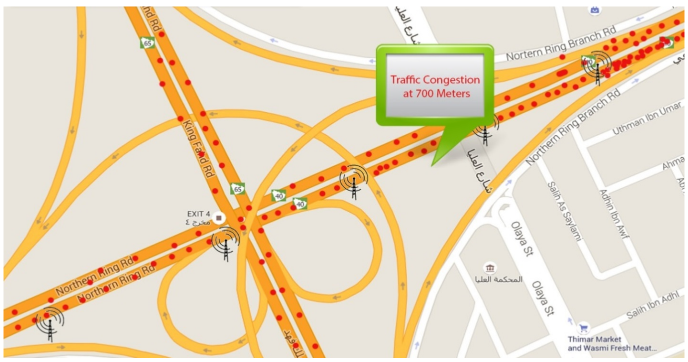


Fig. 12. Messages Signs service output.

information from the RFID tags which the vehicles on the road are equipped with and then the flow of RFID information is assigned to the reader. Then, Northern Ring Road RFID reader and King Abdullah Road RFID reader assigns their information as tag numbers, location coordinates and time-stamps to cloud storage by invoking the internet service. The second service is CCTV services, responsible for getting the live video of Riyadh roads. By assuming in one flow there are two CCTV cameras, one in Northern Ring Road and another in King Abdullah Road, both cameras take live videos of each road and then the video data are assigned to the CCTV camera in the same road. The flow of the CCTV video data is assigned to cloud storage by invoking the internet service.

The RFID and CCTV video data are stored in cloud storage. From the cloud, each application carries out its functionality to give an appropriate service. The proposed services are processed in parallel, therefore it is required to use the flow activity to explain the process of each service.

Firstly, the vehicle counting service as an application in cloud processes a list of activities in ordered sequence. The vehicle counting service receives the Northern Ring Road RFID information and King Abdullah Road RFID Reader information from the cloud storage to initiate the processing and calculating of the average number of RFID tags during specific period of time for each RFID reader and then assign the average number of vehicles for each RFID reader. Secondly, the CCTV live streaming application in cloud is composed from ordered activities, starting by receiving the video data and then assigning them to the CCTV application. Thirdly, the dynamic messages signs application aims to display some warning messages on the DMS boards. This application in cloud is programmed to receive only the RFID reader's information located after the DMS board from the cloud storage and then identify the congestion of vehicles by comparing the results of each RFID reader.

When there are two RFID readers in Northern Ring Road, the Northern Ring Road RFID reader (2) tagged 55 vehicles and the Northern Ring Road RFID reader (3) tagged 110 vehicles, the DMS application compares them and the highest RFID reader which is Northern Ring Road RFID reader (3) is assigned to the DMS application and considered as the more congested reader. Lastly, the alternative rerouting application in the cloud aims to enable Riyadh residents to take another path to avoid road congestion. This application starts with receiving the RFID readers' information from cloud storage. By assuming there are two RFID readers, one in Northern Ring Road and the other in King Abdullah Road. The application is programmed to compare them and take the lowest number of RFID tags in each RFID reader - for example, the RFID readers in Northern Ring road receive 300 RFID tags and the RFID readers in King Abdullah road receive 33 RFID tags. The lowest one which is in King Abdullah road has the data assigned to the alternative rerouting application.

The output from each application in the cloud is processed in a parallel way based on the requests from the users. When the users need to see the warning messages of DMS boards on the roads, the DMS board must invoke the internet service to be able to receive RFID readers' data from the DMS application. The DMS receives the Northern Ring road RFID reader's (3) information as the more congested area, therefore it assigns this information to the DMS board and then the application displays a warning message on the board to the users that there is congestion in this area.

Regarding the proposed mobile application in the users' devices, it has four proposed services that benefit Riyadh residents by identifying and detecting traffic congestion on Riyadh roads. If residents need to access the proposed mobile application, their application must invoke the internet service to be able to display the proposed services. Then, the mobile application receives the processed RFID and video information from the cloud. The application also invokes public map to retrieve the data and services from the public cloud and integrate them with the processed RFID and video data from the cloud. The services are processed in parallel to respond to residents' requests and the values of services are assigned to the map. Hence, the proposed services are shown on the map to enable Riyadh residents to choose and display one or more than one service at the same time, such as number of vehicles at each RFID reader, warning messages on the DMS boards, alternative rerouting. Additionally, residents can choose the CCTV cameras of the needed location to view a live stream of the road. Finally, residents' traffic congestion service request is replied to, since XaaS technology is utilized.

V. Quantitative Evaluation

A quantitative experimental correlational study was designed to measure the impact of this research. The population

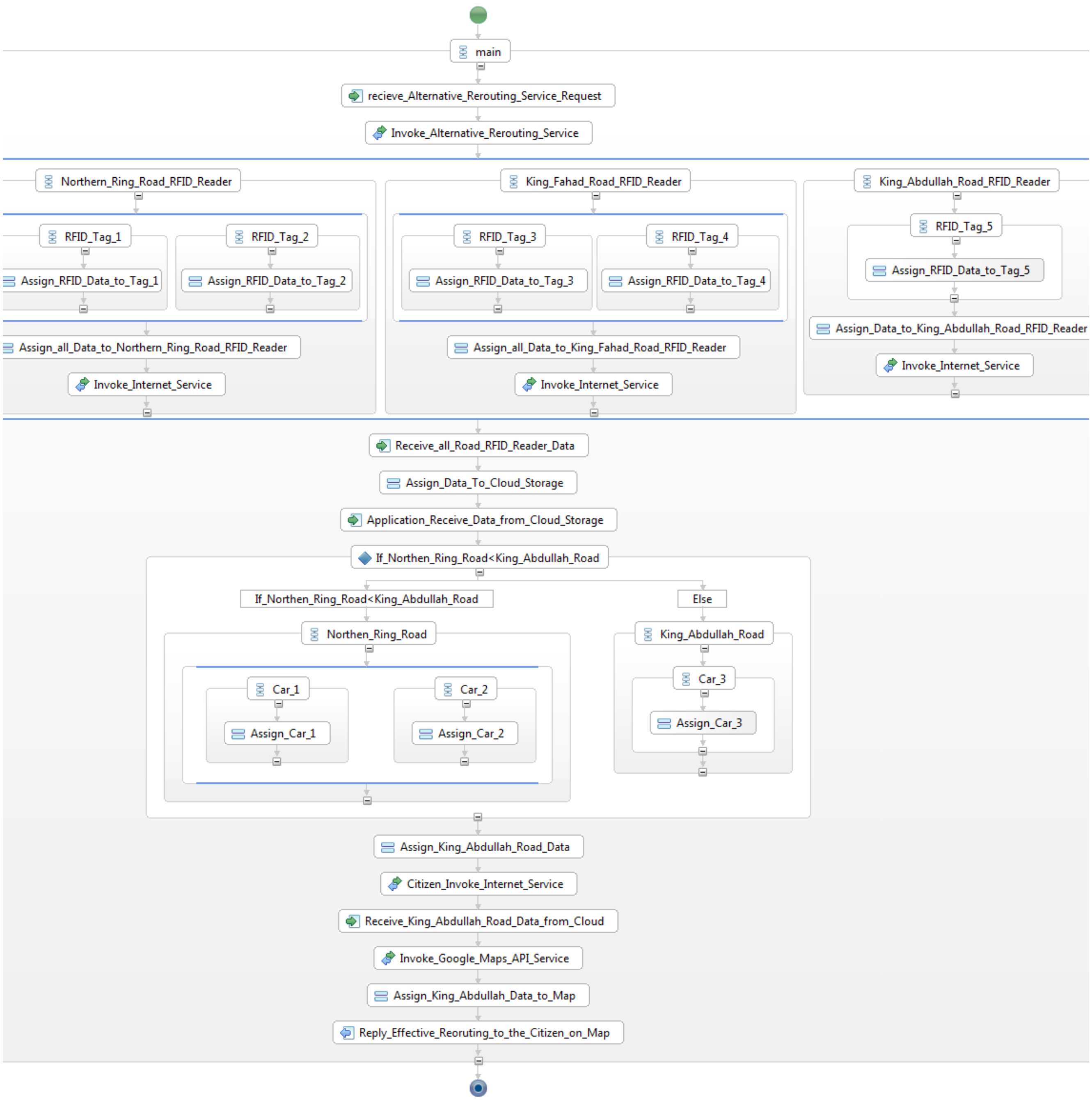


Fig. 13. Alternative rerouting service BPEL Model.

of this study is Riyadh adult users of internet connected devices. To evaluate the usability of the research, a ten-item questionnaire was randomly devised and distributed to seventy participants, who willingly decided to take part in the survey. The participants were provided with a list of service output images. The participants were required to answer the following ten questions on a scale from 1 to 5 where 1 - Not at all, and 5 - To a large extent. Table I shows the results of the survey distributed amongst students and staff for a set of 70 surveyed actors. For example, Q2 has a mean score of 4.02 showing that; overall, participants rated the Fig. 9 as highest. Most of the participants gave a score about 4 to all questions with a low standard deviation meaning that there are only small differences between the participants' answers. This result can be used to roughly identify the central tendency and the outlier of the observed data.

VI. Conclusion

The traffic congestion framework of Riyadh city is presented to improve Smart Riyadh City. XaaS is considered as an appropriate approach used in the framework due to enhance the public services and increase transparency between traffic department and residents. XaaS approach advances the framework through BPEL models which are useful for

TABLE I. SUMMARY OF THE RESULTS OF THE CONDUCTED SURVEY

Questions	Mean	Std Dev.
Q1. After you use the Vehicle Counting Service, do you like the service output (Fig. 7)?	3.38	0.957
Q2. After you use the CCTV Live Streaming Service, do you like the service output (Fig. 9)?	4.02	0.885
Q3. After you use the Messages Signs Service, do you like the service output (Fig. 12)?	3.91	0.940
Q4. After you use the Alternative Rerouting Service, do you like the service output (Fig. 14)?	3.82	1.065
Q5. After you use the Traffic Congestion Service, do you like the service output (Fig. 15)?	3.61	1.175
Q6. The message is easy to understand.	3.83	0.814
Q7. The message is useful during a traffic jam.	4.28	0.696
Q8. I would use the system once it is deployed.	4.08	0.790
Q9. It is a good idea to use those maps as an emergency guide.	3.98	0.690
Q10. The provided information is enough.	3.88	0.725

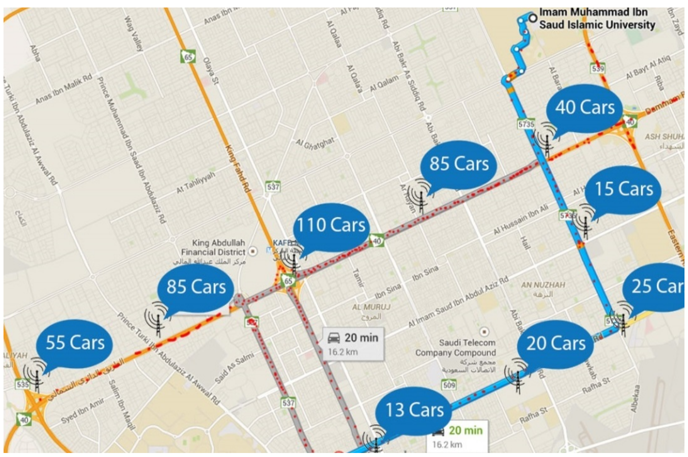


Fig. 14. Alternative rerouting service output.

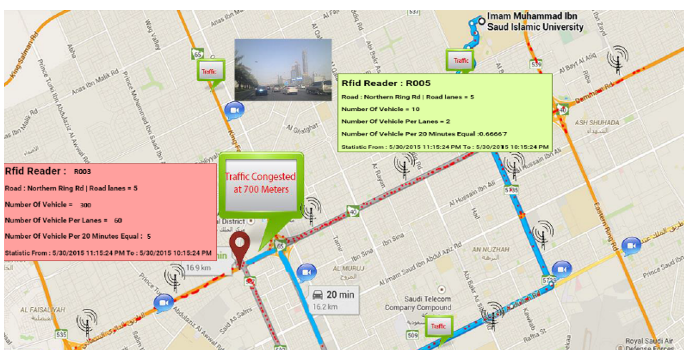


Fig. 15. Traffic congestion service output.

business managers and developers to provide a complete traffic congestion system for Riyadh city to offer their residents effective services. The proposed traffic congestion service are vehicle counting service, which is proposed to detect the total number of vehicles congested in a specific road; dynamic messages signs service which is completely automated, real time message service which is shown via the road's message boards and through mobile devices to residents; CCTV live streaming service which helps residents to know the exact status of traffic by displaying a real video of the roads; and lastly, rerouting service which is more convenient for residents to navigate to another route with less congestion. As services are in the cloud, residents can easily choose the desired services of their choice from the mobile application or from the DMS boards that installed on the roads. As BPEL supports the workflow process to combine atomic services, one BPEL model is designed to combine any IoT services. A quantitative evaluation was conducted to measure the usability impact of the research to residents. The future work includes a stronger comparison analysis between the proposed models by measuring quality of service metrics as key performance indicators and covering more roads in Riyadh.

ACKNOWLEDGMENT

This work was supported by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, Saudi Arabia. The authors, therefore, gratefully acknowledge the DSR technical and financial support.

REFERENCES

[1] Institute of Public Administration. (2015) Traffic congestion.

[2] A. Al-Sibai. (2011) Bottleneck traffic creates a heavy loss of sr81 billion.

[3] F. Al-Naima and H. Hamd, "Vehicle traffic congestion estimation based on rfid," *International Journal of Engineering Business Management*, pp. 1–8, 2012.

[4] C. Campolo, A. Iera, A. Molinaro, S. Paratore, and G. Ruggeri, "Smartcar: An integrated smartphone-based platform to support traffic management applications," Vehicular Traffic Management for Smart Cities (VTM). IEEE, 2012, pp. 1–6.

[5] Grand View Research. (2014) Intelligent transportation system (its) market analysis and segment forecasts to 2020.

[6] Dornier Consulting and Rapp Trans, "Its deployment and integration plan for saudi arabia," Tech. Rep., 2011.

[7] C. Turcu, V. Gaitan, and C. Turcu, "An internet of things-based distributed intelligent system with self-optimization for controlling traffic-light intersections," Applied and Theoretical Electricity (ICATE), 2012 International Conference. IEEE, 2012, pp. 1–5.

[8] D. D. Romero, A. S. Prabuwono, A. Hasniaty *et al.*, "A review of sensing techniques for real-time traffic surveillance," *Journal of applied sciences*, vol. 11, no. 1, pp. 192–198, 2011.

[9] M. Heiskala, J.-P. Jokinen, and M. Tinnilä, "Crowdsensing-based transportation services-an analysis from business model and sustainability viewpoints," *Research in Transportation Business & Management*, vol. 18, pp. 38–48, 2016.

[10] D. Singh, G. Tripathi, and A. J. Jara, "A survey of internet-of-things: Future vision, architecture, challenges and services," in *Internet of things (WF-IoT), 2014 IEEE world forum on*. IEEE, 2014, pp. 287–292.

[11] M. Gigli and S. Koo, "Internet of things: Services and applications categorization," *Scientific Research Journal*, pp. 27–31, 2011.

[12] Riyadh Municipality. (2008) Altahliah's street is the first smart city application in saudi arabia.

[13] High Commission for the Development of Arriyadh. (2014) Extension of abu bakr al-siddiq and oruba street through the air base.

[14] S. Chowdhury, B. Haque, and G. Sarwar, "Traffic information interface development," *Transport and Telecommunication Journal*, pp. 91–96, 2014.

[15] R. Wang, L. Zhang, R. Sun, J. Gong, and L. Cui, "Easitia: A pervasive traffic information acquisition system based on wireless sensor networks," Intelligent Transportation Systems. IEEE, 2011, pp. 615–621.

[16] F. Aloul, A. Sagahyroon, A. Nahle, M. A. Dehn, and R. A. Anani, "Guideme: An effective rfid-based traffic monitoring system," Tech. Rep., 2012.

[17] J. Rijn, "Road capacities," *Journal of Indevelopment*, pp. 6–18, 2004.

[18] A. Bramantoro and T. Ishida, "User-centered qos in combining web services for interactive domain," in *Semantics, Knowledge and Grid, 2009. SKG 2009. Fifth International Conference on*. IEEE, 2009, pp. 41–48.

Selection of Important Sets by using K-Skyband Query for Sets

Md. Anisuzzaman Siddique, Asif Zaman, Yasuhiko Morimoto
Hiroshima University, Japan

***Abstract*—In this paper, we consider "sets" selection problem from a database. In conventional selection problem, which is "objects" selection problem, the skyline query has been utilized, since it can retrieve a set of important objects where each object isn't dominated by another object in a database. However, it is not effective when we have to select important sets, each of which contains more than one objects. Thus, we consider a "set skyline query" that retrieves non-dominated sets of objects from a database, which we call "object sets." The K-skyband query is a popular variant of the skyline query. It retrieves a set of objects, each of which is not dominated by K other objects. In this paper, we propose "K-Skyband set query." It retrieves important sets instead of objects. We investigated the properties of the query, as well as developing pruning strategies to avoid the unnecessary enumeration of objectsets and comparisons among them. Intensive experiments have been performed to examine the implemented algorithm. The results demonstrate the effectiveness and efficiency of the proposed algorithm.**

***Keywords*—Set Selection; Skyline Query; Skyline Set Query; Skyband Query; Skyband-set Query**

I. Introduction

To select important objects from a large-scale database is one of the most important processes to analyze the database. In the database literature, Borzsony et al. have been proposed a query that retrieves a set of objects where each is not dominated by another object in the database [1]. We call it a "skyline query." It has attracted lots of researchers and practitioners due to broad applicability in decision making and analysis tasks [2].

Let us consider an example of a financial investment problem in the table in Fig. 1, which are seven stocks with their corresponding prices (a_1) and risks (a_2). In general, all investors want to invest in stocks with lower commission costs and lower predicted risks. Fig. 1 shows the skyline query result for the table where the result is $\{O_1, O_2, O_3\}$.

In some cases, we may need to select two or more objects in a selection or in a lookup. For example, a user might want to invest in a combination of stocks with low costs and risk. We often call this combination an investment portfolio. Skyline query cannot be utilized effectively in this case because a user may have to choose dominated objects that are not among the skyline objects during multiple object selection in order to obtain their optimal choices.

The investment portfolio selection problem can be analyzed by using sets of objects, which we call "objectsets." We denote s as the number of objects in each objectsets. Let us consider an investor who wants to invest in two stocks, i.e., we assume that $s = 2$. Table I shows the different combinations of two objects from the running example where each record is a set of two stock objects. The attribute values for each objectset are the sum of their corresponding components values. Now, we can formalize the problem where the objectset skyline has lower attribute values a_1 (cost) as well as a_2 (risk). For $s = 2$, an objectset skyline query retrieves the result of $\{OS_{1,2}, OS_{2,3}, OS_{2,6}\}$ because no other objectset can dominate them (see Fig. 2). Notice that the conventional skyline result of $\{O_1, O_2, O_3\}$ does not provide sufficient insights into the selection problem. Investors always want to select non-dominated objectsets to ensure that their portfolio has the minimum cost with minimal risk.

ID	*a_1(cost)*	*a_2(risk)*
O_1	1	7
O_2	3	3
O_3	7	1
O_4	7	3
O_5	5	5
O_6	3	5
O_7	9	9

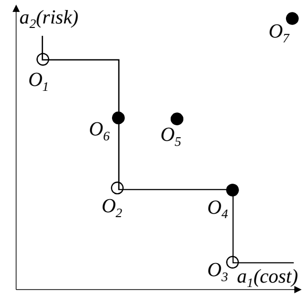


Fig. 1. Non-dominated objects and skyline.

Furthermore, we assume that an investor wants to invest in three stocks. After computing all combinations of three stocks, we need to check the dominance of these combinations to obtain the objectset skyline result. For combinations of three stocks, the objectset skyline query retrieves $\{OS_{1,2,3}$, $OS_{1,2,6}$, $OS_{2,3,4}$, $OS_{2,3,6}\}$ as the output result (see Fig. 3).

To address the issues related to set selection problem, we propose a "K-skyband-set". A K-skyband query, which is a popular variant of skyline query, returns objects that are not dominated by K other objects [6]. A K-skyband-set query retrieves objectsets, each individual objectset of which is not dominated by K other objectsets. In other word, an objectset in a K-skyband-set query's results may be dominated by at most $K - 1$ other objectsets. For example, if we set the objectset size $s = 1$ and the skyband value $K = 1$, then the skyband set query retrieves objectsets $\{O_1, O_2, O_3\}$ for the seven stocks example as in Fig. 1. In addition, for $s = 1$ and $K = 2$, the skyband-set query retrieves $\{O_1, O_2, O_3, O_6\}$. This is because the objectsets comprising O_1, O_2, O_3 are not dominated by any objectset, and objectset O_6 is dominated by only one objectset O_2. For $s = 1$ and $K = 3$, the skyband-set query retrieves $\{O_1, O_2, O_3, O_4, O_5, O_6\}$. Thus, the K-skyband query provides flexibility to increase and decrease the number of objectets by varying the size of K.

TABLE I. SETS OF TWO STOCKS

ID	$a_1(cost)$	$a_2(risk)$	ID	$a_1(cost)$	$a_2(risk)$	ID	$a_1(cost)$	$a_2(risk)$
$OS_{1,2}$	4	10	$OS_{2,4}$	10	6	$OS_{3,7}$	16	10
$OS_{1,3}$	8	8	$OS_{2,5}$	8	8	$OS_{4,5}$	12	8
$OS_{1,4}$	8	10	$OS_{2,6}$	6	8	$OS_{4,6}$	10	8
$OS_{1,5}$	6	12	$OS_{2,7}$	12	12	$OS_{4,7}$	16	12
$OS_{1,6}$	4	12	$OS_{3,4}$	14	4	$OS_{5,6}$	8	10
$OS_{1,7}$	10	16	$OS_{3,5}$	12	6	$OS_{5,7}$	14	14
$OS_{2,3}$	10	4	$OS_{3,6}$	10	6	$OS_{6,7}$	12	14

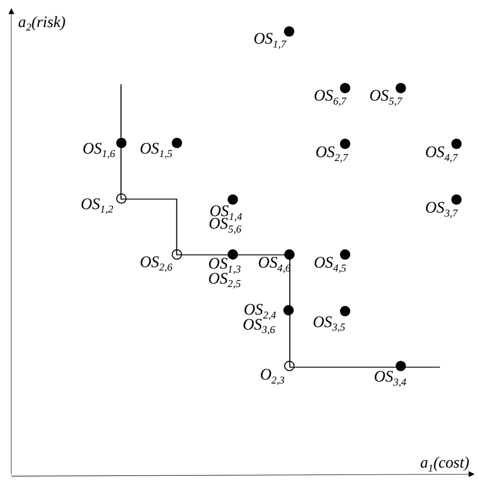


Fig. 2. Skyband-set (s=2, K=1).

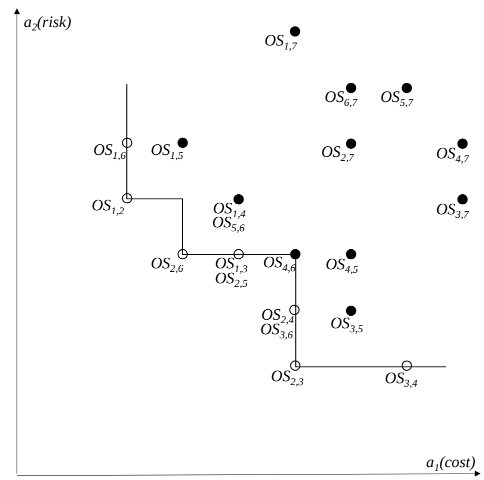


Fig. 4. Skyband-set (s=2, K=2).

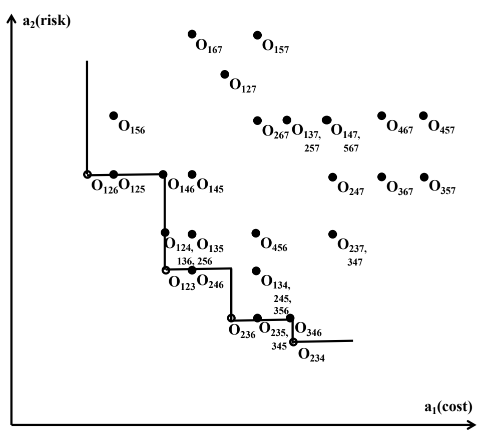


Fig. 3. Three-objects skyline.

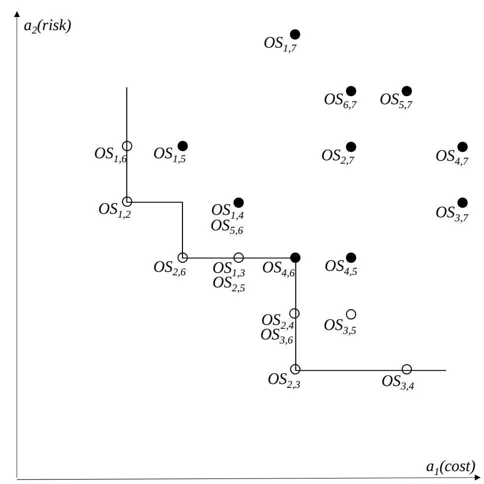


Fig. 5. Skyband-set (s=2, K=3).

Next, let us consider examples for $s = 2$. Table I represents the sets of two objects. According to Fig. 2, it is clear that the objectsets $\{OS_{1,2}, OS_{2,3}, OS_{2,6}\}$ are not dominated by any other objectset. Therefore, they are among the results obtained by the objectset skyband queries for $s = 2$ and $K = 1$. We can increase the number of the results set by increasing K. For $s = 2$ and $K = 2$, the skyband-set query retrieves $\{OS_{1,2}, OS_{1,3}, OS_{1,6}, OS_{2,3}, OS_{2,5}, OS_{2,6}, OS_{3,4}\}$ (see Fig. 4). Similarly, for $s = 2$ and $K = 3$, the skyband-set query retrieves $\{OS_{1,2}, OS_{1,3}, OS_{1,6}, OS_{2,3}, OS_{2,5}, OS_{2,6}, OS_{3,4}, OS_{3,5}\}$ (see Fig. 5).

The main challenge when developing an objectset skyband query is overcoming its large space complexity. For a data set with n records, the number of objectsets of size s is up to ${}_nC_s$. Thus, the time complexity is also high because we need to compute all of the objectsets to obtain the final result. The traditional skyline or skyband algorithm calculates all of the candidate objectsets progressively. Next, it updates the resulting objectset dynamically. Therefore, existing index structures, such as ZBtrees [5] and R-trees [6], are not suitable for objectset skyband computation. So far, there is no existing work that can compute K-skyband set efficiently.

We propose an efficient method that can select the K-skyband set in this paper. We utilize two filtering techniques to prevent computing large volumes of unnecessary objectsets. In addition, we examine and confirm that these pruning strategies are also useful for skyband objectset computation. We empirically verified the efficiency of the proposed algorithm by conducting several experiments with various datasets including synthetic and real datasets.

Organization of the remainder of this paper is as follows. In Section II, we first review the most basic research into skyline queries, before explaining objectset skyline queries. In Section III, we present the problem definition for a skyband-set query and the related definitions. We also discuss the properties of objectsets in this section. We explain the implementation with detailed examples as well as analyzing the proposed method for computing a skyband-set query in Section IV. In addition, we discuss the objectset pruning strategies in this section. We demonstrate the effectiveness and experimentally enumerate our proposed algorithm in a variety of settings in Section V. In Section VI, we conclude the proposals and give some suggestions for future research.

II. Related Work

There are two closely related works. One is "top-k combinatorial skyline queries" [9]. The other is "convex skyline objectsets" [8].

Su et al. [9] examined a method to obtain the optimal top-k combinations according to the preference order of attributes. They retrieve combinations, which are not dominated by another, incrementally according to the preference until the best k results are found. However, their method depends on the preference order of attributes and the required number of combinations (k value). The preference order and the limited number value k help to reduce the search space, which is exponential for combinations. In contrast, our proposed method does not require a attribute preference order nor the number of combinations. Therefore, their method is not suitable for solving our problem. Moreover, it is very difficult to select the appropriate preference order, which restricts the applicability of this method. Siddique et al. introduced the "convex skyline objectset" problem. In the problem, objectsets that are on the upper convex hull are all the skyline objectsets. However, the convex hull is a subset of the skyline. It means some of the skyline may not be on the convex hull. Our previously proposed method depends on the properties of a convex hull. Therefore, it cannot be applied to the skyband-set query retrieval problem.

In addition, "combination skyline queries" were introduced in [4], which described an index-based method called PBP for determining the skyline objectsets where they indexed each individual object. However, PBP also has some limitations. The key problem is that it is very difficult to specify the object selection pattern in advance because the end users are unfamiliar with the PBP algorithm. The pruning capability of the BPB algorithm depends on this pattern selection step. Thus, if a user selects an incorrect pattern, this can exponentially increase the search space for the computation of objectsets. Another limitation of the BPB algorithm is that it does not work well as the cardinality of the objectset size s changes. It only works well for a certain adequate size of s. Thus, it is necessary to start from scratch in order to set s adequately. By contrast, our solution does not require the construction of a pattern in advance. Therefore, there is no possibility of the search space increasing exponentially. Moreover, it is possible to vary the objectset size s in our proposed method. Some previous studies also considered the combination selection problem, but they are not related to our proposed method. In particular, Roy et al. [7] discussed how to choose the "maximal combinations". In the paper, a combination can be considered as "maximal", if it exceeds the pre-specified constraint during incremental adding procedure of an object. Finally, they selected the k most representative maximal combinations, thereby outputting objects with high diversity [7]. Wan et al. [10] studied the issue of constructing k profitable products from a new product set that is non-dominated by other existing products on the market. They constructed non-dominant products by allocating prices to the new products, which differed from the existing products. However, no previous studies considered skyband-set querying. Thus, existing methods are not suitable for solving objectset queries.

III. Preliminaries

Given a data set D with m-attributes $\{a_1, a_2, \cdots, a_m\}$ and n objects $\{O_1, O_2, \cdots, O_n\}$, we use $O_i.a_j$ to denote the j-th attribute's value for object O_i. Without loss of generality, let us assume that the domain of each attribute has positive numerical values and a lower value is desirable for each attribute.

Definition *Dominance*:

An object $O_i \in D$ can dominate another object $O_j \in D$, denoted as $O_i \leq O_j$, if $O_i.a_r \leq O_j.a_r$ $(1 \leq r \leq m)$ for all m attributes and $O_i.a_t < O_j.a_t$ $(1 \leq t \leq m)$ for at least one attribute. We refer to O_i as the *dominant object* and O_j as the *dominated object* between O_i and O_j. If O_i dominates O_j, then O_i is more desirable than O_j.

In the table of Fig. 1, object O_1 dominates object O_7, i.e., $O_1 \leq O_7$ because object O_1 has smaller values for both attributes than object O_7.

Definition *Skyline*:

An object $O_i \in D$ is referred to as a *skyline object* of D if and only if no other object $O_j \in D$ $(j \neq i)$ exists that can dominates O_i. The set of skyline objects in D is denoted by $Sky(D)$. Let us consider the seven stocks example, where object O_2 dominates $\{O_4, O_5, O_6, O_7\}$. No object can dominate objects $\{O_1, O_3\}$ in D. Therefore, a skyline query outputs the result of $Sky(D) = \{O_1, O_2, O_3\}$ (see Fig. 1).

Definition *s-objectset*:

An s-objectset, say OS, is a set of s objects chosen from D, i.e., $OS = \{O_1, \cdots, O_s\}$. To ensure that the representation remains simple, we denote s objects as $OS = OS_{1,\cdots,s}$, where each attribute value for OS is calculated using the following formula:

$$OS.a_j = f_j(O_1.a_j, \cdots, O_s.a_j), (1 \leq j \leq m), \qquad (1)$$

where f_j represent a monotonic aggregate function that receives s parameters as inputs and produces a single value. We can apply any monotonic aggregate function in our proposed method. However, for simplicity, we apply sum as the aggregate function, which returns the aggregate values for s objects as follows:

$$OS.a_j = \sum_{i=1}^{s} O_i.a_j, (1 \leq j \leq m) \tag{2}$$

Notice that the total number of s-objectsets in D is ${}_nC_s = \frac{n!}{(n-s)!s!}$ and we denote the number by $|S|$.

Definition *Dominance Relationship*:

An s-objectset $OS \in D$ will dominate another s-objectset $OS' \in D$, denoted as $OS \leq OS'$, if $OS.a_r \leq OS'.a_r$ $(1 \leq r \leq m)$ for all m attributes and $OS.a_t < OS'.a_t$ $(1 \leq t \leq m)$ for at least one attribute. We refer to this OS as the dominant s-objectset and we also refer to OS' as the dominated s-objectset.

Definition *Objectset Skyline*:

Let OS be an s-objectset in D. If OS is not dominated by any other OS in D, we call it a skyline OS (s-objectset). The skyline of s-objectsets in D, represented by $Sky_s(D)$, is the set of skyline s-objectsets in D. If we consider $s = 2$, for the data set shown in Table I, 2-objectsets $OS_{1,2}, OS_{2,3}$, and $OS_{2,6}$ are not dominated by any other 2-objectsets in D. Therefore, the 2-objectset skyline query outputs $Sky_2(D) = \{OS_{1,2}, OS_{2,3}, OS_{2,6}\}$ (see Fig. 2).

Definition *Skyband-set*:

A skyband-set query retrieves a set of objectsets, where each individual objectset is not dominated by k other objectsets. It also means, an objectset in the skyband-set query can be dominated by at most $k-1$ other objectsets. If we consider $s = 2$ and $k = 2$, a skyband-set query based on data set D outputs $\{OS_{1,2}, OS_{1,3}, OS_{1,6}, OS_{2,3}, OS_{2,5}, OS_{2,6}, OS_{3,4}\}$ as the query results (see Fig. 4).

Recall that any top-k query result (result of a top-k query based on an arbitrary linear function) is contained in the results of the skyband set. Therefore, we can use the skyband set results as a pre-processing step for skyline and top-k query computation.

IV. Skyband-Set Algorithm

In this section, we present the details of the proposed skyband-set method, which is a levelwise search algorithm. First, it calculates the 1-objectsets skyband, before computing all of the 2-objectsets skyband, and this procedure continues up to s-objectsets.

First, we assume that the skyband-set query is for objectset size $s = 1$ and skyband size $K = 1$. This query is similar to a conventional query and it produces similar non-dominant objects as the results set. Any traditional method is suitable for retrieving the results for this initial query. Therefore, we use the SFS algorithm developed in [3] to calculate the skyband-set query with $s = 1$ and $k = 1$. After completing all of the domination checks, we obtain the domination relation table shown in Table II, which we refer to as the $domRelationTable$.

Table II shows that objects O_1, O_2, and O_3 are not dominated by another objects. Therefore, for $s = 1$ and $k = 1$,

TABLE II. Dominance Relationship among 1-Objectsets

Object	*Dominant Object*
O_1	$\varnothing$
O_2	$\varnothing$
O_3	$\varnothing$
O_4	O_2, O_3
O_5	O_2, O_6
O_6	O_2
O_7	$O_1, O_2, O_3, O_4, O_5, O_6$

the skyband-set query output is $\{O_1, O_2, O_3\}$. Next, if we retain the objectset size $s = 1$ and increase the skyband value to $k = 2$, then the skyband-set output becomes $\{O_1, O_2, O_3, O_6\}$. Similarly, a skyband-set query for $s = 1$ and $k = 3$ will retrieve $\{O_1, O_2, O_3, O_4, O_5, O_6\}$, and the query for $s = 1$ and $K = 4$ will retrieve all objects as outputs.

As discussed regarding the skyband-set problem, if we select the objectset size as equal to s, then a data set D with n objects can retrieve the total number objectsets as $|S| = {}_nC_s$. For a large value of n, this represents a severe algorithmic challenge compared with the traditional skyline computation problem. Fig. 2, 4, and 5 show that for $s = 2$, the total number of possible sets is $|S|$ = 21 (${}_7C_2$). Therefore, in order to produce $domRelationTable$ in a similar manner to Table II, we must conduct a domination check based on 420 (21 * 20) comparisons. Thus, even for a small data set containing thousands of objects, the total number of objectsets is remarkably large and vast numbers of comparisons are required. However, given Theorem 1 and Theorem 2, we are free to obtain the dominance relationships for the objectsets without composing them. This also avoids a large number of unnecessary comparisons in the domination check.

Theorem 1: Let OS_1, OS_2, and OS_3 be three s-objectsets in D. If $OS_1 \leq OS_2$, then $OS_1OS_3 \leq OS_2OS_3$, where OS_1OS_3 is the $2s$-objectset that includes the $2s$ objects that are included in OS_1 or OS_3, and similarly, OS_2OS_3 is the $2s$-objectset that includes the $2s$ objects that are included in OS_2 or OS_3.

Proof: Let OS and OS' be two s-objectsets, and let O be an object that is not included in OS or OS' in an m-dimensional database D. Assume that $OS \leq OS'$. If we add O to OS and obtain the superset of OS, which is $O \cup OS$, then by the definition, $(O \cup OS).a_j = O.a_j + \sum_{i=1}^{s} O_i.a_j, (1 \leq j \leq m)$. Similarly, we can make $O \cup OS'$. We can say that $(O \cup OS).a_j \leq (O \cup OS').a_j$ for all j and $(O \cup OS).a_j < (O \cup OS').a_j$ at least one j $(1 \leq j \leq m)$ by the assumption; therefore, $(O \cup OS) \leq (O \cup OS')$.□

Theorem 2: If an objectset OS is dominated by at least k other objectsets, then all supersets that contain OS cannot be a member of the k-skyband-set. Therefore, further skyband-set calculations do not require the composition of the super-objectsets that contain OS.

Proof: Assume that a data set D has four objectsets: OS_1, OS_2, OS_3, and OS_4. If objectset OS_1 is dominated by OS_2 and OS_3, then the proposed algorithm does not require that we compose super-objectset OS_1OS_4 for $K = 2$. By using Theorem 1, if $OS_2 \leq OS_1$, then $OS_2OS_4 \leq OS_1OS_4$. This is also true for $OS_3OS_4 \leq OS_1OS_4$. Thus, two objectsets dominate OS_1OS_4.□

TABLE III. DOMINANCE RELATIONSHIP AMONG 2-OBJECTSETS

Objectset	*Dominant Objectset*	*Objectset*	*Dominant Objectset*
$OS_{1,2}$	$\varnothing$	$OS_{3,4}$	$OS_{2,3}$
$OS_{1,3}$	$\varnothing$	$OS_{3,5}$	$OS_{2,3}, OS_{3,6}$
$OS_{1,4}$	$OS_{1,2}, OS_{1,3}$	$OS_{3,6}$	$OS_{2,3}$
$OS_{1,5}$	$OS_{1,2}, OS_{1,6}$	$OS_{3,7}$	$OS_{1,3}, OS_{2,3}, OS_{3,4}, OS_{3,5}, OS_{3,6}$
$OS_{1,6}$	$OS_{1,2}$	$OS_{4,5}$	$OS_{2,5}, OS_{3,5}, OS_{2,4}, OS_{4,6}$
$OS_{1,7}$	$OS_{1,2}, OS_{1,3}, OS_{1,4}, OS_{1,5}, OS_{1,6}$	$OS_{4,6}$	$OS_{2,6}, OS_{3,6}, OS_{2,4}$
$OS_{2,3}$	$\varnothing$	$OS_{4,7}$	$OS_{1,4}, OS_{2,4}, OS_{3,4}, OS_{4,5}, OS_{4,6}, OS_{2,7}, OS_{3,7}$
$OS_{2,4}$	$OS_{2,3}$	$OS_{5,6}$	$OS_{2,6}, OS_{2,5}$
$OS_{2,5}$	$OS_{2,6}$	$OS_{5,7}$	$OS_{1,5}, OS_{2,5}, OS_{3,5}, OS_{4,5}, OS_{5,6}, OS_{2,7}, OS_{6,7}$
$OS_{2,6}$	$\varnothing$	$OS_{6,7}$	$OS_{1,6}, OS_{2,6}, OS_{3,6}, OS_{4,6}, OS_{5,6}, OS_{2,7}$
$OS_{2,7}$	$OS_{1,2}, OS_{2,3}, OS_{2,4}, OS_{2,5}, OS_{2,6}$		

By utilizing Theorem 1 and 2, we can prune a large number of the unnecessary dominance checks. By checking Table II from top to bottom, we first find that object O_4 is dominated by O_2 and O_3, from which we derive $O_2 \leq O_4$ and $O_3 \leq O_4$. Each of the relationships with $\{O_1, O_5, O_6, O_7\}$ has not yet been examined. By applying Theorem 2 and without performing any comparisons, we can easily compute the following dominance relationship for 2-objectsets.

$$OS_{1,2} \leq OS_{1,4} \quad OS_{1,3} \leq OS_{1,4} \quad OS_{2,5} \leq OS_{4,5}$$
$$OS_{3,5} \leq OS_{4,5} \quad OS_{2,6} \leq OS_{4,6} \quad OS_{3,6} \leq OS_{4,6}$$
$$OS_{2,7} \leq OS_{4,7} \quad OS_{3,7} \leq OS_{4,7}$$

Moreover, we also obtain two more dominance relationship for O_4, as follows:

$$OS_{2,3} \leq OS_{2,4} \quad OS_{2,3} \leq OS_{3,4}$$

Similarly, object O_5 is dominated by $\{O_2, O_6\}$, from which we derive $OS_2 \leq O_5$ and $O_6 \leq O_5$. These relationships can also be used to derive the following relationships for other objects $\{O_1, O_3, O_4, O_7\}$.

$$OS_{1,2} \leq OS_{1,5} \quad OS_{1,6} \leq OS_{1,5} \quad OS_{2,3} \leq OS_{3,5}$$
$$OS_{3,6} \leq OS_{3,5} \quad OS_{2,4} \leq OS_{4,5} \quad OS_{4,6} \leq OS_{4,5}$$
$$OS_{2,7} \leq OS_{5,7} \quad OS_{6,7} \leq OS_{5,7}$$

In addition, we can derive

$$OS_{2,6} \leq OS_{2,5} \quad OS_{2,6} \leq OS_{5,6}.$$

For the sixth relationship $O_2 \leq O_6$ in Table II, the relationships with others are derived as follows:

$$OS_{1,2} \leq OS_{1,6} \quad OS_{2,3} \leq OS_{3,6} \quad OS_{2,4} \leq OS_{4,6}$$
$$OS_{2,5} \leq OS_{5,6} \quad OS_{2,7} \leq OS_{6,7}$$

Similarly, for the last relationship in Table II, the following relationships are derived.

$$OS_{1,2} \leq OS_{1,7},\ OS_{1,3} \leq OS_{1,7},\ OS_{1,4} \leq OS_{1,7},\ OS_{1,5} \leq OS_{1,7},\ OS_{1,6} \leq OS_{1,7}$$
$$OS_{1,2} \leq OS_{2,7},\ OS_{2,3} \leq OS_{2,7},\ OS_{2,4} \leq OS_{2,7},\ OS_{2,5} \leq OS_{2,7},\ OS_{2,6} \leq OS_{2,7}$$
$$OS_{1,3} \leq OS_{3,7},\ OS_{2,3} \leq OS_{3,7},\ OS_{3,4} \leq OS_{3,7},\ OS_{3,5} \leq OS_{3,7},\ OS_{3,6} \leq OS_{3,7}$$
$$OS_{1,4} \leq OS_{4,7},\ OS_{2,4} \leq OS_{4,7},\ OS_{3,4} \leq OS_{4,7},\ OS_{4,5} \leq OS_{4,7},\ OS_{4,6} \leq OS_{4,7}$$
$$OS_{1,5} \leq OS_{5,7},\ OS_{2,5} \leq OS_{5,7},\ OS_{3,5} \leq OS_{5,7},\ OS_{4,5} \leq OS_{5,7},\ OS_{5,6} \leq OS_{5,7}$$
$$OS_{1,6} \leq OS_{6,7},\ OS_{2,6} \leq OS_{6,7},\ OS_{3,6} \leq OS_{6,7},\ OS_{4,6} \leq OS_{6,7},\ OS_{5,6} \leq OS_{6,7}$$

Thus, we can easily construct another "Dominance Relationship Table" for objectset with $s = 2$ using Table II without comparing objectsets according to Theorem 1 and Theorem 2. The new "Dominance Relationship Table" with $s = 2$ is shown in Table III.

Table III is used to retrieve candidates for the objectset skyband queries with $s = 2$ (the objecetset size is two). For example, if a user specifies $s = 2$ and $k = 1$ for a skyband-set query, then the proposed algorithm will select the candidate objectsets comprising $\{OS_{1,2}, OS_{1,3}, OS_{2,3}, OS_{2,6}\}$ from Table III. Next, it will compose the four objectsets and perform the domination checks among them. After the domination checks, we find that $OS_{1,3}$ is dominated by objectset $OS_{2,6}$. Thus, the proposed algorithm will output $\{OS_{1,2}, OS_{2,3}, OS_{2,6}\}$ as the skyband-set.

Next, if the user specifies $s = 2$ and $k = 2$ for a skyband-set query, then the proposed algorithm will first select the candidate objectsets $\{OS_{1,2}, OS_{1,3}, OS_{1,6}, OS_{2,3}, OS_{2,4}, OS_{2,5}, OS_{2,6}, OS_{3,4}, OS_{3,6}\}$, before performing domination checks among these candidate objectsets. Finally, it retrieves $\{OS_{1,2}, OS_{1,3}, OS_{1,6}, OS_{2,3}, OS_{2,5}, OS_{2,6}, OS_{3,4}\}$ as the skyband-set query result.

Similar to the example above, we can retrieve candidate objectsets for any skyband-set query with $s = 2$ from the dominance relations in Table III.

Now, if the user wants to select the top-3 objectsets, our proposed algorithm will examine the μ score of each objectset in the skyband set results and select $OS_{1,2}$, $OS_{2,3}$, and $OS_{2,6}$ with μ scores of 10, 12, and 11, respectively, which are the top-3 scores in the database. In general, the proposed algorithm can retrieve the top-k query from the k-skyband-set results, which can be computed efficiently by the algorithm.

Next, to construct "Dominance Relationship Table" for $s = 3$, we follow the same procedure again and by utilizing Theorems 1 and 2, we can prune a large number of unnecessary dominance checks. After checking Table III, we first find that objectset $OS_{1,4}$ is dominated by $OS_{1,2}$ and $OS_{1,3}$, from which we can derive $OS_{1,2} \leq OS_{1,4}$ and $OS_{1,3} \leq OS_{1,4}$. Each of the relationships with $\{O_5, O_6, O_7\}$ has not yet been examined. By using Theorem 2, we can find the following dominance relationship for 3-objectsets without making any comparisons.

$$OS_{1,2,5} \leq OS_{1,4,5} \quad OS_{1,3,5} \leq OS_{1,4,5}$$

TABLE IV. DOMINANCE RELATIONSHIP AMONG 3-OBJECTSETS

Obj.set	*Dominant Objectset*	*Obj.set*	*Dominant Objectset*
$OS_{1,2,3}$	$\varnothing$	$OS_{2,3,7}$	$OS_{1,2,3}, OS_{2,3,4}, OS_{2,3,5}, OS_{2,3,6}$
$OS_{1,2,4}$	$OS_{1,2,3}$	$OS_{2,4,5}$	$OS_{2,3,5}, OS_{2,4,6}$
$OS_{1,2,5}$	$OS_{1,2,6}$	$OS_{2,4,6}$	$OS_{2,3,6}$
$OS_{1,2,6}$	$\varnothing$	$OS_{2,4,7}$	$OS_{1,2,4}, OS_{2,3,4}, OS_{2,3,7}, OS_{2,4,5}, OS_{2,4,6}$
$OS_{1,2,7}$	$OS_{1,2,3}, OS_{1,2,4}, OS_{1,2,5}, OS_{1,2,6}$	$OS_{2,5,6}$	$\varnothing$
$OS_{1,3,4}$	$OS_{1,2,3}$	$OS_{2,5,7}$	$OS_{1,2,5}, OS_{2,3,5}, OS_{2,4,5}, OS_{2,5,6}, OS_{2,6,7}$
$OS_{1,3,5}$	$OS_{1,2,3}, OS_{1,3,6}$	$OS_{2,6,7}$	$OS_{1,2,6}, OS_{2,3,6}, OS_{2,4,6}, OS_{2,5,6}$
$OS_{1,3,6}$	$OS_{1,2,3}$	$OS_{3,4,5}$	$OS_{2,3,4}, OS_{2,3,5}, OS_{3,4,6}$
$OS_{1,3,7}$	$OS_{1,2,3}, OS_{1,3,4}, OS_{1,3,5}, OS_{1,3,6}$	$OS_{3,4,6}$	$OS_{2,3,4}, OS_{2,3,6}$
$OS_{1,4,5}$	$OS_{1,2,4}, OS_{1,2,5}, OS_{1,3,5}, OS_{1,4,6}$	$OS_{3,4,7}$	$OS_{1,3,4}, OS_{2,3,4}, OS_{2,3,7}, OS_{3,4,5}, OS_{3,4,6}$
$OS_{1,4,6}$	$OS_{1,2,4}, OS_{1,2,5}, OS_{1,2,6}, OS_{1,3,6}$	$OS_{3,5,6}$	$OS_{2,3,5}, OS_{2,3,6}$
$OS_{1,4,7}$	$OS_{1,2,4}, OS_{1,2,7}, OS_{1,3,4}, OS_{1,3,7}, OS_{1,4,5}, OS_{1,4,6}$	$OS_{3,5,7}$	$OS_{1,3,5}, OS_{2,3,5}, OS_{2,3,7}, OS_{3,4,5}, OS_{3,5,6}, OS_{3,6,7}$
$OS_{1,5,6}$	$OS_{1,2,5}, OS_{1,2,6}$	$OS_{3,6,7}$	$OS_{1,3,6}, OS_{2,3,6}, OS_{2,3,7}, OS_{3,4,6}, OS_{3,5,6}$
$OS_{1,5,7}$	$OS_{1,2,5}, OS_{1,2,7}, OS_{1,3,5}, OS_{1,4,5}, OS_{1,5,6}, OS_{1,6,7}$	$OS_{4,5,6}$	$OS_{2,4,5}, OS_{2,4,6}, OS_{2,5,6}, OS_{3,5,6}$
$OS_{1,6,7}$	$OS_{1,2,6}, OS_{1,2,7}, OS_{1,3,6}, OS_{1,4,6}, OS_{1,5,6}$	$OS_{4,5,7}$	$OS_{1,4,5}, OS_{2,4,5}, OS_{2,4,7}, OS_{2,5,7}, OS_{3,4,5}, OS_{3,5,7}, OS_{4,5,6}, OS_{4,6,7}$
$OS_{2,3,4}$	$\varnothing$	$OS_{4,6,7}$	$OS_{1,4,6}, OS_{2,4,6}, OS_{2,4,7}, OS_{2,6,7}, OS_{3,4,6}, OS_{3,6,7}, OS_{4,5,6}$
$OS_{2,3,5}$	$OS_{2,3,6}$	$OS_{5,6,7}$	$OS_{1,5,6}, OS_{2,5,6}, OS_{2,5,7}, OS_{2,6,7}, OS_{3,5,6}, OS_{4,5,6}$
$OS_{2,3,6}$	$\varnothing$		

$$OS_{1,2,6} \leq OS_{1,4,6} \quad OS_{1,3,6} \leq OS_{1,4,6}$$
$$OS_{1,2,7} \leq OS_{1,4,7} \quad OS_{1,3,7} \leq OS_{1,4,7}$$

Moreover, we also obtain two more dominance relationship for $OS_{1,4}$, as follows.

$$OS_{1,2,3} \leq OS_{1,2,4} \quad OS_{1,2,3} \leq OS_{1,3,4}$$

Similarly, objectset $OS_{1,5}$ is dominated by $\{OS_{1,2}, OS_{1,6}\}$, from which we derive $OS_{1,2} \leq OS_{1,5}$ and $OS_{1,6} \leq OS_{1,5}$. These relationships are used to derive the following relationships with other objects $\{O_3, O_4, O_7\}$.

$$OS_{1,2,3} \leq OS_{1,3,5} \quad OS_{1,3,6} \leq OS_{1,3,5}$$
$$OS_{1,2,4} \leq OS_{1,4,5} \quad OS_{1,4,6} \leq OS_{1,4,5}$$
$$OS_{1,2,7} \leq OS_{1,5,7} \quad OS_{1,6,7} \leq OS_{1,5,7}$$

In addition, we can derive:

$$OS_{1,2,6} \leq OS_{1,2,5} \quad OS_{1,2,6} \leq OS_{1,5,6}.$$

For $OS_{1,2} \leq OS_{1,6}$ and the relationships with $\{O_3, O_4, O_5, O_7\}$, we can derive the following:

$$OS_{1,2,3} \leq OS_{1,3,6} \quad OS_{1,2,4} \leq OS_{1,4,6}$$
$$OS_{1,2,5} \leq OS_{1,5,6} \quad OS_{1,2,7} \leq OS_{1,6,7}$$

Next, for $\{OS_{1,2}, OS_{1,3}, OS_{1,4}, OS_{1,5}, OS_{1,6}\} \leq OS_{1,7}$, we derive the following.

$$OS_{1,2,3} \leq OS_{1,3,7} \quad OS_{1,2,4} \leq OS_{1,4,7}$$
$$OS_{1,2,5} \leq OS_{1,5,7} \quad OS_{1,2,6} \leq OS_{1,6,7}$$
$$OS_{1,2,3} \leq OS_{1,2,7} \quad OS_{1,3,4} \leq OS_{1,4,7}$$
$$OS_{1,3,5} \leq OS_{1,5,7} \quad OS_{1,3,6} \leq OS_{1,6,7}$$
$$OS_{1,2,4} \leq OS_{1,2,7} \quad OS_{1,3,4} \leq OS_{1,3,7}$$
$$OS_{1,4,5} \leq OS_{1,5,7} \quad OS_{1,4,6} \leq OS_{1,6,7}$$
$$OS_{1,2,5} \leq OS_{1,2,7} \quad OS_{1,3,5} \leq OS_{1,3,7}$$
$$OS_{1,4,5} \leq OS_{1,4,7} \quad OS_{1,5,6} \leq OS_{1,6,7}$$
$$OS_{1,2,6} \leq OS_{1,2,7} \quad OS_{1,3,6} \leq OS_{1,3,7}$$
$$OS_{1,4,6} \leq OS_{1,4,7} \quad OS_{1,5,6} \leq OS_{1,5,7}$$

For $OS_{2,3} \leq OS_{2,4}$ and the relationships with $\{O_1, O_5, O_6, O_7\}$, we can derive the following.

$$OS_{1,2,3} \leq OS_{1,2,4} \quad OS_{2,3,5} \leq OS_{2,4,5}$$
$$OS_{2,3,6} \leq OS_{2,4,6} \quad OS_{2,3,7} \leq OS_{2,4,7}$$

For $OS_{2,6} \leq OS_{2,5}$ and the relationships with $\{O_1, O_3, O_4, O_7\}$, we can derive the following.

$$OS_{1,2,6} \leq OS_{1,2,5} \quad OS_{2,3,6} \leq OS_{2,3,5}$$
$$OS_{2,4,6} \leq OS_{2,4,5} \quad OS_{2,6,7} \leq OS_{2,5,7}$$

Next, for $\{OS_{1,2}, OS_{2,3}, OS_{2,4}, OS_{2,5}, OS_{2,6}\} \leq OS_{2,7}$, we can derive the following relationships:

$$OS_{1,2,3} \leq OS_{2,3,7} \quad OS_{1,2,3} \leq OS_{1,2,7}$$
$$OS_{1,2,4} \leq OS_{2,4,7} \quad OS_{2,3,4} \leq OS_{2,4,7}$$
$$OS_{1,2,5} \leq OS_{2,5,7} \quad OS_{2,3,5} \leq OS_{2,5,7}$$
$$OS_{1,2,6} \leq OS_{2,6,7} \quad OS_{2,3,6} \leq OS_{2,6,7}$$
$$OS_{1,2,4} \leq OS_{1,2,7} \quad OS_{1,2,5} \leq OS_{1,2,7}$$
$$OS_{2,3,4} \leq OS_{2,3,7} \quad OS_{2,3,5} \leq OS_{2,3,7}$$
$$OS_{2,4,5} \leq OS_{2,5,7} \quad OS_{2,4,5} \leq OS_{2,4,7}$$
$$OS_{2,4,6} \leq OS_{2,6,7} \quad OS_{2,5,6} \leq OS_{2,6,7}$$
$$OS_{1,2,6} \leq OS_{1,2,7} \quad OS_{2,3,6} \leq OS_{2,3,7}$$
$$OS_{2,4,6} \leq OS_{2,4,7} \quad OS_{2,5,6} \leq OS_{2,5,7}$$

Similarly, for $OS_{2,3} \leq OS_{3,4}$ and the relationships with $\{O_1, O_5, O_6, O_7\}$, we can derive the following.

$$OS_{1,2,3} \leq OS_{1,3,4} \quad OS_{2,3,5} \leq OS_{3,4,5}$$
$$OS_{2,3,6} \leq OS_{3,4,6} \quad OS_{2,3,7} \leq OS_{3,4,7}$$

Objectset $OS_{3,5}$ is dominated by $\{OS_{2,3}, OS_{3,6}\}$, from which we derive $OS_{2,3} \leq OS_{3,5}$ and $OS_{3,6} \leq OS_{3,5}$. These relationships can be used to derive the following relationships with other objects comprising $\{O_1, O_4, O_7\}$.

$$OS_{1,2,3} \leq OS_{1,3,5} \quad OS_{1,3,6} \leq OS_{1,3,5}$$
$$OS_{2,3,4} \leq OS_{3,4,5} \quad OS_{3,4,6} \leq OS_{3,4,5}$$
$$OS_{2,3,7} \leq OS_{3,5,7} \quad OS_{3,6,7} \leq OS_{3,5,7}$$

In addition, we can derive:

$$OS_{2,3,6} \leq OS_{3,5,6} \quad OS_{2,3,6} \leq OS_{2,3,5}.$$

Again, for $OS_{2,3} \leq OS_{3,6}$ and the relationships with $\{O_1, O_4, O_5, O_7\}$, we can derive the following:

$$OS_{1,2,3} \leq OS_{1,3,6} \quad OS_{2,3,4} \leq OS_{3,4,5}$$
$$OS_{2,3,5} \leq OS_{3,5,6} \quad OS_{2,3,7} \leq OS_{3,6,7}$$

Then, for $\{OS_{1,3}, OS_{2,3}, OS_{3,4}, OS_{3,5}, OS_{3,6}\} \leq OS_{3,7}$, we can derive the following relationships:

$$OS_{1,2,3} \leq OS_{2,3,7} \quad OS_{1,2,3} \leq OS_{1,3,7}$$
$$OS_{1,3,4} \leq OS_{3,4,7} \quad OS_{2,3,4} \leq OS_{3,4,7}$$
$$OS_{1,3,5} \leq OS_{3,5,7} \quad OS_{2,3,5} \leq OS_{3,5,7}$$

$$\begin{array}{ll}
OS_{1,3,6} \leq OS_{3,6,7} & OS_{2,3,6} \leq OS_{3,6,7} \\
OS_{1,3,4} \leq OS_{1,3,7} & OS_{1,3,5} \leq OS_{1,3,7} \\
OS_{2,3,4} \leq OS_{2,3,7} & OS_{2,3,5} \leq OS_{2,3,7} \\
OS_{3,4,5} \leq OS_{3,5,7} & OS_{3,4,5} \leq OS_{3,4,7} \\
OS_{3,4,6} \leq OS_{3,6,7} & OS_{3,5,6} \leq OS_{3,6,7} \\
OS_{1,3,6} \leq OS_{1,3,7} & OS_{2,3,6} \leq OS_{2,3,7} \\
OS_{3,4,6} \leq OS_{3,4,7} & OS_{3,5,6} \leq OS_{3,5,7}
\end{array}$$

For $OS_{2,4}, OS_{2,5}, OS_{3,5}, OS_{4,6} \leq OS_{4,5}$ and the relationships with $\{O_1, O_7\}$, we can derive the following:

$$\begin{array}{ll}
OS_{1,2,4} \leq OS_{1,4,5} & OS_{2,4,7} \leq OS_{4,5,7} \\
OS_{1,2,5} \leq OS_{1,4,5} & OS_{2,5,7} \leq OS_{4,5,7} \\
OS_{1,3,5} \leq OS_{1,4,5} & OS_{3,5,7} \leq OS_{4,5,7} \\
OS_{1,4,6} \leq OS_{1,4,5} & OS_{4,6,7} \leq OS_{4,5,7}
\end{array}$$

In addition, we can derive the following:

$$\begin{array}{ll}
OS_{2,3,4} \leq OS_{3,4,5} & OS_{2,3,5} \leq OS_{3,4,5} \\
OS_{2,4,6} \leq OS_{4,5,6} & OS_{2,5,6} \leq OS_{4,5,6} \\
OS_{2,3,5} \leq OS_{2,4,5} & OS_{2,4,6} \leq OS_{2,4,5} \\
OS_{3,5,6} \leq OS_{4,5,6} & OS_{3,4,6} \leq OS_{3,4,5}
\end{array}$$

For $OS_{2,4}, OS_{2,6}, OS_{3,6} \leq OS_{4,6}$ and the relationships with $\{O_1, O_5, O_7\}$, we can derive the following:

$$\begin{array}{ll}
OS_{1,2,4} \leq OS_{1,4,6} & OS_{2,4,5} \leq OS_{4,5,6} \\
OS_{1,2,6} \leq OS_{1,4,6} & OS_{2,5,6} \leq OS_{4,5,6} \\
OS_{1,3,6} \leq OS_{1,4,6} & OS_{3,5,6} \leq OS_{4,5,6}
\end{array}$$

In addition, we can derive the following:

$$\begin{array}{ll}
OS_{2,4,7} \leq OS_{4,6,7} & OS_{2,3,4} \leq OS_{3,4,6} \\
OS_{2,6,7} \leq OS_{4,6,7} & OS_{2,3,6} \leq OS_{3,4,6} \\
OS_{3,6,7} \leq OS_{4,6,7} & OS_{2,3,6} \leq OS_{2,4,6}
\end{array}$$

For $OS_{1,4}, OS_{2,4}, OS_{2,7}, OS_{3,4}, OS_{3,7}, OS_{4,5}, OS_{4,6} \leq OS_{4,7}$, the following relationships can be derived.

$$\begin{array}{ll}
OS_{1,2,4} \leq OS_{2,4,7} & OS_{1,3,4} \leq OS_{3,4,7} \\
OS_{1,4,5} \leq OS_{4,5,7} & OS_{1,4,6} \leq OS_{4,6,7} \\
OS_{1,2,4} \leq OS_{1,4,7} & OS_{2,3,4} \leq OS_{3,4,7} \\
OS_{2,4,5} \leq OS_{4,5,7} & OS_{2,4,6} \leq OS_{4,6,7} \\
OS_{1,2,7} \leq OS_{1,4,7} & OS_{2,3,7} \leq OS_{3,4,7} \\
OS_{2,5,7} \leq OS_{4,5,7} & OS_{2,6,7} \leq OS_{4,6,7} \\
OS_{1,3,4} \leq OS_{1,4,7} & OS_{2,3,4} \leq OS_{2,4,7} \\
OS_{3,4,5} \leq OS_{4,5,7} & OS_{3,4,6} \leq OS_{4,6,7} \\
OS_{1,3,7} \leq OS_{1,4,7} & OS_{2,3,7} \leq OS_{2,4,7} \\
OS_{3,5,7} \leq OS_{4,5,7} & OS_{3,6,7} \leq OS_{4,6,7} \\
OS_{1,4,5} \leq OS_{1,4,7} & OS_{2,4,5} \leq OS_{2,4,7} \\
OS_{3,4,5} \leq OS_{3,4,7} & OS_{4,5,6} \leq OS_{4,6,7} \\
OS_{1,4,6} \leq OS_{1,4,7} & OS_{2,4,6} \leq OS_{2,4,7} \\
OS_{3,4,5} \leq OS_{3,4,7} & OS_{4,5,6} \leq OS_{4,5,7}
\end{array}$$

For $OS_{2,5}, OS_{2,6} \leq OS_{5,6}$ and the relationships with $\{O_1, O_3, O_4, O_7\}$, we can derive the following:

$$\begin{array}{ll}
OS_{1,2,5} \leq OS_{1,5,6} & OS_{1,2,6} \leq OS_{1,5,6} \\
OS_{2,3,5} \leq OS_{3,5,6} & OS_{2,3,6} \leq OS_{3,5,6} \\
OS_{2,4,5} \leq OS_{4,5,6} & OS_{2,4,6} \leq OS_{4,5,6} \\
OS_{2,5,7} \leq OS_{5,6,7} & OS_{2,6,7} \leq OS_{5,6,7}
\end{array}$$

For $OS_{1,5}, OS_{2,5}, OS_{2,7}, OS_{3,5}, OS_{4,5}, OS_{5,6}, OS_{6,7} \leq OS_{5,7}$, the following relationships can be derived:

$$\begin{array}{ll}
OS_{1,2,5} \leq OS_{2,5,7} & OS_{1,2,5} \leq OS_{1,5,7} \\
OS_{1,2,7} \leq OS_{1,5,7} & OS_{1,3,5} \leq OS_{1,5,7} \\
OS_{1,4,5} \leq OS_{1,5,7} & OS_{1,5,6} \leq OS_{1,5,7} \\
OS_{1,6,7} \leq OS_{1,5,7} & OS_{1,3,5} \leq OS_{3,5,7} \\
OS_{2,3,5} \leq OS_{3,5,7} & OS_{2,3,7} \leq OS_{3,5,7} \\
OS_{2,3,5} \leq OS_{2,5,7} & OS_{2,4,5} \leq OS_{2,5,7} \\
OS_{2,5,6} \leq OS_{2,5,7} & OS_{2,6,7} \leq OS_{2,5,7} \\
OS_{1,4,5} \leq OS_{4,5,7} & OS_{2,4,5} \leq OS_{4,5,7} \\
OS_{2,4,7} \leq OS_{4,5,7} & OS_{3,4,5} \leq OS_{4,5,7} \\
OS_{3,4,5} \leq OS_{3,5,7} & OS_{3,5,6} \leq OS_{3,5,7} \\
OS_{3,6,7} \leq OS_{3,5,7} & OS_{1,5,6} \leq OS_{5,6,7} \\
OS_{2,5,6} \leq OS_{5,6,7} & OS_{2,6,7} \leq OS_{5,6,7} \\
OS_{3,5,6} \leq OS_{5,6,7} & OS_{4,5,6} \leq OS_{5,6,7} \\
OS_{4,5,6} \leq OS_{4,5,7} & OS_{4,6,7} \leq OS_{4,5,7}
\end{array}$$

Finally, for $OS_{1,6}, OS_{2,6}, OS_{2,7}, OS_{3,6}, OS_{4,6}, OS_{5,6} \leq OS_{6,7}$, the following relationships can be derived:

$$\begin{array}{ll}
OS_{1,2,6} \leq OS_{2,6,7} & OS_{1,2,6} \leq OS_{1,6,7} \\
OS_{1,2,7} \leq OS_{1,6,7} & OS_{1,3,6} \leq OS_{1,6,7} \\
OS_{1,4,6} \leq OS_{1,6,7} & OS_{1,5,6} \leq OS_{1,6,7} \\
OS_{1,3,6} \leq OS_{3,6,7} & OS_{2,3,6} \leq OS_{3,6,7} \\
OS_{2,3,7} \leq OS_{3,6,7} & OS_{2,3,6} \leq OS_{2,6,7} \\
OS_{2,4,6} \leq OS_{2,6,7} & OS_{2,5,6} \leq OS_{2,6,7} \\
OS_{1,4,6} \leq OS_{4,6,7} & OS_{2,4,6} \leq OS_{4,6,7} \\
OS_{2,4,7} \leq OS_{4,6,7} & OS_{3,4,6} \leq OS_{4,6,7} \\
OS_{3,4,6} \leq OS_{3,6,7} & OS_{3,5,6} \leq OS_{3,6,7} \\
OS_{1,5,6} \leq OS_{5,6,7} & OS_{2,5,6} \leq OS_{5,6,7} \\
OS_{2,5,7} \leq OS_{5,6,7} & OS_{3,5,6} \leq OS_{5,6,7} \\
OS_{4,5,6} \leq OS_{5,6,7} & OS_{4,5,6} \leq OS_{4,6,7}
\end{array}$$

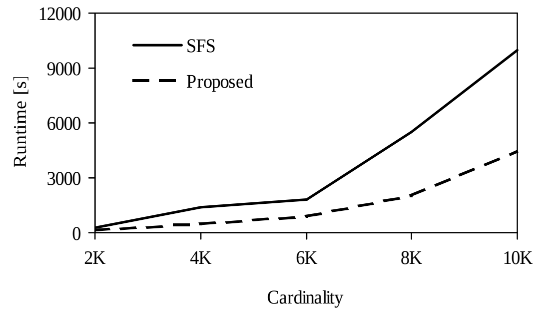
Now, by using all of the relationships given above, we can compute "Dominance Relationship Table" for $s = 3$ from "Dominance Relationship Table" for $s = 2$, as shown in Table IV.

To compute the skyband for $s = 3$ and $k = 1$, it is sufficient to compare $\{OS_{1,2,3}$, $OS_{1,2,6}$, $OS_{2,3,4}$, $OS_{2,3,6}$, $OS_{2,5,6}\}$. We find that $OS_{1,2,6} \leq OS_{2,5,6}$ and the skyband set query retrieves the results of $\{OS_{1,2,3}$, $OS_{1,2,6}$, $OS_{2,3,4}$, $OS_{2,3,6}\}$. Similarly, for $s = 3$ and $k = 2$, the proposed method will retrieve $\{OS_{1,2,3}, OS_{1,2,5}, OS_{1,2,6}, OS_{2,3,4}, OS_{2,3,5}, OS_{2,3,6}\}$. Again, for $k = 3$, the proposed method retrieves $\{OS_{1,2,3}$, $OS_{1,2,5}$, $OS_{1,2,6}$, $OS_{2,3,4}$, $OS_{2,3,5}$, $OS_{2,3,6}$, $OS_{2,4,6}\}$ as the objectset skyband results.

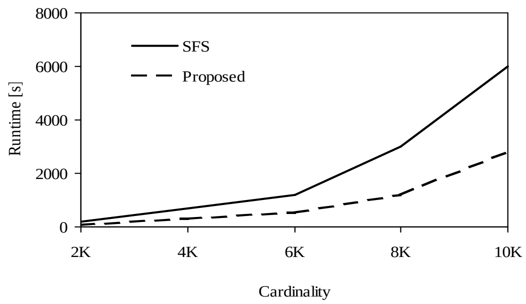
Now, if a user wants to select the top-3 objectsets for $s = 3$, then our proposed algorithm will examine the μ score of each objectset in the skyband set results and select $OS_{1,2,3}$, $OS_{2,3,6}$, and $OS_{1,2,6}$ with μ scores of 8, 7, and 6, respectively, which are the top-3 scores in the database. By applying the same procedure, the proposed method constructs the dominance relation table for higher objectsets, i.e., for higher values of s, in order to answer skyband-set queries and to provide the top-k objectset for any value of s.

V. Performance Evaluation

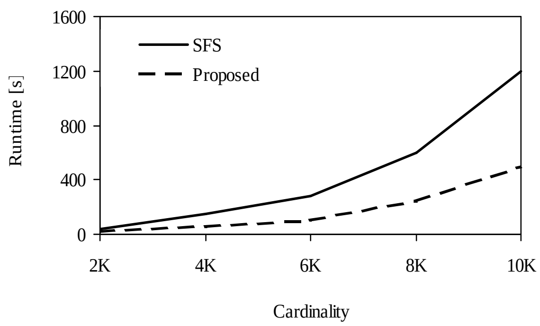
We have examined the proposed method through intensive experiments in a simulated environment on a computer with the Windows 64-bit operating system. The configuration of this computer comprised a core i7 CPU 3.4 GHz, 4 GB RAM, and 250 GB SATA disk. No existing approach can discover the skyband set, so we adapted the SFS method to calculate



a) Anti-correlated

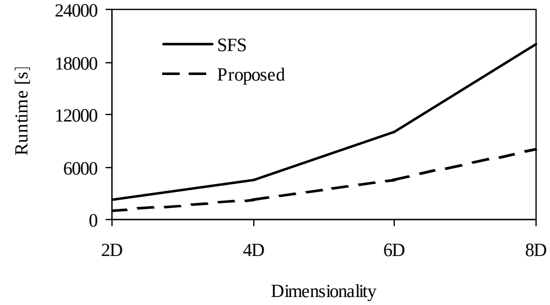


b) Independent

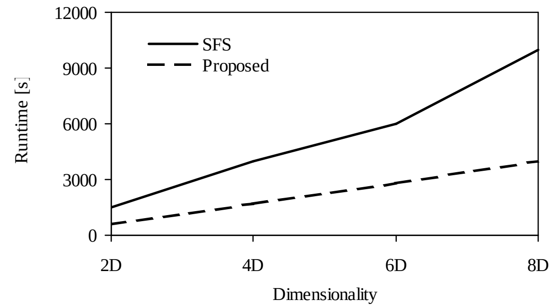


c) Correlated

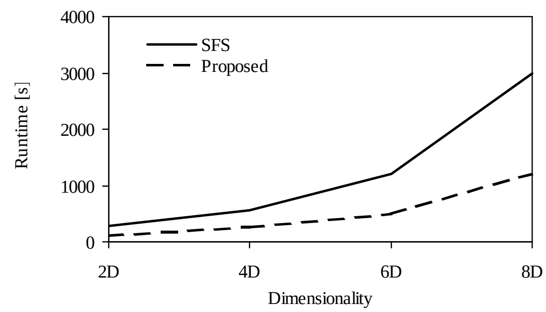
Fig. 6. Performance with different cardinalities.



a) Anti-correlated



b) Independent



c) Correlated

Fig. 7. Performance with different data dimensions.

the skyband set and compared the performance of our method with the skyband-set algorithm proposed in [3].

To ensure that the comparison was fair, we ignored the objectset computation cost for the SFS algorithm. We performed the experiments with different data cardinalities (n), objectset sizes (s), and dimensionalities (m) in order to compare the efficiency and performance of our proposed method. Each experiment was conducted five times. We determined the average output to assess the performance.

A. Performance Based on Synthetic Data Sets

We used benchmark synthetic data sets with three different data distributions, i.e., anti-correlated, correlated, and independent, as proposed in [1].

Correlated: This type of phenomenon occurs in an environment where an objectset is good in one dimension but it also has a better value in other dimensions. In this relationship, a small number of objectsets can dominate a large volume of other objectsets.

Anti-Correlated: In this relationship, an objectset with small coordinate values on some dimensions is expected to have higher coordinate values on other dimensions or at least in another dimension.

Independent: In this data distribution, the values of all dimensions are generated independently. In this relationship, the total non-dominant objectsets lie between the resulting objectsets of the correlated and anti-correlated data sets. All of the experimental results obtained based on the synthetic data set are shown in Fig. 6, 7, and 8. The total number of objectsets for synthetic data sets ranged from 2 k to 10 k depending on the data cardinality (n).

1) Cardinality Effect: We studied the impact of the cardinality by keeping the data dimensionality m as 6 and the objectset size s as 3, but we changed the data set cardinality n from 2 k to 10 k. Fig. 6 (a), (b), and (c) shows the result obtained with the anti-correlated, independent, and correlated data sets, respectively.

The x-axis and y-axis represent the data cardinality and execution time, respectively. The graphs show clearly that both techniques were affected by the data cardinality n. As the data cardinality value increased, the execution time increased considerably with both methods. However, our method performed better compared with SFS. For the anti-correlated data distribution, the performance of SFS declined dramatically as the data set size increased. Thus, the proposed method did not require the composition of all the objectsets and it also avoided many unnecessary comparisons.

2) Dimensionality Effect: We also assessed the impact of the dimensionality on our proposed method. The dataset dimensionality m was varied from 2 to 8, but we kept the data cardinality n as 10 k and the objectset size s as 3. Fig. 7 (a), (b), and (c) show the runtime requirements for the anti-correlated, independent, and correlated data sets, respectively. The performance of both methods became slower as the dimension size increased because the non-dominant objectset increased with higher dimensions. The graphs show that the proposed method performed better than SFS in all cases, where it was 15 and 10 times faster than SFS with the anti-correlated and independent data sets, respectively.

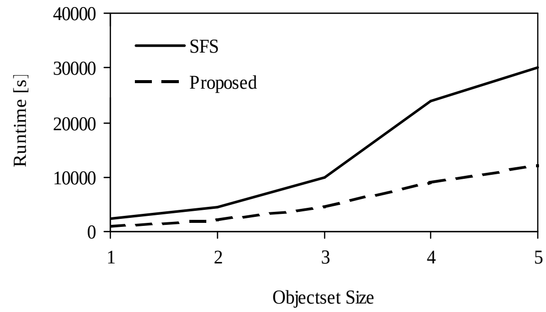
3) Objectset Size Effect: We examined the performance of the proposed method with different objectset size of s in this experiments. We fixed the data set dimensionality m to 6 and the data cardinality n to 10 k. The results are shown in Fig. 8 (a), (b), and (c), where the x-axis represents the objectset size s and the y-axis represents the runtime. The performance of both methods decreased as the objectset size became higher. The results indicated that when the objectset size became $s > 1$, the performance of the SFS method was much worse than our method. This is because when $s = 1$, the proposed method needed to construct $domRelationTable$ as well as performing domination checks for all of the objects. However, it was not necessary to calculate all of the objectsets subsequently, thereby avoiding numerous dimension checks.

The experimental results suggested that for synthetic data sets, our proposed method was efficient in terms of the runtime as well as being highly scalable. The results indicated that our method was 4-10 times faster than SFS. This is because the skyband set number for anti-correlated data sets was greater than that for independent data sets, which was also true for the correlated data sets. Therefore, both techniques required more computational time for the anti-correlated distribution than other two distributions.

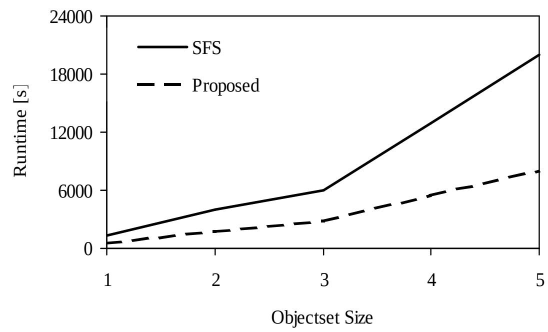
B. Performance Based on a Real Data Set

We also observed the impact of the proposed algorithm using a real data set, where we selected the FUEL data set, which can be downloaded from "`www.fueleconomy.gov`". The volume of the FUEL data set was 24 k and it contained six-dimensional objects, where each object represented the performance attribute of a vehicle, e.g., mileage/gallon gasoline in a city and on a highway. The domain of each attribute varied from 8–89.

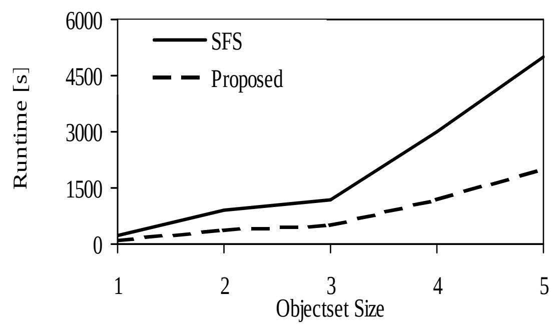
We conducted experiments similar to those using the synthetic data set based on the FUEL data set. Initially, we observed the effect of the dimension. We set the data cardinality n to 10 k, the objectset size s to 3, and varied the data set dimensionality m from 2–8. Fig. 9(a) shows the results



a) Anti-correlated



b) Independent



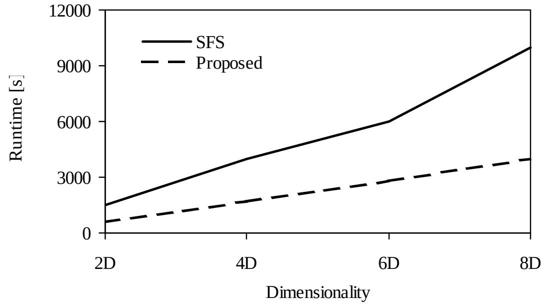
c) Correlated

Fig. 8. Performance with different objectset sizes.

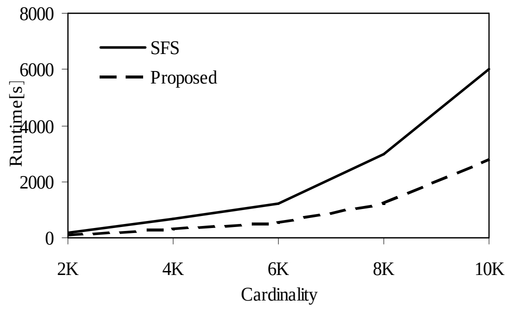
obtained for the skyband set queries with various dimension sizes. There was a positive correlation between the dimension size and runtime. However, the proposed method had a shorter runtime than the SFS technique.

Next, we conducted an experiment where we varied the data cardinality n. We fixed the dimensionality m to 4, the objectset size s to 3, and varied the data set cardinality n from 2 k to 10 k. Fig. 9(b) shows the results obtained. The computational time required by both algorithms increased rapidly with the data cardinality. The results in Fig. 9(b) show that our method performed better than the SFS method.

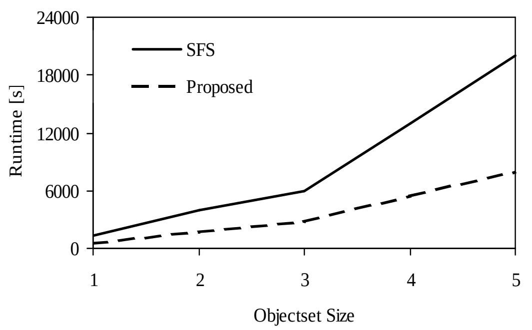
In the final experiment, we studied the performance by varying the objectset size s. We set the dimensionality m to 4 and the data cardinality n to 10 k. The outputs are shown in Fig. 9(c). For both methods, the execution time increased with the objectset size s. The computational time required by the proposed method was much shorter than that by SFS. We obtained similar standardized outputs such as an independent data distribution, thereby confirming the scalability of our technique with the FUEL data set, which was also the case



a) Varying Dimensionality



b) Varying Cardinality



c) Varying Objectset Size

Fig. 9. Experiments based on the FUEL data set.

in all of the experiments.

All of the experiments confirmed that the proposed method was much more efficient than the SFS technique for synthetic and real data sets. Thus, the experiments confirmed the effectiveness and scalability of the proposed method.

VI. Conclusion

In this study, we proposed a query method for sets of objects called a "skyband-set." We proposed an efficient method for calculating the skyband-set output. Our proposed method allows the objectset size to be varied from 1 to s and the skyband size from 1 to k. Furthermore, in order to filter out a large search space and to enhance efficiency, we also developed domRelationTable, which allows the proposed method to prune out large portions of the search space. We validated our method using synthetic and real data sets. The results in Fig. 6, 7, 8, and 9 confirm the usefulness and superior performance of the proposed method.

In addition to the contributions above, the proposed method has preferable property in privacy-aware environment. In privacy-aware environment, we are not allowed to disclose individual record's values. This is a significant restriction to analyze a database. Our set-based database analysis does not have to disclose individual record's values, which implies that the proposed method can be an effective database analysis measure in privacy-aware environment.

In the future, we aim to enhance our proposed algorithm in various areas. In particular, more effective calculations of the objectsets can be performed with a parallel distributed architecture such as MapReduce.

Acknowledgments

Md. Anisuzzaman Siddique was supported by "JSPS postdoctoral fellowship for foreign researchers program (25.03040)" Japan. Asif Zaman was supported by Japan's MEXT scholarship program. This work is supported by KAKENHI (16K00155) Japan.

References

[1] S. Borzsony, D. Kossmann, and K. Stocker. The skyline operator. In *Proceedings of IEEE International Conference on Data Engineering (ICDE)*, pages 421–430, 2001.

[2] S. Bothe, P. Karras, A. Vlachou. eSkyline: Progressing Skyline Queries over Encrypted Data. In *Proceedings of International Conference on Very Large Data Bases (VLDB)*, pages 1338–1341, 2013.

[3] J. Chomicki, P. Godfrey, J. Gryz, and D. Liang. Skyline with presorting. In *Proceedings of IEEE International Conference on Data Engineering (ICDE)*, pages 717–719, 2003.

[4] X. Guo, C. Xiao, Y. Ishikawa. Combination Skyline Queries. *Transactions on Large-Scale Data- and Knowledge-Centered Systems VI*, Vol. 7600, pages 1–30, 2012.

[5] K. C. K. Lee, B. Zheng, H. Li, W.-C. Lee. Approaching the skyline in Z order. In *Proceedings of International Conference on Very Large Data Bases (VLDB)*, pages 279–290, 2007.

[6] D. Papadias, Y. Tao, G. Fu, B. Seeger. Progressive skyline computation in database systems. *ACM Transactions on Database Systems*, Vol. 30, No. 1, pages 41–82, 2005.

[7] S. B. Roy, S. Amer-Yahia, A. Chawla, G. Das, C. Yu. Constructing and exploring composite items. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 843–854, 2010.

[8] M. A. Siddique, Y. Morimoto. Algorithm for computing convex skyline objectsets on numerical databases. *IEICE TRANSACTIONS on Information and Systems*, Vol. E93-D, No. 10, pages 2709–2716, 2010.

[9] I.-F. Su, Y.-C. Chung, C. Lee. Top-k combinatorial skyline queries. *Proceedings of DASFAA*, pages 79–93, 2010.

[10] Q. Wan, R. C.-W. Wong, Y. Peng. Finding top-k profitable products. In *Proceedings of IEEE International Conference on Data Engineering (ICDE)*, pages 1055–1066, 2011.

Simulated Annealing with Levy Distribution for Fast Matrix Factorization-Based Collaborative Filtering

Mostafa A. Shehata, Mohammad Nassef, Amr A. Badr
Department of Computer Science
Faculty of Computers and Information
Cairo University

***Abstract*—Matrix factorization is one of the best approaches for collaborative filtering because of its high accuracy in presenting users and items latent factors. The main disadvantages of matrix factorization are its complexity, and are very hard to be parallelized, especially with very large matrices. In this paper, we introduce a new method for collaborative filtering based on Matrix Factorization by combining simulated annealing with levy distribution. By using this method, good solutions are achieved in acceptable time with low computations, compared to other methods like stochastic gradient descent, alternating least squares, and weighted non-negative matrix factorization.**

***Keywords*—Simulated annealing; levy distribution; matrix factorization; collaborative filtering; recommender systems; metaheuristic optimization**

I. Introduction

The objective of Recommender Systems is to recommend new products or items for users based on their history [1]. There are two major approaches to create a Recommender System. The first one is the *Content Filtering* or (*Content Based*). This approach tries to create a profile for each user and item, and then tries to match these profiles [3]. The second approach is the *Collaborative Filtering*. It uses the rating history of users and items, and creates a large sparse matrix called Rating Matrix. This matrix usually contains ratings from 1 to 5. The 0's are for the incomplete ratings. The objective of the *Collaborative Filtering* is to predict these missing ratings. One of the most successful method for *Collaborative Filtering* is the *Latent Factor Model*[3]. This method tries to learn the latent features of each user and item in a fixed number of dimensions. Then represent each of them in a latent feature vector that can be used to predict the incomplete ratings or measure the similarity.

Matrix Factorization is one of the best techniques used for *Latent Factor Model*. The basic idea is to construct the *low-dimensional* matrices to approximate the original rating matrix [2], [3], [7], [12], [13].

$$R \approx U \cdot I \tag{1}$$

where $R_{M,N}$ is the rating matrix, $U_{M,K}$ is the users matrix, $I_{K,N}$ is the items matrix. M and N are the number of users and items respectively, K is the number of latent feature that represent each user and item. Where $K \ll \min(M, N)$. Row m in matrix U represents user number m in the rating matrix, whereas column n in matrix I represents item number n in the rating matrix. So, in the rating matrix, the rating of user m for item n can be calculated by the *dot product* of row m of matrix U by the column n of matrix I.

$$r_{m,n} \approx u_m \cdot i_n^T \tag{2}$$

The rating matrix is very sparse, because it contains a few users ratings. The objective of this paper is to use the known ratings to construct the *low-rank* matrices, to predict the unknown or incomplete ratings.

One of the most common evaluation metrics for *Collaborative Filtering* is *RMSE* (root mean squared error). We calculate *RMSE* only for the known rating using the following equation:

$$RMSE = \sqrt[2]{\left(\sum_{m,n \in KR} (r_{m,n} - (u_m \cdot i_n^T))^2\right)/|KR|} \tag{3}$$

where KR is the list of known ratings.

There is a lot of work done on Matrix Factorization and Collaborative Filtering. Here we discus three of the most popular methods.

Stochastic Gradient Descent (SGD): SGD is one of the popular Matrix Factorization methods [3]. The idea is to minimize the following cost equation:

$$\begin{aligned} \min_{u^*, i^*, b^*} \sum_{(m,n) \in KR} & (r_{m,n} - \mu - b_m - b_n - u_m \cdot i_n^T)^2 + \\ & \lambda(\|u_m\|^2 + \|i_n\|^2 + b_m^2 + b_n^2) \end{aligned} \tag{4}$$

where λ is a regularization term, μ is the overall average rating, b_m and b_n are the user and item bias, respectively.

$$e_{m,n} = r_{m,n} - \mu - b_m - b_n - u_m \cdot i_n^T \tag{5}$$

To minimize the squared-error (4), the algorithm iterates over all ratings in the training set. Then, it computes the associated prediction error in (5). Next, the error value is used to compute the gradient. The algorithm finally uses the gradient to update user bias, item bias, user matrix U, and item matrix I.

$$b_m = b_m + \gamma(e_{m,n} - \lambda b_m) \tag{6}$$

$$b_n = b_n + \gamma(e_{m,n} - \lambda b_n) \tag{7}$$

$$u_m = u_m + \gamma(e_{m,n} \cdot i_n - \lambda \cdot u_m) \tag{8}$$

$$i_n = i_n + \gamma(e_{m,n} \cdot u_m - \lambda \cdot i_n) \tag{9}$$

where γ is the learning rate. The learning rate determines the moving speed towards the optimal solution. If γ is very large, we might skip the optimal solution. If it is too small, we may need too many iterations to reach the optimal solution. So using an appropriate γ is very important.

Alternating Least Squares (ALS): ALS is very good for parallelization [11]. When we have large data, and need to distribute the computations over cluster of nodes. ALS objective is to minimize the following equation:

$$\min_{u^*,i^*,b^*} \sum_{(m,n)\in KR} (r_{m,n} - u_m \cdot i_n^T)^2 + \lambda(\|u_m\|^2 + \|i_n\|^2) \quad (10)$$

It is the same like (4), but without the bias terms. The basic idea can be summarized as follows:

1) Initialize U, and I matrices.
2) Fix I, solve for U by minimizing (10).
3) Fix U, solve for I by minimizing (10).
4) Repeat the previous two steps until converging or reaching the max iteration.

to solve the user U and item I matrices we use the following two equations, respectively:

$$u_m^T = (r_m \cdot I) \cdot (I^T \cdot I + \lambda Eye)^{-1} \quad (11)$$

$$i_n^T = (r_n^T \cdot U) \cdot (U^T \cdot U + \lambda Eye)^{-1} \quad (12)$$

where Eye is the Identity matrix.

Weighted Non-Negative Matrix Factorization (WNMF): Here we present a special type of Matrix Factorization called *Non-Negative Matrix Factorization* (NMF). The only difference is the non-negativity constraint for the input matrix, and the low rank matrices as well. The problem can be formulated as an optimization problem:

$$\begin{aligned} \min_{A,H} \quad & \|V - A \cdot H\|^2 \\ \text{subject to} \quad & A, H \geq 0 \end{aligned} \quad (13)$$

where V is the original matrix, $WandH$ are the two factorized matrices. One of the most simplest methods for NMF is *Multiplicative Update Rules* [12] [2]. It is a good compromise between speed and ease of implementation. So NMF objective (13) can be optimized using the following update rules:

$$A^{(t+1)} = A^{(t)} \frac{V \cdot H^T}{A \cdot H \cdot H^T} \quad (14)$$

$$H^{(t+1)} = H^{(t)} \frac{A^T \cdot V}{A^T \cdot A \cdot H} \quad (15)$$

The original version of *Multiplicative Update Rules* will not fit in our problem. The two rules will not be able to differentiate between the true ratings, and the incomplete ratings. So we need to modify the original rules, to be able to learn from the true ratings, then predict the incomplete. In [10] they could modify the original *Multiplicative Update Rules* to be able to do Incomplete Matrix Factorization. This method called *Weighted Non-Negative Matrix Factorization (WNMF)*. So now the new objective function is:

$$\min_{u^*,i^*} \sum_{(m,n)\in KR} (r_{m,n} - u_m \cdot i_n^T)^2 \quad (16)$$

It is similar to (4), (10), but without the bias b or regularization λ. Now we can optimize function (16) using the following two rules:

$$U^{(t+1)} = U^{(t)} \frac{(W * R) \cdot I^T}{(W * (U \cdot I)) \cdot I^T} \quad (17)$$

$$I^{(t+1)} = I^{(t)} \frac{U^T \cdot (W * R)}{U^T \cdot (W * (U \cdot I))} \quad (18)$$

where $W_{M,N}$ is a matrix which its elements are equal to 1 if the corresponding entry in R is known rating, and 0 otherwise. $(*)$ denotes to the element wise multiplication.

In this paper we focus on efficiency more than effectiveness. We assume that we have a very large data, and limited time. So we need an acceptable solution in reasonable time. So we chose the Metaheuristic algorithms for this problem, because of its ability to scape from the local optimal, and reaching good solutions in reasonable time. We used Simulated Annealing algorithm, with Levy Flight as a random walk operator.

The rest of the paper is organized as follows: in Section (II) we briefly describe the prerequisite topics that are needed before going through the proposed method. In Section (III) we describe our proposed method. In Section (IV) we discuss the experimental results, effect of each parameter, and compare our proposed method against others. Finally, Section (V) concludes the paper.

II. Preliminaries

A. *Metaheuristic Optimization*

There are two types of optimization algorithms, *Deterministic* and *Stochastic*. Deterministic algorithms usually focus on optimal solution, like Simplex method in linear programming, some of these algorithms use the derivative of the objective function, these algorithms are called *gradient based algorithms*.

In *Stochastic Optimization* we will talk about *Metaheuristic Algorithms*, we can divide *Metaheuristic* into two parts, META and HEURISTIC, META means "beyond" or "higher level", and HEURISTIC means "to find" or "to discover by trial and error". This type of algorithms depends on randomization and local search to find the optimal solution iteratively, whereas each iteration tries to improve the current solutions from previous iteration. Also *Metaheuristic* doesn't guarantee the optimal solution, but it gives good quality solutions in a reasonable time. *Metaheuristic* achieves its goal by making a good balance between two major components, intensification and diversification. Intensification is to search for a better solution within the local area of the current solution. Diversification is to use the randomization to escape from the local optimum, and explore all the search space [4].

There are many types of *Metaheuristic* algorithms, like single solution, or population based, in this paper we use *Single Solution*.

B. Levy Distribution and Random Walk

We presented randomization techniques for exploring the search space (*Diversification*), local search for optimizing the current solution, and searching within the local area of it (*Intensification*). In (19) x^t is the current solution state, s is a new step or random number drawn from a probability distribution, we add s to x to move it from state t to $t+1$.

$$x^{t+1} = x^t + s \tag{19}$$

Levy Flights are a random walk that their steps are drawn from *Levy Distribution*. *Mantegna* algorithm is the best and easiest way to generate random numbers from *Levy Distribution* [4], [5], so the random walk can be achieved using the following equations:

$$x_i^{t+1} = x_i^t + \alpha L(s, \lambda), \tag{20}$$

Where α is the step size.

$$L(s, \lambda) = \frac{\lambda \Gamma \sin(\pi\lambda/2)}{\pi} \frac{1}{s^{1+\lambda}}, \tag{21}$$

$$s = \frac{U}{|V|^{1/\alpha}} \tag{22}$$

$$U\ N(0, \sigma^2), V\ N(0, 1) \tag{23}$$

Where N is a Gaussian normal distribution

$$\sigma^2 = \left[\frac{\Gamma(1+\lambda)}{\lambda\Gamma((1+\lambda)/2)} . \frac{\sin(\pi\lambda/2)}{2^{(\lambda-1)/2}} \right]^{1/\lambda} \tag{24}$$

C. Simulated Annealing

SA is one of the most popular metaheuristic algorithms. It simulates the annealing process for solids by cooling to reach the crystal state. Reaching the crystal state is like reaching the global optimum in optimization. It is a single solution algorithm. The basic idea is to perform a random walk, but with some probability called *Transition Probability* that may accept new solutions that do not improve the objective function, see (25). Accepting bad solutions with *Transition Probability* gives more exploration for the search space (*Diversification*). *Transition Probability* decreases gradually during the iterations to decrease the *Diversification* and increase the *Intensification*. This means that the algorithm will end up with accepting only better solutions [4] [6].

$$p = \exp\left[-\frac{\Delta f}{T} \right] > r \tag{25}$$

In (25) f is the difference between the two evaluation function values of the current solution and the new one. T is the current temperature which is decreased iteratively by the cooling rate. r is a random number. So the algorithm accepts bad solution if the *Transition Probability* p is greater than r.

One of the common cooling schedules is linear cooling schedule, in (26).

$$T = T_0 - \beta t \tag{26}$$

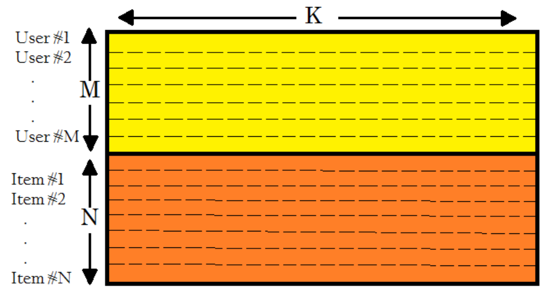


Fig. 1. Representing U matrix and I matrix in one matrix, the number of rows is equal $M + N$, and the number of columns is equal K.

T_0 is the initial temperature, β is the cooling rate, t is the pseudo time for iterations. The following pseudo code demonstrate the basic implementation of Simulated Annealing:

Algorithm 1 Simulated Annealing

Objective function $f(x), x = (x1, ..., xd)^T$
Initialize the initial temperature T_0 and initial guess $x(0)$
Set the final temperature T_f and the max number of iterations N
Define the cooling schedule $T \mapsto \alpha T, (0 < \alpha < 1)$
while $(T > T_f$ and $t < N)$ **do**
  Drawn ϵ from a Gaussian distribution
  Move randomly to a new location: $x_{t+1} = x_t + \epsilon$(random walk)
  Calculate $\Delta f = f_{t+1}(x_{t+1}) - f_t(x_t)$
  Accept the new solution if better
  if not improved **then**
    Generate a random number r
    Accept if $p = \exp[\Delta f / T] > r$
  end if
  Update the best $x\star$ and $f\star$
  $t = t + 1$
end while

III. Proposed Method

In this section we introduce our new method for solving Matrix Factorization. In our method we use *Simulated Annealing* based on *Levy Flight* as a random walk, instead of *Gaussian distribution*, see Section (II-A). We called it *SA-Levy*. We choose *Simulated Annealing* because its low computations, as it is a single solution *Metaheuristic* algorithm. So it will be very fast compared to other population based *Metaheuristic* algorithms. Also compared to the state of the art methods like *ALS*, and *WNMF Simulated Annealing* is much faster, because it needs only one matrix multiplication per iteration. Regarding *SGD*, *Simulated Annealing* is easier to be parallelized.

In *Simulated Annealing* we just need to represent the solution, and implement the evaluation function to compare the current solution against others. We use *RMSE* as an evaluation function.

Fig. 1 shows the representation of the solution, we put the two matrices users and items in one matrix to simplify the

TABLE I. SHOWS THE EFFECT OF THE NUMBER OF ITERATION ON RMSE

Iterations	5	**10**	25	50	100	200
RMSE	1.112	**1.118**	1.118	1.118	1.118	1.118

TABLE II. SHOWS THE EFFECT OF NUMBER OF LATENT FEATURES ON RMSE

Latent Features	10	**20**	30	40	50
RMSE	1.120	**1.118**	1.118	1.118	1.118

TABLE III. SHOWS THE EFFECT OF STEP SIZE OF LEVY FLIGHT ON RMSE

Step Size	0.1	**0.01**	0.001
RMSE	1.119	**1.118**	1.145

solution and the calculation [8], the number of rows is equal the number of users M plus the number of items N, and the number of columns is equal the number of latent features K.

IV. EXPERIMENTAL RESULTS

In this section we show the effect of the parameters on the *RMSE* results, we use MovieLens 1M dataset [9] in our experiments, 80% of the dataset is used for training and 20% for testing.

Tables I, II and III show how *RMSE* can be affected by the number of iteration, number of latent features, and step size, see (20). In Table I, we can see that good *RMSE* can be achieved by few number of iterations, so there is no need for many iteration to converge. In Table II we can see that best *RMSE* can be achieved starting from 20 latent features. In Table III we found that the best value for the step size is 0.01. We can say that the step size is the most important parameter in our method. It manages the balance between Intensification and Diversification, see Section (II-A). Small values of step size give more intensification, and large values give more Diversification.

Table IV shows the difference between using Gaussian distribution and Levy distribution as a random walk. Levy distribution outperform Gaussian because of its ability to escape from the local minimum [4][5].

Table V shows that SA-Levy can be outperformed by other methods in terms of effectiveness. But SA-Levy can outperform all other methods in terms efficiency, because of its low computations, where it needs only one matrix multiplication in each iteration. Unlike WNMF or ALS which need many matrix multiplications or calculating matrix inversion in each iteration. Also it is much easier than SGD to be parallelized because it doesn't need huge amount of data to be shuffled between the cluster nodes. So SA-Levy can be a good choice if we have limited time or resources and large amount of data.

TABLE IV. COMPARES LEVY AGAINST GAUSSIAN DISTRIBUTION AS A RANDOM WALK FOR SIMULATED ANNEALING

Distribution	**Levy**	Gaussian
RMSE	**1.118**	1.168

TABLE V. COMPARES SA-LEVY WITH OTHER METHODS (SGD, WNMF AND ALS)

System	SA-Levy	SGD	WNMF	ALS
RMSE	1.118	0.871	0.943	1.007

Choice of Parameters

We conducted these experiments using **Simanneal**. It is a python module for simulated annealing optimization[1], also the project source code can be found here[2]. Based on the MovieLens 1M dataset [9] we found that the best parameters are 10 Iteration, 20 latent features, 0.01 step size. For the temperature parameter we found that the best values for maximum and minimum temperature are 25000 and 2.5 respectively. To focus more on Diversification at the beginning then decrease it gradually to increase the Intensification.

V. CONCLUSION AND FUTURE WORK

We presented in this work a new method for matrix factorization based Collaborative filtering. We achieved a significant improvement in Simulated Annealing, by using Levy distribution as a random walk, instead of Gaussian distribution. We expect this contribution could fit in many optimization problems, not only matrix factorization. We think that SA-Levy is a good choice for complex matrix factorization problems. When we have a very large data, and limited time for computation. We expect that SA-Levy can be easily implemented on any distributed system, that has basic linear algebra operations, like *Apache Spark*[3], and *Hadoop*[4].

REFERENCES

[1] Kantor, Paul B. Recommender systems handbook. Eds. Francesco Ricci, Lior Rokach, and Bracha Shapira. Berlin, Germany:: Springer, 2015.

[2] Duan, Liang, et al. "Scaling up link prediction with ensembles." Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. ACM, 2016.

[3] Koren, Yehuda, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems." Computer 42.8 (2009).

[4] Yang, Xin-She. Nature-inspired optimization algorithms. Elsevier, 2014.

[5] Mantegna, Rosario Nunzio. "Fast, accurate algorithm for numerical simulation of Levy stable stochastic processes." Physical Review E 49.5 (1994): 4677.

[6] Van Laarhoven, Peter JM, and Emile HL Aarts. "Simulated annealing." Simulated annealing: Theory and applications. Springer Netherlands, 1987. 7-15.

[7] Luo, Xin, et al. "An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems." IEEE Transactions on Industrial Informatics 10.2 (2014): 1273-1284.

[8] Salehi, Mojtaba. "Latent feature based recommender system for learning materials using genetic algorithm." Information Systems & Telecommunication (2014): 137.

[9] Harper, F. Maxwell, and Joseph A. Konstan. "The movielens datasets: History and context." ACM Transactions on Interactive Intelligent Systems (TiiS) 5.4 (2016): 19.

[1] https://github.com/perrygeo/simanneal
[2] https://github.com/mostafaashraf413/MF_SA_Levy
[3] https://spark.apache.org/
[4] http://hadoop.apache.org/

[10] Zhang, Sheng, et al. ”Learning from incomplete ratings using non-negative matrix factorization.” Proceedings of the 2006 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2006.

[11] Zhou, Yunhong, et al. ”Large-scale parallel collaborative filtering for the netflix prize.” Lecture Notes in Computer Science 5034 (2008): 337-348.

[12] Lee, Daniel D., and H. Sebastian Seung. ”Algorithms for non-negative matrix factorization.” Advances in neural information processing systems. 2001.

[13] Hernando, Antonio, Jess Bobadilla, and Fernando Ortega. ”A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model.” Knowledge-Based Systems 97 (2016): 188-202.

Outsourcing of Secure k-Nearest Neighbours Interpolation Method

Muhammad Rifthy Kalideen
Department of Islamic Studies
South Eastern University of Sri Lanka
Oluvil, Sri Lanka, 32360

Bulent Tugrul
Department of Computer Engineering
Ankara University
Ankara, Turkey, 06100

***Abstract*—Cloud computing becomes essential in these days for the enterprises. Most of the large companies are moving their services and data to the cloud servers which offer flexibility and efficiency. Data owner (DO) hires a cloud service provider (CSP) to store its data and carry out the related computation. The query owner (QO) sends a request which is crucial for its future plans to the CSP. The CSP computes all necessary calculations and returns the result back to the QO. Neither the data nor query owners want to reveal their private data to anyone. k-Nearest Neighbour (k-NN) interpolation is one of the essential algorithms to produce a prediction value for an unmeasured location. Simply, it finds k number of nearest neighbours around the query point to produce an output. Oblivious RAM (ORAM) has been used to protect the privacy in cloud computing. In our work, we will perform the k-NN method using the kd-tree and ORAM without revealing both the data-owner's and query owner's confidential data to each other or to third parties. The proposed solution will be analysed to ensure that it provides accurate and reliable predictions while preserving the privacy of all parties.**

Keywords*—Cloud computing; k-Nearest neighbour; spatial interpolation

I. Introduction

Governments, companies and institutions save data for a variety of reasons. Some examples of such data are medical, insurance, banking and geographic. Data should be analysed to obtain useful information. Data mining methods are conducted to extract useful information from the stored data. The biggest concern in data mining is privacy. Two major problems may arise during data mining applications. First one, the data required for mining may have been collected by two different institutions. When executing the data mining algorithm on the combined databases, each party does not want to share its private data or laws may prohibit data sharing. The second issue is that the calculated/resulting data must be transferred to others for further activities. Data transfer should ensure the privacy of the data owner and produce the correct output. The banking sector is one of the examples regarding the privacy of data while running data mining algorithm. Nowadays, multiple different banks need to merge their data to mine for further activities. Due to law and privacy policies, the data cannot be pooled in one place. In these cases, solutions that provide data privacy are preferred instead of traditional data mining algorithms [1], [2].

Information Technology moved to a new era in this decade known as cloud computing. Cloud computing provides access to resources such as servers, storage and applications. Cloud computing is divided into different models according to the services they offer [3].

a) Software as a Service (SaaS)
b) Platform as a Service (PaaS)
c) Infrastructure as a Service (IaaS)
d) Data as a Service (DaaS)
e) Database as a Service (DBaaS)

All models have the following characteristics: shared infrastructure, dynamic provisioning, network access and managed metering. Furthermore, deployment of the cloud computing is divided into four categories. They are private, community, public and hybrid cloud. Some of the challenges associated with cloud computing are privacy, lack of standards, constantly evolving, and compatibility concerns [3].

Spatial data should also be analysed to find concealed patterns and properties. Spatial interpolation methods (SIM) try to analyse and interpret spatial variability using statistics, mathematics, and geographic assumptions. Retrieving the values for the unknown new points from the well-known stored points without a big deviation from the original value is called as spatial interpolation. There are about 42 different spatial interpolation methods such as Nearest Neighbour (NN), Triangular Irregular Network (TIN), Inverse Distance Weighted (IDW), Kriging, etc. These methods are divided into three categories [4].

a) Non-geostatistical methods
b) Geostatistical methods
c) Combined methods

A. Problem Description

There are three parties involved in this scheme. The first party is the data owner which collects data for a specific subject. The second party is the cloud service provider which stores data and conducts necessary computations in place of the data owner. CSP can be a public cloud like Google, Amazon, etc. The CSP also can provide services like SaaS, PaaS, IaaS, DaaS, and DbaaS. Final party is the query owner. Query owner can be an organization, a company or a single user. Database (referred as D hereafter) which is stored in CSP in an encrypted form with the key owned by DO. Data stored in D are points (e.g. location) and the requests are geometric operations like finding a point or calculating nearest neighbours. Stored database on CSP is encrypted with the public key (pk) of DO.

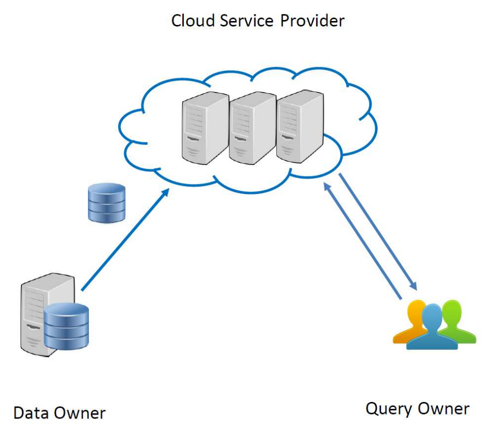


Fig. 1. The system model includes a data owner, query owner and a cloud service provider.

The public and private key pair of DO are generated using a homomorphic encryption scheme.

(i) DO has a spatial database D.
(ii) DO hires a CSP to store the D in encrypted form.
(iii) QO sends a request to CSP to get a prediction value for an unmeasured location.
(iv) k-NN spatial interpolation method is employed to predict the value by the CSP
(v) Finally, CSP sends the prediction value to the QO.

$$\text{Data stored by DO} \quad \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ . & . & . \\ . & . & . \\ . & . & . \\ x_k & y_k & z_k \end{bmatrix} \qquad \text{Data stored by CSP} \quad \begin{bmatrix} \xi_{pk}(x_1) & \xi_{pk}(y_1) & \xi_{pk}(z_1) \\ \xi_{pk}(x_2) & \xi_{pk}(y_2) & \xi_{pk}(z_2) \\ . & . & . \\ . & . & . \\ . & . & . \\ \xi_{pk}(x_m) & \xi_{pk}(y_m) & \xi_{pk}(z_m) \end{bmatrix}$$

In this framework, nobody wants to reveal their private data to others. In other words, DO does not want to reveal his database D to neither CSP nor C. Additionally, both DO and CSP cannot learn the query point which is the private data of QO. The secure outsourcing of k-NN method will be utilized under these data privacy conditions.

B. Security Definition

We consider semi-honest (honest-but-curious) model which means that each party follows the steps of the proposed protocol in the right manner, but tries to obtain clues about the data of the other party from the messages sent to it.

The system model includes a data owner, query owner and a cloud service provider (Fig. 1).

This paper is constructed as follows: Background of this work explained in Section 2. We proposed a solution in Section 3. Next we analyse the solution in Section 4 and Section 5 concludes the paper and suggests the future work.

II. Related Work

The increase in the amount and variety of the data causes the costs of the institutions to increase day by day. This situation necessitates transferring the storage and computation complexity of institutions from client-server architecture to cloud infrastructure. However, the concern that data privacy may be violated is seen as the biggest obstacle in front of the institutions [5]. Health and financial data are protected by laws in many countries. Conscious or unconscious violations of these laws cause both financial and prestige losses of institutions [6]. Protocols that provide data privacy between parties will abolish such concerns.

Encrypting data before transferring to cloud servers provides privacy. However, encryption does not allow some calculations to be made on joint data. Secure multi-party computations (SMC) initiated by Yao [7] are able to calculate a common function using private inputs of two or more parties. Yao employs garbled circuits to reach his objective. On the other hand, other protocols use secret sharing [8], [9]. Many applications that require privacy have been solved with SMC protocols. The first real application of SMC proposed by Bogetoft et al. [10]. They implemented a secure auction system between many sellers and buyers. Andrychowicz et al. [11] presented how SMC protocols can adapt essential Bitcoin properties to their solutions.

Fix and Hodges [12] suggested a non-parametric method for classification algorithms which has since been known as k-nearest neighbours. It has a significant role in machine learning, image processing and spatial analysis. The k-NN is a method that is frequently used in the analysis of minerals in the soil, determining the position of the interior space, and in image processing applications. In machine learning applications, k-NN is used as a classifier method. Joachims [13] used Support Vector Machines and many other classification algorithms to categorize texts. There are various spatial interpolation methods used in geostatistics. Xin et al. [14] implemented a spatial interpolation method based k-NN and IDW to analyse soil nutrient.

III. Background

A. k-NN Method

There are many algorithms used to find similarity metrics in data mining [15]. One of the most commonly used algorithms is k-NN. Because it is easy to apply, fast to train new data and effective for large amounts of data. Generally, points are represented as distance metrics in the k-NN algorithm. It finds the k number of nearest points for a given point using several distance metrics like Euclidean, Manhattan and so on. The steps of k-NN algorithm are as follows:

(i) Compute the distances from all training points to the query point.
(ii) Choose the k nearest points around the query point.
(iii) Class labels are used to predict the class of query point or average of k points is calculated for interpolation methods.

The value of k should be decided carefully. A small k value takes immediate measurements into account. If large k values are selected, it is possible to include the locations where the similarity is low in the calculation process. Therefore, the optimal k value is one of the research topics. kd-tree is one of the binary trees used to represent the spatial data in a multidimensional space. Points are divided in two ways to build a kd-tree. First one by a median of the x-axis. A vertical line divides the points into two halves. Half of the points are on the left side of the line and another half on the right side. A horizontal line is drawn by the median of the y-axis to divide the points in two. One of the halves is in top the line and remaining in the bottom.

B. Oblivious RAM (ORAM)

In a client-server architecture model, a client wants to perform an operation on the data which is stored in a server where the client does not want to share any information about the operation. Because it does not trust the server. Because of this neither the stored data nor the access pattern of the data would be revealed to the server. Goldreich and Ostrovsky [16] suggest a solution to overcome this called as ORAM. ORAM works as follows: all the data were stored as blocks (N blocks). These blocks are encrypted before sending to the server. If the client wants to access one of the stored blocks (i^{th} block), client access all the blocks one by one and when it gets the i^{th} block, it works on it. Then re-encrypts all the block and send back to the server. The server cannot identify any common patterns to identify i^{th} block if it accessed by the client several times.

The very first idea of constructing hierarchical data-structure modelling for the RAM was introduced by Goldreich and Ostrovsky [16]. The idea is ORAM stored in buckets, size of the buckets increases in a geometrical series. Smallest bucket in the top and it increases in downwards. In worst case scenario it needs $O(nlog^2n)$. A tree-based memory structure ORAM is suggested by Shi et al. [17].

The structure of constructing the tree is:

(i) The client wants to store n number of array blocks $[M_1, M_2 \ldots M_n]$ in a server.
(ii) Size of a block B = $log^c n$ where c is a constant.
(iii) Binary tree with n leaves and the height of $logn$ is constructed to store the memory.
(iv) A node in the tree considered as a bucket and a bucket contains Z number of blocks.
(v) A path $p(i)$ in the tree belongs to block i.
(vi) Check each and every bucket which is a match to M_i. If it is not matched, it will be discarded.

The accessed block needs to be stored again in the memory. There is a possibility of leakage of information if the block stored in the previous place when reading the block again and again. To avoid this situation the block is re-encrypted and send it back to the top node.

C. Homomorphic Encryption (HE)

A special type of cryptography that provides mathematical operation on cipher-texts called homomorphic encryption. Many researchers suggested several algorithms for homomorphic encryption likes Paillier cryptosystem [18], Goldwasser-Micali [19], ElGamal [20], Boneh-Goh-Nissim [21] encryption schemes and so on. Paillier encryption system provides two algebraic operations on encrypted text. They are additive and multiplicative operation. Assume that ξ_{pk} is the encryption function and ξ_{qk} is the decryption function where pk is the public key and the private key is qk in this scenario. To retrieve x from $\xi_{pk}(x)$ you need to know about qk, without any knowledge about sk nobody cannot decrypt the encryption function.

Homomorphic Addition: $D_{sk}((\xi_{pk}(a) \cdot \xi_{pk}(b))modN^2) = (a+b)modN$

Homomorphic Multiplication: $D_{sk}((\xi_{pk}(a)^b)modN^2) = (a \cdot b)modN$

where a and b are plain-texts, N is a product of two large prime numbers.

IV. Proposed Solution

As explained above, there are three parties involved in the outsourcing of secure k-NN interpolation method: Data Owner, Cloud Service Provider and Query Owner. The database which stores all measured data is possessed by DO. CSP provides storage and computation services. QO is interested in the query point where it may plan their future investments. Our solution follows the steps explained below (Fig. 2):

(i) First, DO builds a kd-tree based on the points and corresponding measurements in encrypted form. It sends the kd-tree to CSP which acts as an ORAM server.
(ii) QO sends its query point to CSP. However, query owners must send their inputs in encrypted form to hide coordinates where they need prediction values.
(iii) CSP processes the query and determines the k-NN points around the query coordinate. CSP cannot compute the prediction value because all data are encrypted with the key of DO. Therefore, CSP calculates $\prod_{i=1}^{k} \xi(z_i) = \xi(z_1) \cdot \xi(z_2) \cdots \xi(z_k)$ which equals $\sum_{i=1}^{k} z_i$ in decrypted form and sends to DO.
(iv) DO has the required key to decrypt the coming value from CSP. It decrypts and gets the prediction value. The prediction value, alone does not reveal information about query coordinate. Therefore, DO cannot learn the coordinate where the query owner is interested.
(v) DO encrypts the prediction value with the public key of the query owner to hide the prediction value from CSP and sends to CSP.
(vi) CSP forwards the prediction value in cipher-text form to the query owner.
(vii) Query owner has the corresponding private key to open cipher-text and gets the value where it needs a prediction.

V. Analysis of Proposed Solution

A. Supplementary Cost Analysis

DO provides all functionalities like storing the data and computing the prediction value for the location requested by a query owner in the traditional spatial interpolation architecture. However, the private data of both parties are at risk. On

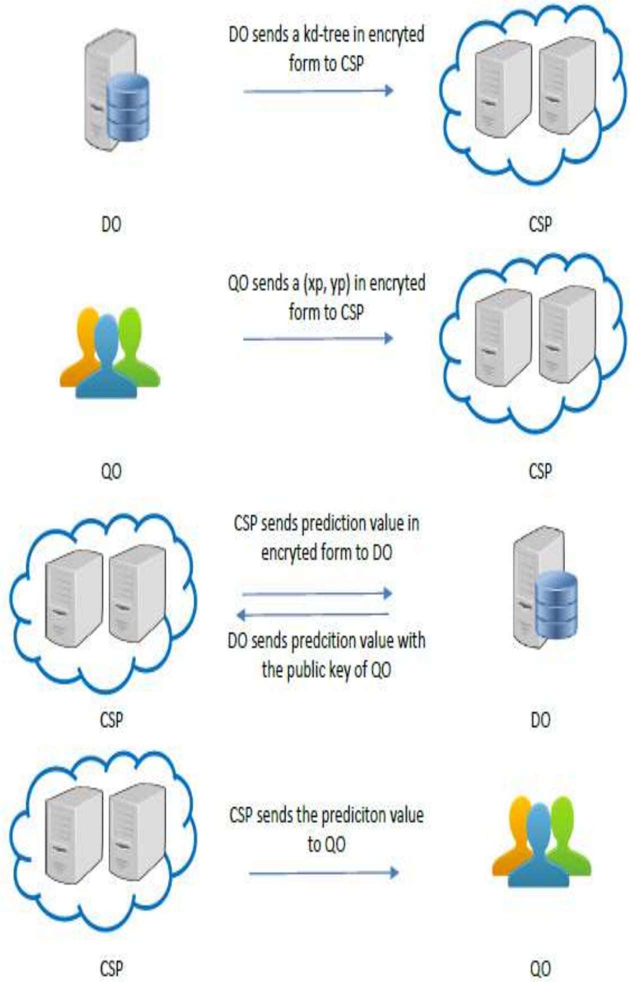


Fig. 2. Framework of secure k-NN spatial interpolation method.

the other hand, our idea of using privacy-preserving scheme protects the privacy of both parties. Besides, our work cause additional computation, storage and communication cost at negligible level than the traditional scheme. These additional costs do not affect the overall performance of spatial interpolation process. Due to the fact that spatial interpolation methods are not a time critical process as in real-time systems, researchers may prefer our solution when their privacy is important.

1) Computational Cost Analysis: Cloud computing is more efficient and profitable than the traditional client-server architecture. However, the privacy of the parties' data should be treated more carefully. Our solution which ensures the privacy with the help of homomorphic encryption and secure multi-party computation methods proposes that CSP has to carry out all necessary calculations instead of DO. Using privacy-preserving methods may increase computation time, but the resulting delay will be negligible for all parties.

DO constructs a kd-tree based on the spatial data in encrypted form and sends the tree to the CSP. CSP stores the received data as an ORAM server. Meanwhile, QO sends a query q to CSP where the q is encrypted with DO's public key. CSP finds the k number of nearest neighbours around the query. Then CSP calculates $\prod_{i=1}^{k} \xi(z_i) = \xi(z_1) \cdot \xi(z_2) \cdots \xi(z_k)$ and sends it to DO. Because, the stored data is encrypted with the DO's public key. So CSP cannot calculate the prediction value on its own. DO decrypts the cipher-texts coming from CSP and gets the prediction value. The final prediction value does not give any information about the query point of QO. After that, prediction value is encrypted by DO with QO's public key and sends it to CSP. CSP forward the encrypted value coming from DO to QO. Finally, QO decrypts the message and learns the prediction value. In the traditional method, DO receives a query location (x_p, y_p) from QO and computes the prediction value and sends the result back to the QO. But in our scenario there are three encryption, two decryption and two mathematical calculation in overall. These additional operations increase the computational cost.

2) Communication Cost Analysis: As expected, communication cost may increase when outsourcing data to cloud servers. DO sends its whole data in encrypted form to CSP. The CSP computes the nearest points and sends them to DO. DO has the required secret key to obtain the prediction value in plain text. DO encrypts the prediction value with the public key of the query owners and sends it back to CSP. QO receives the encrypted prediction value from CSP. The query owner has the necessary key to decrypt the cipher-text. Overall, there are five extra communication activities between DO, CSP and QO. However, in the traditional system there is only two communication between DO and QO.

3) Storage Cost Analysis: In a traditional system, DO is the only party that stores the database. However, in our proposed solution both DO and CSP have to store the database. As a result, the storage capacity requirement of the proposed solution is twice as much as the traditional system and negligible amount of space to store variables during communication and computation process.

B. Accuracy Analysis

Agrawal and Srikant [22] suggested a privacy-preserving technique called data perturbation. Their technique adds additional noise to ensure the privacy of the owner's data. But, this lead to the loss in the accuracy of the prediction model. However, we used a special type of encryption system called Paillier system [18]. In Paillier, the mathematical calculation can be done on the cipher-text, which does not affect the result of the operation. Therefore, our work produces the same prediction values as in traditional scheme. Beyond that it guarantees the confidentiality of the private data of all participants.

C. Privacy Analysis

In our work, data need to be protected in four different places from all three participants. CSP cannot learn the data which is stored in CSP in encrypted form with the key of DO. The second step, when QO sends his request in an encrypted form to hide its points from the CSP and DO. The third one is, CSP process the query to find the k-NN points for the requested coordinate. At this time CSP cannot predict the value. Because all the data were encrypted using the DO's key. After determining the k nearest points, CSP sends the prediction value in encrypted form to DO. DO decrypts and gets the prediction value. In the fourth step, DO cannot get any information about the query coordinate. Therefore, DO cannot predict the location of the requested coordinate. Finally, DO encrypts the predicted value with the public key of QO and send it to CSP. Only QO can decrypt with its private key. CSP does not have any way to learn the predicted value. CSP forwards the cipher text to QO. QO decrypts the received cipher-text and get the predicted value. As a summary, DO cannot predict the query point from QO. CSP cannot predict and/or measure the stored data by the DO, query point from QO and predicted value by the DO. Finally, QO can only get the final predicted value from the DO.

VI. Conclusion

Spatial interpolation is one of the essential operations of geographic information systems. There is a variety of interpolation methods used by researchers. k-NN interpolation method is preferred due to its simplicity and efficiency in computation. Recently, companies tend to move their data and computation needs to cloud servers. However, their privacy may be at risk. Therefore, we propose a k-NN spatial interpolation method which protects privacy of all participants during the process. We are aware of that privacy-preserving methods lead to additional storage, communication and computation cost. We analysed our solution in terms of these costs and presented that they are not that critical from the perspective of all participants.

References

[1] T. Tassa, "Secure mining of association rules in horizontally distributed databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 4, pp. 970–983, 2014.

[2] L. Sun, W.-S. Mu, B. Qi, and Z.-J. Zhou, "A new privacy-preserving proximal support vector machine for classification of vertically partitioned data," *International Journal of Machine Learning and Cybernetics*, vol. 6, no. 1, pp. 109–118, 2015.

[3] S. Kumar and R. Goudar, "Cloud computing-research issues, challenges, architecture, platforms and applications: A survey," *International Journal of Future Computer and Communication*, vol. 1, no. 4, p. 356, 2012.

[4] J. Li and A. D. Heap, "Spatial interpolation methods applied in the environmental sciences: A review," *Environmental Modelling & Software*, vol. 53, pp. 173–189, 2014.

[5] R. L. Krutz and R. D. Vines, *Cloud Security: A Comprehensive Guide to Secure Cloud Computing*. Wiley Publishing, 2010.

[6] S. Pearson and A. Benameur, "Privacy, security and trust issues arising from cloud computing," in *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on*. IEEE, 2010, pp. 693–702.

[7] A. C. Yao, "Protocols for secure computations," in *23rd Annual Symposium on Foundations of Computer Science*. IEEE, 1982, pp. 160–164.

[8] A. Shamir, "How to share a secret," *Communications of the ACM*, vol. 22, no. 11, pp. 612–613, 1979.

[9] M. Ito, A. Saito, and T. Nishizeki, "Secret sharing scheme realizing general access structure," *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, vol. 72, no. 9, pp. 56–64, 1989.

[10] P. Bogetoft, D. L. Christensen, I. Damgård, M. Geisler, T. Jakobsen, M. Krøigaard, J. D. Nielsen, J. B. Nielsen, K. Nielsen, J. Pagter *et al.*, "Secure multiparty computation goes live," in *International Conference on Financial Cryptography and Data Security*. Springer, 2009, pp. 325–343.

[11] M. Andrychowicz, S. Dziembowski, D. Malinowski, and L. Mazurek, "Secure multiparty computations on bitcoin," in *2014 IEEE Symposium on Security and Privacy*, 2014, pp. 443–458.

[12] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

[13] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*. Springer, 1998, pp. 137–142.

[14] X. Xu, H. Yu, G. Zheng, H. Zhang, and L. Xi, "The soil nutrient spatial interpolation algorithm based on knn and idw," in *Computer and Computing Technologies in Agriculture IX*. Springer International Publishing, 2016, pp. 412–424.

[15] S. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 1, no. 4, pp. 300–307, 2007.

[16] O. Goldreich and R. Ostrovsky, "Software protection and simulation on oblivious rams," *Journal of the ACM (JACM)*, vol. 43, no. 3, pp. 431–473, 1996.

[17] E. Shi, T.-H. H. Chan, E. Stefanov, and M. Li, "Oblivious ram with o ((logn) 3) worst-case cost," in *International Conference on The Theory and Application of Cryptology and Information Security*. Springer, 2011, pp. 197–214.

[18] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 1999, pp. 223–238.

[19] S. Goldwasser and S. Micali, "Probabilistic encryption," *Journal of computer and system sciences*, vol. 28, no. 2, pp. 270–299, 1984.

[20] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," *IEEE transactions on information theory*, vol. 31, no. 4, pp. 469–472, 1985.

[21] D. Boneh, E.-J. Goh, and K. Nissim, "Evaluating 2-dnf formulas on ciphertexts," in *Theory of Cryptography Conference*. Springer, 2005, pp. 325–341.

[22] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *ACM Sigmod Record*, vol. 29, no. 2. ACM, 2000, pp. 439–450.

Modeling of Quadrotor Roll Loop using Frequency Identification Method

Mizouri Walid, Najar Slaheddine, Aoun Mohamed, Bouabdallah Lamjed
National Engineering School of Gabes
University of Gabes
Gabes, Tunisia

***Abstract*—Model estimation is an important step in quadrotor control design because model uncertainties can cause unstable behavior especially with non-robust control methods. In this paper, a modeling approach of a quadrotor prototype has been proposed. First an initial dynamic model of quadrotor UAV based on Euler-Lagrange formalism was developed. Then the roll system has been estimated using closed loop identification method and frequency domain analysis. An experimental tests has been performed for the roll system to validate the estimated dynamic model.**

***Keywords*—Quadrotor; modelling; frequency domain; closed loop identification**

I. Introduction

Quadrotor is a Vertical Take-Off and Landing VTOL aircraft consists on two arms on four rotors placed on the extremities of its arms as is shown in Fig. 1. Quadrotor motion is controlled through motors speed variation. In fact changing simultaneously the speed of all motors with the same average produces vertical motion. Rotation around y_B axis (Pitch) is provided by adjusting speed of motor (1) or (3), which results a forward/backward motion. Roll motion is a rotation around x_B axis obtained by varying motor (2) or (4) speeds, thus a right/left translation motion is acquired. Rotation around z_B axis (Yaw) is given by changing velocity of motor (1) and (3) against (2) and (4). Modelling of quadrotor UAV has attracted a lot of interest and there are several methods proposed in literature. In fact mathematical model of the quadrotor UAV was developed [1] and [2] using Newton-Euler and Lagrange formalism resulting a six degree of freedom DOF equations system describing quadrotor motion. Whereas in [3] quaternion method was used to get the quadrotor model. In [4] a non linear model was developed using Euler-Lagrange formalism, then the unknown model parameters has been identified using mathematical calculations and experimental tests. The quadrotor system was described with an ARMAX model obtained by the closed loop identification method [5]. Authors in [6] have used Black-box approach for a variable-pitch quadrotor identification was also developed and explained. In [7] a non linear dynamic model of a quadrotor UAV was detailed by considering the aerodynamic effects into account, to cover a wide flight regime not only hovering position. [8] used the Prediction Error Method PEM with blak box model approach to identify Linenar Time Invariant LTI model exploring experimental flight data. In [9] identification of linear parametric model of quadrotor UAV is developed using frequency domain method. [10] developed system identification of translation for a cost open-source quadrotor prototype MikroKopter.

This paper presents an identification method to estimate dynamic model of quadrotor roll loop basing on initial double integration model. This paper is organized as follows: Mathematical modeling of quadrotor attitude is elaborated in Section II. Section III describes Sunbird quadrotor prototype. Section IV presents the frequency identification method for quadrotor roll loop. Simulation results and discussion is shown in Section V. Finally, conclusion and future work are given in Section VI.

II. Mathematical Modelling

This section describes the quadrotor mathematical model developed using Euler-Lagrange formalism.

Denoting q the generalized coordinates of the quadrotor

$$q = (x, y, z, \phi, \theta, \psi) \in \Re^6 \tag{1}$$

where $\xi = (x, y, z)$ is the positions of the quadrotor center of mass relative to the fixed inertial frame ***E***, the orientation of the quadrotor are expressed by $\vartheta = (\phi, \theta, \psi)$, where ϕ is the roll angle around the x_B axis, θ is the pitch angle around the y_B axis and ψ is the yaw angle around the z_B axis. The

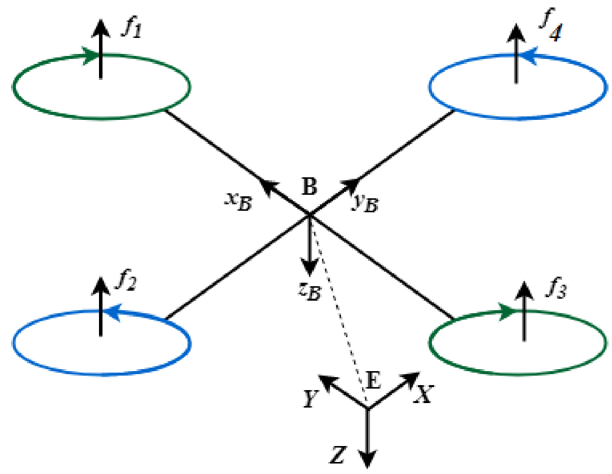


Fig. 1. Quadrotor UAV coordinate system.

transformation matrix from inertial frame ***E*** to the body frame ***B*** is given by

$$R = \begin{bmatrix} c\psi\ c\theta & c\psi\ s\theta\ s\phi - s\psi\ c\phi & c\psi\ s\theta\ c\phi + s\psi\ s\phi \\ s\psi\ c\theta & s\psi\ s\theta\ s\phi + c\psi\ c\phi & s\psi\ s\theta\ c\phi - s\phi\ c\psi \\ -s\theta & c\theta\ s\phi & c\theta\ c\phi \end{bmatrix} \tag{2}$$

where $s\theta = \sin(\theta)$ and $c\theta = \cos(\theta)$.

Lagrangian equation is defined as follow

$$L(q, \dot{q}) = E_p - E_c \tag{3}$$

where E_c and E_p are respectively the potential and kinetic energy.

$$\begin{cases} E_c = mgz \\ E_p = \frac{m}{2}\dot{\xi}^T\dot{\xi} + \frac{1}{2}\dot{\upsilon}^T I \dot{\upsilon} \end{cases} \tag{4}$$

where I is the inertia matrix, m is the total mass of the quadrotor and g is the gravity acceleration.

Euler-Lagrange formulas is

$$\frac{d}{dt}(\frac{\delta L}{\delta \dot{q}}) - \frac{\delta L}{\delta q} = \begin{bmatrix} F \\ \tau \end{bmatrix} \tag{5}$$

where F is the force of translation, τ is the total torque. Let's start with the rotational dynamic of the quadrotor. where τ is the total torques produced by quadrotor system. We can write

$$\tau = \tau_1 + \tau_2 \tag{6}$$

where τ_1 is the roll, pitch and yaw torques produced by the quadrotor motors

$$\tau_1 = \begin{bmatrix} bl(\Omega_2^2 - \Omega_4^2) \\ bl(\Omega_1^2 - \Omega_3^2) \\ d(\Omega_1^2 - \Omega_2^2 + \Omega_3^2 - \Omega_4^2) \end{bmatrix} \tag{7}$$

τ_2 is the gyroscopic effects given by

$$\tau_2 = \begin{bmatrix} \dot{\theta} I_{rotor} \Omega_g \\ \dot{\phi} I_{rotor} \Omega_g \\ 0 \end{bmatrix} \tag{8}$$

where Ω_i is the angular speed of the i^{th} motor, b is the thrust coefficient, d is the drag coefficient, l is the distance between the motors and the quadrotor center of gravity and I_{rotor} is the rotor inertia moment.

$$\Omega_g = (\Omega_1 - \Omega_2 + \Omega_3 - \Omega_4) \tag{9}$$

The generalized moments are

$$\begin{cases} \frac{d}{dt}(\frac{\delta L}{\delta \dot{\phi}}) - \frac{\delta L}{\delta \phi} = I_{xx}\ddot{\phi} + (I_{zz} - I_{yy})\dot{\theta}\dot{\psi} \\ \frac{d}{dt}(\frac{\delta L}{\delta \dot{\theta}}) - \frac{\delta L}{\delta \theta} = I_{yy}\ddot{\theta} + (I_{xx} - I_{zz})\dot{\phi}\dot{\psi} \\ \frac{d}{dt}(\frac{\delta L}{\delta \dot{\psi}}) - \frac{\delta L}{\delta \psi} = I_{zz}\ddot{\theta} + (I_{yy} - I_{xx})\dot{\phi}\dot{\theta} \end{cases} \tag{10}$$

Where I is the inertia matrix:

$$I = \begin{bmatrix} I_{xx} & 0 & 0 \\ 0 & I_{yy} & 0 \\ 0 & 0 & I_{zz} \end{bmatrix}$$

The rotational dynamic equations can deduced from (4), (5) and (7)

$$\begin{cases} \ddot{\phi} = \dot{\theta}\dot{\psi}(\frac{I_{yy} - I_{zz}}{I_{xx}}) + \frac{I_{rotor}\theta\dot{\Omega}_g}{I_{xx}} + \frac{1}{I_{xx}}U_2 \\ \ddot{\theta} = \dot{\phi}\dot{\psi}(\frac{I_{zz} - I_{xx}}{I_{yy}}) + \frac{I_{rotor}\phi\dot{\Omega}_g}{I_{yy}} + \frac{1}{I_{yy}}U_3 \\ \ddot{\psi} = \dot{\phi}\dot{\theta}(\frac{I_{xx} - I_{yy}}{I_{zz}}) + \frac{1}{I_{zz}}U_4 \end{cases} \tag{11}$$

where (U_1, U_2, U_3, U_4) are the model inputs which given by the following expressions

$$\begin{cases} U_1 = b\ (\Omega_1^2 + \Omega_2^2 + \Omega_3^2 + \Omega_4^2) \\ U_2 = bl\ (\Omega_2^2 - \Omega_4^2) \\ U_3 = bl\ (\Omega_1^2 - \Omega_3^2) \\ U_4 = d\ (\Omega_1^2 - \Omega_2^2 + \Omega_3^2 - \Omega_4^2) \end{cases} \tag{12}$$

In order to reduce complexity of calculus, gyroscopic effects and moments of inertia terms can be neglected in the motion (8). Thus the roll

$$\begin{cases} \ddot{\phi} = \frac{1}{I_{xx}}U_2 \\ \ddot{\theta} = \frac{1}{I_{yy}}U_3 \\ \ddot{\psi} = \frac{1}{I_{zz}}U_4 \end{cases} \tag{13}$$

To estimate moment of inertia, the quadrotor can be decomposed into several parts (Arm, Battery, Card, Rotors). Then using the formulas (14) the inertia moment of each part relative to his axis can be calculated. After that by applying the Huygens-Steiner theorem (15) we can get the moment of inertia of each part through the quadrotor center of gravity. The quadrotor moment of inertia is given by the sum of the moment of each parts.

$$I_0 = \int x^2 dm \tag{14}$$

$$Ig = I_0 + md^2 \tag{15}$$

Denoting $G(s)$ as the dynamic model of the roll axis

$$G(s) = \frac{1}{I_{xx}s^2} \tag{16}$$

III. Sunbird Quadrotor Prototype

In this paper a quadrotor prototype is used as a test platform named Sunbird Shown in Fig. 2. Sunbird is a home-made quadrotor platform designed at MACS research laboratory.

Sunbird quadrotor is composed as follows:

• Arduino DUE card used as flight controller. It based on a 32-bit ARM core micro-controller with 54 digital input/output, 2 analog output and 12 analog inputs.

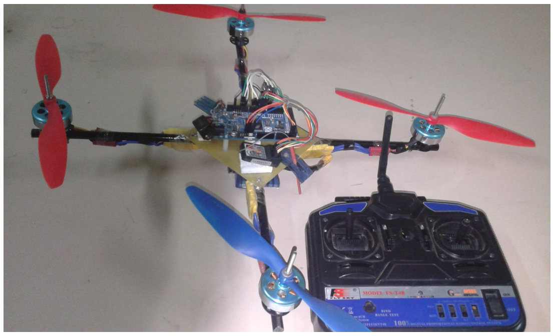


Fig. 2. Sunbird quadrotor prototype.

• FlySky 2.4 Ghz 4 channels Radio Control Transmitter and Receiver used to remotely control the quadrotor.

• Rotor composed with Hextronik DT750 brushless DC motor, 10×4.5 propellers and Electronic Speed Controller ESC 30A operating with Pulse Width Modulation (PWM) signal generated by Arduino Due card.

• Ultrasonic sensor HC-SR 05 for altitude measuring with a 4.5 meter of measurement range.

• Measurement Processing Unit MPU6050 is also used to estimate attitude. MPU-6050 is a six axis IMU sensor containing 3 axis gyroscope ITG 3200 giving the angular velocities and 3 axis accelerometer ADXL 345 measuring linear accelerations. Arduino Due board receives data from MPU-6050 via I2C-bus. Then a complementary filter is used to attenuate high frequency signals affecting accelerometer like vibration and compensates the drifts affecting gyroscope measurements. The composition of quadrotor prototype is described in Fig. 3.

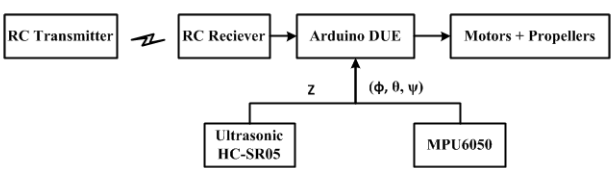


Fig. 3. Sunbird quadrotor prototype architecture.

IV. Roll System Frequency Identification

In this section we interested on roll system modeling using the closed loop identification method. It is based on frequency domain analysis in order to estimate the process model.

The roll system described in (13) is unstable, hence the necessity of a control loop.

A. Control loop identification

A Proportional Integral Derivative controller was used to ensure the roll loop stability. The PID control loop used in this section was described in Fig. 4.

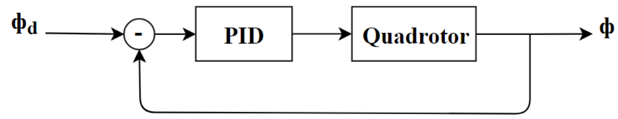


Fig. 4. Process model in a PID control loop.

The PID controller is described by the following expression:

$$C(s) = k_p + \frac{k_i}{s} + k_d s \tag{17}$$

where (k_p, k_i, k_d) are respectively the proportional, integral and derivative factors.
The control system is based on reverse model technique. In fact the control signals was used to calculate the desired motors speed $(\Omega_{d1}, \Omega_{d2}, \Omega_{d3}, \Omega_{d4})$ from (12).

$$\begin{cases} \Omega_{d1} = \sqrt{\frac{U_1}{4b} + \frac{U_\theta}{2bl} + \frac{U_\psi}{4d}} \\ \Omega_{d2} = \sqrt{\frac{U_1}{4b} + \frac{U_\phi}{2bl} - \frac{U_\psi}{4d}} \\ \Omega_{d3} = \sqrt{\frac{U_1}{4b} - \frac{U_\theta}{2bl} + \frac{U_\psi}{4d}} \\ \Omega_{d4} = \sqrt{\frac{U_1}{4b} - \frac{U_\phi}{2bl} - \frac{U_\psi}{4d}} \end{cases} \tag{18}$$

Then using the desired motor speed, we can deduce the PWM signal. The relation between PWM and motor speed was experimentally estimated. Fig. 5 illustrates the relation of PWM signal over the motor speed, which can be expressed by the following second order equation using "$polyfit$" Matlab function

$$PWM = a_2\Omega^2 + a_1\Omega + a_0 \tag{19}$$

where $a_2 = 1.2093\ 10^{-5}$, $a_1 = 6.5776\ 10^{-2}$ and $a_0 = 948.2$.

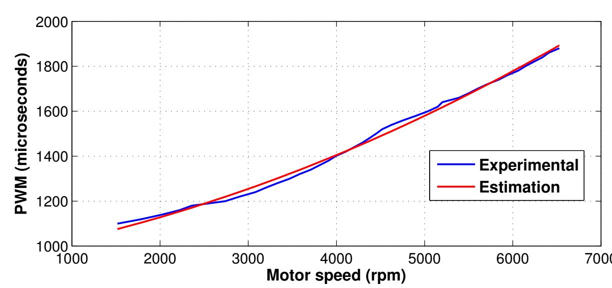


Fig. 5. Measured and approximated PWM signal and motor speed characteristics.

The PID controller was implemented on the Arduino DUE board and then was experimentally adjusted. The chosen controller parameters are

$$k_p = 0.35\ k_i = 0.28,\ \ k_d = 0.46$$

B. Frequency Identification

The The closed loop system in Fig. 4 was excited by a varied sine setpoint with a magnitude of $5°$ and a frequency between $[10^{-2},\ 10^{2}]\ Hz$. The PID control loop is implemented in the Arduino due card operating with a sample time of 0.01

second. The quadrotor was placed on a test bench ensuring rotation around the roll axis with minimum friction effects shown in Fig. 6.

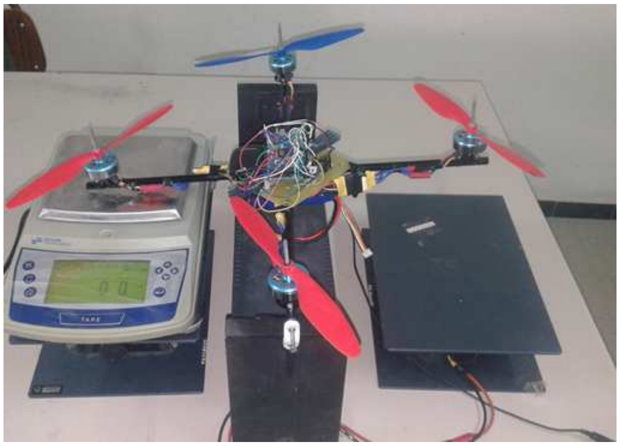


Fig. 6. Quadrotor prototype placed on a test bench ensuring rotation around roll axis.

Fig. 7, 8 and 9 illustrate the setpoint and the output signals with a frequency of respectively 10^{-2}, 10^{-1} and $1Hz$. As

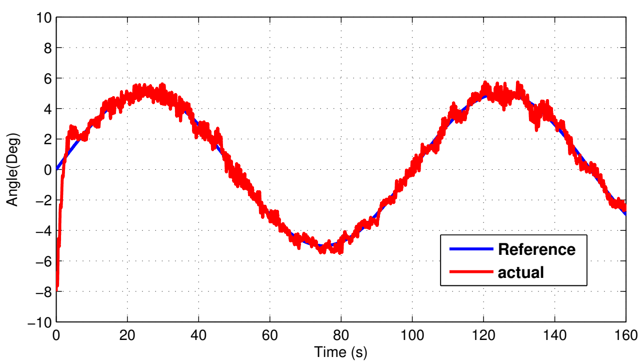


Fig. 7. Roll angle response for a sine setpoint with a frequency of $0.01Hz$.

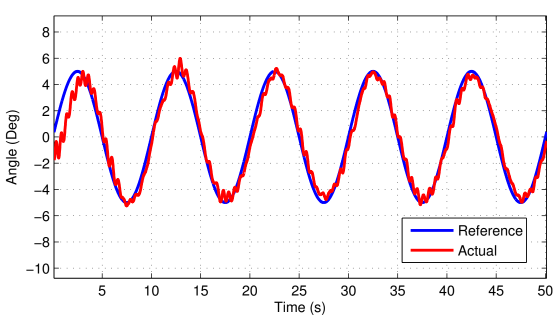


Fig. 8. Roll angle response for a sine setpoint with a frequency of $0.1Hz$.

can be seen, for low frequency the input and the output signals are quietly non phases and gains are observed. However for frequency around $1Hz$ the output amplitude increases locally then decreases for a highest frequencies, and a lag behavior

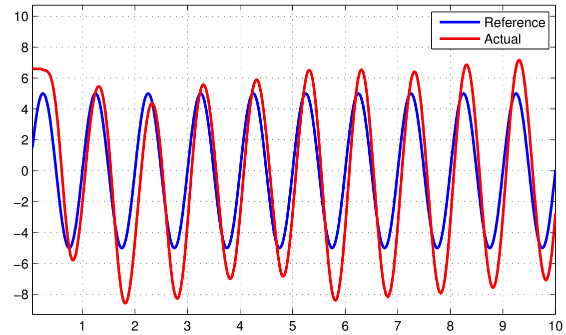


Fig. 9. Roll angle response for setpoint frequency $f = 1Hz$.

is also observed. From several sine responses, the phase and magnitude can be measured for a each frequency. Then the Bode diagram can be plotted. Bode plot of the quadrotor proposed control loop is shown in Fig. 10.

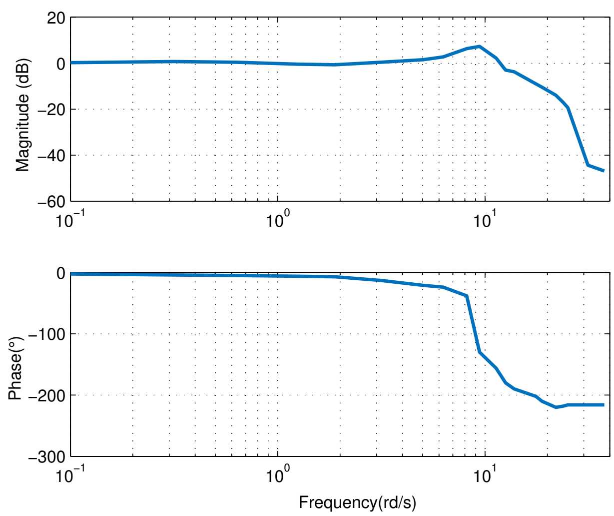


Fig. 10. Bode plot of the closed loop.

Basing on bode plot in Fig. 10 the closed loop transfer function can be calculated. From the Bode diagram the closed loop system can be estimated as a second order system with resonance. Denoting $F(s)$ the control loop transfer function described in Fig. 4.

$$F(s) = \frac{1}{\frac{1}{\omega_n^2}s^2 + \frac{2\xi}{\omega_n}s + 1} \tag{20}$$

where ω_n is the normal frequencies and ξ is the damping factor.
Denoting Q the resonance factor of the closed loop system which can be described by the following expression:

$$Q = \frac{1}{2\xi\sqrt{1-\xi^2}} \tag{21}$$

Thus

$$4Q^2\xi^2(1-\xi^2) = 1 \tag{22}$$

We can write also

$$4Q^2\xi^4 - 4Q^2\xi^2 + 1 = 0 \quad (23)$$

The damping factor ξ is given by solving the second order differential equation (23).
We denotes ω_r as the resonance frequency. Equation (24) describes the relation between the resonance ω_r and normal frequencies ω_n.

$$\omega_n = \frac{\omega_r}{\sqrt{1-2\xi^2}} \quad (24)$$

We obtained the following normal frequency and damping factor:

$$\xi = 0.24, \ \omega_n = 10rad/s \quad (25)$$

The transfer function control loop can be written as

$$F(s) = \frac{C(s)G(s)}{1 + C(s)G(s)} \quad (26)$$

where $C(s)$ is the PID controller and $G(s)$ is the process model transfer function.

$$G(s) = \frac{F(s)}{C(s)(1 - F(s))} \quad (27)$$

Replacing $C(s)$ and $F(s)$ with their expression we can write

$$G(s) = \frac{1}{(\frac{k_d s^2 + k_p s + k_i}{s})(\frac{s^2}{\omega_n^2} + \frac{2\xi s}{\omega_n})} \quad (28)$$

Thus the open loop roll model is following third order transfer function

$$G(s) = \frac{1}{\frac{k_d s^3}{\omega_n^2} + s^2(\frac{2\xi k_d}{\omega_n} + \frac{k_p}{\omega_n^2}) + s(\frac{k_i}{\omega_n^2} + \frac{2\xi k_p}{\omega_n}) + \frac{2\xi k_p}{\omega_n}} \quad (29)$$

The deduced process model is a third order transfer function system

$$G(s) = \frac{1}{c_3 s^3 + c_2 s^2 + c_1 s + c_0} \quad (30)$$

where

$$\begin{cases} c_3 = \frac{k_d}{\omega_n^2} \\ c_2 = \frac{2\xi k_d}{\omega_n} \\ c_1 = \frac{k_i}{\omega_n^2} + \frac{2\xi k_p}{\omega_n} \\ c_0 = \frac{2\xi k_p}{\omega_n} \end{cases} \quad (31)$$

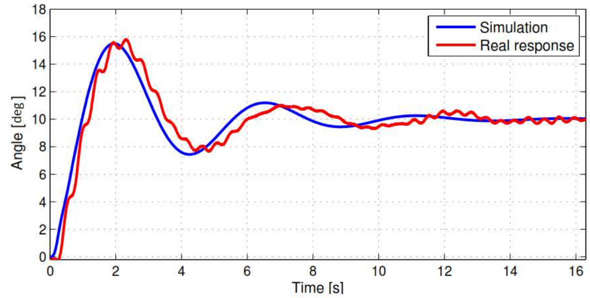


Fig. 11. Step responses of estimated dynamic model and experimental prototype.

V. Results and Discussion

To evaluate the performances of the proposed identification method. The quadrotor has been placed in a PID control loop to compare the real prototype responses and simulations of the estimated process model. Fig. 11 illustrates the experimental simulation step responses of the roll system. We can remark that both of these responses are very close, and the rise and settling time are nearly the same.

In addition the quadrotor roll axis was tested with sine response. Fig. 12 shows the measured response and simulation of the identified dynamic model for sine setpoint. We can see that the responses are approximately superposed. Basing on these results we can judge the validity of estimated process model by the proposed identification method.

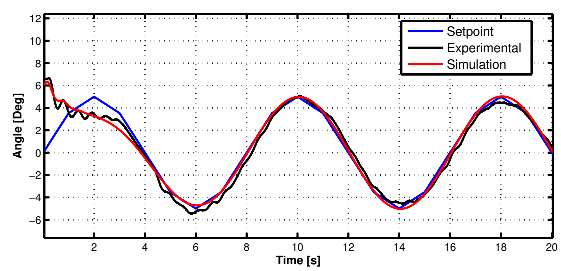


Fig. 12. Roll response of experimental prototype and simulation of process model.

VI. Conclusion

In this work the frequency identification method is proposed to estimate the quadrotor roll model. A non linear model is firstly developed using Euler-Lagrange formalism. Then the proposed identification method is applied to Sunbird quadrotor prototype using a PID control loop. Frequency domain analysis is then used for roll loop model. Finally, a comparison of experimental and simulated results in order to validate the estimated model. For the future work, the other quadrotor components $(\theta, \ \psi, x, \ y, \ z)$ should be estimated using the frequency identification to validate the effectiveness of this method in quadrotor system full modeling.

References

[1] S. Bouabdallah, "Design and control of quadrotors with application to autonomous flying," 2007.

[2] P. Castillo, R. Lozano, and A. E. Dzul, *Modelling and control of mini-flying machines*. Physica-Verlag, 2006.

[3] E. Fresk and G. Nikolakopoulos, "Full quaternion based attitude control for a quadrotor," in *Control Conference (ECC), 2013 European*. IEEE, 2013, pp. 3864–3869.

[4] A. Chovancová, T. Fico, L. Chovanec, and P. Hubinsk, "Mathematical modelling and parameter identification of quadrotor (a survey)," *Procedia Engineering*, vol. 96, pp. 172–181, 2014.

[5] R. Schreurs, S. Weiland, H. Tao, Q. Zhang, J. Zhu, Y. Zhu, and C. Xu, "Open loop system identification for a quadrotor helicopter system," in *Control and Automation (ICCA), 2013 10th IEEE International Conference on*. IEEE, 2013, pp. 1702–1707.

[6] P. Panizza, F. Riccardi, and M. Lovera, "Black-box and grey-box identification of the attitude dynamics for a variable-pitch quadrotor," *IFAC-PapersOnLine*, vol. 48, no. 9, pp. 61–66, 2015.

[7] Y.-R. Tang, X. Xiao, and Y. Li, "Nonlinear dynamic modeling and hybrid control design with dynamic compensator for a small-scale uav quadrotor," *Measurement*, vol. 109, pp. 51–64, 2017.

[8] M. G. Yoo and S. K. Hong, "System identification of the quadrotor flying robot in hover using prediction error method," *Electrical and Electronic Engineering*, vol. 118, pp. 21–26, 2015.

[9] P. Niermeyer, T. Raffler, and F. Holzapfel, "Open-loop quadrotor flight dynamics identification in frequency domain via closed-loop flight testing," *AIAA Guidance, Navigation, and Control Conference*, vol. 118, pp. 21–26, 2015.

[10] I. Sa and P. Corke, "System identification, estimation and control for a cost effective open-source quadcopter," in *Robotics and automation (icra), 2012 ieee international conference on*. IEEE, 2012, pp. 2202–2209.

A Comparison of Near-Hidden and Information Asymmetry Interference Problems in Wireless Mesh Networks

Sadiq Shah*, Muhammad Atif†, Samiullah Khan‡, Misbah Daud§ and Fahim Khan Khalil ¶
*Department of Computer Science, FATA University, FR Kohat,
†‡§¶Institute of Business Management Sciences, The University of Agriculture Peshawar-Pakistan

***Abstract*—Multi-radio Multi-channel (MRMC) Wireless Mesh Networks (WMNs) have made quick progress in current years to become the best option for end users due to its reliability and low cost. Although WMNs have already been used still the capacity of WMNs is limited due to information asymmetry and near hidden interference among frequency channels. In the past, various research studies have been done to investigate both these issues. To minimise both these interference types and maximise network capacity; channel assignment has always been a critical area of research in WMNs. In this research, a comparative analysis is done between NH and IA interference based on their impact on network capacity. This comparison is made using the existing Optimal Channel Assignment Model (OCAM). After extensive simulations, it is figured out that NH interference performs a major role in degrading overall network capacity up to 4% in comparison to IA interference. Further, in this research an optimal channel assignment model Information Asymmetry and Near Hidden Minimization (INM) model is proposed that collectively minimises both NH and IA interference problems. The proposed model considers three non-overlapping channels 1, 6 and 11 from IEEE802.11b standard. For simulations, four different Multi-radio Multi-channel Wireless mesh topologies have been considered to find out the average network capacity. An extensive simulation in OPNET shows that the proposed INM model performs 7% better than the existing OCAM model in maximising the WMN net capacity.**

Keywords*—Wireless Mesh Network (WMN); near hidden and information asymmetry interference; non-overlapping channels

I. Introduction

All types of Wireless networks develop into the next generation to offer improved and attractive services. Among those networks, Wireless Mesh Network (WMN) has emerged in recent times. Three different types of nodes are used in a wireless mesh network that is mesh router, mesh client and gateway router [1]. In WMN nodes are connected with one another, and these nodes operate as a router as well as host, sending data to the node which are not in transmission range with the help of intermediate nodes. Due to its different features, i.e. reliability, redundancy and low cost the WMNs are deployed most quickly. In WMN when one link is down or not working, communication can be done with the help of other links. Due to self-forming and self-configuration, a WMN convey several advantages to its users, i.e. low costs, network maintenance, reliable coverage and robustness [3].

For Commercial users, WMN offers a guarantee for their applications. Both conventional and mesh clients used the WMNs to connect to the backbone or mesh routers, and these routers are fixed, or minimum mobility and these nodes operate as a router as well as host. In WMN nodes send data to the destination directly or with the help of intermediate nodes [12]. A wireless mesh network made from multiple wireless nodes. In a wireless mesh network, every node can transmit information directly to single or multiple peer nodes. In a wireless mesh network, new links are created, and the nodes are connected to each other with the help of these links. When the numbers of interconnected nodes are increased at that time the possible path and the total bandwidth is also increased [5]. Dynamic channel assignment permits that different channel may be assigned to different interfaces and these interfaces may regularly change between multiple channels [18].

In WMN routers are usually fixed or minimal mobility, and there is no power issue unlike in sensor network. Usually, the structure design of WMN consists of three levels. First one is mesh gateways second one is mesh router, and the third one is mesh, client. In WMN the internet backbone is connected with mesh backbone with the help of mesh gateways [4]. In a wireless mesh network, the second level consists of mesh router/backbone. The mesh backbone consists of mesh routers, and these mesh routers are fixed sites. These mesh router collect data from mesh client and send to gateway router and from gateway to the internet. The third and last level of WMN is mesh client. The mesh clients are the actual transmission and reception of wireless mesh network [7].

In Fig. 1 the mesh backbone receives data which are sending by mesh client. The mesh backbone then forwards data to mesh gateway and then to the Internet. Internet is shown in the form of cloud. These mesh routers have no mobility or fixed, and multiple interfaces are assigned to them. In Fig. 1 the mesh client is laptops, computers and mobiles as shown. The links are wireless links the works as backbone among all the mesh routers. The mesh routers have multiple interfaces.

A. Applications of WMN

Wireless mesh networks consist of fixed or mobile devices. The solutions are as different as communication needs, for example in unmanageable environments such as emergency situations, oil rigs, tunnels, battlefield surveillance, high-speed mobile video applications on board public transport. An important possible application for wireless mesh networks is VoIP. By using a Quality of Service scheme, some current applications are mentioned below [8].

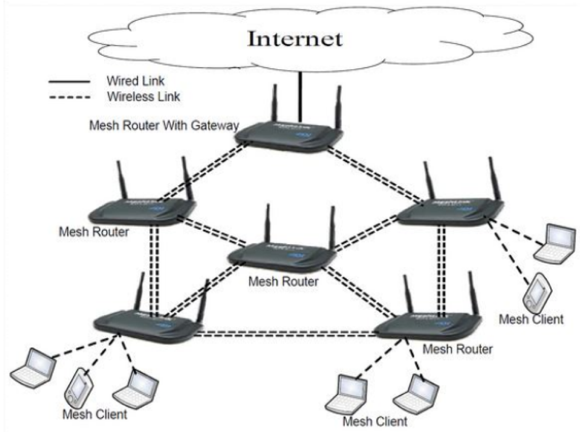


Fig. 1. Diagram of wireless mesh network.

- U.S. military forces connect their computers through wireless mesh networking field operations.
- Now in houses, electric meters are used that transmit their readings from one to another and lastly to the main office for billings without the requirement of human meter readers.

B. Multi-Radio Multi-Channel Wireless Mesh Network

"In MR-MCWMN have multiple interfaces and more than one frequency channels may be assigned to them. In this network particular wireless mesh routers work as a gateway and these gateways routers are connected to the internet or another network that may be heterogeneous. MRMC-WMNs have got development quickly in current years and have become the best option for end users as it is trusty, scalable and can expend network connectivity on the last mile. WMN provide excessive throughput because of its multi-radio and multi-channel technology [10]. A WMN can be categorized into three sections. The first section includes gateways. Through these gateways routers, the wireless mesh network is connected to the internet as discussed in Fig. 1. The second section consists of mesh router. These mesh router forward information inside the WMNs on the interest of end users and these mesh router is also called nodes. The third section consists of mesh client or wireless LANs. In wireless mesh network end users are the real sender and receiver of data. These mesh routers send and receive data if these routers work on the same frequency channel. Due to multiple channel and multiple interfaces use interference is tack place that causes network throughput degradation [18]."

Single-radio single-channel (SR-SC) is a type of wireless mesh network (WMN). Which consist of mesh router and every mesh router consist of only one network interface card. Due to frequent backoff and packet collisions, the capacity and throughput of the network are suffering and low. In this research, multi-radio multi-channel WMN will be used for exploration. The main issue that is to be solved is channel interference [15].

In multiple simultaneous transmissions, the capacity of a wireless network is limited due to interference which is a major problem. In WMNs this problem is removed by mesh routers which have multiple radios. In multi-radios, nodes can send data on multiple channels at the same time and also send and receive data simultaneously. However, the interference is not eliminated due to a specific number of channels. In a wireless mesh network, two nodes can only send and receive data if these nodes are within the transmission range and working on the same frequency channel. In a wireless mesh network at a distance of transmission range, every mesh router has carrier sensing range. Within receiving range, mesh router can generate interference if they are transmitting information and working on the identical frequency channel. When interference arises, it degrades WMN performance, or it can cause transmission losses. In a wireless mesh network, two different types of interference occur. The first one is coordinated and the second one is non-coordinated interference. The non-coordinated interference is categorized as information asymmetry (IA), near-hidden (NH) and far-hidden (FH). If source nodes of multiple links are within receiving a range of each other, then these links are coordinated (CO) interfering links. In this research information asymmetry and near-hidden interference will be minimised and their effect will be compared [20].

C. Information Asymmetry

In the event of information asymmetry (IA) interference, both links L1 (A, B) and L2 (C, D) are operating on the same frequency channel. In IA both source nodes A and C are outside each other carrier sensing ranges (Cs). Similarly, A and D are also outside each other carrier-sensing range, but C and B are inside each other carrier-sensing ranges. Let d represents the physical distance between nodes. For IA interference the following terms need to be fulfilled.

$d(A, C) > Cs$

$d(B, C) < Cs$

$d(A, D) > Cs$

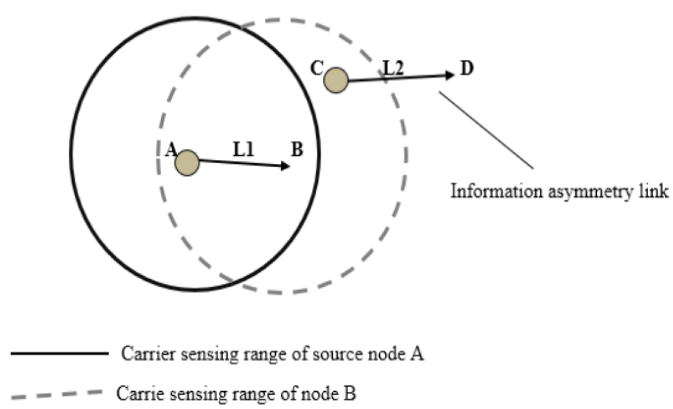


Fig. 2. Information asymmetry interference in wireless mesh network.

In Fig. 2 both source nodes A and C are far away from each other carrier sensing ranges (Cs). Similarly, A and D are also far away from each other receiving range, but source C and receiver B are inside each other carrier-sensing ranges. In these conditions if C starts data transmission to D then B will sense all the data sent by C. In this case B will sense data from both A and C and data collision is experienced. The flow on

L1 (A, B) can be minimised due to interference from L2(C, D).

D. Near-Hidden Terminal

In the event of near hidden (NH) interference, the source nodes of both links are far away from each other carrier sensing range. However, the receivers of both links are inside each other carrier sensing range. When both links L1 (A, B) and L2 (C, D) are working on the same frequency channel. Then for near-hidden (NH) interference, the following terms need to be fulfilled:

$d(A, C) > Cs$

$d(B, C) < Cs$

$d(A, D) < Cs$

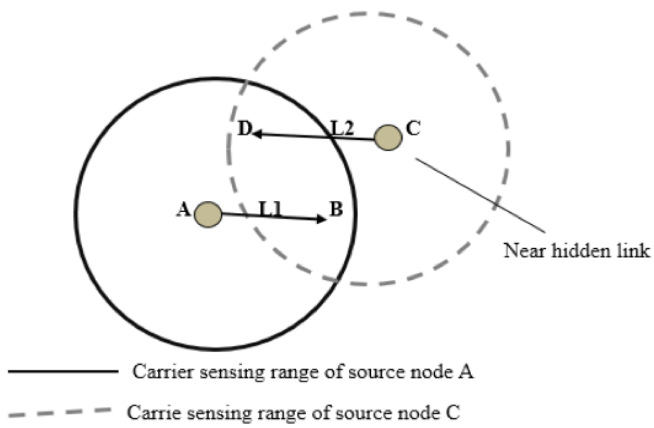


Fig. 3. Near-hidden terminals in a wireless mesh network.

In Fig. 3 source nodes A and C are outside each other carrier-sensing range Cs. However, the receivers of both the links are inside the sensing range of each others carrier sensing range. In these links interference occur when node A sends data to B. In the middle of this transmission, node C also has data to send to node D at the same frequency channel. The nodes A and D are hidden from each other concerning the destination node. Hidden nodes can reduce the capacity of the network because of the possibility of interference. Minimization of IA and NH interference, an optimal channel will be proposed because channel assignment is a major area for dealing channel interference.

E. Channel Assignment in Wireless Mesh Network

There are different problems confronted by an MRMC-WMNs, i.e. channel assignment, node deployment and link scheduling. In MRMC -WMN channels assigning to a radio interface to minimise the interference and attain efficient channel utilisation is done by channel assignment (CA). Channel assignments are used to perform an appropriate mapping among channels and the radios at specific nodes. To minimise interference between WMN links channel assignment are used. The purpose of CA scheme in MRMC-WMN is to improve the combined network capacity and reduced interference and to keep the network connectivity. Channel assignment in MRM-WMN will allot channels to the free radio interfaces which are free to get best uses of the channel and for minimisation of interference. The CA schemes are commonly categorised as Fixed (static)and dynamic channel assignment. For CA the technology used will be IEEE 802.11b. The frequency range of 802.11b is 2.4 GHz, 11 frequency channels are used in IEEE 802.11b frequency band. Non-overlapping channel or orthogonal are those channels which are separated by at least 25 MHz non-overlapping channel do not interfere, in IEEE 802.11b only three channels (1,6,11) are currently used that do not overlap with each other. By allotting these non-overlapping channels through the optimum way, it minimises the channel interference. Channel assignment scheme is categorised as static (fixed), dynamic and hybrid channel assignment. In static (fixed) channel assignments different channels can be assigned to a specific node permanently or for a long time with the help of static assignment algorithm [7].

II. Literature Review

The basic idea familiarised by the authors to transmit data in a bottleneck collision domain in the overall geographic area is discussed in [9]. The authors displayed that in WMN the capacity of every node is reduced as O(1/n). Where n denote the entire mesh nodes in the wireless mesh network. The authors equate existing proposed function, and they delivered the exact higher boundary on the capacity of any node.

An algorithm of load-aware in a wireless mesh network is suggested in [17], which assigns channels dynamically to multi-radio WMN. The author suggestion needs that after channel assignment that expected traffic loads and paths traversed are take place. The essential, unique feature in his suggestion is that to allot channels entirely based on awareness of interference in the wireless mesh network.

The available network capacity could be increased by using multiple channels. However, it required new protocols fulfil the capacity [11]. The author further discussed that by using multiple interfaces the capacity of multi-channel wireless mesh network could be increased. The author considered the scenario in which multiple interfaces are presented. However, the numbers of the available interfaces are more than the available channels. He offers different interface assignment strategies and proposes new strategies without the adaptations to 802.11, and for those strategies, he also recognises the routing heuristics. The author discussed that in wireless mesh network the frequent channel switching arises many problems. Just like performance degradation and end-to-end delay performance. In this research paper, the capacity of multi-channel WMN is enhanced with the help of multiple interfaces.

Author in [13] supposed that for practical purposes in IEEE 802.11a the overlapping among adjacent channels is so small which can be ignored. In this research, the authors displayed that by using 802.11b and calculating the throughput, three channels are separated, i.e. the free distance between two non-overlapping neighbouring channels, and a 10m distance is sufficient for two links to send and receive actual amount of data without interference with each other. However, only calculating the network throughput is not enough but also the interference is decreased to come to the results. The supposition of the author is entirely incorrect for 802.11a. It is answered merely that more susceptible spreading technique and higher modulation scheme in 802.11a.

In WMN every node consists of more than one radio interfaces, and more than one channel can be assigned to that radio for communication [19]. To assign channel through proper manner minimise the overall network interference. On any node number of channel is more than the numbers of radios. The constraint must follow by channel assignment that different channel assigned to the links on any node. This process is called NP-hard problem. For channel assignment distributed and centralised algorithm are designed. Through this algorithm, he developed a semidefinite program formulation of the optimisation problem to achieve edge boundaries on overall network interference.

Author in [7] studied the different constraints of CA in WMN and show the main issues in channel assignment schemes for multipath connectivity, traffic patterns and interference situation. Three different channel assignment schemes are used, i.e. fixed, dynamic and hybrid channel assignment schemed were examined. The main objective of CA algorithm to reduced interference and improve the overall network capacity and to manage the hole network connectivity in WMN. They presented a new MESTIC scheme to associates the mesh connectivity issues and mesh traffic issues to decline interference in a multi-radio mesh network.

In MR-MC WMN multiple channels are assigned to radio interfaces for communication. The objective of the authors is to minimise the network interference through allocating channels to communication links in a network [2]. The number of available channels is more than the number of radios on every node and constraint must be followed by channel assignment. When the number or radios channel is more than the radio links then the interference occurs on that node, and this problem is called NP-hard, and for this problem, the author developed distributed and centralised algorithms for channel assignment. To measure the output achievement by these algorithms, the author developed liner program and semidefinite program formulation to achieve the edge boundaries on network interference.

Author in [6] discussed that in wireless mesh network there are two main types of interference. First one is coordinated (CO) and the second one is non-coordinated (n-CO) interfering links. He further divides the non-coordinated (n-CO) interference links into three types. First one is information asymmetry (IA) the second one is near hidden (NH), and the third one is far hidden (FH) interference (Garettoet al., 2008). The author further discussed these interference links on an MRMC -WMN, and find out all the probabilities of packet loss in these links. Then compare the coordinated (CO) and non-coordinated (n-CO) interference links, and prove that non-coordinated links have high transmission loss as compared to coordinated interference. After that he proposed an analytical model that minimizes this interference.

It was shown in [14] that the non-interfering links such as radio frequency band, traffic pattern and physical modulation (PHY), as well as a multi or single radio system. By using more than one non-overlapping radio frequency channels to improve the capacity of WMNs, and several devices can transmit data on a different channel within a collision domain, and adjacent channel interference (ACI) problem is discussed.

Author in [16] stated the suggestion of dynamic channel-assignment algorithm LYCAS. He delivered an optimisation model to minimising the non-coordinated interference and maximising the network capacity. The authors discussed coordinated and non-coordinated interference and displayed that non-coordinated interference is more destructive than coordinated interference. The author used non-overlapping channel assignment approach to minimise non-coordinated interference. In this paper, decision variable is used for non-overlapping channel assignment. His objective is to minimise interference and maximise network capacity.

Author in [18] proposed optimisation model that concentrates on the maximisation of network capacity and minimisation of non-coordinated interference operating the effectiveness of non-overlapping channel assignment in MRMC WMNs by offering an optimised channel assignment approach. Through simulation, both spare and dense channel assignment are compared in MRMC-WMN topologies. Non-overlapping channels which are three (1, 6, 11) of 802.11b were considered for channel assignment strategy. In last with the help of OPNET simulator both dense and sparse topologies were concluded. Through simulation, the author shows that proposed model gives 19% better results as compared to the dense network.

Author in [21] suggested that in this scenario non-overlapping channels are less than the wireless mesh nodes, so these non-overlapping channels repeatedly assigned to nodes and shared between these nodes. Due to this failure, the channel interference occurs and also degrades the performance of the network. The author has presented a mathematical Linear Programming (LP) model to minimise the interference through a group based channel assignment method. This model determines the channel assignment strategy for the removing of interference. For solving the optimisation function, the Discrete Particle Swarm Optimization (DPSO) are used.

The authors [22] considered a memoryless high power amplifier PAPR-efficiency model, and the auxiliary channel model under low complexity symbol-bysymbol receiver is also derived. Moreover, the achievable spectral efficiency (ASE) which can be viewed as the low bound for any modulation and coding schemes is taken as a figure of merit throughout this paper. By jointly optimizing the time-frequency spacing and clipping ratio to maximization the ASE under the given shaping pulse and modulation format, we show that the ASE of clipping-based MFTN substantially outperforms conventional Nyquist signaling schemes.

III. Proposed INM Model

In this proposed model decision variable, objective function single channel per link constraint, coordinated interference constraint, near-hidden and information asymmetry interference constraints are explained which are given in details. The proposed model that is information asymmetry and near hidden minimisation (INM) Model consist of Near-Hidden and information asymmetry interference constraint. The objectives of this proposed model to improve the channel assignment mechanism. INM model minimise interference and improve the overall capacity of a network.

A. Decision Variable

In this proposed model the decision variable allots an IEEE 802.11b channel cj to a link ei. In this optimisation model, the following decision will be used. This model state that if on any channel cj the directed link ei is active then it is equal to 1 otherwise 0. Such type of decision variable is represented as a binary variable [18].

$$x(e_i.c_j) = \begin{cases} 1 \; directed\ link\ e_i\ operates\ channel\ c_j \\ \quad 0 \qquad\qquad otherwise \end{cases} \tag{1}$$

B. Objective Function

The objective is to maximise the MRMC-WMN capacity through this proposed model. To achieve the objectives, all the constraints must be considered. Equation (2) shows the objective of the proposed model. For the purpose of the objective function, all the objective are used, and all the constraints must be measured. All the link flows fulfilled over all the links and channels are added.

$$max \sum_{Q_i \epsilon E} \sum_{c_i \epsilon M} x(e_i, c_j).\lambda(e_i).f(e_i) \tag{2}$$

C. Constraints

Constraints are the limitations of an optimization model, and they display the unwanted results. Below is the set of constraints consist of this proposed channel assignment model.

1) Single Channel per Link Constraint: In this model, the constraint will make sure that each link in the set Q must be allotted a single channel. This constraint states that if is, a link and cj is a channel assign to a link then the summation of all channels evaluates to 1.

$$\sum_{c_j \epsilon M} x(e_i, c_j) = 1 \qquad \forall e_i \epsilon Q, c_j \epsilon M \tag{3}$$

2) Coordinated Interference Constraint: In this constraint, if more than one coordinated links operating on the identical frequency channel it does not form intense interference and network performance is not affected or minimised. Capacity is shared among all the interfering links. The following constraints will display in this optimisation model that on which method the capacity is distributed when more than on coordinated links are working on same channel (Shah et al., 2013).

$$\begin{aligned} x(e_i, c_j).\lambda(e_i).f(e_i) + \sum e_k \epsilon co(e_i).x(e_k, c_j). \\ \lambda(e_k).f(e_k) \leq Cc_j \qquad \forall e_i \epsilon Q, \quad c_j \epsilon M \end{aligned} \tag{4}$$

Fig. 4 consist of coordinated links. The solid line circle shows the transmission range and the upper dotted line circle display the carrier sensing range. There are three links represented in a figure. The source node A, C and E are inside the carrier sensing range of each other. Moreover, these three links share the capacity of a same frequency channel Cj.

3) Near hidden interference constraint: The second constraint in the optimisation model is near hidden. NH links do not work on a fully overlapping channel. Equation (5) shows that if there are more than one near hidden interfering links, then only one channel cj will be assigned to them. Two

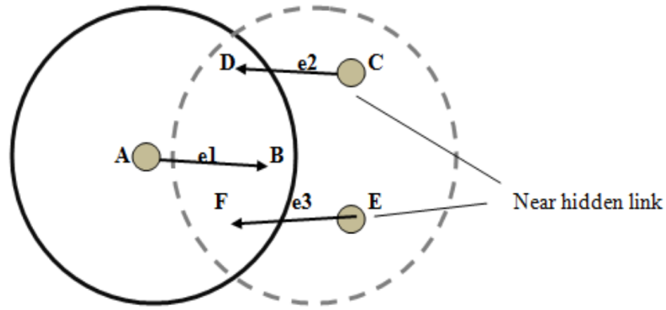


Fig. 4. Coordinated interfering links.

different links ei and ek working on the same channel cj. Then in interfering range, only one of them will be active.

$$x(e_i, c_j)max \sum_{Q_i \epsilon E} \sum_{c_i \epsilon M} x(e_i, c_j).\lambda(e_i).f(e_i) \tag{5}$$

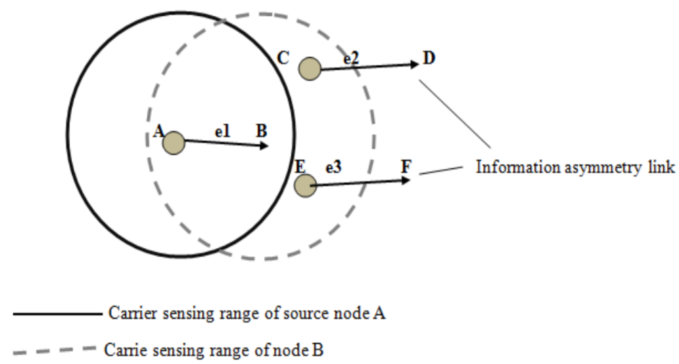


Fig. 5. Near hidden links.

Fig. 5 explains that e2 and e3 are NH links. The receiver node D and F are inside the carrier sensing of source A which cause near hidden interference. The objective of the model is to minimize this interference by assigning certain channel ei which cannot assign to e2 and e3.

4) Information asymmetry interference Constraint: Equation (6) shows that if there is more than one information asymmetry (IA) interfering links, then only one channel will be assigned to them. Information asymmetry links do not operate on a fully overlapping or common channel. Two different links ei and ek are operational on the identical channel cj. Then in interfering range, only one of them will be active.

$$\begin{aligned} x(e_i, c_j) + \sum e_k \epsilon IA(e_i) x(e_k c_j) \leq 1 \\ \forall e_i \epsilon Q, \quad \forall c_j \epsilon M, ek \epsilon Q \end{aligned} \tag{6}$$

Fig. 6 shows that e2 and e3 are Information asymmetry (IA) links. The links e2 and e3 are far away from the receiving range of source node of e1. However, in the range of destination node B of link e1 and information asymmetry interference may occur.

5) Near hidden and Information asymmetry interference constraint: This equation shows that if there are multiple near hidden, information asymmetry (NHIA) interfering links

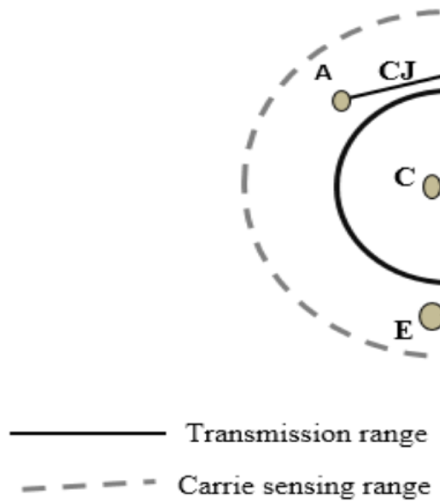


Fig. 6. Information asymmetry links.

then they are combined into one set NHIA. Only one channel will be assigned to them which means that NH and IA links do not work on a fully overlapping or same channel. In these constraints, two different links ek belongs to the set of interfering link of ei and they will not operate on the identical channel. Then in interfering range only a single link are active on channel cj.

$$x(e_i, c_j) + \sum e_k \epsilon NHIA(e_i) x(e_k c_j) \leq 1 \quad \forall e_i \epsilon Q, \quad \forall c_j \epsilon M, ek\epsilon Q \tag{7}$$

IV. Result and Discussion

In this chapter, a detailed discussion is made on the results taken from this research study. This chapter consists of three main sections. In first part different MRMC, wireless mesh topologies were constructed with the help of Matrix Laboratory (MATLAB). Then "A Mathematical programming language"' (AMPL) is used for getting near-optimal channel assignment strategy. In the third section, OPNET simulations are used for result verification and analysis.

A. Simulation Parameters

Before simulations, some research parameters have been considered that are present in Table I. The radio technology that is used in this research is 802.11b where frequency channels are 11. The total number of topologies which are used in this research is 4, and every topology contains 30 mesh nodes. Every node contains three numbers of radios. The transmission capacity of a link is 11mbps. Each node has a transmission range which is 10 meters. The carrier sensing range of each node is 2.6*10 meters. The flow demand is changed from 50 to 500 Packets/sec. Poisson traffic generator is used for traffic generation, and that is flow. For all network topologies, the total simulation time is 3 simulation minutes.

B. MATLAB Topology Construction

In MATLAB 4 different topologies of MRMC are generated. Each topology consists of 30 nodes. In these four topologies, the interference effect of IA and NH is to be

TABLE I. Parameters used During Simulation Results

Parameter	Value
Radio Technology	IEEES02.11b
Numbers of Topology	4
Number of Nodes	30
Radios Per Node	3
Transmission Capacity	11 Mbps
Transmission Range	10 meters
Carrier Sensing Range	2.6 * 10 meters
Simulation Time	3 minutes
Terrain area	80 m x 80 m

checked. Every node has a transmission range and carrier sensing range which are 10 meters and 26 meters, respectively. In transmission, range node can send and receive information successfully. While in sensing range nodes can only sense the other nodes. Interference may be generated when more than one node tries to communicate on the same channel within carrier-sensing range of each other.

In coordinate system 80 by 80-meter area is taken and construct four different topologies. Then find out information asymmetry (IA) and near-hidden (NH) links in these topologies.

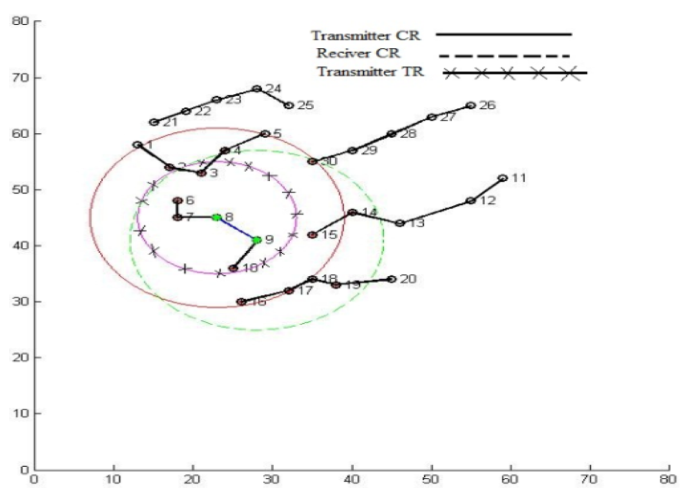


Fig. 7. MRMC-WMN MATLAB generated topology 1.

Fig. 7, 8, 9 and 10 consist of WMN topologies which are constructed in MATLAB. In these four figures, the x and y-coordinates are 80 x 80 meters. These four figures consist of three types of circles cross line, solid line and dotted line circle. Which represent the transmission range of source node, carrier sensing range of source node, and carrier sensing range of receiver node, respectively.

Fig. 7 consist of WMN topology which is created in MATLAB. In Fig. 7 (8, 9) link have the following coordinated links i.e. (2, 3), (3, 4), (4, 5), (6, 7), (7, 8), (9, 10), (16, 17), (17, 18). While a set of information asymmetry (IA) links is (19, 20) and set of near hidden (NH) links is (14, 15). Those links which are present in the carrier sensing range of a source node is said to be coordinated links. Those links which are far away from the sensing range of source node but present in the receiver sensing range is said to be information asymmetry links, i.e. (19, 20). Near hidden are those links in which source nodes are far away from the sensing range, but its receiver is present in the carrier sensing range, i.e. (14, 15).

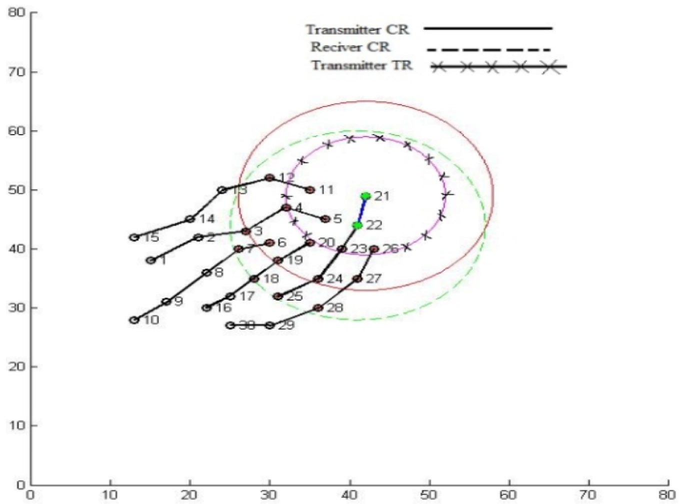


Fig. 8. MRMC-WMN MATLAB generated topology 2.

In Fig. 8 (21, 22) link have the following coordinated links. i.e. (4,5), (6,7), (11, 12), (12, 13), (19, 20), (22, 23), (23, 24), (24, 25), (26, 27), (27, 28) while set of IA links is (7, 8), (28, 29) and the set of NH consist of (3, 4), (18, 19). Similarly in Fig. 9 (6,7) links have the following coordinated links. i.e. (2, 3), (3, 4), (4, 5), (7, 8), (8, 9), (11, 12), (12,13), (13,14), (22, 23), (23, 24), (24, 25), (26, 27), (27, 28), (28, 29) while the set of information asymmetry (IA) links are (9, 10), (29, 30) and the set of near hidden links are (1, 2), (19, 20), (21, 22).

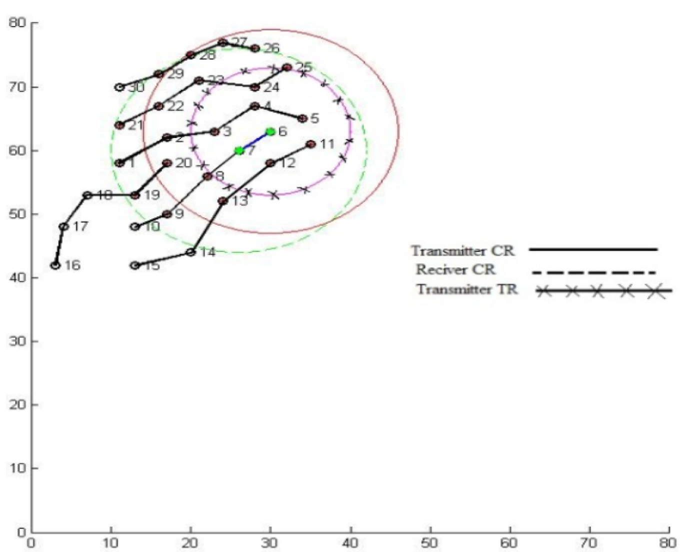


Fig. 9. MRMC-WMN MATLAB generated topology 3.

In Fig. 10 (11, 12) links have the following coordinated links i.e. (12, 13), (13, 14), (19, 20), (26, 27), (27, 28). While set of information asymmetry (IA) links are (14, 15), (28, 29) and the set of near hidden (NH) links is (18, 19).

The Table II consists of coordinated links, information asymmetry and near hidden links. In this table, the first column consists of considered links. The second column shows information asymmetry (IA) links and the third column consist of near hidden (NH). From Fig. 9 these IA and NH links are taken. For example consider link (1, 2) have the following information asymmetry links, i.e. (4, 5), (24, 25). And near hidden (NH) links consists of (6, 7), (12, 13), (28, 29). Due to the effect, IA and NH interfering links the flow of (1, 2) link will be reduced.

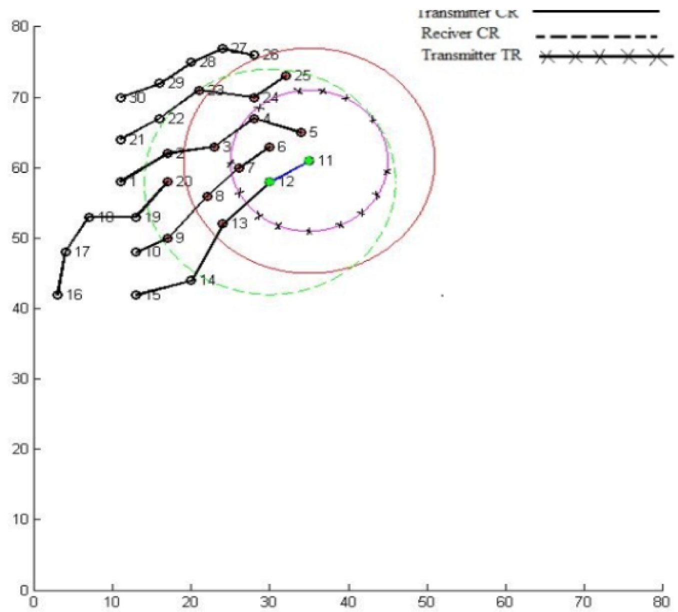


Fig. 10. MRMC-WMN MATLAB generated topology 4.

Table II consist of 30 nodes in which IA and NH nodes links of a consider link are shown.

C. AMPL Results

The result obtained from MATLAB should be given to a mathematical programming language (AMPL) solver, and that gives us channel assignment information. In MATLAB software four different topologies are generated through which all the coordinated, information asymmetry and near hidden links have been finding out. The obtaining results given to AMPL which perform the linear programming (LP) proposed a model. To assign channels to different network nodes is performed by AMPL gurobi solver. The channel allocation is near optimal, and the flow demand is changed from 50-500 packets per seconds on every link.

Fig. 11 is one of the topologies that has been taken. This topology consists of 30 nodes. The above figure consists of snapshots. During execution of AMPL software, these snapshots are taken. The binary variables are used, 0 is used for no channel, and one is used for the presence of channels. Same traffic load is assigned to every link through AMPL, and the flow demand is changed from 50-500 packets per second. Due to channel sharing, some channels carry more data traffic than other channels. In the proposed model, the objective functions are used to achieve an aggregate capacity of all links. In IEEE 802.11b for channel assignment, three non-overlapping channels (1, 6, and 11) are used. For a solution, the gurobi solver is preferred. For solving linear programming models, the gurobi solver is used. For the actual implementation of the proposed model, the “mm.txt” file is used. For output of the proposed model the “solve” command issued. The “display”

TABLE II. SET OF NEAR HIDDEN AND INFORMATION ASYMMETRY LINKS OF 30 NODES WMN TOPOLOGY

Links	Information Asymmetry Links (IA)	Near Hidden Links
(1,2)	(4,5)(24,25)	(6,7)(12,13)(28,29)
(2,3)	NIL	(11,12)(27,28)
(3,4)	NIL	NIL
(4,5)	NIL	NIL
(6,7)	(9,10)(29,30)	(1,2)(19,20)(21,22)
(7,8)	(14,15)	(18,19)
(8,9)	NIL	(17,18)
(9,10)	NIL	(16,17)
(11,12)	(9,10)	(2,3)(23,24)
(12,13)	NIL	(1,2)(18,19)
(13,14)	NIL	(18,19)
(14,15)	NIL	(17,18)
(16,17)	(1,2)	(9,10)
(17,18)	(2,3)(21,22)	(8,9)(14,15)
(18,19)	(3,4)(22,23)	(7,8)(13,14)
(19,20)	(4,5)(23,24)	(6,7)(12,13)
(21,22)	(4,5)(24,25)	(6,7)(27,28)
(22,23)	(12,13)	NIL
(23,24)	NIL	(11,12)
(24,25)	NIL	NIL
(26,27)	NIL	NIL
(27,28)	NIL	(2,3)(21,22)
(28,29)	(7,8)	(1,2)
(29,30)	NIL	NIL

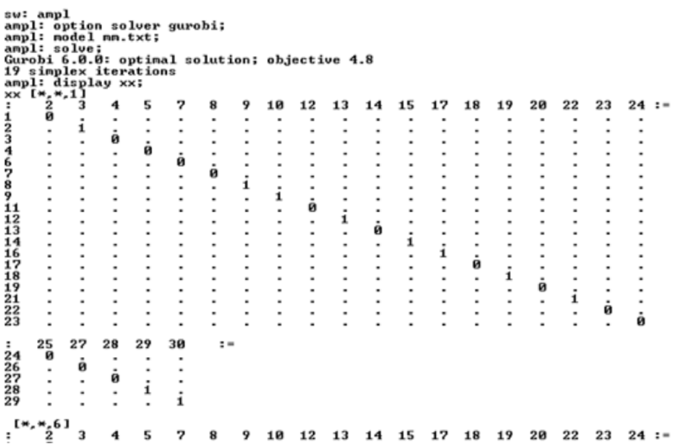


Fig. 11. 11AMPL channel assignment results.

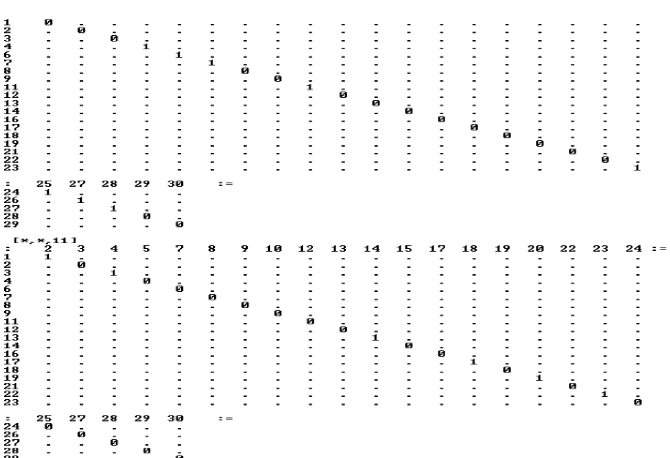


Fig. 12. 11AMPL channel assignment results.

command is used to show the channel assignment results which are 1 or 0.

TABLE III. PROPOSED MODEL CHANNELS ASSIGNMENT

Node	Channel Assigned
(1,2)	11
(2,3)	1
(3,4)	11
(4,5)	6
(5,6)	6
(6,7)	6
(7,8)	1
(8,9)	1
(9,10)	6
(10,11)	1
(11,12)	11
(12,13)	1
(13,14)	1
(14,15)	6
(15,16)	1
(16,17)	11
(17,18)	1
(19,20)	11
(20,21)	1
(21,22)	11
(22,23)	1
(23,24)	11
(24,25)	6
(25,26)	6
(26,27)	6
(27,28)	6
(28,29)	1
(29,30)	1

The gurobi solver is selected for a solution. This solver is used for linear programming models. The implementation of the proposed model is stored in “mm.txt” file. Solve command gives output or objective of the proposed model. The channel assignment results that are 0 or 1is assigned by “display” command. Binary variables that show links or channel binding is represented by “xx”. For example xx [*,*,*] shows xx [sender link, receiving link, channel assigned]. For xx [2, 3, 11] show channel 11 is assigned to a link (2, 3).

Table III is made on the bases of Fig. 11. The table consists of node value and the assigned channel value. The table consists of two columns, the first column shows the number of nodes and the second column represent a number of the assigned channel to each link, and these are non-overlapping channels, i.e. (1, 6, and 11). These channels are assigned randomly to the nodes through optimal channel assignment strategy.

D. Simulation in OPNET

In this thesis, all the scenario are designed and created by OPNET modeller. Many choices are given by OPNET modeller, i.e. selecting area and placement of nodes in that area. For example, in OPNET modeller a scenario is built which consist of 30 nodes and area selected is 80*80 meter as indicated in Fig. 12. In this research, four different scenarios are discussed and their information asymmetry (IA) and near hidden (NH) results are compared.

In Fig. 12 shows simulation scenario for 30 nodes. In which every node is linked to its adjacent node and these nodes have no mobility and does not change its positions. Each node has its particular properties. For example, channel assigning to a radio nodes, power level, and packet size and internet protocol (IP) address. In this topology, coordinates are identical to MATLAB topology.

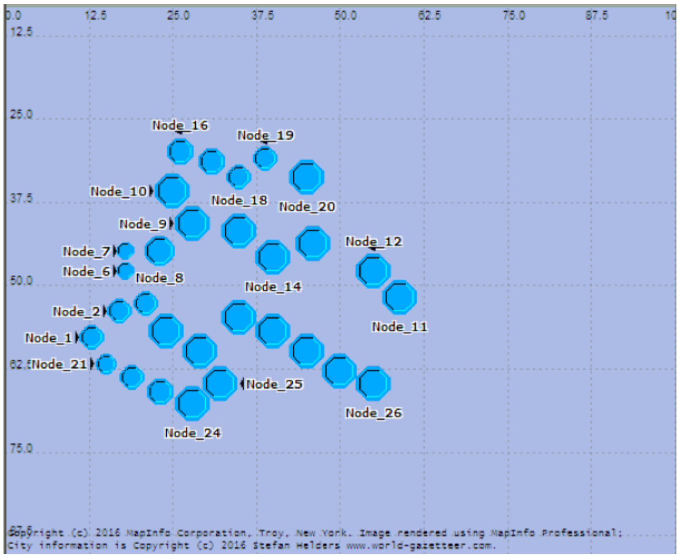


Fig. 13. OPNET 30 node topology of proposed (INM) models.

TABLE IV. SIMULATION RESULTS OF IA

Flow demand	Total capacity Of a network	Network Capacity(30N)	Network Capacity(30N)	Network Capacity(30N)	Network Capacity(30N)	Average network capacity
Packet/sec	Packet/sec	Packet/sec	Packet/sec	Packet/sec	Packet/sec	Packet/sec
50	1250	1133	1008	1145	1168	1113.5
100	2500	2161	1970	2136	2183	2132.5
150	3750	2980	2503	2616	2726	27068.25
200	5000	3302	2587	2664	2962	2878.75
250	6250	3530	2588	2667	3146	2982
300	7500	3653	2589	2668	3261	3042.75
350	8750	3848	2590	2669	3306	3103
400	10000	3953	2592	2670	3338	3138.25
450	11250	4121	2595	2673	3339	3182
500	12500	4240	2598	2675	3347	3215

E. Simulations Results in OPNET

Table IV displays the information asymmetry (IA) summary results for each topology. The data in each column consists the results of IA. The Table IV consists of seven columns. The first columns consist of flow demand that varies from 50 to 500 packet/sec for every topology. The second column consists of total network capacity. The third, fourth, fifth and six columns show the results of four different topologies. Each topology consists of 30 nodes. The last column shows the average IA results for all the four topologies.

Table V displays the near-hidden (NH) summary results for each topology. The data in each column consist the results of NH. The table 4.5 consists of seven columns. The first columns consist of flow demand that varies from 50 to 500 packet/sec for every topology. The second consist of the total capacity of a network. The third, fourth, fifth and six columns show the results of four different topologies. Each topology consists of 30 nodes. The last column shows the average NH results for all the four topologies.

Fig. 13 presents the comparison of 30 nodes for information asymmetry (IA) and near hidden (NH) topology. The grey line graph represents the network capacity of near hidden (NH). Blackline graph represents the network capacity of information asymmetry (IA). In Fig. 13 the horizontal line values represent the flow demand in packet/sec, which are varying from 50 to 500 packet/sec. Vertical line values shown on the graph represent for both IA and NH topologies. Whenever the number of flow demand changes the network capacity of both IA and NH is change.

TABLE V. SIMULATION RESULTS OF NH

Flow demand	Total capacity Of a network	Network Capacity(30N)	Network Capacity(30N)	Network Capacity(30N)	Network Capacity(30N)	Average network capacity
Packet/sec	Packet/sec	Packet/sec	Packet/sec	Packet/sec	Packet/sec	Packet/sec
50	1250	1180	1036	533	1124	968.25
100	2500	2234	2129	1087	2251	1925.25
150	3750	2916	2600	1480	2779	2443.75
200	5000	3270	2610	1673	2969	2630.5
250	6250	3509	2613	1927	3082	2782.75
300	7500	3605	2615	2058	3150	2857
350	8750	3758	2617	2075	3240	2922.5
400	10000	3839	2618	2131	3316	2976
450	11250	3951	2620	2144	3398	3028.25
500	12500	4082	2621	2257	3475	3108.75

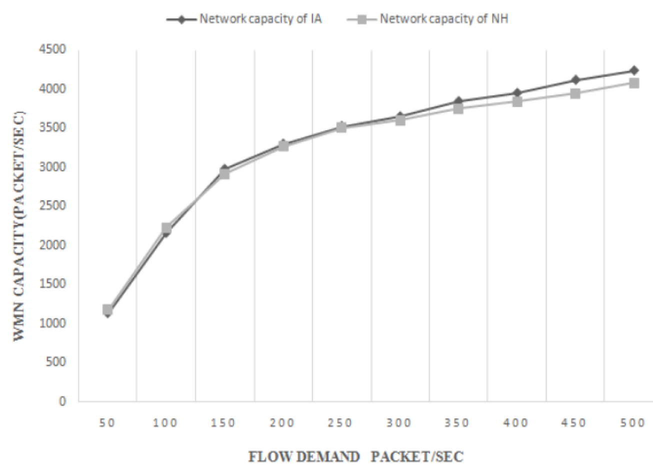


Fig. 14. Capacity comparison of IA and NH in WMN Topology I.

Fig. 14 shows the comparison of 30 nodes for the information asymmetry (IA) and near hidden (NH) topologies. In Fig. 14 all the values are taken from table 4 and 5. The grey line represents the near-hidden (NH) network capacity graph, and the black line represents the information asymmetry (IA) network capacity graph. The x-axis represents the flow demand in packet/sec. The y-axis represents wireless mesh network capacity for both IA and NH topologies. The behaviour of a graph shows that the flow demand change the network capacity of both IA and NH are also changed.

Fig. 15 shows the compression of Information asymmetry (IA) and near hidden (NH) topologies. These results have been taken from wireless mesh topology which consists of 30 nodes. For Fig. 15 the entire values have been selected from Tables IV and V. The lower grey line indicates the result of NH and the upper black line indicate the result of IA. On the x-axis, the value represents the flow demand, and on the y-axis, the values represent the network capacity of IA and NH. The graph shows that the flow demand changed the network capacity also changed for both IA and NH topologies.

Fig. 16 shows the comparison of information asymmetry (IA) and near hidden (NH) topologies. The topology consists

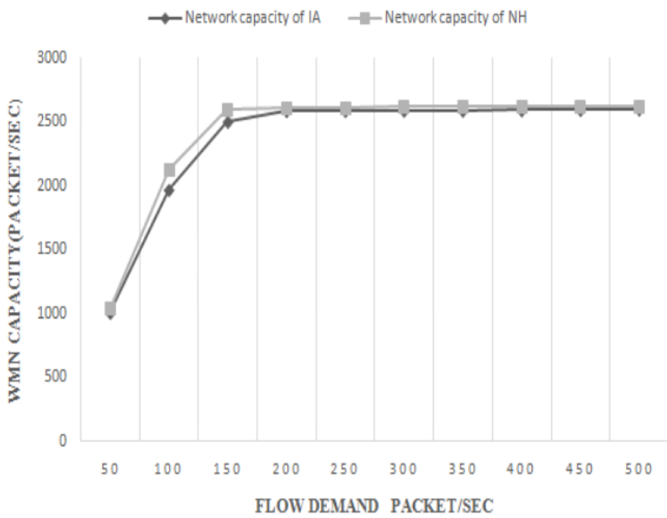


Fig. 15. Capacity comparison of IA and NH in WMN Topology 2.

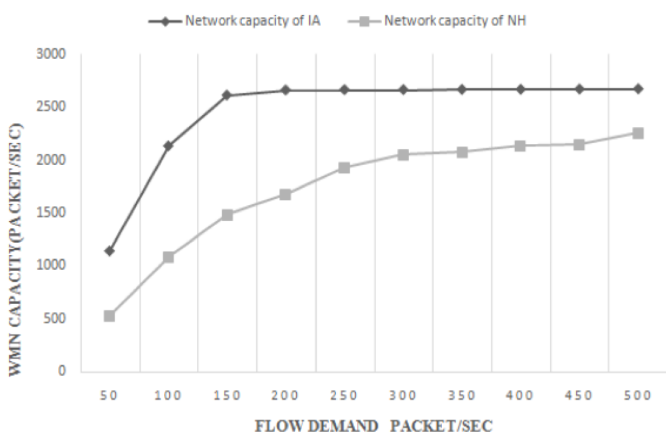


Fig. 16. Capacity comparison of IA and NH in WMN Topology 3.

of 30 nodes. In Fig. 16 the graph consists two lines. The black line represents the IA network capacity, and the grey line indicates the NH network capacity. On x-axis, the values represent the flow demand. Moreover, the flow demand varies from 50 to 500 packet/sec. On y-axis, the value shows the network capacity. The behaviour of graph represents that the flow demand varies the network capacity of IA and NH also change.

Fig. 17 is an average capacity comparison of information asymmetry (IA) and near hidden (NH) topologies. This figure graphically indicates NH and information asymmetry topologies by two line graphs the grey and black line. The grey line indicates the average network capacity of NH and the black line indicates the average capacity of IA. In Fig. 17 x-axis values represent the average flow demand for four different topologies. The flow demand value varies from 50 to 500 packet/sec, while the y-axis values represent the average network capacity for both IA and NH topologies. Fig. 17 shows that NH interference is more concerning IA interference.

Table VI displays average results for both information asymmetry (IA) and near hidden (NH) topology. Table VI consists of seven columns. The first columns consist of flow

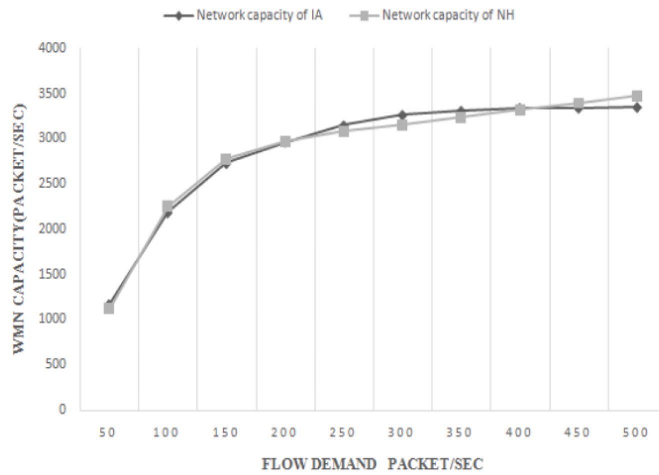


Fig. 17. Capacity comparison of IA and NH in WMN Topology 4.

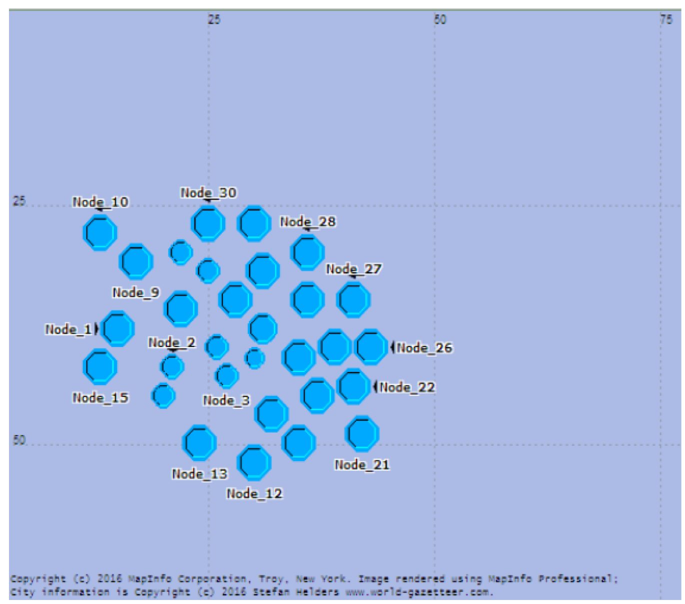


Fig. 18. Average capacity comparison of IA and NH.

demand that varies from 50 to 500 packet/sec for every topology. The second column shows the network flow demand. The third column shows the fulfil network capacity of IA and the fourth one consist of a percentage of IA. The fifth column shows the NH fulfil network capacity and the percentage of NH is shown in six columns, and the seven columns consist of the difference of IA and NH results.

F. INM (Proposed) Model Results

In this section results are taken from the proposed model (INM) are presented. The results are compared with the existing model (OCAM). For comparison, a new WMN topology is constructed in OPNET that is given in Fig. 18.

Fig. 19 shows a capacity comparison of INM (information asymmetry, near-hidden minimization) proposed model topology and OCAM (optimized channel Assignment) existing model. This figure graphically indicates INM and OCAM topologies by two line graphs the grey and black line. The grey

TABLE VI. AVERAGE RESULT OF IA AND NH

Packet/sec	Network Flow demand	IA fulfill network capacity	Percentage of IA	NH fulfill network capacity	Percentage of NH	Difference between IA and NH
50	1250	1113.5	89.08%	968.25	77.46%	11.62%
100	2500	2112.5	84.5%	1925.25	77.01%	7.49%
150	3750	2706.25	72.16%	2443.75	65.16%	7%
200	5000	2878.75	57.57%	2630.5	52.61%	4.96%
250	6250	2982	47.71%	2782.75	44.52%	3.19%
300	7500	3042.75	40.57%	2857	38.09%	2.48%
350	8750	3103	35.46%	2922.5	33.4%	2.06%
400	10000	3138.25	31.38%	2976	29.76%	1.62%
450	11250	3182	28.28%	3028.25	26.91%	1.37%
500	12500	3215	25.72%	3108.75	24.87%	0.85%

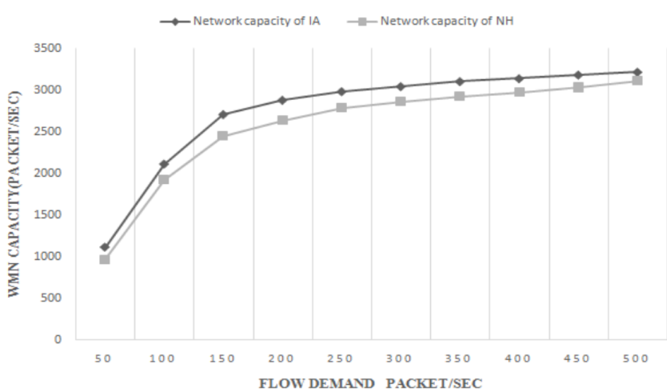


Fig. 19. OPNET 30 node topology for comparison of INM and OCAM.

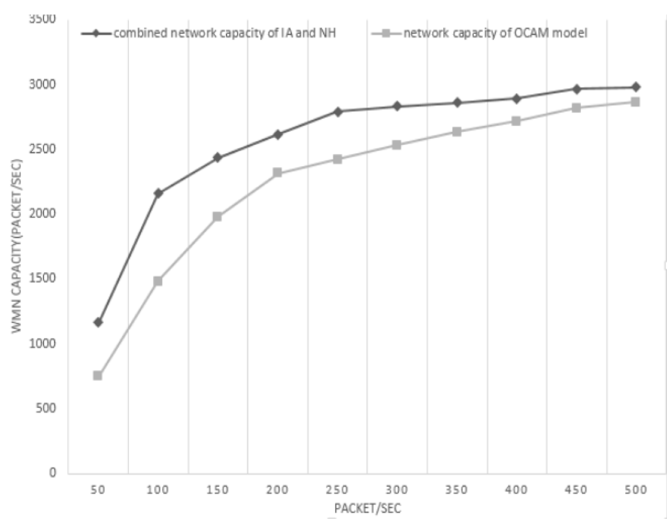


Fig. 20. Comparison of INM and OCAM model of 30 nodes WMN.

line indicates the network capacity of OCAM (existing model) and the black line indicates the capacity of INM (proposed model). X-axis values represent the flow demand. The flow demand value varies from 50 to 500 packet/second. While the y-axis values represent the network capacity for both NHIA and OCAM topologies.

Table VII shows the achieved capacity of both INM and OCAM model. The results are given in percentage format to shows the difference in results. The average is taken from all the flow demands that show that INM performs 7% better that OCAM model. The results are also given in Fig. 19 which

TABLE VII. RESULTS OF INM AND OCAM MODEL FOR WMN TOPOLOGY 1

Packet/sec	Total capacity Of a network	Network Capacity of INM Proposed model	Network Capacity of OCAM (Existing) model	Percentage Capacity of INM proposed model	Percentage Capacity of OCAM (Existing) model	Percentage improvement of INM over OCAM
50	1250	1165	752	93.2%	75.2%	18%
100	2500	2159	1486	86.36%	74.3%	12.06%
150	3750	2438	1982	65.01%	51.41%	13.6%
200	5000	2616	2315	52.32%	46.3%	6.02%
250	6250	2789	2423	44.62%	38.76%	5.86%
300	7500	2831	2535	37.74%	33.8%	3.94%
350	8750	2860	2636	32.68%	30.12%	2.56%
400	10000	2893	2715	28.93%	27.15%	1.78%
450	11250	2964	2818	26.34%	25.04%	1.3%
500	12500	2980	2865	23.84%	22.27%	1.55%

shows that after 350 packets per second which is a high data rate the results are consistent. It shows that the proposed model is more consistent in case of high data rate demand.

V. CONCLUSION

This research has been successfully done to compare the interference effect of information asymmetry (IA) and near hidden (NH) wireless mesh network over OCAM (existing) model. Then the combined interference effect of IA and NH has been presented on the INM (proposed) model. Moreover, then compare the INM (proposed) model and OCAM (existing) model.

The OPNET results show that the near-hidden and information asymmetry (INM) gives better performance than OCAM model. Both of the models consist of 30 nodes. Where the Near Hidden (NH) and information asymmetry (IA) interference is high between the nodes, the proposed optimisation model gives improved results regarding an increase in network capacity. The results are given in percentage format to shows the difference in results. The average is taken from all the flow demands that show that INM performs 7% better that OCAM model. The results are also given in Fig. 20 which shows that after 350 packets per second which is a high data rate the results are consistent. It shows that the proposed model is more consistent in case of high data rate demand.

A. Future Work

In future work the research work can be expanded to large wireless mesh networks, i.e. the number of wireless mesh nodes may increase to hundreds. The performance can be checked over 802.11g and 802.11n radios technologies instead of 802.11b. In this research far-hidden terminals are not considered so in the future far-hidden interference can also be considered.

REFERENCES

[1] Ajayi, A, O., A, A, Adigun., And I, W, Oladimeji. (2015). A Review of Routing Protocols for Practical Rural Wireless Mesh Networks (WMNs).International Journal of ComputerApplications, 114(16).

[2] Angelakis, V., S, Papadakis.,V,Siris., and A,Traganitis.(2008). Adjacent channel interference in 802.11a: Modeling and testbed validation. In Proc. of IEEE Radio and Wireless Symposium.IEEE, no. 2, 591594.

[3] Bansal, D., S,Sofat., and A,K,Gankotiya. (2010). Selfish MAC Misbehavior Detection in Wireless Mesh Network.International conference on advances in Computer engineering.IEEE computer society,doi:10.1109/ACE.2010.91. 130-133

[4] Bukkapatanam,V., A,A,Franklin., and C,S,R,Murthy.(2009).Using PartIAlly Overlapped Channels for End-to-End Flow Allocation and Channel Assignment in Wireless Mesh Networks, in Proc. of IEEE ICC, 2009.10.1109/ICC.2009.5199559.1 - 6.

[5] Fu,y.(2013). Research of Wireless MESH Network Performance based on NSTUns. Computational and Information Sciences (ICCIS), 2013 Fifth International Conference on 1EEE.10.1109/ICCIS.2013.354. 1339 - 1342.

[6] Garetto,M., T,Salonidis., and E,W,Knightly. (2008). Modeling per-flow through-put and capturing starvation in CSMA multi-hop wireless networks, IEEE/ACM Transactions on Networking (TON), 16(4): 864-877.

[7] Huang,Y., X,Yang., S,Yang., W,Yu., and X,Fu.(2011). A Cross-Layer Approach Handling Link Asymmetry for Wireless Mesh Access Networks. IEEE transactions on vehicular technology, 60(3):10451058.

[8] Iqbal,F., M,Y,Javed., and A,Naveed.(2014).Interference-aware multipath routing in wireless mesh network. EURASIP Journal on Wireless Communications and Networking, doi:10.1186/1687-1499-2014-140

[9] Jun, j., and M, L,Sichitiu. (2003).The nominal capacity of wireless mesh networks, IEEE Wireless Communications 10 (5) pp. 8, 2003.

[10] Kaur, N., and J, S,Saini. (2013). Performance ENHancement of 802.11 based Wireless Mesh Network by usingMulti- Radio Multi-Channel. International Conference on Green Computing, Communication and Conservation of Energy -(ICGCE), 10.1109/ICGCE.2013.6823402.71-76.

[11] Kyasanur, P., N, H,Vaidya.(2005). Routing and interface assignment in multi-channel multi-interface wireless networks. Wireless Communications and Networking Conference, IEEE .4: 20512056.

[12] Mahmud,S,A., S,Khan., S,Khan., and H,Al-Raweshidy.(2006). A Comparison of MANETs and WMNs: CommercIAl Feasibility of Community Wireless Networks and MANETs. In Proceedings of the First International Conference on Access Networks (Acess-Nets'06),10.1145/1189355.1189373. 18-24.

[13] Mishra,A., V,Shrivastava., S, Banerjee., and W,Arbaugh. (2006). partIAllyoverlappedchannels not considered harmful .In ACM Sig Metrics, 6:6374.

[14] Nachtigall,J., A,Zubow., and J,P,Redlich.(2008).the impact of adjacent channel interference in multi-radio systems using ieee 802.11.in Proc. IWCMC '08, Crete Island, Greece, Aug. 2008.10.1109/IWCMC .151 874-881.

[15] Naz, K., (2016). Minimization of near hidden interferance in multi radio multi-channel wireless mesh network. (Unpublished doctoral dissertation). The University of Agriculture, KPK, Pakistan.

[16] Naveed,A., S,S,KaNHere.(2008).Channel Assignment in Multi-Radio Multi-Channel Wireless Mesh Networks, School of Computer Science and Engineering, University of South Wales, October 2008.

[17] Raniwala, A., K,Gopalan., and T,Chiueh.(2010).Centralized Channel Assignment and Routing Algorithms for Multi-Channel Wireless Mesh Networks. In Mobile Computing and Communications Review, 8: 5065.

[18] Shah, S., H, Hussain. and M,Shoaib.(2013). Minimizing non-coordinated interference in multi-radio multi-channel Wireless Mesh Networks (MRMC -WMNs),24-28.

[19] Sombrutzki,R.,A,Zubow., M,Kurth., and J,P,Redlich. (2006). Self-organization in community mesh networks the berlin roofnet in Operator-Assisted (Wireless Mesh) Community Networks, 1st Workshop on, sept.10.1109/WOACN.2006.337188 . 1 -11.

[20] Subramanian, A,P., H, Gupta., S,R,Das., and J,Cao.(2008). Minimum Interference Channel Assignment in Multi radio Wireless Mesh Networks. IEEE transactions on mobile computing, 7(12):1459-1473.

[21] Manikandan, A., and Y,Palanichamy.(2015).Optimized Group Channel Assignment Using Computational Geometry over Wireless Mesh Networks.Hindawi Publishing Corporation,Mobile Information Systems, Volume 2015, Article ID 718079, 18 pages.

[22] Peng, S., Liu, A., Song, L., Memon, I., Wang, H. (2018). Spectral Efficiency Maximization for Deliberate Clipping-Based Multicarrier Faster-Than-Nyquist Signaling. IEEE Access, 6, 13617-13623.

Convex Hybrid Restoration and Segmentation Model for Color Images

Matiullah*, Samiullah Khan†, Noor Badshah‡, Fahim Ullah§ and Ziaullah¶
*‡§Department of Basic Sciences and Islamiat, ¶ Department of Computer System Engineering
University of Engineering and Technology, Peshawar-Pakistan
†Institute of Business Management Sciences, The University of Agriculture Peshawar-Pakistan

***Abstract*—Image restoration and segmentation are important areas in digital image processing and computer vision. In this paper, a new convex hybrid model is proposed for joint restoration and segmentation during the post-processing of colour images. The proposed Convex Hybrid model is compared with the existing state of the art variational models such as Cai model, Chan-Vese Vector-Valued (CV-VV) model and Local Chan-Vese (LCV) model using noises such as Salt & Pepper and Gaussian. Additional four experiments were performed with increasing combination of noises such as Salt & Pepper, Gaussian, Speckle and Poisson to thoroughly verify the performance of Convex Hybrid Model. The results revealed that the Convex Hybrid model comparatively outperformed qualitatively and has successfully removed the noises and segment the required object properly. The Convex Hybrid model used the colour Tele Vision (TV) as a regularizer for denoising of the corrupt image. The Convex Hybrid Model is convex and can get global minima. The PDEs obtained from the minimisation of the Convex Hybrid Model are numerically solved by using explicit scheme.**

***Keywords*—Color images; image restoration; image segmentation; noise**

I. Introduction

Image restoration works to restore an image in digital image processing. Noise in an image is the random variation of intensity or colour in images (unwanted signals), which is usually produced by the sensor or digital camera. Restoration is the process of denoising an image. Many variational models are proposed for restoration of images. Fundamental variational model for denoising of grey level images is proposed by Rudin et al. [10]. Later on, Blomgren et al. [1] proposed a model restoration of a colour image which is based on colour total variation (CTV). The advantage of the TV-based model is the preserving edges during denoising. TV-based model creates the staircase effects. Much higher order's derivatives based models are proposed in [10]–[14] to avoid the staircase effect.

Segmentation is divided into distinct subdomains based on some criterion like homogeneity in intensity, colour and texture. Segmentation of noisy images is an issue in the field of computer vision. Many variational models are developed for segmentation of images to solve the issue of segmentation of noisy images. These models are classified into two categories, i.e. Edge-Based Models and Region-Based Models. Edged based models use edge information for segmentation of images, such as snake model and geodesic active contour model [9]. Snake model and geodesic active contour model use edge information due to which they are not applicable in noisy images. Region-based models use region information for segmentation of images such as Mumford-Shah (MS) model [6]. Chan Vese (CV) model [4] is the special case of MS model. The CV model is not convex, and it might stick to local minima. CV model cannot segment noisy images due to non-convexity. Bresson et al. [2] proposed Fast Global Minimization (FGM) model for the convex formulation of the active contour model for grey level images. This model also does not segment images having heavy noise and is more dependent on the tradeoff parameters [15]–[17].

Image restoration is the first field of image processing technology which improves the quality of an image. X-Cai [3] proposed a Cai model which works jointly for image restoration and segmentation. Cai models restore the degraded image and then segment the restored image. Cai model is a non-convex. Segmentation results of Cai model are not very satisfactory in an image having intensity inhomogeneity.

In this paper, a new Convex Hybrid model is proposed which works jointly for restoration and segmentation of colour images. The Convex Hybrid model is based on colour total variation (CTV) of restoration and global minimization for segmentation [2].

II. Related Variational Models

A brief overview of some image denoising and segmentation models are presented in this section.

A. Rudin, Osher and Fatimi (ROF) Model

ROF model is one of the famous and commonly used model for the restoration and de-noising of corrupted images which was designed by Rudin, Osher and Fatemi (ROF) in 1992. Let $\Omega \subset \mathbb{R}^2$ be a bounded, open, connected set, and $v_0(x, y) : \Omega \longrightarrow \mathbb{R}$ be given corrupted image and $v(x, y) : \Omega \longrightarrow \mathbb{R}$ is the desired clean image. Then the energy functional of ROF model can be written as:

$$E^{ROF}(v) = \gamma \int_\Omega (v_0 - v)^2 dxdy + \int_\Omega |\nabla v| dxdy, \qquad (1)$$

where the regularization parameter is represented by γ and the task of first term in (1) is that the clean image must be close to the observed image. While in (1) the last term is the smoothing or regularization term which is actually l_1 norm. The corresponding Euler-Lagrange equation for ROF model is as:

$$\frac{\partial}{\partial x}\left(\frac{v_x}{\sqrt{v_x^2+v_y^2}}\right)+\frac{\partial}{\partial y}\left(\frac{v_y}{\sqrt{v_x^2+v_y^2}}\right)$$
$$+\gamma(v-v_0)=0 \;\; in \;\; \Omega, \tag{2}$$

with $\frac{\partial v}{\partial n}=0$ on the boundary of $\Omega=\partial\Omega$. For solving (2) the following parabolic equation is considered:

$$\frac{\partial v}{\partial t}=\frac{\partial}{\partial x}\left(\frac{v_x}{\sqrt{v_x^2+v_y^2}}\right)+\frac{\partial}{\partial y}\left(\frac{v_y}{\sqrt{v_x^2+v_y^2}}\right)$$
$$+\gamma(v-v_0)=0 \;\; for \;\; t>0. \tag{3}$$

The solution of (2) can be obtained from the steady state solution of (3). ROF model uses l_1 norm instead of l_2 norm as a result this model perform better than other models which utilizes l_2 norm. The advantage of l_1 norm over l_2 norm is that l_1 norm smooth the image but keeps the boundaries of object sharp, in contrast l_2 norm also smooth the edges of the objects which is one of the main drawback to l_2 norm. Moreover, the results of ROF model is batter than linear smoothing but the main drawback of ROF model is that it create staircase effects. In order to overcome the limitation of ROF model, many other models have been developed [6].

B. Peter Blomgren TV Color Model

As ROF model uses the one dimensional TV norm for the restoration of scalar images. Therefore, to extend this model for the restoration of vector-valued images, Blomgren and Chan proposed a new definition of TV norm for the restoration of color and vector-valued images. Which posses the properties of not penalizing discontinuities (edges) in the image and is rotational invariant [1]. TV-VV model proposed the following energy functional:

$$\min_{v\in BV(\Omega)} ||v||_{TV-VV}+\frac{\gamma}{2}||v-v_0||_2^2, \tag{4}$$

where γ is Lagrange multiplier, $|v|_2^2=\sum_{j=1}^{n}||v^j||_2^2$ and $||.||_{TV-VV}=TV_{m,n}(v)$ and is defined as:

$$TV_{m,n}(v)=\sqrt{\sum_{j=1}^{n}[TV_{m,1}(v^j)]^2}. \tag{5}$$

Now clearly for $n=1$ (14) turn out to TV norm for scalar valued function. Now the corresponding Euler-Lagrange equation for the minimization problem (13) with $||.||_{TV-VV}=TV_{m,n}(v)$ can be obtained as:

$$\frac{TV_{m,1}(v_j)}{TV_{m,n}(v)}\nabla.\left(\frac{\nabla v_j}{||\nabla v_j||}\right)-\gamma(v_j-v_{0,j})=0. \tag{6}$$

The solution can be computed by explicit time marching scheme by considering the following PDE:

$$\frac{\partial v_j}{\partial t}=\frac{TV_{m,1}(v_j)}{TV_{m,n}(v)}\nabla.\left(\frac{\nabla v_j}{\sqrt{\gamma_1+||\nabla v_j||^2}}\right)-\gamma(v_j-v_{0,j}), \tag{7}$$

where γ_1 is used to avoid division by zero and is regarded as small regularization parameter.

C. Active Contour without Edges (CV) Model

CV model was introduced by Chan and Vese, which is actually a special case of the Mumford-Shah (MS) energy functional [6] when restricted to only two phases [4], [7]. CV model tries to demonstrate the minimization energy functional as:

$$E^{CV}(e_1,e_2,\Gamma)=\nu.(Length(\Gamma))$$
$$+\gamma_1\int_{inside(\Gamma)}|v_0(x,y)-e_1|^2dxdy$$
$$+\gamma_2\int_{outside(\Gamma)}|v_0(x,y)-e_2|^2dxdy, \tag{8}$$

where the segmented and smooth curve is represented by Γ. ν, γ_1 and γ_2 are tuning positive parameters. While e_1 and e_2 denotes the mean intensities of v_0 inside and outside of Γ respectively. CV model can be written in level set formulation as follow:

$$E^{CV}(e_1,e_2,\varphi)=\nu\int_\Omega\delta(\varphi)|\nabla\varphi|dxdy$$
$$+\gamma_1\int_\Omega|v_0-e_1|^2H(\varphi)dxdy$$
$$+\gamma_2\int_\Omega|v_0-e_2|^2(1-H(\varphi))dxdy, \tag{9}$$

where the level set function is represented by φ, $H(\varphi)$ and $\delta(\varphi)$ are the Heaviside and dirac delta functions, respectively. CV model used the regularized form of Heaviside and dirac delta functions instead of the original ones because Heaviside function is not differentiable at the origin. The regularized versions of Heaviside and dirac delta functions are denoted by $H_\epsilon(\varphi)$ and $\delta_\epsilon(\varphi)$, respectively and is defined So (9) is written as:

$$E^{CV}(\varphi,e_1,e_2)=\nu\int_\Omega\delta_\epsilon(\varphi)|\nabla\varphi|dxdy$$
$$+\gamma_1\int_\Omega|v_0-e_1|^2H_\epsilon(\varphi)dxdy$$
$$+\gamma_2\int_\Omega|v_0-e_2|^2(1-H_\epsilon(\varphi))dxdy. \tag{10}$$

Minimizing (10) with respect to e_1 and e_2 we have:

$$e_1=\frac{\int_\Omega v_0H_\epsilon(\varphi)dxdy}{\int_\Omega H_\epsilon(\varphi)dxdy},$$
$$e_2=\frac{\int_\Omega v_0(1-H_\epsilon(\varphi))dxdy}{\int_\Omega(1-H_\epsilon(\varphi))dxdy}, \tag{11}$$

whereas, Euler-Lagrange equation for model (10) can be obtained by its minimization and leads to the following PDE:

$$\begin{cases} \delta_\epsilon\Big[\nu \text{div}\Big(\frac{\nabla\varphi}{|\nabla\varphi|}\Big) \\ -\gamma_1(v_0-e_1)^2+\gamma_2(v_0-e_2)^2\Big]=0 & \text{in } \Omega, \\ \dfrac{\delta_\epsilon(\varphi)}{|\nabla\varphi|}\dfrac{\partial\varphi}{\partial\vec{n}}=0 & \text{on } \partial\Omega. \end{cases} \tag{12}$$

The above PDE can be solved by numerical methods and the values of e_1 and e_2 are updating at every iteration from (11).

D. Chan-Vese Vector-Valued (CV-VV) Model

CV model is capable to segment only gray scale images. Therefore, in [5] Chan, Sandberg and Vese extended the CV model to vector-valued images. As their are many images which can be properly segmented in their vector representation rather than in its scalar representation. For instance, objects with different missing parts in different channels are completely detected (such as occlusion). Also, in color images, objects which are invisible in each channel or intensity can be detected by our algorithm. Let $v_{0,j}$ be the given vector-valued image over the domain Ω, with $j=1,2,...,M$ channels, and Γ be the evolving curve. Then the energy functional of CV-VV model is written as:

$$\begin{aligned} E^{CV-VV}(\overline{e_1},\overline{e_2},\Gamma) &= \nu.(Length(\Gamma)) \\ +\int_{inside(\Gamma)}&\frac{1}{M}\sum_{j=1}^{M}\gamma_{1,j}|v_{0,j}(x,y)-e_{1,j}|^2dxdy \\ +\int_{outside(\Gamma)}&\frac{1}{M}\sum_{j=1}^{M}\gamma_{2,j}|v_{0,j}(x,y)-e_{2,j}|^2dxdy, \end{aligned} \tag{13}$$

where $\overline{e_1}=(e_{1,1},e_{1,2},\ldots,e_{1,M})$ and $\overline{e_2}=(e_{2,1},e_{2,2},\ldots,e_{2,M})$ are two unknown constant vectors and $\gamma_{1,j}$ and $\gamma_{2,j}$ are positive parameters for each channel. In level set formulation model 13 can be written as:

$$\begin{aligned} E^{CV-VV}(\overline{e_1},\overline{e_2},\varphi) &= \nu.\int_\Omega \delta(\varphi(x,y))|\nabla\varphi(x,y)|dxdy \\ +\int_\Omega\frac{1}{M}&\sum_{j=1}^{M}\gamma_{1,j}|v_{0,j}(x,y)-e_{1,j}|^2H(\varphi(x,y))dxdy+ \\ \int_\Omega\frac{1}{M}\sum_{j=1}^{M}&\gamma_{2,j}|v_{0,j}(x,y)-e_{2,j}|^2(1-H(\varphi(x,y)))dxdy, \end{aligned} \tag{14}$$

where $H(\varphi)$ and $\delta(\varphi)$ are Heaviside and dirac delta functions, respectively as defined in CV model. The values of constants $e_{1,j}$ and $e_{1,j}$ can be obtain by from (14) by taking its partial derivatives.

The corresponding Euler-Lagrange equation can be find out by keeping $e_{1,j}$ and $e_{1,j}$ fixed and minimizing (14) with respect to φ we have:

$$\begin{aligned} \frac{\partial\varphi}{\partial t} &= \delta_\epsilon\Big[\nu.div\left(\frac{\nabla\varphi}{|\nabla\varphi|}\right)-\frac{1}{M}\sum_{j=1}^{M}\gamma_{1,j}(v_{0,j}-e_{1,j})^2 \\ &+\frac{1}{M}\sum_{j=1}^{M}\gamma_{2,j}(v_{0,j}-e_{2,j})^2\Big] \ \ in \ \ \Omega, \end{aligned} \tag{15}$$

where δ_ϵ is the regularized version of the dirac delta function and with the boundary condition:

$$\frac{\delta_\epsilon(\varphi)}{|\nabla\varphi|}\frac{\partial\varphi}{\partial\vec{n}}=0,$$

on $\partial\Omega$, and $\vec{n}$ represent the unit normal at the boundary of Ω. The advantage of CV-VV model [5] over CV model [4] is that it can also segment those images which can not be segmented in any scalar representation. CV-VV is also able to detect those objects in vector valued images that are not visible in each channel. However, the limitation of CV-VV model can be seen in intensity inhomogeneous images because CV-VV model utilized the averages of CV model.

E. Fast Global Minimization (FGM) Model

Bresson et al. proposed a fast global minimization of the CV model [2]. In this model, first, they made CV model convex and then used dual formulation for minimization. As the CV model is non-convex, so local minima exist, and therefore, a smooth approximation of the Heaviside function is selected in this model. Thus the steady state solution of (12) is same as:

$$\frac{\partial\varphi}{\partial t} \ = \ \nu\nabla.\left(\frac{\nabla\varphi}{|\nabla\varphi|}\right)-|v_0-e_1|^2+|v_0-e_2|^2 \tag{16}$$

Moreover, (16) is the partial differential equation obtained from the model (8). It is non-convex and stuck at local minima. FGM model incorporated an edge detector function into the CV model (8) in order to make it convex and proposed the following energy functional:

$$\begin{aligned} E^{FGM}(e_1,e_2,\varphi) &= \int_\Omega g(x)|\nabla\varphi|dxdy \\ &+\int_\Omega(|v_0-e_1|^2+|v_0-e_2|^2)\varphi dxdy, \end{aligned} \tag{17}$$

where $g(x)$ is an edge indicator function and the above model is convex and provide us a global minima. Actually, it is homogeneous of degree 1 in φ and therefore, it has no stationary solution. To get the optimal solution, some constrained must be imposed on φ, then the model (17) can be written as follows:

$$\begin{aligned} \min_{0\le\varphi\le1}\{E^{FGM}(e_1,e_2,\varphi) &= \int_\Omega g(x)|\nabla\varphi|dxdy \\ &+\int_\Omega(|v_0-e_1|^2+|v_0-e_2|^2)\varphi dxdy\}. \end{aligned} \tag{18}$$

The unconstrained form of minimization problem (18) is:

$$\min_{\varphi}\{E^{FGM}(e_1,e_2,\varphi)=\int_{\Omega}g(x)|\nabla\varphi|dxdy$$
$$+\int_{\Omega}\{(|v_0-e_1|^2+|v_0-e_2|^2)\varphi+\gamma v(\varphi)dxdy\}\}, \qquad (19)$$

where $v=\max\left\{0,2|\varphi-\frac{1}{2}|-1\right\}$ is an exact penalty function provided that the constant γ is chosen a large number such that:

$$\gamma>\frac{1}{2}||(v_0-e_1^2)+(v_0-e_2)^2||.$$

Furthermore, using dual formulation the variational problem (19) is regularized and minimized as in CV model. This model improved the performance of the CV model when there is little bit intensity change in the object to be detected and the background.

F. Xiaohao Cai Model

Image segmentation and restoration are closely related problems to each other. Therefore, in [3] Cai proposed a joint image segmentation and restoration model which is capable to restore as well as segment corrupted and degraded images. Cai et al. also provided a link between image restoration and segmentation in [17] while, in [18] the author proved that the solution of CV model [4] for a certain γ can be obtained by thresholding the minimizer of the ROF model [8] using a proper threshold, which provides a clear link between image segmentation and restoration. Cai model proposed the following energy functional as:

$$E^{Cai}(v,e_j,\phi_j)=\nu\Phi(v_0,\mathcal{A}w)+\gamma\psi(v,e_j,\phi_j)$$
$$+\sum_{j=1}^{K}\int_{\Omega}|\nabla\phi_j|dx, \qquad (20)$$

such that

$$\sum_{j=1}^{K}\phi_j(x)=1,\phi_j(x)\in\{0,1\},\forall x\in\Omega.$$

In model (20) the aim of first term is to restore the given corrupted image and comes from image restoration methods while the second term is designed for segmentation of images and the third term is the regularization term. Whereas, $\mathcal{A}$ represent the problem related operator. Cai model is designed for the segmentation of three types of images namely, blurry images, noisy images and images with missing pixels. Therefore the operator $\mathcal{A}$ is to be tuned according to the given image. Now by incorporating the data terms of ROF and CV model in model (20) i.e., by setting $\Phi(v_0,\mathcal{A}v)=\int_{\Omega}(v_0-\mathcal{A}v)^2dxdy$ and $\psi(v,e_j,\phi_j)=\sum_{j=1}^{K}\int_{\Omega}(v-e_j)^2\phi_jdxdy$, we have:

$$E^{Cai}(v,e_j,\phi_j)=\nu\int_{\Omega}(v_0-\mathcal{A}v)^2dxdy$$
$$+\gamma\sum_{j=1}^{K}\int_{\Omega}(v-e_j)^2\phi_jdxdy$$
$$+\sum_{j=1}^{K}\int_{\Omega}|\nabla\phi_j|dxdy, \qquad (21)$$

where $\sum_{j=1}^{K}\phi_j(x)=1,\phi_j(x)\in\{0,1\}$. Now the above energy functional (21) is minimized by the alternating minimization technique. For finding v keeping ϕ_j and e_j fixed and minimizing (21) with respect to v, we have:

$$v=(\nu\mathcal{A}^T\mathcal{A}+\gamma)^{-1}(\nu\mathcal{A}^Tv_0+\gamma\sum_{j=1}^{K}e_j\phi_j), \qquad (22)$$

where $\mathcal{A}$ is considered to be linear operator. For finding e_j keeping ϕ_j and v fixed and minimize (21) with respect to e_j, thus:

$$e_j=\frac{\int_{\Omega}v\phi_jdxdy}{\int_{\Omega}\phi_jdxdy}. \qquad (23)$$

Furthermore, for the minimization of energy functional (21) with respect to ϕ_j optimization methods can be used like ADMM (Alternating Direction Method of Mutiplying) method [19], [20] the max-flow approach [21] or the primal-dual algorithm [22], [23]. Cai model is able to segment images corrupted from noise, blur or having missing pixels but on the other hand the results of Cai model are not promising in images suffered from intensity inhomogeneity. This is due to the fact that Cai model used the data term of CV model, which uses the global intensity information of images rather than local one.

This section was devoted to the study of literature review of variational models used for image segmentation and restoration. Each of the above discussed models play a crucial role in image segmentation and restoration fields. However, these models have also some limitations as discussed above. Based on these techniques in the next chapter, we develop a new variational model for the joint image restoration and segmentation which is capable to segment images from intensity inhomogeneity.

III. Convex Hybrid Model

In this paper, a novel convex hybrid model is proposed for both denoising and segmentation of colour images. In the convex hybrid model, vector-valued TV is used for denoising. The energy functional of the convex hybrid model is given below:

A. Numerical Scheme

The discretization of (8) using explicit finite difference scheme becomes:

$$\frac{\partial \varphi_{i,j}^{k}}{\partial t} = E_{i,j} - \nabla \cdot \Big(\frac{g_{i,j}\nabla\varphi_{i,j}}{|\nabla\varphi_{i,j}|}\Big), \tag{24}$$

where $E_{i,j} = \sum_{j=1}^{M}(|v_j - e_{1,j}|^2 - |v_j - e_{2,j}|^2)$

$$\frac{\varphi_{i,j}^{k+1} - \varphi_{i,j}^{k}}{\triangle\ t} = E_{i,j} - \nabla \cdot \Big(\frac{g_{i,j}\nabla\varphi_{i,j}}{|\nabla\varphi_{i,j}|}\Big),$$

Simplifying we get,

$$\begin{aligned}\frac{\varphi_{i,j}^{k+1} - \varphi_{i,j}^{k}}{\triangle\ t} &= E_{i,j}\\ -\frac{\triangle\ t}{h_1^2} g_{i,j}\ \triangle_{-}^{x} &\left\{\frac{\triangle_{+}^{x}\ \varphi_{i,j}^{k}}{\sqrt{\Big(\frac{\triangle_{+}^{x}\varphi_{i,j}^{k}}{h_1}\Big)^2 + \Big(\frac{\triangle_{+}^{y}\varphi_{i,j}^{k}}{h_2}\Big)^2 + \beta}}\right\}\\ +\frac{\triangle\ t}{h_2^2} g_{i,j}\ \triangle_{-}^{y} &\left\{\frac{\triangle_{-}^{y}\ \varphi_{i,j}^{k}}{\sqrt{\Big(\frac{\triangle_{+}^{x}\varphi_{i,j}^{k}}{h_1}\Big)^2 + \Big(\frac{\triangle_{+}^{y}\varphi_{i,j}^{k}}{h_2}\Big)^2 + \beta}}\right\},\end{aligned} \tag{25}$$

where the differences $\triangle_{+}^{x}, \triangle_{-}^{x}, \triangle_{+}^{y}, \triangle_{-}^{y}$ are given by,

$$\begin{aligned}\triangle_{+}^{x}\varphi_{i,j}^{k} &= \varphi_{i+1,j}^{k} - \varphi_{i,j}^{k}, &\quad \triangle_{-}^{x}\varphi_{i,j}^{k} &= \varphi_{i,j}^{k} - \varphi_{i-1,j}^{k},\\ \triangle_{+}^{y}\varphi_{i,j}^{k} &= \varphi_{i,j+1}^{k} - \varphi_{i,j}^{k}, &\quad \triangle_{-}^{y}\varphi_{i,j}^{k} &= \varphi_{i,j}^{k} - \varphi_{i,j-1}^{k},\end{aligned} \tag{26}$$

$$\begin{aligned}\varphi_{i,j}^{k+1} &= \triangle\ tE_{i,j} + \varphi_{i,j}^{k}\\ -\frac{\triangle\ t}{h_1^2} g_{i,j}\ \triangle_{-}^{x} &\left\{\frac{\triangle_{+}^{x}\ \varphi_{i,j}^{k}}{\sqrt{\Big(\frac{\triangle_{+}^{x}\varphi_{i,j}^{k}}{h_1}\Big)^2 + \Big(\frac{\triangle_{+}^{y}\varphi_{i,j}^{k}}{h_2}\Big)^2 + \beta}}\right\}\\ +\frac{\triangle\ t}{h_2^2} g_{i,j}\ \triangle_{-}^{y} &\left\{\frac{\triangle_{-}^{y}\ \varphi_{i,j}^{k}}{\sqrt{\Big(\frac{\triangle_{+}^{x}\varphi_{i,j}^{k}}{h_1}\Big)^2 + \Big(\frac{\triangle_{+}^{y}\varphi_{i,j}^{k}}{h_2}\Big)^2 + \beta}}\right\},\end{aligned} \tag{27}$$

$$\begin{aligned}\varphi_{i,j}^{k+1} &= \triangle\ tE_{i,j} + \varphi_{i,j}^{k}\\ -\frac{\triangle\ t}{h_1^2} g_{i,j}\ \triangle_{-}^{x} &\left\{\frac{\varphi_{i+1,j}^{k} - \varphi_{i,j}^{k}}{\sqrt{\Big(\frac{\triangle_{+}^{x}\varphi_{i,j}^{k}}{h_1}\Big)^2 + \Big(\frac{\triangle_{+}^{y}\varphi_{i,j}^{k}}{h_2}\Big)^2 dx + \beta}}\right\}\\ +\frac{\triangle\ t}{h_2^2} g_{i,j}\ \triangle_{-}^{y} &\left\{\frac{\varphi_{i,j+1}^{k} - \varphi_{i,j}^{k}}{\sqrt{\Big(\frac{\triangle_{+}^{x}\varphi_{i,j}^{k}}{h_1}\Big)^2 + \Big(\frac{\triangle_{+}^{y}\varphi_{i,j}^{k}}{h_2}\Big)^2 + \beta}}\right\},\end{aligned}$$

where

$$D_{i,j} = \frac{1}{\sqrt{\Big(\frac{\triangle_{+}^{y}\varphi_{i,j}^{k}}{h_1}\Big)^2 + \Big(\frac{\triangle_{+}^{y}\varphi_{i,j}^{k}}{h_2}\Big)^2 + \beta}}$$

$$\begin{aligned}varphi_{i,j}^{k+1} &= \triangle\ tE_{i,j} + \varphi_{i,j}^{k}\\ &-\frac{\triangle\ t}{h_1^2} g_{i,j}\Big(D_{i,j}\varphi_{i+1,j}^{k}\\ &-D_{i-1,j}\varphi_{i,j}^{k} - D_{i,j}\varphi_{i,j}^{k}\\ &+D_{i-1,j}\varphi_{i-1,j}^{k}\Big)\\ +\frac{\triangle\ t}{h_2^2} g_{i,j}\Big(D_{i,j}\varphi_{i,j+1}^{k} &- D_{i,j-1}\varphi_{i,j}^{k} - D_{i,j}\varphi_{i,j}^{k} + D_{i,j-1}\varphi_{i,j-1}^{k}\Big)\end{aligned} \tag{28}$$

$$\begin{aligned}\varphi_{i,j}^{k+1} &= \triangle\ tE_{i,j} + \varphi_{i,j}^{k}\\ -\frac{\triangle\ t}{h_1^2} g_{i,j}\Big(D_{i,j}\varphi_{i+1,j}^{k} &- (D_{i-1,j} + D_{i,j})\varphi_{i,j}^{k} + D_{i-1,j}\varphi_{i-1,j}^{k}\Big)\\ +\frac{\triangle\ t}{h_2^2} g_{i,j}\Big(D_{i,j}\varphi_{i,j+1}^{k} &- (D_{i,j-1} + D_{i,j})\varphi_{i,j}^{k} + D_{i,j-1}\varphi_{i,j-1}^{k}\Big)\end{aligned} \tag{29}$$

$$\begin{aligned}\varphi_{i,j}^{k+1} &= \triangle\ tE_{i,j} + \varphi_{i,j}^{k} - \frac{\triangle\ t}{h_1^2} g_{i,j}\\ \Big(D_{i,j}\varphi_{i+1,j}^{k} &- (D_{i-1,j} + D_{i,j})\varphi_{i,j}^{k} + D_{i-1,j}\varphi_{i-1,j}^{k}\Big)\\ +\frac{\triangle\ t}{h_2^2} g_{i,j}\Big(D_{i,j}\varphi_{i,j+1}^{k} &- (D_{i,j-1} + D_{i,j})\varphi_{i,j}^{k} + D_{i,j-1}\varphi_{i,j-1}^{k}\Big),\end{aligned} \tag{30}$$

where

$$D_{i-1,j} = \frac{1}{\sqrt{\Big(\frac{\triangle_{+}^{x}\varphi_{i-1,j}^{k}}{h_1}\Big)^2 dx + \Big(\frac{\triangle_{+}^{y}\varphi_{i-1,j}^{k}}{h_2}\Big)^2 dx + \beta}}$$

$$D_{i,j-1} = \frac{1}{\sqrt{\Big(\frac{\triangle_{+}^{x}\varphi_{i,j-1}^{k}}{h_1}\Big)^2 dx + \Big(\frac{\triangle_{+}^{y}\varphi_{i,j-1}^{k}}{h_2}\Big)^2 dx + \beta}}$$

$$\begin{aligned}\varphi_{i,j}^{k+1} &= \triangle\ tE_{i,j} + \varphi_{i,j}^{k} - \frac{\triangle\ t}{h_1^2} g_{i,j}\\ D_{i,j}\varphi_{i+1,j}^{k} &+ (D_{i-1,j} + D_{i,j})\frac{\triangle\ t}{h_1^2} g_{i,j}\varphi_{i,j}^{k} +\\ D_{i-1,j}\frac{\triangle\ t}{h_1^2} g_{i,j}\varphi_{i-1,j}^{k} &+ \frac{\triangle\ t}{h_2^2} g_{i,j} D_{i,j}\varphi_{i,j+1}^{k}\\ -(D_{i,j-1} + D_{i,j})\frac{\triangle\ t}{h_2^2} g_{i,j}\varphi_{i,j}^{k} &+ D_{i,j-1}\\ \frac{\triangle\ t}{h_2^2} g_{i,j}\varphi_{i,j-1}^{k}\end{aligned} \tag{31}$$

This

$$\begin{aligned}\varphi_{i,j}^{k+1} =& \triangle\, tE_{i,j} + \varphi_{i,j}^{k} - A_1\varphi_{i+1,j}^{k} + A\varphi_{i,j}^{k} - A_2\varphi_{i-1,j}^{k} \\ & +B_1\varphi_{i,j+1}^{k} - B\varphi_{i,j}^{k} + B_2\varphi_{i,j-1}^{k} \quad (32)\end{aligned}$$

where

$$A_1 = D_{i,j}\frac{\triangle\, t}{h_1^2}g_{i,j}, \quad A = (D_{i-1,j} + D_{i,j})\frac{\triangle\, t}{h_1^2}g_{i,j}, \\ A_2 = D_{i-1,j}\frac{\triangle\, t}{h_1^2}g_{i,j}, \quad (33)$$

and

$$B_1 = D_{i,j}\frac{\triangle\, t}{h_2^2}g_{i,j}, \quad B = (D_{i,j-1} + D_{i,j})\frac{\triangle\, t}{h_2^2}g_{i,j}, \\ B_2 = D_{i,j-1}\frac{\triangle\, t}{h_2^2}g_{i,j}, \quad (34)$$

$$\begin{aligned}\varphi_{i,j}^{k+1} \quad =& \triangle\, tE_{i,j} - A_1\varphi_{i+1,j}^{k} + (I + A - B)\varphi_{i,j}^{k} + A_2\varphi_{i-1,j}^{k} \\ & +B_1\varphi_{i,j+1}^{k} + B_2\varphi_{i,j-1}^{k} \quad (35)\end{aligned}$$

The matrix form of the above equation becomes:

$$\varphi^{k+1} = E^k + M\varphi^k, \quad (36)$$

where φ^{k+1},φ^k and E^k are column vectors of size $1 \times n$. M is a block Tri-diaognal matrix.

In this paper, a novel convex hybrid model is proposed for both denoising and segmentation of colour images. In the convex hybrid model, vector-valued TV is used for denoising. The energy functional of the convex hybrid model is given below.

IV. Results and Discussion

In this section, the proposed Convex Hybrid model is compared with X. Cai [3] and CV vector-valued model [5] using synthetic and real images. The results of these experiments are compared using a qualitative approach.

Fig. 1 show ideal noise-free synthetic images of flower. Two different noise, i.e. “Salt and Pepper” and Gaussian were added to synthetic images as shown in Fig. 1 to evaluate the proposed Convex Hybrid model thoroughly. Fig. 2 shows salt and pepper noised synthetic image with Gaussian noise (variance = 0.002).

In next phase, Cai, CV-VV, LCV and Convex Hybrid models are applied to the noised synthetic images of bottles as shown in Fig. 2. Fig. 3 display the results of Cai model. Fig. 4 revealed the results of CV-VV model. Fig. 5 shows the result of LCV model. The balancing parameters such as $\mu_1 = 10$, $\mu = 0.0001$, $\lambda_1 = 1$ $\lambda_2 = 5$ and $\lambda_3 = 15$ are used in the Convex Hybrid Model.

Finally, the Convex Hybrid model results are revealed by Fig. 5. It is clear from the results of the above figures that Convex Hybrid model outperformed the Cai, CV-VV and LCV model as mentioned in Fig. 6. Convex Hybrid model segment the noised synthetic images of flower and bottle (Fig. 2) properly and successfully removed the noise as shown in Fig. 6. The Convex Hybrid model does not depend on the position of initial contour due to convexity.

Fig. 1. Original image of a bottle.

Fig. 2. Noisy image (having Salt and Pepper noise) of a bottle.

For extensive evaluation, the performance of Convex Hybrid model is verified using images with increasing combination of noises such as Salt & Pepper, Gaussian, Speckle and Poisson. Four further experiments are performed as mentioned below.

A. Test Case 1: Salt & Pepper, Gaussian, Speckle and Poisson

In test case 4, four noises i.e. salt and pepper, Gaussian, Speckle and Poisson noises are added having variance 0.02

Fig. 3. Joint results of restoration and segmentation for colour images using Cai model.

Fig. 4. Joint results of restoration and segmentation for colour images using Chan-Vese Vector-valued model.

as shown in Fig. 7. Fig. 8 is the result of Convex Hybrid model with smooth/denoised image having four noises. Convex Hybrid model with a segmented image is shown in Fig. 9.

V. Conclusion

In this paper, the Convex Hybrid model is proposed for restoration and segmentation of images. Initially, Convex Hybrid model is compared with three other models, i.e. Cai, CV-VV and LCV using noises such as Salt & Pepper and Gaussian. The results revealed that the Convex Hybrid model comparatively outperformed qualitatively. Four additional experiments were performed with increasing combination of noises such as Salt & Pepper, Gaussian, Speckle and Poisson. The convex

Fig. 5. Joint results of restoration and segmentation for colour images using Local Chan-Vese Model.

Fig. 6. Joint results of restoration and segmentation for colour images using Convex Hybrid model.

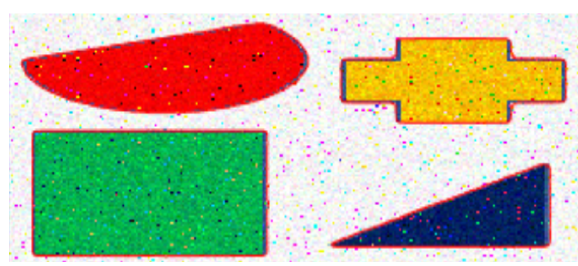


Fig. 7. Test Case 4: Noisy image having four noises, i.e. Salt & Pepper noise, Gaussian, Speckle noise and Poisson noise of variance=0.02.

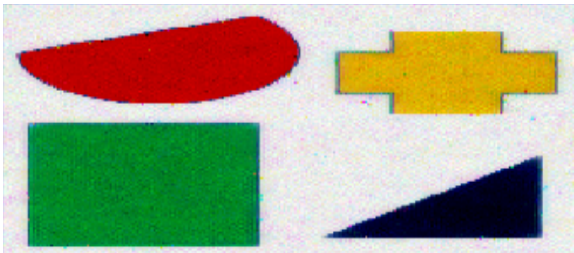


Fig. 8. Test Case 4: Convex Hybrid model with smooth/denoised image having four noises.

Fig. 9. Test Case 4: Convex Hybrid model with a segmented image having four noises.

Hybrid model successfully removed the noises and segmented the required object accurately. The Convex Hybrid model used the colour TV as a regularizer for denoising of the corrupt image. The Convex Hybrid Model is convex and can get global minima. The PDEs obtained from the minimisation of the Convex Hybrid Model are numerically solved by using explicit scheme.

In future, this research work will be extended to 3D image reconstruction and segmentation. There is a need to deeply study and understand the relationship among image segmentation, restoration and enhancement to achieve more robust models for image segmentation. Developing of new and robust numerical techniques for variational image segmentation models are also our research field of interest for future work.

References

[1] P. Blomgren and T.F. Chan, "Colour TV: Total variation methods for restoration of vector-valued images", *IEEE Transactions on Image Processing*, Vol. 7, No. 3, pp. 304-309, March, 1998.

[2] X. Bresson, S. Esedoglu, P.Vandergheynst, J.P. Thiran and S. Osher, "Fast global minimization of the active contour/snake model", *Journal of Mathematical Imaging and Vision,* Vol.28, No. 2, pp. 151-167, June, 2007.

[3] X. Cai, "Variational image segmentation model coupled with image restoration achievements", *Pattern Recognition,* Vol.48, No. 6, pp. 2029-2042, June, 2015.

[4] T.F. Chan and L. A. Vese, "Active contours without edges", *IEEE Transactions on Image Processing,* Vol.10, No. 2, pp. 266-277, February, 2001.

[5] T. F. Chan, B.Y. Sandberg and L.A. Vese, "Active contours without edges for vector-valued images", *Journal of Visual Communication and Image Representation,* Vol.11, No.2, pp. 130-141, June, 2000.

[6] D. Mumford, and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems", *Communications on Pure and Applied Mathematics,* Vol.42, No. 5, pp. 577-685, July,1989.

[7] R. Ronfard, "Region-based strategies for active contour models", *International Journal of Computer Vision,* Vol.13, No. 2, pp. 229-251, October, 1994.

[8] L.I. Rudin, S. Osher and E. Fatemi, "Nonlinear total variation based noise removal algorithms", *Physica D: Nonlinear Phenomena,* Vol.60 No. 4, pp. 259-268, November, 1992.

[9] N. Paragios and R. Deriche, "Geodesic active regions: A new framework to deal with frame partition problems in computer vision", *Journal of Visual Communication and Image Representation,* Vol.13, No.2, pp. 249-68, March, 2002.

[10] J.F. Cai, R.H. Chan, Z.A. Shen, "framelet-based image inpainting algorithm", *Applied and Computational Harmonic Analysis,* Vol. 24, No.2, pp. 131-149, March, 2008.

[11] T. Chan, A. Marquina and P. Mulet, "High-order total variation-based image restoration", *SIAM Journal on Scientific Computing,* Vol. 22, No.2, pp. 503-516, July, 2006.

[12] X.F. Wang, D.S. Huang and H. Xu, "An efficient local ChanVese model for image segmentation", *Pattern Recognition,* Vol. 43, No. 3, pp.603-618, March, 2010.

[13] S.I. Setzer and G.A. Steidl, "Variational methods with higher order derivatives in image processing", *Approximation*, Vol. 12, pp.360-86, January, 2008.

[14] X. Zhu, Z. Chen, C. Tang, Q. Mi and X. Yan, "Application of two oriented partial differential equation filtering models on speckle fringes with poor quality and their numerically fast algorithms", *Applied Optics,* Vol. 52, No.9, pp.814-23, March, 2013.

[15] I. Csisz and G. Tusnady, "Information geometry and alternating minimization procedures", *Statistics and Decisions,* January, 1984.

[16] H. Ali, G. Murtaza and N. Badshah, "Covariance Based Image Selective Segmentation Model",*Journal of Information & Comunication Technology*, Vol.4, No.2, pp. 11-19, 2010.

[17] X. Cai, R. Chan, and T. Zeng. "A two-stage image segmentation method using a convex variant of the mumfordshah model and thresholding". *SIAM Journal on Imaging Sciences*, Vol.6, NO.1, pp. 368390, 2013.

[18] X. Cai and G. Steidl. "Multiclass segmentation by iterated rof thresholding". *In International Workshop on Energy Minimization Methods Computer Vision and Pattern Recognition*, pages 237250. Springer, 2013

[19] S. Boyd, N. Parikh, E. Chu, B. eleato, and J. Eckstein. "Distributed optimization and statistical learning via the alternating direction method of multipliers". *Foundations and Trendsr in Machine Learning*, Vol. 3, No. 1, pp. 1122, 2011

[20] Y.He, M. Y.Hussaini, J. Ma, B. Shafei, and G. Steidl. "A new fuzzy c-means method with total variation regularization for segmentation of images with noisy and incomplete data", *Pattern Recognition*, Vol. 45, No. 9, pp.34633471, 2012

[21] J. Yuan, E. Bae, X. Tai, and Y. Boykov. "A continuous maxflow approach to potts model". *Computer Vision*, pages 379392, 2010.

[22] T. Pock, A. Chambolle, D. Cremers, and H. Bischof. "A convex relaxation approach for computing minimal partitions". *In Computer Vision and Pattern Recognition*. pages 810817. 2009.

[23] A. Chambolle and T. Pock. "A first-order primal-dual algorithm for convex problems with applications to imaging". *Journal of mathematical imaging and vision*, Vol.40, No.1, pp.120145, 2011.

Object-Oriented Context Description for Movie Based Context-Aware Language Learning

Hazriani*†, Tsuneo Nakanishi‡, Kenji Hisazumi§, and Akira Fukuda¶
*Department of Computer System, School of Management Informatics and Computer (STMIK) Handayani
Makassar 90231, Indonesia
†School of Information Science and Electrical Engineering, Kyushu University
Fukuoka 819–0395, Japan
‡Department of Electronic Engineering and Computer Science, Faculty of Engineering, Fukuoka University
Fukuoka 814–0180, Japan
§System LSI Research Center, Kyushu University,
Fukuoka 819–0395, Japan
¶Faculty of Information Science and Electrical Engineering, Kyushu University
Fukuoka 819–0395, Japan

Abstract—**Context-aware ubiquitous learning is a promising way to learn languages; however, it requires developers and operators of much effort to construct, deploy, and use the specialized system. As its alternative, this paper proposes movie based context-aware language learning (MBCALL) that enables learners to learn languages through quizzes generated along virtual contexts occurring in the movie to be replayed. Since full automatic context capturing from the movie is impossible, the authors define an object-oriented context model (OOCM) and also a textual context description language subject to the OOCM to describe the movie context easily by human work. The OOCM introduces the case grammar concept of natural language processing. This enables quiz generation based on types of the words for objects, actions, and modes found in the movie. Evaluation with a small movie by three subjects shows that the OOCM can guide them to enrich information included in the movie context; therefore, we can generate more types of quizzes based on the movie context.**

Keywords—Movie based context-aware language learning (MB-CALL); object-oriented context model (OOCM); context description language; case grammar

I. Introduction

So far many context-aware ubiquitous learning systems have shown that learning from real environment is attractive and effective for learners. However, it requires developers and operators to make much effort for using context-aware ubiquitous learning systems in practice. Developers have to develop complicated learning applications interacting with real environment through various sensors and smart devices. Operators have to prepare contextual learning materials that can provide learning tasks reacting to dynamic situations. Physical devices deployed in real environment can be out of order by long time operation under harsh environment such as outdoors. Availability of real time network connection is another important issue to be concerned. These matters will lead more development and operation cost or limited functionality of the learning system.

The movie can be a place in where context-aware ubiquitous learning is virtually practiced. The authors propose the concept of movie based context-aware learning (or MBCAL for short) as an alternative approach for context-aware ubiquitous learning to overcome the above-mentioned difficulties on development and operation [1]. MBCAL utilizes movies to facilitate learner's proactive learning from the context situated in the scenes as ubiquitous learning systems try to do in real environment. Moreover, it does not require any physical equipment to be deployed in real environment and any specific learning applications interacting with them, since every learning is performed in the movie player with learning functions. Considering the movie has been utilized to teach languages so far, one of the most suitable MBCAL applications will be language learning, which is hereinafter called as Movie Based Context-Aware Language Learning (or MBCALL for short). This paper focuses solely on MBCALL.

In MBCALL, the learner watches the movie and the movie player with learning functions generates quizzes automatically based on both movie's and learner's contexts [2]. The MBCALL system generates different quizzes to different learners depending on their different contexts even for the same movie.

The movie context is the current scene of the replayed movie and any information derived from it. It is desirable to compose the movie context description automatically from the movie; however, considering state-of-the-art image and phonetic recognition and understanding technologies, only a limited class of automatic composition is possible. Moreover, the movie sometimes includes the context which is not represented explicitly and is deduced from its story or implication. Therefore, human work is needed for composition of the movie context. The instructor describes the movie context by language independent context description and associates them with relating movie frames. This paper defines and evaluates a language for movie context description.

On the other hand, the learner context is information relating to the learner such as his/her learning level, completed learning tasks, watched/unwatched movies, and preferred movies. It is initially given by the learner him/herself and updated by the system along with his/her learning activity.

Followed by related work in Section II, the authors show the standing point of the movie based context-aware learning as well as its comprehensive comparison to existing learning paradigms in Section III. Section IV presents a scheme to describe the movie context with the object-oriented context model (or OOCM for short) and the textual context description based on the model, which are key conceptual contribution of this paper. In Section V, the authors evaluate how the OOCM can guide us to enrich information included in the movie context. Finally, Section VI concludes the paper.

II. Related Work

Development of information and communication technology (ICT) has derived a series of emerging learning paradigms. Ogata and Uosaki presented a research map on use of computers in education that identifies six emerging learning paradigms including CAI (computer assisted instruction) / ITS (intelligent tutoring system), game based learning (GBL), computer supported collaboration learning (CSCL), web based learning (or e-learning), mobile learning (or m-learning), and ubiquitous learning (or u-learning), along with involved ICTs and their contribution to user behaviors [3].

CAI is a form of computer supported learning and teaching using fixed PC, which was introduced in 1960s. Later, researchers incorporated artificial intelligence into CAI to improve learning and problem solving [4], which is known as ITS or ICAI (Intelligent CAI).

GBL, which was introduced in 1980s along with emergence of multimedia technologies, is another type of learning incorporating the subject matter into game play with defined learning outcomes. The learner can apply what they have learned to the real world through the game play.

CSCL, which is derived by invention of the world wide web (WWW) in 1990, facilitates interaction and collaboration among a group of learners connected each other.

E-learning facilitates learning on web based services in the internet.

M-learning facilitates learning in mobility, which is realized by mobile and wireless network and technologies.

U-learning facilitates learning through learner's surroundings captured by mobile, wireless, and ubiquitous network and technologies such as sensors and RFIDs.

E-, m-, and u-learnings have similar potential in enhancing social interaction and cooperative learning among learners through network. Moreover, u-learning enables learning from experience through interaction between cyber and physical spaces. Various research and commercial applications of these learning paradigms are available in various learning domains.

Movie based learning is an alternative way to achieve learning from experience, although the experience is virtual. Therefore, the movie has been practically used for learning so far. Kerber *et al.* encouraged students to learn the medical theory from corresponding provided movies [5]. Lumlertgul *et al.* conducted a cinemeducation project, which aimed to help students learn medical professionalism using movies [6]. Jungraithmayr and Weder improved the learning process on technique of orthotropic mouse lung transplantation with full-length movie visualization [7].

Furthermore, the movie can facilitate to learn context dependent aspects of the language. Chan and Herrero used movie clips to teach various languages in the classroom in their project and stated that plentiful contexts in the movie can be utilized for language learning as same as real environment, as it is known that the movie can facilitate comprehension activities perceived as real [8]. Ismaili showed effectiveness of using movies in the EFL classroom [9].

It is demonstrated by these researchers that movies improve student's understanding significantly and attract student's attention successfully. Additionally, in the field of language learning, the movie presents languages naturally as well as shows quite rich linguistic and environmental contexts that can be perceived as real. Considering the above-mentioned facts, the authors are motivated to incorporate the movie into a learning platform and develop an MBCALL system.

III. Movie based Context-Aware Learning

Movie based context-aware learning, which the authors propose, is a concept that introduces context-awareness into movie based learning. MBCAL enables learning from virtual experience in the movie as efficiently as u-learning that enables learning from experience in real environment through ubiquitous networks and technologies.

To compare MBCAL with e-, m-, and u-learnings and clarify differences, the authors surveyed existing work in the field of e-, m-, and u-learnings and adopt six features for comparison including permanency, accessibility, immediacy, interactivity, context-awareness, and content of the learning portfolio. The features except the content of the learning portfolio are adopted by many researchers as presented in [10], [11], [12] and [13]. The content of the learning portfolio is solely mentioned in [11]. These features are too abstract to differentiate the learning paradigms; thus, the authors define some levels of functionality with explanation for each feature.

As shown in Table I, these four learning paradigms can be accessed through the wireless network, show context-awareness based on the learning portfolio, and record learner's online behavior. Both e-learning and MBCAL can be accessed also via the wired network, while m-learning and u-learning must be accessed via the wireless network for its inherent characteristics. Both u-learning and MBCAL are able to keep learner's work permanently and provide dynamic adaptive services. In terms of interactivity, u-learning shows effective interactivity with peers, teachers, and experts. While both m-learning and MBCAL may have active interactivity, e-learning has interactivity in a limited form. Furthermore, the most recognized advantage of u-learning is context-awareness by both learner's physical context and real environment context. On the other hand, MBCAL shows context-awareness by utilizing virtual environment; *i.e.* movie scenes but with some technical limitations in performing real/physical context in u-learning. Lastly, u-learning records all aspects as the learning portfolio; namely, on-line behaviors, real-world behaviors, and corresponding environmental information of the learner. The others record online behaviors only.

In short, MBCAL has closest functionality to u-learning. Although it has functionality almost common to u-learning,

TABLE I. COMPARISON OF LEARNING PARADIGMS

Features	Functionality	e-learning	m-learning	u-learning	MBCAL
Permanency	The leaner cannot keep learning record.	X			
	The learner can keep learning record partially.		X		
	The learner can keep every learning record.			X	X
Accessibility	The learner can access the system via the computer network.	X			X
	The learner can access the system via the wireless network.	X	X	X	X
	The learner and related devices can communicate with ubiquitous technologies.			X	
Adaptive Services	The learner cannot get services immediately.	X			
	The learner can get information immediately in fixed environment.		X		X
	The learner can get information immediately in dynamic environment (as active services).			X	
Interactivity	The learner can interact with peers, teachers, and experts in a limited form.	X			
	The leaner can interact with peers, teachers, and experts actively.		X		X
	The learner can interact with peers, teachers, and experts effectively.			X	
Context-Awareness	The system can provide context-aware learning based on the learning portfolio.	X	X	X	X
	The system can provide context-aware learning based on learner's personal context (physical context).			X	
	The system can provide context-aware learning based on the context of real environment.			X	
	The system can provide context-aware learning based on the context of virtual environment.				X
Content of Learning Portfolio	The system records the online behaviors.	X	X	X	X
	The system records the real-world behaviors.			X	
	The system records the corresponding environmental information of the learner.			X	

MBCAL has one unique functionality; that is, context-awareness in virtual environment. Note that Table I lists possible aspects of functionality that each learning paradigm may show generally; yet, in practice, an application of each learning paradigm may not show all of them. Applications of a certain learning paradigm may show slightly different functionality depending on their purposes.

IV. OBJECT-ORIENTED CONTEXT MODEL AND TEXTUAL CONTEXT DESCRIPTION

In the MBCALL system, learners learn vocabularies and language expressions along the context in the movie. In contrast to the context-aware u-learning system which requires sensing devices and other data sources to acquire the context of physical environment, MBCALL requires none of them. Instead, MBCALL requires context description of movie scenes to be prepared before learning. Quizzes and their expected answers are generated based on the context description prepared for the movie.

The context description must be uniform, universal, and natural language independent to guarantee its reusability for different languages [14]. The authors have introduced the object-oriented paradigm and defined an object-oriented context model (or OOCM for short) to achieve this goal. The OOCM introduces the case grammar concept, which is a concept in the field of natural language processing, for its definition. The textual description based on the OOCM is defined as a means to describe the movie scenes.

A. Case Grammars

Fillmore proposed the case grammar concept [15]. The basic structure of the sentence can be divided into a couple of constituents: proposition and modality. The proposition consists of a tenseless set of relationships involving verbs and cases. The verb represents an event or activity, while cases are nouns having semantical relationships to the verb. Fillmore defines the following cases:

- Agentive: Who causes the event or activity?
- Instrument: What is used for the event or activity?
- Dative: Who experienced the event or activity?
- Factitive: What is the result of the event or activity?
- Locative: Where the event or activity happens?
- Objective: What is the target of the event or activity?

The modality includes such modalities on the sentence as-a-whole as negation, tense, mood, and aspect. For example, "Alice slices the bread with the knife" can be broken into *Alice* as agentive, *cut* as verb, *bread* as objective, and *knife* as instrument.

Followed by Fillmore, many linguists have refined his case grammar concept and introduced more cases. The authors surveyed those works[15], [16], [17], [18] and summarized as shown in Table II. Note that there are some cases which have the same or similar meaning but are named differently by different linguists; for example, Fillmore's *agentive* is renamed as *agent* by Jurafsky and Martin[16] and as *cAgent* by Parunak[17].

The cases in the table have the following meanings:

- An *agent* means a perceived instigator of the action of the verb.

- An *object* means a thing affected by the action or state of the verb.
- An *instrument* means force or an inanimate object casually involved in the state or action of the verb.
- An *experiencer* means an entity whose mental or emotional state is affected by the action of the verb.
- A *factitive* means a thing produced by the action of the verb.
- A *goal* means a place to where something moves or thing toward which an action is directed.
- A *beneficiary* means an entity that possesses an object or participates with an agent in transfer of an object.
- A *manner* means how the action, experiences, or process of an event is carried out.
- A *time* indexes the action of the verb in time.
- A *location* identifies the place or spatial orientation of the state or action.
- A *measure* means quantification of the event.
- A *source* means the place of the origin, entity from which a physical sensation emanates, or original owner in a transfer.
- A *commutative* refers to somebody else who performs an action with the agent.
- A *complement* means the new state of a factitive case.
- A *causer* means the referent which instigates an event rather than doing it actually.

B. Object-Oriented Context Model based on the Case Grammar Concept

Since the cases represent what the human being can cognize and verbalize, the case grammar concept can be used to describe movie scenes in a semantically formalized manner. The authors defined the OOCM integrating the case grammar concept for this purpose [14]. In this paper the OOCM of the previous work is improved for more comprehensive context description after writing movie contexts. The classes are defined by selecting cases from the summary shown in Section IV-A and mapping them on the OOCM as roles of the class. Fig. 1 shows the OOCM in the class diagram.

The OOCM has 11 classes. Classes *Thing*, *Action*, and *Mode* are the most fundamental ones.

- Class *Thing* means a thing appearing in the movie.
- Class *Action* means an action performed in the movie.
- Class *Mode* means a mode of the thing or the action.

Thing has the following descendant classes:

- Class *Animate*, a sub class of *Thing*, has its own intention.
- Class *Inanimate*, a sub class of *Thing*, does not have any intention.
- Class *Object*, a sub class of *Inanimate*, is something passively involved in the action.
- Class *Space*, a sub class of *Inanimate*, represents a location.
- Class *Time*, a sub class of *Inanimate*, represents a time instant of a duration.

Action has the following sub-classes:

- Class *Transitive Action* is an action that finishes immediately.
- Class *Continuous Action* is an action that continues for a certain duration.

C. Context Description Language

The movie context can be represented by the object diagram instantiated from the OOCM explained in the previous subsection. Since it is time consuming to edit the object diagram, the authors define the context description language (or CDL for short) to compose the movie context easily and quickly in a textual form. Fig. 2 shows the syntax of the CDL in the extended BNF in accordance with the current version of the OOCM.

The start symbol of the grammar is `ContextScript`. Symbol `ClassName` is the name of a class defined in the OOCM. Symbol `AssocName` is the name of an association between classes defined in the OOCM. `ClassName` and `AssocName` must be prefixed by the backslash (\). Symbol `LanguageCode` is a language code specified by ISO 639-1[19] and 639-3[20] such as `en` for English, `fr` for French, `zh` for Chinese, *etc.* Symbol `DQ` represents a double quotation mark (`"`).

For example, the movie scene shown in Fig. 3 can be modeled by the OOCM as shown in Fig. 4 and described in CDL as shown in Fig. 5.

V. Evaluation

In order to evaluate the descriptive power of the OOCM, the authors conducted an experiment by asking three volunteer subjects to describe the context of a public domain movie from *Popeye the Sailor* Series which is entitled as *Insect to Injury* [21]. The length of the movie clip is 6 minutes 7 seconds. The authors asked the subjects to describe the movie scenes in the free text without any more direction. Then the authors gave a short lecture on the OOCM and asked them to do the same thing again but with referencing the OOCM. All the subjects are Indonesian and they describe the movie scenes either in English or in Bahasa Indonesia, which is the common language in Republic of Indonesia.

Descriptions by the subjects were transformed into the OOCM for comparison. Movie contexts included in a single sentence can vary significantly depending on forms on writing. A complex sentence with more modifiers can have more meaning than a simple sentence. It enables fair comparison in a normalized form to transform a sentence into the OOCM and compare them based on the associations included in the OOCM. In comparison, the authors checked if the same type of association exists in the OOCMs describing the same movie

TABLE II. SURVEY ON THE CASE GRAMMAR [14]

Cases	Fil. [15]	Long. [18]	Lar. [18]	Par. [17]	Berk. [18]	Jur. [16]
Agentive (A) / Agent	X	X	X	X	X	X
Instrumental (I) / Instrument	X	X	X	X	X	X
Dative (D) / Experiencer (E)	X	X	-	X	X	X
Factitive (F) / Result / Range	X	X	X	-	-	-
Locative (L) / Location	X	X	-	X	X	X
Objective (O) / Object / Patient / Theme	X	X	X	X	X	X
Goal (G)	-	X	X	-	-	-
Measure	-	X	X	-	-	-
Path	-	X	-	-	-	-
Source	-	X	-	-	-	-
Accompaniment / Comitative	-	-	X	X	-	-
Beneficiary / Benefacte	-	-	X	X	-	X
Causer / Force	-	-	X	-	-	X
Manner	-	-	X	-	-	-
Time	-	-	X	X	-	-
Recipient	-	-	-	-	X	X
Possessor	-	-	-	-	X	X
Content	-	-	-	-	-	X

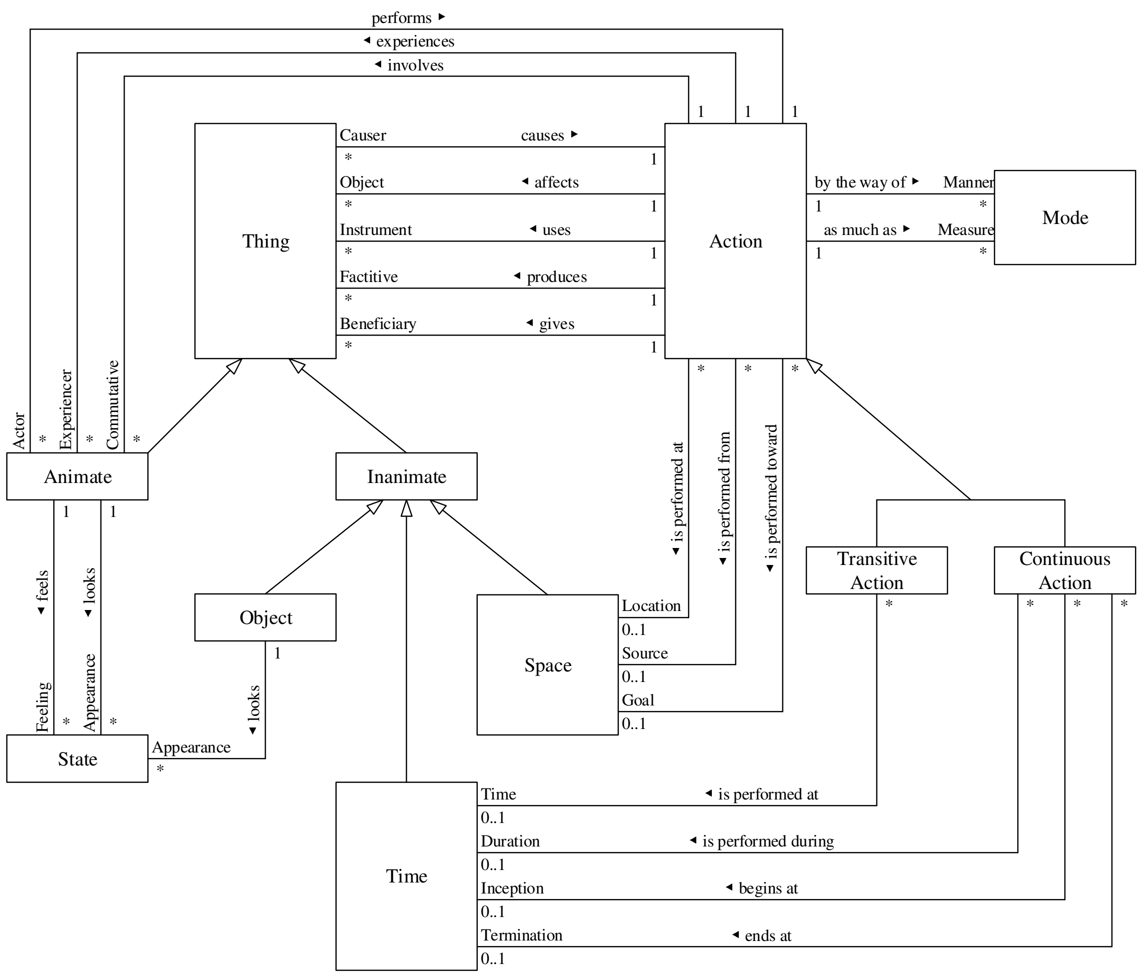


Fig. 1. Object oriented context model.

```
ContextScript
  = (InstanceDesc | ContextDesc)* .
InstanceDesc
  = ClassName
      "{" InstanceName "}"
      "{" WordDescList "}" .
WordDescList
  = WordDesc
  | WordDescList "," WordDesc .
WordDesc
  = LanguageCode "=" DQ Word DQ .
ContextDesc
  = "\context" "{" AssocDesc* "}" .
AssocDesc
  = AssocName
      "{" SourceInstance "}"
      "{" SinkInstance "}" .
SourceInstance = InstanceName .
SinkInstance = InstanceName .
InstanceName = [A-Za-z0-9]+ .
```

Fig. 2. Syntax of the context description language.

Fig. 3. An example movie scene.

scene by the subjects with ignoring trivial difference of words. Rather, the authors focused on difference of meanings.

In this evaluation, the authors collected forty-two (42) equivalent sentences from different subjects. Eighty-five (85) OOCM associations were identified from these sentences. The sentences and associations are listed in the appendix. The authors checked if these OOCM associations appear in each subject's description one by one. Table III shows the results.

The subjects could identify more associations of the OOCM describing the movie scenes with referencing OOCM than without referencing the OOCM. That means the OOCM can guide the composer to enrich information included in the movie context; therefore, we can generate more types of quizzes based on the movie context.

TABLE III. COMPARISON RESULTS

Approaches	Subjects		
	#1	#2	#3
Without Referencing OOCM	46	47	18
With Referencing OOCM	63	55	53

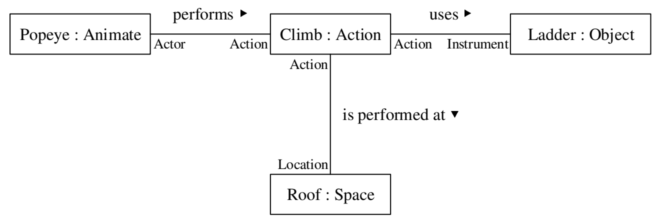


Fig. 4. An example OOCM.

```
\context{
  \Actor{Popeye}{en="Popeye", id=Popeye}
  \Action{Climb}{en="climb", id=menaiki}
  \Space{Roof}{en="roof", id=atap}
  \Object{Ladder}{en="ladder", id=tangga}
  \performs{Popeye}{Climb}
  \isPerformedToward{Climb}{Roof}
  \uses{Climb}{Ladder}
}
```

Fig. 5. An example CDL.

VI. Conclusion

This paper presented the concept of the movie based context-aware learning (MBCAL). The learner learns something through quizzes generated depending on the movie context with replaying the movie. In contrast to context-aware ubiquitous learning, the MBCAL system does not require to develop learning applications interacting with physical sensors or smartphones as well as to operate physical systems. Language learning is a promising application of MBCAL.

MBCAL requires less effort for its development and operation than context-aware ubiquitous learning. On the other hand, it requires effort to describe the movie context by human work, since state-of-the-art image and phonetic recognition and understanding technologies can recognize a limited class of the movie context. The authors presented the object-oriented context model (OOCM) which was constructed based on the case grammar concept of natural language processing. The OOCM can guide us to enrich description of the movie context. In the evaluation conducted in this paper, with referencing the OOCM, the subjects can identify and describe more OOCM associations describing the contexts that occur in the movie. Moreover, the authors defined a textual context description language (CDL) that can describe an instance of the OOCM. We can describe movie contexts instantly by using the textual CDL.

Acknowledgment

This work is partially supported by Directorate of Higher Education (DGHE) of Indonesian Government.

References

[1] Hazriani, T. Nakanishi, K. Hisazumi, and A. Fukuda, A, "A Movie Based Context-Aware Learning System: Its Concept and System," *Proc. IEEE 6th Int. Conf. on Technology for Education (T4E)*, pp. 164–167, December 2014.

[2] Hazriani, T. Nakanishi, and A. Fukuda, "Architecture, Textual Context Description, and Quiz Generation Scheme for the Movie Based Context-Aware Learning System," *Proc. 2016 IEEE Region 10 Int. Conf. (TENCON)*, pp.2412–2415, November 2016.

[3] H. Ogata and N. Uosaki, "A New Trend of Mobile and Ubiquitous Learning Research: Toward Enhancing Ubiquitous Learning Experiences," *Int. J. Mobile Learning and Organisation (IJMLO)*, Vol. 6, No. 1 pp. 64-78, March 2012.

[4] A. Molnar, "Computers in Education: A Brief History," *Int. J. Transforming Education through Technology*, Vol. 24, No. 11, pp. 63–68, June 1997.

[5] C. H. S. Kerber, D. Clemens, and W. Medina, "Seeing is Believing: Learning about Mental Illness as Potrayed in Movie Clips," *J. of Nursing Education (JNE)*, Vol. 43, No. 10, p. 479, October 2004.

[6] N. Lumlertgul, N. Kijpaisalratana, N. Pityaratstian, and D. Wangsaturaka, "Cinemeducation: A Pilot Student Project Using Movies to Help Stident Learn Medical Professionalism, Medical Teacher," *Informa Healtcare LTd.* Vol. 31, pp. e327–e332, August 2009.

[7] W. Jungraithmayr and W. Weder, "The Technique of Orthotopic Mouse Lung Transplantation as a MovieImproved Learning by Visualisation," *American J. of Transplantation*, Vol. 12, No. 6 pp. 1624–1626, March 2012.

[8] D. Chan and C. Herrero, *Using Film to Teach Languages*, Cornerhouse, 2010.

[9] M. Ismaili, "The Effectiveness of Using Movies in the EFL Classroom: A study Conducted at South East European University," *Academic Journal of Interdisciplinary Studies*, Vol. 2, No. 4 pp. 121–132, May 2013.

[10] P. S. Chiu, Y. H. Kuo, Y. M. Huang, and T. S. Chen, "The Ubiquitous Learning Evaluation Method Based on Meaningful Learning," *Proc. 3rd Int. Conf. on Computer Education (ICCSE)*, pp.257–264, July 2008.

[11] G. J. Hwang, C. C. Tsai, and S. J. H. Yang, "Criteria, Strategies and Research Issues of Context-Aware Ubiquitous Learning," *J. of Education Technology & Society (ET & S)*, Vol. 11, No. 2 pp.81–91, April 2008.

[12] H. Ogata, R. Akamatsu, and Y. Yano, "Computer Supported Ubiquitous Learning Environment for Vocabulary Learning Using RFID Tags," *Proc. IFIP TC3 Technology Enhanced Learning Workshop (TeL04)*, pp. 121-130, August 2004.

[13] S. Yahya, E. A. Ahmad, and K. A. Jalil, "The Definition and Characteristics of Ubiquitous Learning: A Discussion," *International Journal of Education and Development Using ICT (IJEDICT)* Vol. 6, No. 1 pp. 117–127, January 2010.

[14] Hazriani, T. Nakanishi, and A. Fukuda, "Introducing the Case Grammar Concept to Object Oriented Movie Description," *Proc. 2016 IEEE Int. Conf. on Teaching, Assesment, and Learning for Engineering (TALE)*, pp. 261–265, December 2016.

[15] J. Fillmore, *The Case for Case*, *Proc. of Texas Syposium on Language Universals* April 1967.

[16] D. Jurafsky and J. H. Martin, "Speech and Language Processing: Semantic Role Labeling," 3rd Edition draft, 2017, pp. 377–395.

[17] V. Parunak, "Case Grammar: A Linguistic Tool for Engineering Agent-Based System," *Journal or Conference*, Doi: 10.1.1.58.7325, 1995.

[18] SIL International, "Semantic Role," https://glossary.sil.org/term/semantic-role/, Last Accessed on April 1st, 2018.

[19] ISO 639-1:2002, Codes for the Representation of Names of Languages / Part 1: Alpha-2 Code, 2002.

[20] ISO 639-3:2007, Codes for the Representation of Names of Languages / Part 3: Alpha-3 Code for Comprehensive Coverage of Languages, 2007.

[21] D. Tendlar (direction), M. Reden, T. Moore (animation), I. Klein (story), A. Loeb (scenics), and W. Sharples (music), *Insect to Injury*, *Popeye the Sailor* Series, 1956.

APPENDIX

In the evaluation conducted in Section V, the authors collected fourty-two (42) equivalent sentences from three subjects. Each item in the enumeration below shows an equivalent sentences in the first line and the OOCM associations derived from the sentence in the following line(s). The fourty-two equivalent sentences derives eighty-five (85) OOCM associations.

1) *The surroundings of the house looks shady and cool.*
Object(*the surroundings of the house*) **looks** Appearance(*shady*)
Object(*the surroundings of the house*) **looks** Appearance(*cool*)

2) *Popeye climbs to the roof by a ladder.*
Actor(*Popeye*) **performs** Action(*climb*)
Action(*climb*) **isPerformedToward** Goal(*the roof*)
Action(*climb*) **uses** Instrument(*a ladder*)

3) *Popeye lays bricks very quickly.*
Actor(*Popeye*) **performs** Action(*lay*)
Action(*lay*) **affects** Object(*bricks*)
Action(*lay*) **byTheWayOf** Manner(*quickly*)

4) *Popeye feels satisfied.*
Animate(*Popeye*) **feels** Feeling(*satisfied*)

5) *Popeye slides down the stair.*
Actor(*Popeye*) **performs** Action(*slide down*)
Action(*slide down*) **isPerformedAt** Location(*the stair*)

6) *Popeye puts in the mailbox in front of the fence.*
Actor(*Popeye*) **performs** Action(*put in*)
Action(*put in*) **affects** Object(*the mailbox*)
Action(*put in*) **isPerformedAt** Location(*in front of the fence*)

7) *Popeye wears a yellow T-Shirt.*
Actor(*Popeye*) **performs** Action(*wear*)
Action(*wear*) **affects** Object(*a yellow T-Shirt*)

8) *Popeye writes his name on the mailbox.*
Actor(*Popeye*) **performs** Action(*write*)
Action(*write*) **affects** Object(*his name*)
Action(*write*) **isPerformedAt** Location(*on the mailbox*)

9) *The insects eat up the mailbox.*
Actor(*the insects*) **performs** Action(*eat up*)
Action(*eat up*) **affects** Object(*the mailbox*)

10) *Popeye closes the gate.*
Actor(*Popeye*) **performs** Action(*close*)
Action(*close*) **affects** Object(*the gate*)

11) *Popeye is in a panic.*
Animate(*Popeye*) **looks** Appearance(*in a panic*)

12) *Popeye blocks the insects.*
Actor(*Popeye*) **performs** Action(*block*)
Action(*block*) **affects** Object(*the insects*)

13) *Popeye catches the insects into the trash can.*
Actor(*Popeye*) **performs** Action(*catch*)
Action(*catch*) **affects** Object(*the insects*)
Action(*catch*) **isPerformedToward** Goal(*the trash can*)

14) *The insects escape from the trash can.*
Actor(*the insects*) **performs** Action(*escape*)
Action(*escape*) **isPerformedFrom** Source(*the trash can*)

15) *Popeye repairs the fence by the hammer.*
Actor(*Popeye*) **performs** Action(*repair*)
Action(*repair*) **affects** Object(*the fence*)
Action(*repair*) **uses** Instrument(*the hammer*)

16) *Popeye looks stressful.*
Animate(*Popeye*) **looks** Appearance(*stressful*)

17) *Popeye casts out the insects with wheels.*
Actor(*Popeye*) **performs** Action(*cast out*)
Action(*cast out*) **affects** Object(*the insects*)
Action(*cast out*) **uses** Instrument(*wheels*)

18) *The insects run towards the river.*
Actor(*the insects*) **performs** Action(*run*)
Action(*run*) **isPerformedToward** Goal(*the river*)

19) *The insects ruin the bridge.*
Causer(*the insects*) **causes** Action(*ruin*)
Action(*Action*) **affects** Object(*the bridge*)

20) *Popeye falls into the river.*
Experiencer(*Popeye*) **experiences** Action(*fall*)
Action(*fall*) **isPerformedToward** Goal(*the river*)

21) *The insects rush for the house.*
Actor(*the insects*) **performs** Action(*rush*)
Action(*rush*) **isPerformedToward** Goal(*the house*)

22) *Popeye digs a trench by the shovel.*
Actor(*Popeye*) **performs** Action(*dig*)
Action(*dig*) **affects** Object(*a trench*)
Action(*dig*) **uses** Instrument(*the shovel*)

23) *Popeye fills the trench with water.*
Actor(*Popeye*) **performs** Action(*fill*)
Action(*fill*) **affects** Object(*the trench*)
Action(*fill*) **uses** Instrument(*the water*)

24) *The insects look angry.*
Animate(*the insects*) **looks** Appearance(*angry*)

25) *Popeye feels secure.*
Animate(*Popeye*) **feels** Feeling(*secure*)

26) *Popeye enters to the house.*
Actor(*Popeye*) **performs** Action(*enter*)
Action(*enter*) **isPerformedToward** Goal(*the house*)
27) *Popeye locks the door.*
Actor(*Popeye*) **performs** Action(*lock*)
Action(*lock*) **affects** Object(*the door*)
28) *Popeye reads a book.*
Actor(*Popeye*) **performs** Action(*read*)
Action(*read*) **affects** Object(*a book*)
29) *The insects cross the trench with a can.*
Actor(*the insects*) **performs** Action(*cross*)
Action(*cross*) **isPerformedAt** Location(*the trench*)
Action(*cross*) **uses** Instrument(*a can*)
30) *The insects enter to popeye's house.*
Actor(*the insects*) **performs** Action(*enter*)
Action(*enter*) **isPerformedToward** Goal(*Popeye's house*)
31) *Popeye looks shocked.*
Animate(*Popeye*) **looks** Appearance(*shocked*)
32) *The insects destroy the house.*
Causer(*the insects*) **causes** Action(*destroy*)
Action(*destroy*) **affects** Object(*the house*)
33) *Popeye falls into the bathtub.*
Actor(*Popeye*) **performs** Action(*fall*)
Action(*fall*) **isPerformedToward** Goal(*the bathtub*)
34) *The insects look satiate.*
Animate(*the insects*) **looks** Appearance(*satiate*)
35) *Popeye eats spinach.*
Actor(*Popeye*) **performs** Action(*eat*)
Action(*eat*) **affects** Object(*spinach*)
36) *Popeye looks strong.*
Animate(*Popeye*) **looks** Appearance(*strong*)
37) *Popeye rebuilds his house with the metal.*
Actor(*Popeye*) **performs** Action(*rebuild*)
Action(*rebuild*) **affects** Object(*his house*)
Action(*rebuild*) **uses** Instrument(*the metal*)
38) *Popeye rebuilds the house in a moment.*
Actor(*Popeye*) **performs** Action(*rebuild*)
Action(*rebuild*) **affects** Object(*his house*)
Action(*rebuild*) **byTheWayOf** Manner(*in a moment*)
39) *The insects look tired.*
Animate(*the insects*) **looks** Appearance(*tired*)
40) *Popeye triumphs over the insects.*
Actor(*Popeye*) **performs** Action(*triumph over*)
Action(*triumph over*) **affects** Object(*the insects*)
41) *The insects eat a cigar.*
Actor(*the insects*) **performs** Action(*Action*)
Action(*eat*) **affects** Object(*a cigar*)
42) *Popeye feels relieved.*
Animate(*insects*) **feels** Feeling(*relieved*)

Smart Jamming Attacks in Wireless Networks During a Transmission Cycle: Stackelberg Game with Hierarchical Learning Solution

Moulay Abdellatif LMATER*, Majed Haddad†, Abdelillah Karouit† and Abdelkrim Haqiq*
*Computer, Networks, Mobility and Modeling laboratory
FST, Hassan 1st University, Settat, Morocco
†LIA/CERI University of Avignon Agroparc,
BP 1228, 84911, Avignon, France

Abstract*—Due to the broadcast nature of the shared medium, wireless communications become more vulnerable to malicious attacks. In this paper, we tackle the problem of jamming in wireless network when the transmission of the jammer and the transmitter occur with a non-zero cost. We focus on a jammer who keeps track of the re-transmission attempts of the packet until it is dropped. Firstly, we consider a power control problem following a Nash Game model, where all players take action simultaneously. Secondly, we consider a Stackelberg Game model, in which the transmitter is the leader and the jammer is the follower. As the jammer has the ability to sense the transmission power, the transmitter adjusts its transmission power accordingly, knowing that the jammer will do so. We provide the closed-form expressions of the equilibrium strategies where both the transmitter and the jammer have a complete information. Thereafter, we consider a worst case scenario where the transmitter has an incomplete information while the jammer has a complete information. We introduce a Reinforcement Learning method, thus, the transmitter can act autonomously in a dynamic environment without knowing the above Game model. It turns out that despite the jammer ability of sensing the active channel, the transmitter can enhance its efficiency by predicting the jammer reaction according to its own strategy.

Keywords*—*Wireless networks; jamming attacks; game theory; reinforcement learning

I. Introduction

Technology and system requirements in the telecommunications domain are changing very rapidly. Over the previous years, since the transition from analogue to digital communications, and from wired to wireless networks, different standards and solutions have been adopted, implemented and modified, often to deal with new and different business requirements. However, in the development of the wireless Next Generation Networks (NGNs) in which the layered architecture is adopted the common challenge of how further improve the resource utilization efficiency and provide better quality-of-service (QoS) is conditioned by the capacity of systems to accommodate changes quickly and with minimum impact on the services already implemented. Furthermore, the flexible topology and the low cost in term of use and setup have motivated the exploration of the wireless NGNs with increasingly higher data rates to meet the rapidly growing demand for wireless access.

Distributed protocols would be required to improve the radio resource utilization and provide high performance for wireless NGNs. In particular, an integrated design of Medium Access Control (MAC) based on Wireless Random Access (WRA) mechanism may lead to an efficient solution. This is why it is important to design distributed algorithms which can be used by the mobiles to compute the equilibrium strategy and simultaneously achieve the optimal operation points. On the other hand, the basic underlying assumption in legacy WRA protocols is that any concurrent transmission of two or more users causes all transmitted packets to be lost [2]. However, this model does not reflect the actual situation in many practical wireless networks where some information can be received correctly from a simultaneous transmission of several packets. This result is due to the fact that the packet arriving with the highest power has a good chance to be detected accurately, even when other packets are present. The effect of capture on Aloha [9], [10], [11], [18] and on IEEE 802.11 protocol (Carrier Sense Multiple Access-Collision Avoidance (CSMA/CA)) [19], [20], [21] has been studied extensively in the literature and new MAC protocols for channels with capture have been proposed. Furthermore, the full system utilization requires coordination among users which may be impractical given the distributed nature and arbitrary topology changes of wireless collision channels.

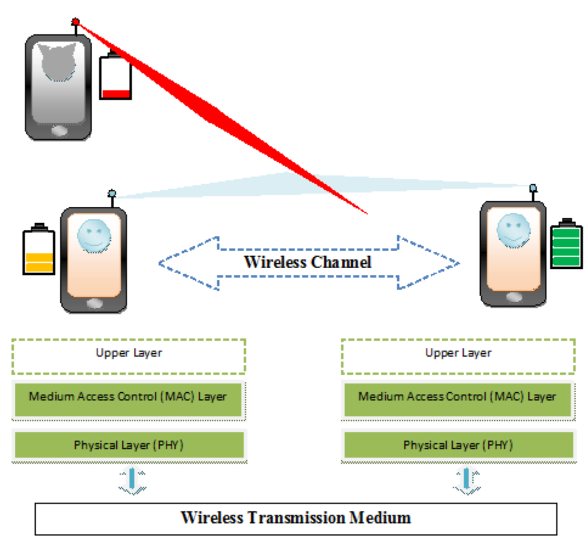


Fig. 1. Layered architecture for wireless networks.

However, while seeking ways to increase the performance of wireless network, there are increasing number of critical security issues that need to be addressed in order to make these wireless NGNs safer [7], [24], [25] (e.g., time-critical services, military operations, etc.). Note that wireless networks are vulnerable to security threats such as distributed denial of service attacks (DoS), spoofing attacks, Sybil attacks, faked sensing attacks and smart jamming attacks [7]. Thus the study of jamming problem in the context of wireless networks is an important challenge since it's easy to destroy communications due to the fact that the jammer can create dynamic and intelligent jamming attacks [23], [5].

The Game theory provides a convenient framework for approaching the power control in wireless based distributed MAC protocols. In fact, given the broadcast nature of the wireless MAC, the users are considered as selfish transmitters [2], and each transmitter seeks to maximize its payoff, while a malicious user tries to degrade the performance of the whole system. In this paper, we consider the IEEE 802.11 MAC CSMA/CA mechanism which is used by a large number of wireless systems, therefore, the problem of jamming can occurs during the transmission duty. In addition, the adversary or the jammer has to expend a significant amount of energy to jam the selected frequency bands, also the continuous presence of unusual high interference levels makes these attacks easy to detect. Thus, the main challenge in this paper is to derive the optimal strategy defense against the DoS attacks [16], [3], [4], [8], knowing the fact that the behavior of a malicious user may jam the network by sending abnormal packets to another user to block the channel from doing any things useful (Fig. 1).

It is well known that the Game theoretical approach is an appropriate concept to dealing with the competitive situation. Compared to the approaches used in previous works [12], [13], [14], [15], [17], etc. we are interested here in the impact of a smart jammer on the transmitter power levels during the period that starts at the first attempt of a packet transmission until the next packet transmission first attempt, due to the fact that when re-transmissions are used, the jammers cause the effective network activity factor (and hence the interference among the Receiver Sides (RSs) to be doubled [24]. In particular, we consider a scenario where a single transmitter (player 1) and a single jammer (player 2) coexist. The case of several transmitters/jammers is a subject of future research. Namely, the strategies of both the jammer and the transmitter are their transmission power levels during the packet transmission cycle. Since each packet transmission attempt incurs a cost in term of power, we consider that the Game objective utilities of both players are functions of the Signal to Interference plus Noise Ratio (SINR) value and the transmission cost. Under this anti-jamming Game based on power control problem, we propose two Game formulations, Nash Game where all players act simultaneously and Stackelberg Game where the transmitter is considered as leader (i.e. first to determine its transmit power) while jammer is considered as follower. At first, we derive the Nash Equilibrium (NE) expression, thereafter, we prove the existence of the Stackelberg Equilibrium (SE) and by using the Simulated Annealing Algorithm we sort out the SE measurement. From the comparison of the two schemes, we deduce that the transmitter can efficiently enhance the system performance. The main limitation with regard to the proposed power control-based anti-jamming problem is that there may be information loss for unknown jamming patterns. Thus, we consider a worst case scenario where the transmitter has an incomplete information while the jammer has a complete information. We introduce a Reinforcement Learning method, thus, the transmitter can act autonomously in a dynamic environment without knowing neither the estimating jamming patterns and parameters nor the above Game model.

The rest of this paper is outlined as follows. We briefly describe the related work in Section II. Then, we introduce the system model and the Game formulation in Section III. In Sections IV and V, we analyze the system in the presence of a regular and a smart jammer. In Section VI we propose a hierarchical learning solution. Simulation results are provided in Section VII. Finally, we conclude the paper and give some perspectives for future research.

II. Related Work

Designing mechanisms that can be able to detect wireless network jamming as well as avoid it has been widely studied under several works. In [26], authors investigate the anti-jamming problem with discrete power strategies, they formulate a Stackelberg Game to model the competitive interactions between the user and jammer. Then, they analyzed the asymptotic convergence by proposing a hierarchical power control algorithm (HPCA). In [27], a smart jammer can quickly learn the transmission strategies of the legitimate transmitters, and then he would adjust his strategy to damage the legitimate transmission. Meanwhile, the transmitters are aware of the existence of the smart jammer. The difference from [28] is that they consider relay nodes which help the source counteract a smart jammer. Furthermore, in [29] reinforcement learning can be applied to determine transmission powers against a jammer in a dynamic environment without knowing the underlying Game model. In [1], authors propose an anti-jamming Bayesian Stackelberg Game with incomplete information. In all the previous works on anti-jamming, the authors consider the problem transmitter-jammer during only one transmission attempt.

In this paper, we study the power control problem during a packet transmission cycle in the presence of a smart jammer, which has energy-efficiency and keeps track of the re-transmission attempts of the packet until that it is dropped. We suppose that the power level set is continuous and we consider a non-zero Game by introducing a transmission power cost.

III. System Model

Let a mobile use IEEE 802.11 CSMA/CA standard which is the most widely known standard in wireless networks. We assume that a transmission fails with probability that depends on the SINR. If a transmission fails then it is attempted again after some back-off time. After a certain number of attempts K the packet is dropped. Let's assume that the power is controlled. Hence, the power of the mobile user used at the ith transmission attempt can be denoted by $T_i \in [0, \bar{T}]$. Assume that:

$$T_i = p_0 x^{i-1} \tag{1}$$

where $p_0 \geq 0$ is the initial transmission power and $x > 1$ is the power multiplier factor for each re-transmission attempt. In this paper, we examine a scenario with one transmitter, which has its own traffic to send, and one jammer, which does not have its own traffic and simply wants to jam the transmitter attempts. As the mobile user spreads its signal over a common frequency band and treats interference as noise, thus, the signal to interference plus noise ratio at the ith transmission attempt $SINR_i$ at the receiver side is given by

$$SINR_i = \frac{\alpha T_i}{N + \beta J_i} \tag{2}$$

where N is the background noise level on the channel, $J_i \in [0, J^{max}]$ is the jammer power at the ith transmission attempt, $\alpha > 0$ and $\beta > 0$ denote the fading channel gain of the mobile user and the jammer, respectively.

Since a jammer chooses which transmission or re-transmission to jam, we assume that it jams all packets that are in the back-off stage $k \geq K_2$, where K_2 is an integer, that means that the competition starts from the back-off stage K_2. Since the quick detection of the start of a packet is becoming very harder for the jammer and this is due to the large bandwidths and the widely spread signals, we assume the worst situation in which the jammer can jam the communication from the first transmission attempt despite the fact that it arrives at a completely unpredictable time and frequency.

On the other hand, Let's define a cycle as the period that starts from the first attempt of a packet transmission to the first attempt of the next packet transmission. During a cycle, we consider a Game in which the two mobiles are players.

Moreover, we consider that each transmission occurs a certain cost and let $C > 0$ and $D > 0$ be the transmission costs per unit power of the mobile user and jammer respectively. We assume that players have perfect knowledge of the environment state and costs constraint at the beginning of each cycle.

Let $S_t = \{(p_0, x)|\ 0 < p_0 \leq p_0^{max};\ 1 \leq x \leq x^{max}\}$ the feasible set of the power multiplier and the initial transmission power of the mobile user and $S_j = \{(J_1, J_2, ...J_K)|\ J_i \geq 0;\ J_i \leq J^{max}\}$ the feasible set of the jammer power vector. We consider the following power control problem where (T, J) is to be determined, where $J = (J_1, J_2, ...J_K)$ and $T = (p_0, x)$.

The mobile user objective is to achieve the maximum $\sum_{i=1}^{K} SINR_i$ with the minimum cost. Intuitively, from (1) and (2), the utility function of the mobile user during a cycle denoted as $U(T, J)$ is given by:

$$U(T, J) = \sum_{i=1}^{K} \left(\frac{\alpha p_0 x^{i-1}}{N + \beta J_i} - C p_0 x^{i-1}\right) \tag{3}$$

The jammer objective is to achieve the minimum $\sum_{i=1}^{K} SINR_i$ with the minimum cost. From (1) and (2), the utility function of the jammer during a cycle denoted as $V(T, J)$ is given by:

$$V(T, J) = \sum_{i=1}^{K} \left(-\frac{\alpha p_0 x^{i-1}}{N + \beta J_i} - D J_i\right) \tag{4}$$

IV. Nash Game

In this section, we assume the presence of a regular jammer, and we consider a Game G_n = ({Transmitter, Regular jammer}, {T, J}, {U, V}). Since the regular jammer does not have the capability to sense the ongoing transmission power, all players take actions simultaneously. We focus on finding a Nash equilibrium in which neither the transmitter nor the jammer can increase its utility function by unilaterally changing its strategy. we define the Nash Equilibrium by the following formulation:

$$\begin{aligned} T^{NE} &= Arg\max\nolimits_{T \in S_t} U(T, J^{NE}) \\ J^{NE} &= Arg\max\nolimits_{J \in S_j} V(T^{NE}, J) \end{aligned} \tag{5}$$

Theorem 1: Let a jammer without the intelligence of learning the transmitter strategy. There exists a NE (T^{NE}, J^{NE}) in the Game, in addition,

$$C > \alpha/N \begin{cases} T^{NE} = (0, 1) \\ J^{NE} = 0 \end{cases} \tag{6}$$

$$C < \alpha/(N + \beta J^{max}) \begin{cases} T^{NE} = (p_0^{max}, x^{max}) \\ J^{NE} = (Min(J^{max}, \\ Max(0, \frac{1}{\beta}(\sqrt{\frac{p_0^{max}\alpha\beta}{D}} x^{max i-1} \\ -N))))_{i \in [1,K]} \end{cases} \tag{7}$$

$$ow \begin{cases} T^{NE} = (\frac{\alpha D}{\beta C^2}, 1) \\ J^{NE} = (\frac{1}{\beta}(\frac{\alpha}{C} - N))_{i \in [1,K]} \end{cases} \tag{8}$$

Proof: By (3) we have:

$$\frac{\partial U(T, J)}{\partial x} = p_0 \sum_{i=2}^{K} (i-1)\left(\frac{\alpha}{N + \beta J_i} - C\right) x^{i-2} \tag{9}$$

$$\frac{\partial U(T, J)}{\partial p_0} = \sum_{i=1}^{K} \left(\frac{\alpha}{N + \beta J_i} - C\right) x^{i-1} \tag{10}$$

The first order partial derivative of $V(T, J)$ with respect to J_i for $i \in [1, K]$, is

$$\frac{\partial V(T, J)}{\partial J_i} = \frac{\alpha\beta p_0 x^{i-1}}{(N + \beta J_i)^2} - D \tag{11}$$

The second order partial derivatives of the jammer objective function are:

$$\frac{\partial^2 V(T, J)}{\partial J_i \partial J_j} = \begin{cases} -\frac{2\alpha\beta^2 p_0 x^{i-1}}{(N+\beta J_i)^3} & i = j \\ 0 & ow \end{cases} \tag{12}$$

Therefore, the Hessian matrix of $V(T, J)$ with respect to the vector J is negative and $V(T, J)$ is strictly concave in J. Thus we consider the following cases:

- $C > \alpha/N$:
  As $\frac{\partial U}{\partial T} < 0\ \forall T \in S_t$, thus $x^{NE} = 1$ and $p_0^{NE} = 0$ yielding $T^{NE} = (0, 1)$. By using the concavity of V in J and setting $\frac{\partial V(T,J)}{\partial J_i}$ to zero, we have $J^0{}_i = \frac{1}{\beta}(\sqrt{\frac{p_0\alpha\beta}{D}} x^{i-1} - N)$. Since $0 \leq J_i \leq J^{max}$, let $J^{'} =$

$(Min(J^{max}, Max(0, J^0{}_i)))_{i\in[1,K]}$. According to Fig. 1, we have $\forall J \in S_j : V(T^{NE}, J) \leq V(T^{NE}, J^{'})$. Thus $J^{NE} = J^{'} = 0$.

- $C < \alpha/(N + \beta J^{max})$:
As $\frac{\partial U}{\partial T} > 0 \ \forall T \in S_t$, then, $p_0^{NE} = p_0^{max}$ and $x^{NE} = x^{max}$. By using the concavity of V in J and setting $\frac{\partial V(T,J)}{\partial J_i}$ to zero, we have $J^0{}_i = \frac{1}{\beta}(\sqrt{\frac{p_0^{max}\alpha\beta}{D}x^{max i-1}} - N)$. Since $0 \leq J_i \leq J^{max}$, let $J^{'} = (Min(J^{max}, Max(0, J^0{}_i)))_{i\in[1,K]}$. According to Fig. 1, we have $\forall J \in S_j : V(T^{NE}, J) \leq V(T^{NE}, J^{'})$. Thus $J^{NE} = J^{'}$.

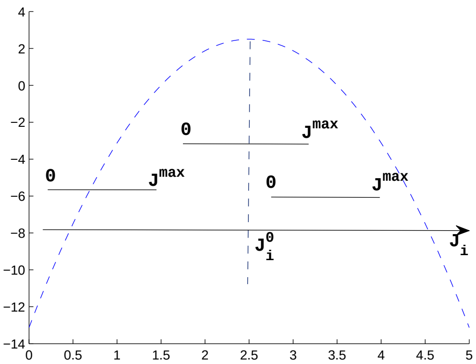


Fig. 2. The assumption of $V_i(T^{NE}, J_i)$ with respect to J_i.

- $\alpha/(N + \beta J^{max}) \leq C \leq \alpha/N$:
let $J^{'} = (\frac{1}{\beta}(\frac{\alpha}{C} - N))_{i\in[1,K]}$, then $U(T, J^{'}) = 0$ $\forall T \in S_t$. By using the concavity of V in J and setting $\frac{\partial V(T,J)}{\partial J_i}$ to zero, we have $J_i^0 = \frac{1}{\beta}(\sqrt{\frac{p_0\alpha\beta}{D}x^{i-1}} - N)$. In order to have $J_i^0 = J_i^{'}$ for $i \in [1, K]$ we must have $x = 1$ and $p_0 = \frac{\alpha D}{\beta C^2}$, without loss of generality we assume that $\frac{\alpha D}{\beta C^2} \leq p_0^{max}$. As result, we get, $T^{NE} = (\frac{\alpha D}{\beta C^2}, 1)$ and $J^{NE} = (\frac{1}{\beta}(\frac{\alpha}{C} - N))_{i\in[1,K]}$, That means:
$\forall T \in S_t : U(T, J^{NE}) = U(T^{NE}, J^{NE}) = 0$.
$\forall J \in S_j : V(T^{NE}, J) \leq V(T^{NE}, J^{NE})$.

■

V. Stackelberg Game

We assume the presence of a smart jammer. Since this kind of jammer has the capability to sense the ongoing transmission power, we model this problem as a Stackelberg Game denoted as: G_s = ({Transmitter, Regular jammer}, {T, J}, {U, V}), where the leader is the transmitter and the follower is the jammer. Thus, the leader fixes its optimal strategy based on the reaction of the follower, then the follower optimizes its own utility according to the leader strategy, namely, we define the Stackelberg Equilibrium by the following formulation:

$$\begin{aligned} T^{SE} &= Arg\max_{T\in S_t} U(T, Arg\max_{J\in S_j} V(T, J)) \\ J^{SE} &= Arg\max_{J\in S_j} V(T^{SE}, J) \end{aligned} \tag{13}$$

A. Jammer's Optimal Reaction

Assume that the two players have a complete information about the environment.

Lemma 1: Let T be a given strategy of the transmitter. There exists a unique $J^*(T)$ such that $J^*(T) = Arg\max_j V(T, j)$. In addition, the optimal jammer reaction is given by:

$$J_{i\in[1,K]}{}^*(T) = \begin{cases} 0 & E_1 \\ J^{max} & E_2 \\ \frac{1}{\beta}(\sqrt{\frac{p_0\alpha\beta}{D}x^{i-1}} - N) & ow \end{cases} \tag{14}$$

The conditions are given by:

- $E_1 : x^{i-1} < \frac{N^2 D}{p_0\alpha\beta}$
- $E_2 : x^{i-1} > \frac{(N+\beta J^{max})^2 D}{p_0\alpha\beta}$

Proof: According to (4), $V(T, .)$ is a continuous function on the compact set S_j and it can achieve its maximum value at some point $J \in S_j$. Since the first order partial derivative of the jammer objective function with respect to J_i, $\forall i \in [1, K]$ is:

$$\frac{\partial V(T, J)}{\partial J_i} = \frac{\alpha\beta p_0 x^{i-1}}{(N + \beta J_i)^2} - D \tag{15}$$

and the second order partial derivatives of the jammer objective function are:

$$\frac{\partial^2 V(T, J)}{\partial J_i \partial J_j} = \begin{cases} -\frac{2\alpha\beta^2 p_0 x^{i-1}}{(N+\beta J_i)^3} & i = j \\ 0 & ow \end{cases} \tag{16}$$

Therefore, the Hessian matrix of $V(T, J)$ with respect to the vector J is negative and $V(T, J)$ is strictly concave in J, [30]. Thus there exists a unique solution $J^*(T)$ such that $J^*(T) = Arg\max_{J\in S_j} V(T, J)$.

On the other hand, by resolving the following equation $\frac{\partial V(T,J)}{\partial J_i} = 0$, we have $J^0{}_i = \frac{1}{\beta}(\sqrt{\frac{p_0\alpha\beta}{D}x^{i-1}} - N)$. Since $0 \leq J_i \leq J^{max}$. 1) If $J^0{}_i > J^{max}$, yielding E_2, then $V(T, J)$ increases in S_j and thus $J_{i\in[1,K]}{}^*(T) = J^{max}$. 2) If $J^0{}_i < 0$, yielding E_1, then $V(T, J)$ decreases in S_j yielding $J_{i\in[1,K]}{}^*(T) = 0$. 3) If $0 \leq J^0{}_i \leq J^{max}$, yielding E_3, therefore, $J_{i\in[1,K]}{}^*(T) = J^0{}_i$. Thus, we deduce the property of the optimal jammer strategy given the strategy of the transmitter given in lemma 1. ■

B. Stackelberg Equilibrium

Let's now focus on analyzing the transmitter objective function given the reaction of the jammer.

Theorem 2: There exists $T^{SE} \in S_t$ such that $(T^{SE}, J^*(T^{SE}))$ is a Stackelberg Equilibrium of the Game.

Proof: To do so, we begin by proving the continuity of J^* on S_t. It's obvious that J^* is continuous in $S_t \setminus \bigcup_{i=1}^{K}\{Sa_i, Sb_i\}$, where for each $i \in [1, K]$, Sa_i is the set of couple $(p_0, x) \in S_t$ such that $p_0 x^{i-1} = \frac{N^2 D}{\alpha\beta}$, and Sb_i is the set of couple

$(p_0, x) \in S_t$ such that $p_0 x^{i-1} = \frac{(N+\beta J^{max})^2 D}{\alpha\beta}$. Let now $t_{a_i} \in Sa_i$ and $t_{b_i} \in Sb_i$, it's clear from lemma 2 that $\lim_{T\to t_{a_i}^-} J_i^*(T) = \lim_{T\to t_{a_i}^+} J_i^*(T) = 0$; $\lim_{T\to t_{b_i}^-} J_i^*(T) = \lim_{T\to t_{b_i}^+} J_i^*(T) = J^{max}$. The continuity of J^* on all S_t is proved. Since $U(T, J)$ is a continuous function on $S_t \times S_j$, thus $U(T, J^*(T))$ is continuous in T.

Since the set S_t is compact, $U(T, J^*(T))$ achieves its maximum at some point $T^{SE} \in S_t$. This prove the existence of $T^{SE} \in S_t$ such that $(T^{SE}, J^*(T^{SE}))$ is a Stackelberg Equilibrium of the Game. ∎

$\mathbf{U(T, J^*(T))}$ is not a concave function:: Despite we proved the existence of a SE $(T^{SE}, J^*(T^{SE}))$, calculating the SE is a challenging due to the non-concavity of the function $U(T, J^*(T))$. We use an example to show that there exists $(T_1, T_2, t) \in S_t^2 \times]0, 1[$ where: $U(t.T_1 + (1-t)T_2, J^*(t.T_1 + (1-t)T_2)) < t.U(T_1, J^*(T_1)) + (1-t).U(T_2, J^*(T_2))$

Let $N = 0.2; E = 1; C = 0.1; p_{0_1} = p_{0_2} = 2; k = 10; \alpha = \beta = 0.5; x_1 = 1.05, x_2 = 1.1, t = 0.63$.

In this example we have, $U(t.x_1 + (1-t)T_2, J^*(t.T_1 + (1-t)T_2)) = 13.7463$, $U(T_1, J^*(T_1)) = 13.3062$, $U(T_2, J^*(T_2)) = 14.5018$.

Hence, $U(t.T_1 + (1-t)T_2, J^*(t.T_1 + (1-t)T_2)) - t.U(T_1, J^*(T_1)) + (1-t).U(T_2, J^*(T_2)) = -0.0023$. Thus $U(T, J^*(T))$ is not a concave function on the set S_t. This results proves the complexity of finding a closed form of the global optimum, that's why we propose a simulated annealing technique as shown in Algorithm 2 in order to approximate the global optimum of our given function $U(T, J^*(T))$.

Algorithm 1 Calculate $T^{SE} = Arg \max_{T\in S_t} U(T, J^*(T))$

Require: $T \in S_t$
  Initialize the system parameters.
  Initialize G with a large value.
  T0=[0,1];
  while $(G \neq 0)$ **do**
    while (Accepted states number is below a threshold level) **do**
      Pick a random neighbor, $Tnew \leftarrow neighbour(T)$
      $\delta T = neighbour(T) - Tnew$
      $\delta U = U(Tnew, J^*(Tnew)) - U(T, J^*(T))$
      if $\delta U > 0$ **then**
        $T \leftarrow Tnew$
      else
        $T \leftarrow Tnew + \delta T. \exp(\delta U/G)$
      end if
    end while
    $G \leftarrow G - 1$
  end while

VI. Anti-Jamming with Reinforcement Learning

Reinforcement Learning (RL) is considered as a method in which the player takes action in a current time step and receives the corresponding reward in the next time step to evaluate its previous action [6]. RL is capable of solving more complex problems, specially, as the player does not require knowledge about the environment reaction and the reward function. However, the player learns just from previous experiences by interacting with the environment.

Through the above Game model, where both the transmitter and the jammer have a complete information of each other (i.e., channel gain and transmission cost), the SE strategies are derived. However, in view of the fact that Neither the jammer physical location nor its transmission cost is known by the transmitter due to the assumption that firstly, the jammer can change its physical location in a completely unforeseen time; secondly, the value of the jammer's transmission cost is not shared over the channel. Consequently, we introduce a reinforcement learning technique, especially the Q-learning method, so that the transmitter can act autonomously in a dynamic environment without knowing the above Game model.

We assume that the transmitter can choose its power multiplier and its initial transmission power from M and N levels respectively. Let P and A^{MN} denote the power action taken by the transmitter and the set of power action respectively. Meanwhile, the state observed by the transmitter is denoted by ${s_t}^n$. In each transmission cycle, the transmitter and the jammer take actions sequentially, we denote by J the jammer power action. At the beginning of the n-th transmission cycle, the transmitter first takes action and the decision making of its power action P_n is based on the transmission state in the previous transmission cycle, i.e., ${s_t}^n = (J_{n-1})$. sequentially, based on the observed state ${s_j}^n = (P_n)$, the jammer chooses its optimal power J_n given by (14). The received utility value of the transmitter is denoted by u_n. Let now describe the anti-jamming power control strategy based on Q-learning. Let α_t and β_t denote the learning rate and the discount factor of the transmitter. The Q-function with the power action P in the state s_t is denoted by $Q(s_t, P)$. The maximum Q value in the state s_t is denoted by $V(s_t)$. We define the update rule of the Q-function in the n-th transmission cycle as follows:

$$Q({s_t}^n, P_n) \leftarrow Q({s_t}^n, P_n) + \alpha_t(u_n + \beta_t V_{n+1} - Q({s_t}^n, P_n)) \tag{17}$$

$$V({s_t}^n) \leftarrow \max_{\forall P \in A^{MN}} Q({s_t}^n, P) \tag{18}$$

As a well-known reinforcement learning method, Q-learning should try to balance between exploration and exploitation according to ϵ-greedy policy where the transmitter chooses with a high probability 1-ϵ the power action that maximizes the Q value in the state s_t while other power actions are taken with an equal low probability $\frac{\epsilon}{MN-1}$. Thus, the probability of power action x taken by the transmitter is given by the following formulation:

$$Pr(P = x) = \begin{cases} 1-\epsilon & x = \underset{\forall P \in A^{MN}}{\text{argmax}}\, Q({s_t}^n, P) \\ \frac{\epsilon}{MN-1} & ow \end{cases} \tag{19}$$

Anti-Jamming Strategy of the transmitter with Q-learning is shown in detail as Algorithm 2.

VII. Simulation Results

In this section, numerical results are performed to evaluate the performance of the proposed power control problem during

Algorithm 2 Anti-Jamming Strategy of the transmitter with Q-learning

Require: $P \in S_t$
  Set the system parameters: β_t, α_t, ϵ, episode
  Set s_t, A^{MN}
  Initialize $Q(s_t, P)$,$V(s_t)$ as zero $\forall\ s_t, P \in A^{MN}$
  while $(episode \neq 0)$ **do**
    Set the starting state s_t^1
    For $n = 1, 2, 3, ..$
    Observe the current state s_t^n
    Pick a random power P_n from s_t^n according to Equation (19).
    if $P_n = \underset{\forall P \in A^{MN}}{\operatorname{argmax}}\, Q(s_t^n, P)$ **then**
      Decrease ϵ
    end if
    Observe the next state s_t^{n+1} and u_n
    Update $Q(s_t^n, P_n)$ by Equation (17).
    Update $V(s_t^n)$ by Equation (18).
    Break if convergence: small deviation on Q
    $episode \leftarrow episode - 1$
  end while

a cycle in both scenarios: 1) Transmitter against smart jammer (in which the jammer has the intelligence to quickly learn the transmission power of the transmitter and adjust its own transmission power). 2) Transmitter against regular jammer (in which both players play the Game simultaneously in a non-cooperative manner).

Among all the system variables, only fading channel gains of the transmitter and the jammer, may vary significantly due to the fact that the players can change their physical locations. Thus, we investigate the relations of the utilities of all players in equilibrium with respect to α and β. Let $N = 1$, $D = 0.2$, $C = 0.2$ and $K = 10$.

Fig. 3 shows the impact on the Utility function with respect to α at NE and SE. We observe that, as α increases, transmitter's SE utility increases while jammers' SE utility decreases; this phenomenon is due to the fact that the larger α became, the closer the transmitter became from the receiver. In addition, we depict in Fig. 4 the Utility function at NE and SE of both players with respect to β. As we can remark, the transmitter's utility at the SE decreases with β, while the jammer's utility increases with it; this is due to the fact that the larger β became, the closer the jammer became from the receiver. Moreover, in both Fig. 3 and 4, the transmitter at the NE has a lower utility than that at the SE, because at the latter the transmitter knows the existence of a jammer and utilizes its transmit power more efficiently. Similarly, a jammer obtains a higher utility at the SE than that at the NE, due to its ability to learn and adjust its own power according to the ongoing transmission power. This results proves that despite the jammer ability of sensing the active channel, the transmitter can enhance its efficiency by predicting the jammer reaction according to its own strategy.

Let now consider a scenario with power control strategy based on Q-learning. In this simulation, we set $M = N = 20$ and we set the maximum episode numbers in the learning to 120 in order to ensure the transmitter can learn an optimal action. The learning rate $\alpha_t = 0.8$ which indicates how far the current estimate value of Q is adjusted toward the update target value of Q. The discount factor of the source $\beta_t = 0.8$ that indicates the increasing uncertainty about rewards that will be received in the future. We assume a transmitter that does not have a complete knowledge about the dynamic environment, while the jammer has these knowledge. The initialization of the ϵ value for greedy algorithm is starting from 0.5 to ensure that the transmitter can try all actions in all states repeatedly. The utility of the transmitter received by the receiver according to the learning episodes are shown in Fig. 5. We can remark that the utility of the transmitter converges towards the solution proved in the above model. This result validate the proposed power control model. Note that, as the transmitter is gradually aware of the dynamic environment with the learning episodes increasing, which indicates a well anti-jamming performance. This is due to the fact that the transmitter chooses a more proper power action after has a well knowledge about the environment.

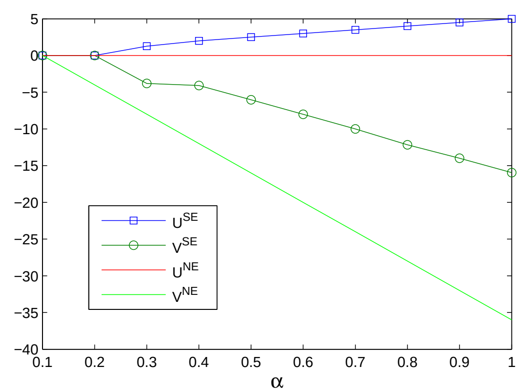


Fig. 3. The impact of α on the utility function of Jammer/transmitter at NE and SE. β=0.5.

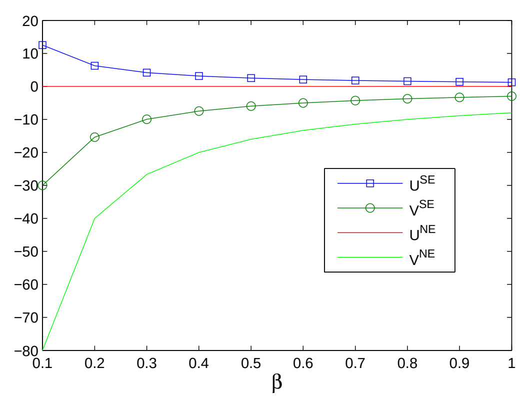


Fig. 4. The impact of β on the utility function of Jammer/Transmitter at NE and SE. α=0.5.

VIII. Conclusion

In this paper we studied denial of service vulnerability in wireless networks in the presence of jamming attacks. We choose a Game theoretical approach which is an abstract concept that indicates how the final outcome of a competitive

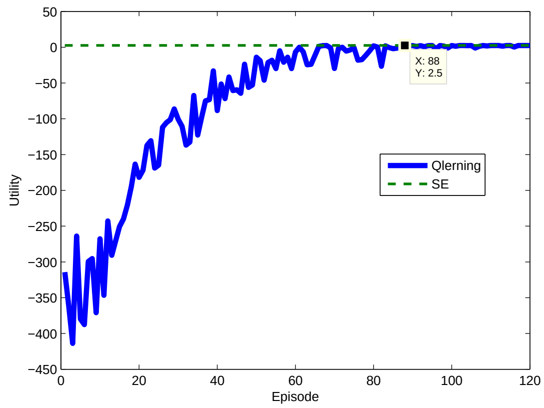


Fig. 5. Utility function of the transmitter, where the transmitter action is chosen based on Q-learning.

situation is dictated by interactions among the players. We considered a jamming during a transmission cycle. We studied the case where all players take action simultaneously and the case where the transmitter is the leader and the jammer is the follower. We proposed a Nash Game in the simultaneous Game and a Stackelberg Game in the hierarchical Game. A closed form of Nash Equilibrium is derived, then, we proved the existence of Stackelberg equilibrium. We sorted out the Stackelberg problem by using a simulated annealing algorithm. Moreover, we studied the relations of the utilities of all players in Nash and Stackelberg equilibrium. In order to validate our Stackelberg model, Q-learning method can is considered to be used by the transmitter to determine their transmission power actions in the presence of a smart jammer in a dynamic environment without knowing the underlying Game model. Simulation results have verified that despite the jammer ability of sensing the active channel, the transmitter can enhance its efficiency by predicting the jammer reaction according to its own strategy. Finally, this work can be extended to the case of several jammers that operate on a single sub-carrier during a single time slot in order to investigate the interaction among jammers who have interest to damage the source node transmission.

REFERENCES

[1] L. Jia, F. Yao, Y. Sun, Y. Niu and Y. Zhu "Bayesian Stackelberg Game for Anti-jamming Transmission with Incomplete Information." IEEE Communications Letters, vol. 20, no. 10, p. 1991-1994, 2016.

[2] Lmater Moulay Abdellatif, Abdelillah Karouit, and Abdelkrim Haqiq. "An efficient pricing mechanism of random access in wireless network with self-interested mobile users." Wireless Networks and Mobile Communications (WINCOM), 2015 International Conference on. IEEE, 2015.

[3] Guan, Yanpeng, and Xiaohua Ge. "Distributed secure estimation over wireless sensor networks against random multichannel jamming attacks." IEEE Access, vol. 5, p. 10858 - 10870, June 2017.

[4] Koh, Jing Yang, and Pengfei Zhang. "Localizing Wireless Jamming Attacks with Minimal Network Resources." International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage. Springer, vol. 10658, p. 322-334, 2017.

[5] Jaitly, Sunakshi, Harshit Malhotra, and Bharat Bhushan. "Security vulnerabilities and countermeasures against jamming attacks in Wireless Sensor Networks: A survey." Computer, Communications and Electronics (Comptelix), 2017 International Conference on. IEEE, 2017.

[6] Watkins, Christopher JCH, and Peter Dayan. "Q-learning." Machine learning 8.3-4 (1992): vol. 8, no. 3-4, p. 279âĂŞ292, May 1992.

[7] Pelechrinis, Konstantinos, Marios Iliofotou, and Srikanth V. Krishnamurthy. "Denial of service attacks in wireless networks: The case of jammers." IEEE Communications surveys and tutorials 13.2, p. 245-257, 2011.

[8] Dhuria, Shivam, and Monika Sachdeva. "Detection and Prevention of DDoS Attacks in Wireless Sensor Networks." Networking Communication and Data Knowledge Engineering. Springer, Singapore, vol. 3, p. 3-13, 2017.

[9] Arnbak, J. E. N. S. C., and Wim Van Blitterswijk. "Capacity of slotted ALOHA in Rayleigh-fading channels." IEEE Journal on Selected Areas in Communications 5.2, p. 261-269, 1987.

[10] Nguyen, Gam D., Anthony Ephremides, and Jeffrey E. Wieselthier. "Comments on âĂIJcapture and retransmission control in mobile radioâĂİ." IEEE Journal on Selected Areas in Communications 24.12, p. 2340-2341 ,2006.

[11] LaMaire, Richard O., Arvind Krishna, and Michele Zorzi. "On the randomization of transmitter power levels to increase throughput in multiple access radio systems." Wireless Networks 4.3, p. 263-277, 1998.

[12] Guan, Yanpeng, and Xiaohua Ge, "Distributed Attack Detection and Secure Estimation of Networked Cyber-Physical Systems Against False Data Injection Attacks and Jamming Attacks". IEEE Transactions on Signal and Information Processing over Networks , vol. 4, no. 1, p. 48-59, March 2018.

[13] Tayebi, Arash, Stevan Berber, and Akshya Swain. "Performance analysis of chaotic DSSS-CDMA synchronization under jamming attack." Circuits, Systems, and Signal Processing , vol. 35, no. 12, p. 4350âĂŞ4371, December 2016.

[14] Kim, Yongchul, Young-Hyun Oh, and Jungho Kang. "Asynchronous Channel-Hopping Scheme under Jamming Attacks." Mobile Information Systems, vol. 98, no. 4, p. 3583âĂŞ3610, February 2018.

[15] Wei, Xianglin, et al. "Collaborative mobile jammer tracking in multi-hop wireless network." Future Generation Computer Systems, vol. 78, no. 3, p. 1027-1039, January 2018.

[16] Li, Li, et al. "Security estimation under Denial-of-Service attack with energy constraint." Neurocomputing , vol. 292, no. 31, p. 111-120, May 2018.

[17] Peng, Lianghong, et al. "Energy efficient jamming attack schedule against remote state estimation in wireless cyber-physical systems" , vol. 272, no. 10, p. 571-583, January 2018.

[18] Sant, Jeetendra, and Vinod Sharma. "Performance analysis of a slotted-ALOHA protocol on a capture channel with fading." Queueing Systems 34.1-4, p. 1-35, 2000.

[19] Manshaei, Mohammad Hossein, et al. "Performance analysis of the IEEE 802.11 MAC and physical layer protocol." World of Wireless Mobile and Multimedia Networks, 2005. WoWMoM 2005. Sixth IEEE International Symposium on a. IEEE, 2005.

[20] Nyandoro, Alfandika, Lavy Libman, and Mahbub Hassan. "Service differentiation in wireless lans based on capture." Global Telecommunications Conference, 2005. GLOBECOM'05. IEEE. Vol. 6. IEEE, 2005.

[21] Hadzi-Velkov, Zoran, and Boris Spasenovski. "Capture effect in IEEE 802.11 basic service area under influence of Rayleigh fading and near/far effect." Personal, Indoor and Mobile Radio Communications, 2002. The 13th IEEE International Symposium on. Vol. 1. IEEE, 2002.

[22] X Ge, J Yang, H Gharavi, Y Sun, "Energy Efficiency Challenges of 5G Small Cell Networks" in IEEE Communications Magazine, vol. 55, no. 5, pp. 184 - 191, May 2017.

[23] Pelechrinis K, Iliofotou M, Krishnamurthy SV, "Denial of Service Attacks in wireless networks: The Case of Jammers" IEEE Communications Surveys and Tutorials , vol. 13, no. 2, pp. 245 - 257, May 2010.

[24] SD Amuru, HS Dhillon, RM Buehrer, "On Jamming Against wireless networks" IEEE Transactions on Wireless Communications, vol. 16, no. 1, pp. 412 - 428, November 2016.

[25] C. Kaufman and R. Perlman, M. Speciner, "Network security: private communication in a public world". Prentice Hall Press, 2002.

[26] L. Jia, et al., "A Hierarchical Learning Solution for Anti-Jamming Stackelberg Game With Discrete Power Strategies" IEEE Wireless Communications Letters , vol. 6, no. 6, pp. 818 - 821, August 2017.

[27] Y. Li, L. Xiao, J. Liu and Y. Tang "Power control stackelberg Game in cooperative anti-jamming communications". In Int. Conf. Game Theory for Networks (GameNETS), pp. 1-6, Nov. 2014.

[28] D. Yang, J. Zhang, X. Fang and A. Richa "Optimal transmission power control in the presence of a smart jammer". In : Global Communications Conference (GLOBECOM), 2012 IEEE, p. 5506-5511, December 2012.

[29] XIAO, Liang, LI, Yan, LIU, Jinliang, et al., "Power control with reinforcement learning in cooperative cognitive radio networks against jamming". The Journal of Supercomputing, vol. 71, no. 9, p. 3237-3257, 2015.

[30] Boyd, Stephen, and Lieven Vandenberghe. "Convex optimization". Cambridge university press, 2004.

A New Hybrid Intelligent System for Prediction of Medical Diseases

Sultan Noman Qasem[1,2]
[1]Computer Science Department, College of Computer and Information Sciences
Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia
[2]Computer Science Department, Faculty of Applied Science
Taiz University, Taiz, Yemen

Monirah Alsaidan
Computer Science Department, College of Computer and Information Sciences
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, Saudi Arabia

***Abstract*—This paper proposes a hybrid intelligent system as medical decision support tool for data classification based on the Neural Network, Galactic Swarm Optimization (NN-GSO), and the classification model. The goal of the hybrid intelligent system is to take the advantages and reduce the disadvantages of the constituent models. The system is capable of learning from data sets and reach great classification performance. Consequently, various algorithms have been developed that include Neural Network based on Galactic Swarm Optimization (NN-GSO), Neural Network based on Particle Swarm Optimization (NN-PSO) and Neural Network based on Genetic Algorithm (NN-GA) to improve NN structure and accuracy rates. For the evaluation process, the hybrid intelligent system has used multiple of benchmark medical data sets to evaluate the effectiveness. These benchmarks were gotten from the UCI Repository of Machine Learning. The three-performance metrics were calculated are accuracy, sensitivity and specificity. These metrics are useful for medical applications. The proposed algorithm was tested on various data sets which represent binary and multi-class medical diseases problems. The proposed algorithm performance analyzed and compared with others using k-fold cross validation. The significance tests results have proven that the proposed algorithm is effective to solve neural networks with good generalization ability and network structure for medical diseases detection.**

Keywords*—*Artificial neural network; galactic swarm optimization; particle swarm optimization; genetic algorithm; hybrid intelligent system; medical decision support

I. Introduction

The medical applications consider an important field for researchers. Because healthcare is one of the reasons that will help to encourage the general health and prosperity of the society. The two significant components in medical sciences are prediction and diagnosis of various diseases. In particular, prediction is the disease symptoms and diagnosis is relying on the experience of the physician. Typically, the physician gains the knowledge depending on the symptoms and the diagnosis. However, decision making problems such as prediction and diagnosis include complexity and uncertainty.

The medical knowledge and the treatments are progress very fast such as the appearance of new diseases and drugs. So, the ability of physicians to be aware of all current knowledge and development it is challenging. That's why the deployment of intelligent systems is helpful as medical decision support tool to help physicians in prediction and increasing diagnosis accuracy. Also, it will assist to arrive to decision quickly.

In machine learning, neural networks (NN) are used widely in medical decision support tools and have important advantages for these systems. Artificial neural networks are nonlinear sophisticated modelling techniques inspired by biological nervous systems. Neural networks capture the patterns in data by iteratively adjusting their synaptic weights in line with the learning algorithm [1]. Neural networks are a useful tool for various fields such as classification, prediction, pattern recognition, system identification, signal processing and function approximation. Classification problems consider the most artificial neural networks applications for medicine. There are multiple advantages of neural networks such as avoiding the time wasting and exacting knowledge gain procedure through learning from the data sets the relationship between patient symptoms (input) and the disease (output). However, choosing an appropriate architecture and learning algorithm is very important to have a high efficiency in ANN. In addition to learning, there are other useful properties for neural networks, which involve dealing with missing or incomplete data and filtering noise.

Feed-forward neural networks, in particular Multi-Layer Perceptron's (MLP) has been used in wide range of science and engineering. Because MLP has a high ability to classification and forecasting, it has been widely used in medical diagnosis, detection, and evaluation of medical conditions. It is composed of fully connected feed forward network with one input layer, one or more hidden layers there is a weighted connection between each neuron and all neurons in the next layer. The input layer neurons compute the NN independent variables and output layer neurons will transfer the results.. Between input and output layers there are hidden layers that can have any number of neurons. In each hidden layer there is defined sequence of activation functions through that the output value will be obtained.

Chitra and Seenivasagam [2] have used a multi-layer feed forward neural network (MLFFNN) and particle swarm optimization (PSO) as a hybrid system for Heart disease prediction at the early stage using the patient medical record. Within specified range this system adopted local and global optimization of the network parameters. Also, Christian and Krzysztof [3] have used feed-forward neural networks for pattern classification with Ant colony optimization (ACO) algorithm as training algorithm. On the other hand, a hybrid Particle Swarm Optimization (PSO) and Gravitational Search Algorithm (GSA) as training algorithm for Feedforward Neural Networks (FNNs) has been done by Seyed Ali Mirjalili et al. [4]. To avoid local minima and enhance the convergence speed. Hamada et al. [5] have used a hybrid system that involves the artificial neural networks (ANN), fuzzy logic, and genetic algorithms (GA). The combination of the neural networks and fuzzy logic helped to improve the performance. The Genetic algorithm have been used to minimize the fuzzy rules and number of features. Moreover, the GA worked on optimizing the initial weights of the artificial neural networks.

This research paper proposes a hybrid intelligent system that consists of artificial neural network (ANN) and the galactic swarm optimization (GSO). This hybrid intelligent capable to learn from data samples to be able to correctly classify the problems. The GSO algorithm is used for NN learning. It is inspired by the motion of stars, galaxies and super clusters of galaxies under the influence of gravity [6]. Comparing to state-of-the-art PSO algorithms the GSO algorithm consider faster in converged to a significantly better solution on a variety of multi-modal and high dimensional benchmark optimization problems. we can conclude from the paper [6] that comparing to other algorithms GSO consider better and gives a good result because of the characteristics that GSO have. On the other hand, a decision support tool with high accuracy, sensitivity, specificity is important for reducing cost and saving time. Furthermore, these metrics are calculated to measure the performance of the classification.

In addition, the aim of the hybrid intelligent system is to incorporate multiple techniques to complement each other and solving each other’s limitation. Because the previous meta-heuristics have several disadvantages, the proposed hybrid system uses a new meta-heuristic GSO that use in each epoch in the explorative phase and exploitative phase to prevent premature convergence and allows multi-modal surfaces to be efficiently explored. Also, GSO superior many multi swarm algorithms. On the other hand, FFNN take most of the research interests because of its ability to map any function to an arbitrary degree of accuracy.

The objectives of this paper are to propose a hybrid intelligent system for the design of neural network for medical data classification, to use benchmark medical data sets for evaluating the effectiveness of the system and to evaluate, validate and compare the performance of the proposed hybrid intelligent system with Neural Network based on Particle Swarm Optimization (NN-PSO) and Neural Network based on Genetic algorithm (NN-GA). The rest of the paper is structured as follows. Background and overview of literature review are provided in Section 2. The proposed methodologies are presented in Section 3. Section 4 shows the experimental studies. The results and the significance of the results are shown in Sections 5 and 6, respectively. Section 7 presents the discussion and analysis of the results. The conclusion and future work are presented in Section 8.

II. Background and Literature Review

This section provides a brief explanation of Artificial Neural Network, Genetic Algorithm, Particle Swarm Optimization and Galactic Swarm Optimization along with some of the key basic concepts.

A. *Artificial Neural Network (ANN)*

An Artificial Neural Network (ANN), also known as a neural network, it is a mathematical model inspired by biological nervous systems that consist of an interconnected group of simulated neurons. Neural networks process information for computation by using connectionist approach and they can model the simple and complex relationships. They are an adaptive system that changes its structure in the learning phase [1].

An ANN can be designed for different type of applications, such as data classification and pattern categorization. Network structures are the arrangement of neurons to form layers and the connection pattern formed within and between layers. There are different types of NN architectures are Multi-Layer Perceptron network, Single-layer Perceptron network, Radial Basis Function network (RBF), Hopfield network and Recurrent network. Multi-Layer Perceptron network (MLP) composed of fully connected feed forward network with one input layer, one or more hidden layers and output layer. There is a weighted connection between each neuron and all neurons in the next layer.

The important aspect of neural network is its capability of learning. The performance of neural network is relying on the success of the training process, and the training algorithm. Training or learning is a procedure of parameter tuning by which a neural network adapts itself to a stimulus and then desired output is produced [23].

B. *Genetic Algorithm*

A genetic algorithm (GA) can be understood as an “intelligent” probabilistic search algorithm which can be applied to a variety of combinatorial optimization problems. The genetic algorithm originally developed by Holland [25] and it is based on principles of natural evolution [11]. The natural populations develop according to the principles of natural selection and survival of the fittest. The Individuals who have a better chance of surviving and reproducing they are the more successful in adapting to their environment while the less fit individuals fit will be removed. So, in each successive generation, the genes from the highly fit individuals will spread to an increasing number of individuals [7].

Therefore, GA will imitate these procedures by taking an initial population of individuals and using genetic operators to every reproduction. Each individual in the population will be encoded into a chromosome or string which represents a possible solution to a given problem. The given objective function evaluated the fitness of an individual. The reproducing process will be done to highly fit individuals by replacing parts of their genetic information with other highly fit individuals, in a crossover procedure. Which will result new offspring solutions which share some characteristics taken from both parents. after crossover, the mutation process is applied to individuals by modifying some genes in the strings. Until a satisfactory solution is found, this evaluation selection-reproduction cycle is repeated [7]. The outline of a GA is shown in Fig. 1.

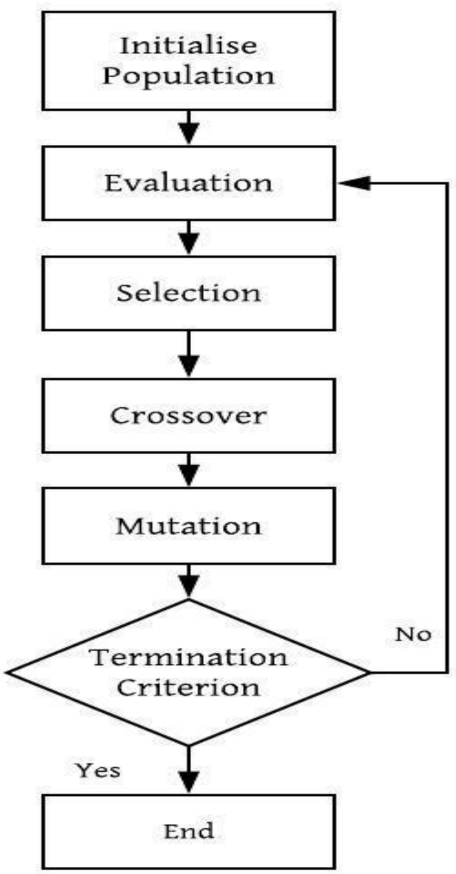


Fig. 1. Outline of GA algorithm [9].

C. Particle Swarm Optimization

Particle Swarm Optimization (PSO) was introduced by Eberhart and Kennedy [24], it is a population-based stochastic optimization technique. PSO doesn't require gradient information of the objective function and with less iteration, PSO can reach the global optimum value [4]. In PSO, each particle in the population has two vectors a velocity vi(t) vector, that enable the particles it to move within the problem space and position xi(t) vector. In the optimization problem, the number of decision variables will identify the dimensions of the two vectors. Updating the particle position is done by using the previous position information and the current velocity of the particle.

$$v_i(t+1) = wv_i(t) + c_1r_1\big(p_i - x_i(t)\big) + c_2r_2\left(p_g - x_i(t)\right) \quad (1)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (2)$$

Where v_i is the velocity of particle i , x_i is the position of particle i , c_1 is a weight applied to the cognitive learning portion, c_2 is a similar weight applied to the influence of the social learning portion, r_1 and r_2 are separately generated random number in the range of zero and one. p_i is the previous best location of particle i also known as p_{best}, p_g is the best location found by the entire population, also known as the G_{best}. Fig. 2 shows the basic pseudo-code for the PSO algorithm [8].

```
For each particle do
    initialize particle position and velocity
end for
while stopping criteria are not fulfilled do
    for each particle do
        calculate fitness value using problem specific objective function
        if fitness value is better than best fitness value pbest in particle
            history then
            set current position as pbest
        end if
    end for
    choose as gbest the particle with best fitness value among all particles in
    current iteration
    for each particle do
        calculate particle velocity based on Eq. (1)
        update particle position based on Eq. (2)
    end for
end while
```

Fig. 2. Standard PSO algorithm [8].

D. Galactic Swarm Optimization (GSO)

The GSO algorithm imitates the attraction of stars within galaxy to large masses and galaxies themselves to different large masses [2] as follows: First, according to the PSO algorithm, in every subpopulation the individuals are attracted to better solutions in the subpopulation. Secondly, the best solution for each sub swarm will be treated as a super swarm. The individuals in the super swarm will also move according to the PSO algorithm. The sub swarm is analogous to a galaxy of stars and the super-swarm is analogous to a cluster of galaxies.

In terms of concept and computation, GSO algorithm consider different than other multi-swarm approaches that run independently and continuously a multi-swarm that share information periodically with the slave swarms. While the super swarm in GSO only exists through level 2.

In GSO algorithm, promote exploration and avoid premature convergence consider the main difference between GSO and lots of alternative multi-swarm PSO because the flow of data between the sub swarms and the super swarm is unidirectional. This unidirectional relation means that the super swarm doesn't influence the exploration of the sub swarms by inserting good solutions just like other multi-swarm approaches. So, the solutions that have been computed by the super swarm not reinserted into the sub-swarms to ensure independent exploration of the sub swarms. Fig. 3 shows the GSO algorithm.

```
Begin PSO Level 1
For each particle
    initialize particle position and velocity
end for
For i=1 to Maximum number of epochs
For i=1 to number of subswarms
  While stopping criteria are not fulfilled do
     For each particle to size of set do
     calculate fitness value using problem specific
                    objective function
        If fitness value is better than the best fitness value pbest in
                    particle history
        then
        set current position as pbest
        end If
     end For
     choose as gbest the particle with best fitness value among all
              particles in current iteration
     For each particle do
       vi = w1 vj + c1 r1 (pj − xj) + c2 r2 (gi − xi)

       xi = xi + vi

     end For
  end while
end For
Begin PSO Level 2
For each particle do
initialize particle position with gbest for each sub swarm
end For
     While stopping criteria are not fulfilled do
       For each particle to number of subswarms do
          calculate fitness value using problem specific objective
                    function
            If fitness value is better than best fitness value pbest in
                    particle history then
                set current position as pbest
            end If
       end For
       choose as global best the gbest with best fitness value among
              all particles in current iteration
         For each particle do
       v = w2 v + c3 r3 (p − y) + c4 r4 (g − y)

       y = y + v

         end For
         end While
end For
```

Fig. 3. GSO algorithm [6].

III. Methodology

A. Neural Network Structure

In this paper, the objective of the hybrid approach is to apply GSO learning algorithm to train the weights of feed forward neural network, a multi-layer perception (MLP) neural network has been used. This network use Three-layers, which the first layer is composed of the input variables, the second layer consists of hidden nodes and the last layer is composed of the output variables [10]. It is built as shown in Fig. 4.

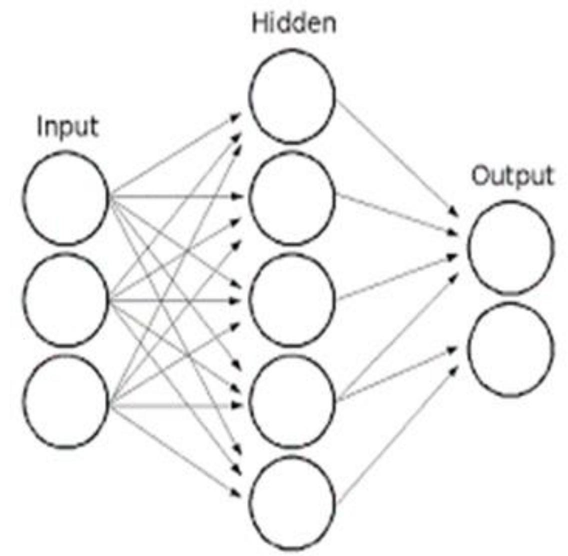


Fig. 4. Structure of neural network [10].

To determine the number of hidden nodes, (3) in [11] has been used as it is shown below:

$$j = \sqrt{n + m + a} \tag{3}$$

Where j is the number of the hidden node, $a \in [1, 10]$, n is the number of input nodes and m represents the number of output nodes.

B. Neural Network Representation

NN-GSO and NN-PSO algorithms will be represented by particles, but NN-GA algorithm will be represented by individuals. The representation is formed by four parts: the connection weights between the input layer and the hidden layer w_{ih}, the weights between the hidden layer and the output layer w_{ho}, and the hidden layer bias weights B_h and the output layer bias weights B_o. NN-GSO, NN-PSO and NN-GA algorithms will be encoded as vectors as in Fig. 5 where vectors are sequence of real numbers each of which belongs to the interval [−0.5, 0.5] where the dimension of individuals is given by (4).

$$(i \times h) + (h \times o) + h + o \tag{4}$$

Where i is the number of nodes in the input layer, h is the number of nodes in the hidden layer and o is the number of nodes in the output layer.

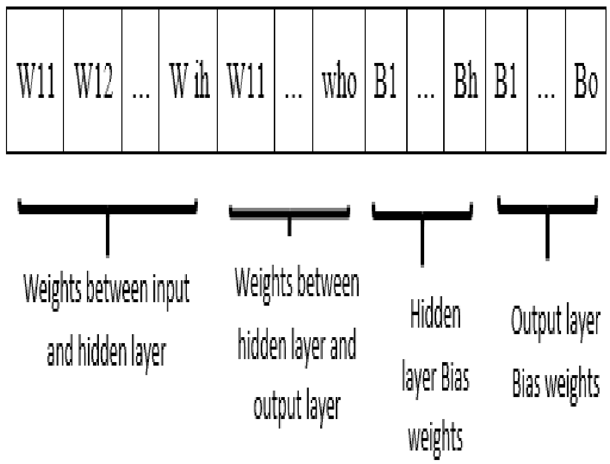


Fig. 5. Representation of NN-GSO, NN-PSO and NN-GA individuals structure.

C. Training Scheme

In this paper, Supervised Learning has been used in which a network defined with a training dataset that contain attributes (input) and output pairs. The learning procedure for the network is done by changing weights at each step of the training in order to reduce the error measure between the network's output and the desired known target value for a given input.

D. Fitness Function

Mean squared error (MSE) is used in this research as performance measure indicators to evaluate the performance of solutions. MSE fitness function considers the most common performance function used to train NN. The weights of the neural network are adjusted to minimize MSE on training set. Equation (5) has been used to calculate MSE.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\| t_j(i) - y_j(i) \|^2 \tag{5}$$

Where N, $t_j(i)$, $y_j(i)$ are maximum number of patterns, desired outputs and estimated outputs of Neural Network respectively.

E. GSO-Based NN Network Training Algorithm

GSO algorithm has been applied to improve Neural Network in various aspects such as network connection (weights, biases), and learning algorithm. The main process in our research is to apply GSO-based training algorithm on biases and weight optimization and investigate the efficiency of GSO in terms of robustness for training NN and the accuracy rate. By applying GSO algorithm the swarm is subdivided in to sub swarms. The GSO algorithm will update the sub swarms as well as the super swarm using the PSO algorithm.

GSO is a population (swarm) based optimization tool. The swarm is subdivided in to sub swarms and GSO algorithm will update the sub swarms as well as the super swarm according to the PSO algorithm. Every single solution (called a particle) which flies over the solution space in search for the optimal solution. Through following the personal best solution of each particle and the global best amount of the entire sub swarm, the algorithm finishes the first level and the best solution of each sub swarm will participate in the second level the super swarm which also move according to PSO. The particles are evaluated using a fitness function to see how close they are to the optimal solution. The output of this algorithm is weights and biases. The particles (weight, bias) values are initialized randomly. The particles are updated according to (6) and (7):

$$W(t+1) = W(t) + \Delta W(t+1) \tag{6}$$

$$\Delta W(t+1) = w.\Delta W(t) + c_1.rand.[pBest\ (t) - W(t)] + c_2.rand[gBest(t) - W(t)] \tag{7}$$

Where w, c_1, c_2 are inertia, cognitive and social acceleration constant respectively. The flowchart procedure for implementing the GSO is given in Fig. 6.

The Pseudo code for NN-GSO algorithm is as follows:

Step 1: Initialize the network Choose the number of nodes for the input, output and hidden layers

Step 2: Determine the initial value of weights between 0.5 and - 0.5

Step 3: Learning step and calculation of the weight values

```
Define PSO parameters (c1, c2, w, r1, r2)
Initialize population
The swarm will be divided to sub swarms
For each sub swarm
Save best position of any particle (global-best)
Loop
  For each particle in sub swarm
    Compute new particle velocity based on Eq. (6)
    Compute new position based on Eq. (7)
    Compute the error of new position based on Eq. (5)
    If new error better than best-error
      Best-position = new position
    If new error better than global-best
      Global-best = new position
  End for
End loop
Return global-best position
End for
The global-bests will participate on the super swarm population
Save best position of any particle (galactic-best)
Loop
  For each particle in super swarm
    Compute new particle velocity
    Compute new position
    Compute error of new position
    If new error better than best-error
      Best-position = new position
    If new error better than galactic -best
      Galactic-best = new position
  End for
```

End loop
Return galactic-best position
Step 4: give a diagnosis

F. PSO-Based NN Network Training Algorithm

Here we will present the procedure of NN-PSO. PSO is a population (swarm) based optimization tool. Every single solution (called a particle) which flies over the solution space in search for the optimal solution. The particles are evaluated using a fitness function to see how close they are to the optimal solution. The output of this algorithm is weights and biases. The particles (weight, bias) values are initialized randomly. The particles are updated according to (6) and (7).

The Pseudo code for NN-PSO algorithm is as follows:

Step 1: Initialize the network

Choose the number of nodes for the input, output and hidden layers

Step 2: Determine the initial value of weights between 0.5 and - 0.5

Step 3: Learning step and calculation of the weight values

Define PSO parameters (c_1, c_2, w, r_1, r_2)
Loop
For each particle in swarm
Compute new particle velocity based on Eq. (6)
Compute new position based on Eq. (7)
Compute error of new position based on Eq. (5)
If new error better than best-error
Best-position = new position
If new error better than global-best
Global-best = new position
End for
End loop
Return global-best position
Step 4: give a diagnosis

G. GA-Based NN Network Training Algorithm

Here we will present the proposed algorithm of NN-GA, to compare the results of this method with NN-GSO and NN-PSO.

The Pseudo code for NN-GA algorithm is as follows:

Step 1: Initialize the network

Choose the number of nodes for the input, output and hidden layers

Step 2: Determine the initial value of weights between 0.5 and - 0.5

Step 3: Learning step and calculation of the weight values

Creation of the initial population of chromosomes.
For each individual in the population
Loop
Evaluate the fitness of all the chromosomes of the population based on (5).
The best chromosomes will be selected to reproduce, using mutation and crossover.
With the new chromosomes created from the fittest of the previous generation, a new generation is created.
end loop
Evaluate the fitness for all the chromosomes of the population.
Select the fittest chromosome of the population as the new weight.
end for
Step 4: Give a diagnosis

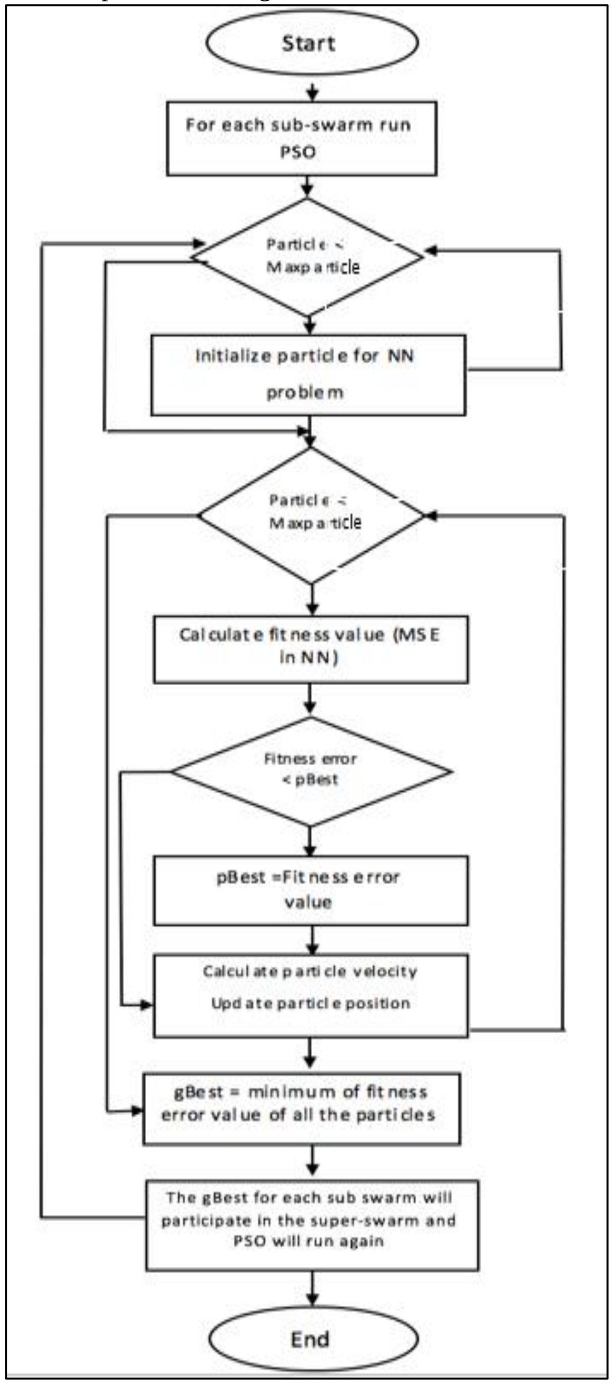


Fig. 6. NN-GSO procedure.

H. Performance Metrics

The performance metrics, accuracy, sensitivity and specificity are used to analyze and compare the outcomes of NN-GSO with NN-PSO and NN-GA algorithms. The accuracy measures how often the classifier makes the correct prediction on a dataset is defined by (8):

$$\text{Accuracy} = (TP + TN) / (TP + FN + FP + TN) \quad (8)$$

Where, TP (True Positive): is the proportion of positive cases that are correctly diagnosed as positive; FP (False Positive): is the proportion of negative cases that are wrongly diagnosed as positive; FN (False Negative): is the proportion of positive cases that are wrongly diagnosed as negative; TN (True Negative): is the proportion of negative cases that are correctly diagnosed as negative. On the other hand, sensitivity and specificity are statistical measures of the performance of a binary classification test, also known in statistics as classification function. Sensitivity which is also called the true positive rate, measures the proportion of positives that are identified correctly. Such as the percentage of sick people who are correctly identified as having the condition. This can be expressed mathematically as (9):

$$\text{Sensitivity} = TP / (TP + FN) \quad (9)$$

Specificity which is also called the true negative rate, measures the proportion of negatives that are correctly identified, such as the percentage of healthy people who are identified correctly as not having the condition. This can be expressed mathematically as (10):

$$\text{Specificity} = TN / (TN + FP) \quad (10)$$

IV. Experimental Studies

Several experiments were performed on nine real-world medical data sets. The results of NN-GSO for each dataset are compared to NN-PSO and NN-GA algorithms based on their overall classification performance.

A. Description of the Datasets

The proposed approach is applied on nine bench mark data sets for medical diseases for training and testing of the NN-GSO algorithm. These are Breast cancer, Diabetes, Heart, Hepatitis, Liver Disorders and Appendicitis datasets which represent binary class classification problems, while Thyroid, Dermatology and Alzheimer represent multiclass classification problems. All datasets are obtained from the University of California at Irvine UCI machine learning databases [21], except Alzheimer dataset which obtained from Open Access Series of Imaging Studies (OASIS) [22]. Table I shows the different dataset characteristics such as number of patterns and features.

1) Breast cancer dataset

This dataset has been collected from Dr. William H. Wolberg in the Wisconsin Hospitals University, Madison. The main purpose of it is to predict if the patient has benign tumour or malignant tumour. This data set includes 683 examples that have nine inputs (attributes) and two output classes.

2) Diabetes dataset

The dataset purpose is to predict a Pima Indian individual either positive or negative, depend on medical examination and personal data. This dataset considers a difficult problem which contains 768 examples with eight inputs (attributes) and two output classes.

3) Heart dataset

The detection of the existence or absence of heart diseases (heart positive or heart negative) is the main objective of this dataset. There are 303 examples in this dataset of which 139 are positive instances and 164 are negative instances. It has thirteen inputs (attributes) and two output classes.

4) Hepatitis dataset

The objective of this dataset is to detect if a hepatitis patient will die or live. However, this dataset considers noisy and complicated data because it has many missing data. It includes 155 examples that have 19 inputs (attributes) and two output classes.

5) Thyroid dataset

The objective of this dataset is to diagnosis if the patient is normal (1) or suffers from hyperthyroidism (2) or hypothyroidism (3). It contains 2069 examples with twenty-one input (attributes) and three output classes.

6) Alzheimer dataset

This dataset obtained from Open Access Series of Imaging Studies (OASIS) that contains cross-sectional MRI data. The objective of this dataset is to detect if the patient has normal control (NC) or mild cognitive impairment (MCI) or Alzheimer’s disease (AD). This dataset consists of 158 subjects aged 30 to 96. It contains fife attributes and three output classes’ normal control (NC), mild cognitive impairment (MCI) and Alzheimer’s disease (AD).

7) Appendicitis dataset

The objective of this dataset is to predict if the patient has appendicitis (1) or not (0). This data represents seven medical measures (inputs) taken over 106 patients and two output class.

8) Liver dataset

This dataset includes 345 examples to diagnose the existence or absence of liver disorders diseases. It has six inputs and two output classes (abnormal or normal). This dataset has 145 are positive examples (abnormal) and 200 are negative examples (normal).

9) Dermatology dataset

In dermatology, the differential diagnosis of erythemato-squamous diseases considers a real problem because they all share the clinical features of erythema and scaling, with very little differences and in the first stages the disease may show the features of another disease and may have the characteristic features at the following stages. This data contains 34 attributes. The diseases in this data set are psoriasis, seboreic dermatitis, lichen planus.

TABLE I. DESCRIPTION OF DATASETS.

Data Set	Attributes	Classes	Samples
Breast cancer	9	2	683
Diabetes	8	2	768
Heart	13	2	297
Hepatitis	19	2	155
Thyroid	21	3	2069
Alzheimer	5	3	158
Appendicitis	7	2	105
Liver	6	2	343
Dermatology	33	3	242

B. Experimental Setup

The Evaluation of the proposed algorithms has been done by random 10-fold cross validation. In 10-fold cross-validation, the dataset is split in to 10 equal parts .one part used as testing dataset and the other parts used as training dataset. This operation will be repeated until all parts used as testing dataset [9]. We use all datasets to evaluate the performance of proposed algorithms and analyze the evolutionary process of it. Table II shows the parameters settings for the proposed algorithms. While the number of input and output nodes is depending on the problem domain. We use the common parameter settings for GSO, PSO and GA algorithms that are recommended in literature [6], [15]. To unify all algorithms the population size is set for 100 and Maximum number of iterations is set to 1000 for all algorithms.

TABLE II. PARAMETERS SETTINGS FOR THE PROPOSED ALGORITHMS

Neural Network Initialization	
Initial weights	[−0.5,0.5]
NN-GSO	
Maximum number of iterations L1	198
Acceleration coefficient (c_1)	1.4
Acceleration coefficient (c2)	1.4
NN-PSO	
Population size	100
Acceleration coefficient (c_1)	1.4
Acceleration coefficient (c2)	1.4
NN-GA	
Crossover Probability (cp)	0.9
Mutation Probability (mp)	0.5

V. RESULTS

The results of study are presented in this section. Table III shows the Mean Square Error (MSE) of the proposed algorithms NN-GSO, NN-PSO and NN-GA on the training and the testing set on Breast cancer, Heart, Hepatitis, Diabetes, Hepatitis, Liver Disorders, Appendicitis, Thyroid, Dermatology and Alzheimer datasets. Also, the average of MSE for all dataset for each of the proposed algorithms is shown in Table III. The Mean, Min, Max and STDV indicate the mean value, minimum value, maximum value and standard deviation, respectively. Ten-fold cross-validation has been used for all datasets to obtain the results. As can be seen from Table III, the testing error values indicate that NN-GSO algorithm has resulted in better convergence, for most of the data sets compared to NN-PSO and NN-GA algorithms. It also shows that NN-GSO algorithm has produced the smallest error in the testing set for Diabetes, Hepatitis, Breast cancer, Liver Disorders and Dermatology datasets with an average error 0.15650, 0.16568, 0.02729, 0.20814, and 0.16665 respectively. However, NN-PSO algorithm outperforms the others with the smallest error for Heart and Thyroid datasets with an average error 0.08763 and 0.04546 respectively. For Alzheimer and Appendicitis datasets NN-GA algorithm produce the smallest error with an average error 0.03851 and 0.11803 respectively. NN-GSO algorithm does not show good performance for Heart and Thyroid dataset the reason for that is the nature of data. While NN-GA algorithm is better for Alzheimer and Appendicitis datasets because of the small number of samples.

Accuracy, sensitivity and specificity consider the most using performance measures for dataset classification. The measure of the ability of the classifier to obtain an accurate diagnosis is the accuracy. Sensitivity will evaluate the performance of classifier through identifying the positive examples which is the number of false negatives and true positives. While specificity will evaluate the performance of classifier through identifying the negative examples which is the number of false positives and true negatives.

Table IV and Fig. 7 to 9 shows the statistical results for accuracy, sensitivity and specificity of NN-GSO, NN-PSO, and NN-GA algorithms for all datasets on both the training and the testing set. Hepatitis, Thyroid and Liver datasets deserve special mention. These datasets represent difficult classification problems for all algorithms because they are very unbalanced datasets. For Hepatitis and Liver datasets NN-GSO algorithm shows the best results in accuracy and sensitivity, while NN-PSO algorithm shows the best results in specificity. However, NN-PSO algorithm shows the best results for Thyroid dataset in accuracy, sensitivity and specificity with 85.68, 77.03 and 91.98 respectively. As per the results, NN-GSO algorithm outperforms other algorithms in Breast cancer, Diabetes, Appendicitis, Hepatitis, Liver, Alzheimer and Dermatology datasets in accuracy and sensitivity with an average accuracy of 97.09, 77.73, 92.13, 77.33, 70.37, 95.59 and 79.97 respectively. NN-GA algorithm outperforms other algorithms with accuracy 80.95 and sensitivity 80.27 for Heart dataset. As can be seen from the Table IV, NN-GSO algorithm has produced the best results in terms of sensitivity specificity and accuracy, on testing datasets and in average as well. Also, it can be noticed that NN-GSO algorithm has small standard deviation compared to other algorithms.

Overall, NN-GSO algorithm shows better result than other algorithms in Breast cancer, Diabetes, Appendicitis, Hepatitis, Liver, Alzheimer and Dermatology datasets because GSO algorithm uses in each epoch explorative and exploitative phase to prevent premature convergence and allows multi-modal surfaces to be efficiently explored. However, NN-GSO

TABLE III. COMPARISON ON MSE RESULTS OF NN-GSO, NN-PSO AND NN-GA ON THE TRAINING AND TESTING SETS

Data Set		**NN-GSO**		**NN-PSO**		**NN-GA**	
		Training Error	**Testing Error**	**Training Error**	**Testing Error**	**Training Error**	**Testing Error**
Breast Cancer	**MEAN**	0.01310	**0.02729**	0.01625	0.03721	0.02283	0.03243
	MIN	0.01290	0.00116	0.01463	0.00146	0.01940	0.00264
	MAX	0.01320	0.06527	0.33498	0.07353	0.02501	0.08697
	STDAV	0.00116	0.02180	0.01461	0.02476	0.00179	0.02524
Diabetes	**MEAN**	0.14838	**0.15650**	0.14473	0.18110	0.17198	0.18864
	MIN	0.14550	0.10363	0.14177	0.14252	0.15823	0.13382
	MAX	0.15410	0.19584	0.22139	0.21819	0.19507	0.24561
	STDAV	0.00280	0.02539	0.00698	0.02256	0.01251	0.03646
Heart	**MEAN**	0.09780	**0.14231**	0.18570	0.08763	0.13060	0.15204
	MIN	0.09520	0.08977	0.12079	0.08017	0.10650	0.09404
	MAX	0.10600	0.24954	0.34086	0.24750	0.15373	0.28738
	STDAV	0.00338	0.04615	0.06862	0.01160	0.01353	0.05803
Hepatitis	**MEAN**	0.08073	**0.16568**	0.05374	0.22422	0.21076	0.25915
	MIN	0.08070	0.08303	0.04286	0.11621	0.19018	0.16649
	MAX	0.08140	0.30362	0.34198	0.47183	0.22937	0.30316
	STDAV	0.00013	0.05891	0.02187	0.10804	0.01407	0.04962
Thyroid	**MEAN**	0.04722	0.05295	0.02036	**0.04546**	0.13991	0.14005
	MIN	0.02630	0.01644	0.01002	0.01032	0.13420	0.10570
	MAX	0.07580	0.08540	0.05374	0.06051	0.14271	0.16201
	STDAV	0.01681	0.02348	0.01292	0.01598	0.00239	0.02186
Alzheimer	**MEAN**	0.02620	0.04633	0.01986	0.06696	0.03765	**0.03851**
	MIN	0.02190	0.00000	0.01537	0.00000	0.02606	0.00439
	MAX	0.03090	0.13514	0.02470	0.18804	0.05996	0.10109
	STDAV	0.00339	0.05291	0.00322	0.06225	0.01013	0.03142
Appendicitis	**MEAN**	0.07323	0.12103	0.02553	0.18110	0.08068	**0.11803**
	MIN	0.05173	0.00861	0.01190	0.14252	0.05474	0.00766
	MAX	0.08887	0.34799	0.08113	0.21819	0.09501	0.36292
	STDAV	0.01276	0.11575	0.02106	0.02256	0.01221	0.12134
Liver	**MEAN**	0.17529	**0.20814**	0.18811	0.21628	0.21842	0.23699
	MIN	0.16796	0.17508	0.18250	0.16903	0.20319	0.21519
	MAX	0.18969	0.23508	0.19688	0.25274	0.23915	0.27375
	STDAV	0.00590	0.02056	0.00461	0.02794	0.01070	0.02411
Dermatology	**MEAN**	0.18748	**0.16665**	0.11720	0.17788	0.25733	0.28699
	MIN	0.16687	0.04269	0.03579	0.03142	0.14592	0.09534
	MAX	0.20753	1.09850	0.15872	1.09239	0.39257	1.18110
	STDAV	0.01128	0.32803	0.03480	0.32330	0.07761	0.31956
Average	**Mean**	0.09438	**0.12076**	0.08572	0.13532	0.14113	0.16143
	STDAV	0.00640	0.07700	0.02097	0.06878	0.01722	0.07640

TABLE IV. COMPARISON OF THE SENSITIVITY (SEN), THE SPECIFICITY (SPE), THE ACCURACY (ACC) OF NN-GSO AND NN-PSO ON THE TRAINING AND TESTING SETS

Data Set		**NN-GSO**						**NN-PSO**					
		Training Set			**Testing Set**			**Training Set**			**Testing Set**		
		ACC	**SEN**	**SPE**	**ACC**	**SEN**	**SPE**	**ACC**	**SEN**	**SPE**	**ACC**	**SEN**	**SPE**
Breast Cancer	**MEAN**	97.85	97.27	98.93	**97.09**	**96.06**	**98.65**	97.51	96.97	98.51	96.21	95.46	97.37
	MIN	97.40	96.52	98.17	92.96	90.48	95.24	97.22	96.48	97.21	92.65	89.36	90.91
	MAX	98.53	97.99	99.54	100	100	100	97.72	97.51	99.09	100	100	100
	STDAV	0.33	0.41	0.49	2.34	3.51	2.17	0.21	0.37	0.57	2.47	4.00	3.14
Diabetes	**MEAN**	77.73	62.38	85.95	**77.73**	**61.70**	**86.37**	74.31	59.65	82.17	72.41	58.01	80.22
	MIN	76.01	59.17	83.00	71.05	51.72	79.59	70.09	32.65	68.94	64.47	26.09	65.22
	MAX	79.62	65.98	87.97	89.47	78.26	94.34	78.07	78.99	93.07	80.26	76.67	92.45
	STDAV	1.12	2.77	1.56	4.69	8.64	5.25	2.83	12.58	7.43	4.71	16.95	9.44
Heart	**MEAN**	86.83	83.44	89.72	80.33	75.91	**84.91**	90.72	87.65	93.34	77.50	76.34	78.96
	MIN	85.08	80.80	86.71	62.07	50.00	70.59	89.18	82.05	91.03	62.07	41.67	61.54
	MAX	88.81	87.20	93.06	89.66	91.67	92.31	93.28	95.04	96.53	86.21	92.31	92.31
	STDAV	1.37	1.85	1.96	7.99	12.91	6.45	1.40	3.97	1.64	7.16	15.19	9.06
Hepatitis	**MEAN**	88.82	96.68	58.67	**77.33**	**88.44**	33.33	93.26	97.37	77.35	74.67	84.48	39.17
	MIN	86.43	90.99	44.83	60.00	75.00	0.00	90.00	91.89	51.72	46.67	58.33	0.00
	MAX	92.14	100	75.86	86.67	100	66.67	95.71	100	93.10	86.67	100	100
	STDAV	1.73	2.77	8.82	7.17	8.42	23.24	1.93	2.64	10.56	12.88	14.37	33.80
Thyroid	**MEAN**	81.78	69.56	85.38	82.59	69.97	85.61	90.72	84.31	96.35	**85.68**	**77.03**	**91.98**
	MIN	66.45	33.18	72.89	66.18	32.98	74.12	68.87	42.80	93.88	66.51	41.67	83.33
	MAX	90.66	86.53	99.00	93.13	92.48	97.92	98.80	98.04	98.25	99.36	99.49	100
	STDAV	8.28	17.39	8.19	7.83	17.45	7.73	9.88	17.11	1.59	9.89	16.06	6.02
Alzheimer	**MEAN**	97.59	86.60	99.83	**95.59**	**84.17**	**98.52**	98.16	89.29	99.75	92.77	74.17	96.20
	MIN	97.18	82.61	99.16	85.71	50.00	92.31	97.20	85.00	99.16	78.57	0.00	84.62
	MAX	97.90	90.91	100	100	100	100	98.59	91.30	100	100	100	100
	STDAV	0.35	2.37	0.35	4.62	21.68	3.13	0.48	2.53	0.40	7.02	40.91	5.41
Appendicitis	**MEAN**	85.75	60.37	93.77	**92.13**	**63.51**	**99.20**	74.31	59.65	82.17	72.41	58.01	80.22
	MIN	60.00	0.00	77.78	89.58	57.90	97.40	70.09	32.65	68.94	64.47	26.09	65.22
	MAX	100	100	100	94.79	72.22	100	78.07	78.99	93.07	80.26	76.67	92.45
	STDAV	14.24	42.44	8.45	1.60	4.43	0.92	2.83	12.58	7.43	4.71	16.95	9.44
Liver	**MEAN**	73.26	87.44	53.82	**70.37**	**87.62**	46.59	75.39	88.95	56.82	68.96	80.03	**52.65**
	MIN	70.65	84.18	47.69	64.71	78.95	31.25	73.23	83.33	42.86	55.88	57.90	31.25
	MAX	75.49	91.53	60.16	76.47	94.74	70.59	77.74	96.05	61.72	76.47	94.74	76.47
	STDAV	1.70	1.93	3.47	3.73	5.85	11.90	1.45	3.30	5.83	6.38	10.71	15.89
Dermatology	**MEAN**	79.98	72.63	80.10	**79.97**	**78.20**	**80.48**	76.66	85.19	76.62	74.23	73.78	74.29
	MIN	67.15	32.56	66.89	66.67	66.33	66.67	68.63	66.67	68.42	66.67	65.67	66.67
	MAX	97.43	100	98.03	100.00	100.00	100.00	94.99	100	95.18	94.44	95.77	95.83
	STDAV	12.04	26.05	12.20	12.60	12.85	12.85	7.98	17.57	8.07	8.80	9.47	9.13
Average	**Mean**	85.51	79.60	82.91	**83.68**	**78.40**	**79.30**	85.67	83.23	84.79	79.43	75.26	76.78
	STDAV	4.57	10.89	5.05	5.84	10.64	8.18	3.22	8.07	4.84	7.11	16.07	11.26

TABLE IV. (CONTINUED)

Data Set		**NN-GA**					
		Training Set			**Testing Set**		
		ACC	**SEN**	**SPE**	**ACC**	**SEN**	**SPE**
Breast Cancer	**MEAN**	97.54	97.23	98.15	96.07	95.65	96.53
	MIN	97.24	96.54	97.24	90.14	89.36	90.48
	MAX	98.05	97.99	99.53	100	100	100
	STDAV	0.23	0.44	0.82	3.01	3.79	3.53
Diabetes	**MEAN**	74.67	57.42	83.92	71.37	53.38	80.97
	MIN	69.94	40.57	77.63	59.21	25.00	71.74
	MAX	78.07	65.98	90.31	78.95	69.57	86.96
	STDAV	2.84	7.88	4.61	6.15	13.47	4.66
Heart	**MEAN**	83.94	83.15	84.64	**80.95**	**80.27**	81.88
	MIN	79.85	75.20	77.08	65.52	50.00	62.50
	MAX	87.31	91.74	93.01	89.66	100	100
	STDAV	2.17	5.72	5.56	8.23	13.00	11.39
Hepatitis	**MEAN**	67.74	43.71	87.07	57.00	36.50	**79.11**
	MIN	62.86	13.56	75.34	26.67	0.00	58.33
	MAX	73.57	71.64	98.77	93.33	80.00	100

	STDAV	3.91	18.09	7.56	17.95	22.69	18.40
Thyroid	**MEAN**	64.37	33.55	66.77	64.51	33.81	66.67
	MIN	64.20	33.31	66.67	63.93	33.33	66.67
	MAX	64.78	34.11	67.16	66.79	38.10	66.67
	STDAV	0.19	0.36	0.17	0.88	1.51	0.00
Alzheimer	**MEAN**	95.98	79.21	99.00	95.49	81.67	98.40
	MIN	93.01	59.09	98.32	85.71	50.00	91.67
	MAX	97.89	90.48	100	100	100	100
	STDAV	1.46	8.60	0.66	5.62	24.15	3.38
Appendicitis	**MEAN**	74.31	59.65	82.17	86.75	60.37	80.22
	MIN	70.09	32.65	68.94	60.00	0.00	65.22
	MAX	78.07	78.99	93.07	100	100	92.45
	STDAV	2.83	12.58	7.43	14.14	42.44	9.44
Liver	**MEAN**	66.04	82.57	43.26	59.61	78.67	33.43
	MIN	59.80	71.51	9.30	50.00	64.71	6.25
	MAX	69.68	96.61	60.87	67.65	95.46	60.00
	STDAV	2.99	7.93	16.07	6.23	11.21	16.78
Dermatology	**MEAN**	79.45	54.33	79.71	77.42	74.72	77.86
	MIN	67.34	31.82	66.69	66.67	66.67	66.67
	MAX	91.72	66.67	92.07	86.08	85.64	85.71
	STDAV	8.21	16.87	8.37	7.00	6.63	6.64
Average	**Mean**	78.23	65.65	80.52	76.57	66.12	77.23
	STDAV	2.76	8.72	5.69	7.69	15.43	8.25

Algorithm doesn't show good results for Thyroid and Heart dataset in accuracy, sensitivity and specificity because of the nature of the data is irrelevant, redundant and has huge features.

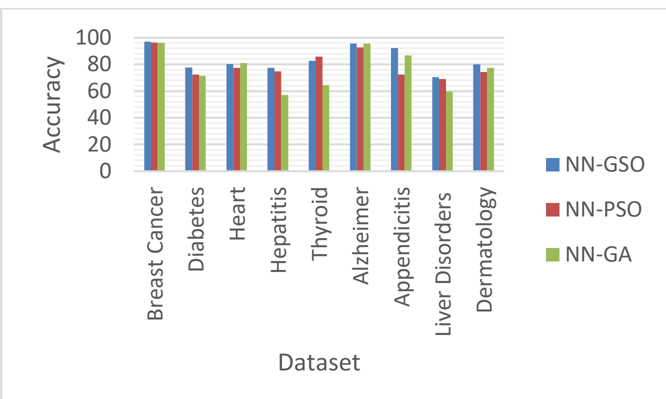


Fig. 7. Comparison of the average accuracy on the testing set.

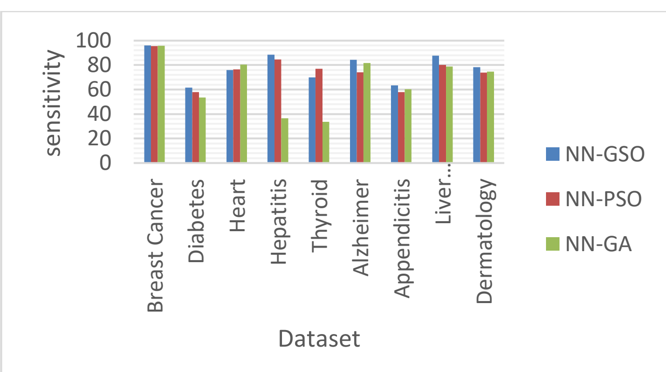


Fig. 8. Comparison of the average sensitivity on the testing set.

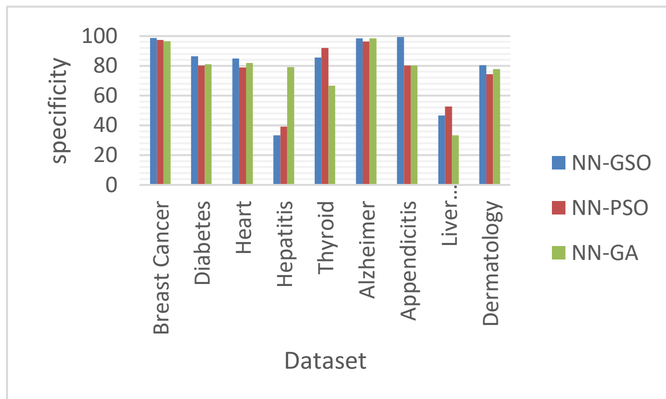


Fig. 9. Comparison of the average specificity on the testing set.

VI. SIGNIFICANCE OF THE RESULTS

As can be seen from the result that NN-GSO algorithm is better in terms of accuracy compared to other algorithms. However, to show that there is a significant difference between the algorithms statistical tests have been applied in this paper. These statistical tests are the Paired t-test test [26] and Wilcoxon's signed-ranks test [27]. They are performed by using 0.05 as a level of confidence (α).

A. Classification Accuracy Rate Analysis

The analysis of the classification accuracy is discussed in this section. Table V presents the average ranks for each algorithm for all datasets. Each algorithm has been ranked by giving the best performance algorithm the rank of 1, the second gets 2 and the third get 3 for each dataset. In case of ties, all tie algorithms get an average rank. Then, we will compare the average ranks between all datasets. As can be seen from Table V, NN-GSO algorithm gets the lower value in the ranking which means that NN-GSO algorithm is the best algorithm. On the other hand, Table VI shows the win, loss and ties count of the number of times that the algorithm is significantly better, loss and tie than other algorithms respectively. As can be seen from this table, NN-GSO algorithm is statistically different from other algorithms in terms of accuracy rate.

The Wilcoxon's signed-ranks test and Paired t-test has been used to analyze the accuracy for the proposed algorithms to determine whether the proposed methods are statistically different, the null hypothesis states that: H_0: there is no difference between the average accuracy of the proposed algorithms. Table VII shows the results of applying Wilcoxon's signed-ranks test and Paired t-test on the proposed algorithms. As can be seen from Table VII, the values of P-value signed rank and P-value t-test of NN-GSO vs. NN-PSO and NN-GSO vs. NN-GA is lower than significance level, a = 0.05. This means that NN-GSO is statistically different from the NN-PSO and NN-GA. Therefore, the null hypothesis, H_0 is rejected. However, the values of the P-value t-test and P-value signed rank of NN-PSO vs. NN-GA are greater than the significance level, a = 0.05 which means that there are no significant differences between the two algorithms. Thus, the null hypothesis is not rejected. The greater than symbol ('>') and equal sign ('=') mean the algorithm on the left side is significantly and no significantly better than the algorithm on the right side respectively.

TABLE V. RANKINGS OBTAINED FOR THE ALGORITHMS CONSIDERING ACCURACY RATE

Algorithm	Ranking
NN-GSO	1.22
NN-PSO	2.33
NN-GA	2.44

TABLE VI. WINE-LOSS-TIE COUNT OBTAINED FOR THE ALGORITHMS TAKING IN TO ACCOUNT THE ACCURACY RATE

Algorithm	Win	Loss	Tie
NN-GSO	7	2	0
NN-PSO	1	8	0
NN-GA	1	8	0

TABLE VII. PAIRED T-TEST AND WILCOXON'S SIGNED-RANKS TEST RESULTS

Algorithm	P-value t-test	P-value signed rank	Significant
NN-GSO vs NN-PSO	0.0408	0.0195	>
NN-GSO vs GAONN	0.0243	0.0117	>
NN-PSO vs NN-GA	0.5201	0.6523	=

B. *Mean Squared Error Analysis*

The analysis of the Mean Squared Error (MSE) is discussed in this section. Table VIII presents the average ranks for each algorithm for all datasets. A smaller value in the ranking represents a better algorithm. Table IX shows the win-loss-tie count for the proposed algorithms in terms of average MSE. The results as illustrated in Table IX show that NN-GSO algorithm is significantly better than other algorithms in terms of MSE.

The null hypothesis is stated as, H_0: there is no difference between the average MSE of the proposed algorithms. Table X shows the results of applying both tests (Paired t-test and Wilcoxon's signed-ranks). The P-value t-test and P-value signed rank values of NN-GSO vs NN-PSO and NN-PSO vs NN-GA are greater than significance level a = 0.05 which means there are no significant differences. Therefore, the null hypothesis H_0 is failed to reject. Otherwise, the P-value t-test and P-value signed rank values of NN-GSO vs NN-GA is lower than a, which means that there are significant differences. Therefore, the null hypothesis is rejected.

TABLE VIII. RANKINGS OBTAINED FOR THE ALGORITHMS CONSIDERING MSE

Algorithm	Ranking
GSONN	1.4
PSONN	2.11
GANN	2.4

TABLE IX. WINE-LOSS-TIE COUNT OBTAINED FOR THE ALGORITHMS CONSIDERING MSE

Algorithm	Win	Loss	Tie
GSONN	5	4	0
PSONN	2	7	0
GANN	2	7	0

TABLE X. RESULTS OF PAIRED T-TEST AND WILCOXON'S SIGNED-RANKS TEST TAKING INTO ACCOUNT MSE RATE

Algorithm	P-value t-test	P-value signed rank	Significant
GSONN vs PSONN	0.1204	0.0977	=
GSONN vs GAONN	0.0164	0.0273	>
PSONN vs GANN	0.1009	0.2031	=

VII. DISCUSSION AND ANALYSIS

Based on the above results, this study has been successfully presented a new hybrid intelligent system for the design of neural network for medical data classification (NN-GSO). By using multiple benchmark medical data sets and many performance metrics for evaluating the effectiveness of the system, the classification results show that NN-GSO algorithm has better and acceptable results compared to NN-PSO and NN-GA algorithms. In addition, the computational results of NN-GSO algorithm have been compared with the results of other algorithms in the literature in terms of classification accuracy.

Table XI and Fig. 10 shows a summary of the comparative results. Note that none of the algorithms that presented in Table XI (MEPGANf1-f3 [12], MLP-BP [13], ISO-FLANN [13], NN-CAPSO [14], NN-GSA [14], NN-ICA [14], NN-BP [15], NN-MVO [15], MODE-ESNN [16], DPM [16], SAE-MR

[17], SAE-ZEROMASK [17], MKSVM [18], PMC [19], DG [19], FA [20] and LFA [20]) were tested on all the datasets that had been used in this study. Also, the results presented here are not fine-tuned in any manner, (i.e., the same parameters and experimental settings are used for all datasets). As can be seen from the table, NN-GSO algorithm outperformed other algorithms for breast cancer, diabetes and appendicitis datasets.

However, for heart and liver disorders datasets NN-GSO algorithm provided comparable results to NN-CAPSO [14] and NN-MVO [15] respectively. For hepatitis dataset NN-MVO [15] outperformed other algorithms. PMC [19] algorithm outperformed other algorithms for thyroid and dermatology datasets, but it performed comparable for appendicitis and liver disorders. However, for Alzheimer dataset, NN-GSO algorithm outperformed DPMS [16], SAE-MR [17], SAE-ZEROMASK [17] and MKSVM [18].

Over all, NN-GSO algorithm is better or at least competitive for breast cancer, diabetes, heart, hepatitis, appendicitis and Alzheimer. On the other hand, NN-GSO performs comparable for the liver, disorders, thyroid and dermatology datasets (with respect to classification accuracy) comparing to other algorithms. In short, it has been shown from the results that the NN-GSO is a suitable algorithm that can be applied to solve classification problems because it shows good performance in terms of classification accuracy for most datasets.

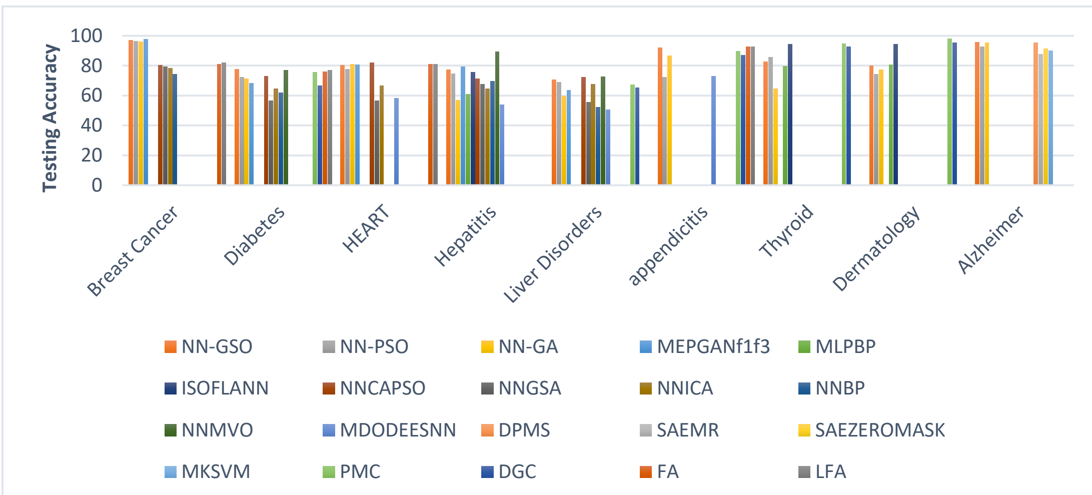


Fig. 10. Performance comparisons of the proposed and existing algorithms on the testing set for all datasets.

TABLE XI. NEURAL NETWORKS CLASSIFICATION ACCURACIES FOR ALL DATASET PROBLEMS WITH CLASSIFICATION ACCURACIES OBTAINED BY OTHER STUDIES

Algorithm / Reference	**Breast Cancer**	**Diabetes**	**Heart**	**Hepatitis**	**Liver**	**Appendicitis**	**Thyroid**	**Dermatology**	**Alzheimer**
NN-GSO	97.09	**77.73**	80.33	77.33	70.73	92.13	82.59	79.97	**95.59**
NN-PSO	96.21	72.41	77.50	74.67	68.96	72.41	85.68	74.23	92.77
NN-GA	96.07	71.37	80.95	57.00	59.61	86.75	64.51	77.42	95.47
MEPGANf1-f3/ [12]	**97.80**	68.35	80.79	79.38	63.50	-	-	-	-
MLP-BP/ [13]	-	-	-	60.83	-	-	79.77	80.63	-
ISO-FLANN/ [13]	-	-	-	75.72	-	-	94.47	94.43	-
NN-CAPSO / [14]	80.25	72.99	**81.85**	71.29	72.32	-	-	-	-
NN-GSA/ [14]	79.25	56.43	56.67	67.74	55.65	-	-	-	-
NN-ICA / [14]	78.25	64.61	66.67	64.52	67.68	-	-	-	-
NN-BP/ [15]	74.41	61.98	-	69.62	52.20	-	-	-	-
NN-MVO/ [15]	-	76.79	-	**89.43**	**72.46**	-	-	-	-
MDODE-ESNN/ [16]	-	-	58.20	54.00	50.57	73.00	-	-	-
DPMS / [16]	-	-	-	-	-	-	-	-	95.35
SAE-MR/ [17]	-	-	-	-	-	-	-	-	87.79
SAE-ZEROMASK / [17]	-	-	-	-	-	-	-	-	91.40
MKSVM [18]	-	-	-	-	-	-	-	-	90.11
PMC/ [19]	-	75.65	-	-	67.25	89.62	**94.67**	**98.09**	-
DGC/ [19]	-	66.62	-	-	65.22	87.13	92.56	95.44	
FA / [20]	80.88	76.04	80.88	-	-	**92.50**	-	-	-
LFA/ [20]	81.94	77.08	80.88	-	-	92.05	-	-	-

VIII. CONCLUSION AND FUTURE WORK

In this paper we have proposed a new a hybrid intelligent system as medical decision support tool for medical diseases predication and classification based on the neural network, galactic swarm optimization (NN-GSO). The effectiveness of the proposed algorithm has been evaluated by using multiple of benchmark medical data sets which are Diabetes, Liver Disorders, Heart, Breast cancer, Hepatitis, and Appendicitis datasets which represent binary class classification problems, while Thyroid, Dermatology and Alzheimer represent multiclass classification problems. Experimental results have shown that the proposed algorithm gets better classification compared with NN-PSO and NN-GA algorithms. Multiple of statistical tests have been used to analyze the accuracy of the proposed algorithm. In conclusion, it has been shown that NN-GSO approach is a suitable algorithm that can be employed to solve complex classification problems. For future work, we will focus on improving NN-GSO by training other types of neural networks such as recurrent neural network and implementation of NN-GSO algorithm with different fitness functions. Furthermore, GSO algorithm can be improved for multi-objective algorithm to optimize the structure, number of connections and learning of ANN simultaneously to avoid the problem of trial and error.

ACKNOWLEDGMENT

This work is supported by Al Imam Mohammad Ibn Saud Islamic University (IMSIU) under Research University Grant Scheme. The authors would like to thank Deanery of Academic Research, IMSIU, for the research activities, and Computer Science Department, College of Computer and Information Sciences, IMSIU for the support and incisive comments in making this study a success.

REFERENCES

[1] Q.Al-Shayea. “Artificial Neural Networks in Medical Diagnosis” IJCSI International Journal of Computer Science Issues, Volume 8, Issue 2, 2011.

[2] R. Chitra and V. Seenivasagam” Risk Prediction of Heart Disease Based on Swarm Optimized Neural Network” Proceedings of International Conference on Computer Science and Information Technology, Advances in Intelligent Syste ms and Computing, 2014.

[3] C.Blum and K.Socha “Training feed-forward neural networks with ant colony optimization An application to pattern classification” Fifth International Conference on Hybrid Intelligent Systems, 2010.

[4] SeyedAli, S. Hashim and H. Sardroudi “Training feedforward neural networks using hybrid particle swarm optimization and gravitational search algorithm” Applied Mathematics and Computation, 2012.

[5] H. R.H.Al-Absi1 , A.Abdullah ,M. I.Hassan, and K. B. Shaban. “Hybrid Intelligent System for Disease Diagnosis Based on Artificial Neural Networks, Fuzzy Logic, and Genetic Algorithms” series Communications in Computer and Information Science, Volume 252, pp 128-139,2011.

[6] V.M.Nakarajan and M.M.Noel. “Galactic Swarm Optimization: A new global optimization metaheuristic inspired by galactic motion” Applied Soft Computing Volume 38, 2016.

[7] Z. Beheshti, S.Hj. Shamsuddin, E.Beheshti and S.S.Yuhaniz.” Enhancement of artificial neural network learning using centripetal accelerated particle swarm optimization for medical diseases diagnosis” Methodologies and Application Soft Computing,2014.

[8] Y.E. Shaoa, C.Houa and C.Chiu. “Hybrid intelligent modeling schemes for heart disease classification” Applied Soft Computing Volume 14, 2014.

[9] S. Qasem, S. Shamsuddin and A. Zain “Multi-objective hybrid evolutionary algorithms for radial basis function neural network design” Knowledge-Based Systems 27, 2012.

[10] Jiake Lv et al “A hybrid approach of neural network with particle swarm optimization for tobacco pests prediction” IFIP International Federation for Information Processing, 2009.

[11] T.Masters “Practical neural network recipes in C++” San Diego [u.a.]: Acad. Press, 1999.

[12] S.Qasem and S.Shamsuddina “Memetic Elitist Pareto Differential Evolution algorithm based Radial Basis Function Networks for classification problems” Applied Soft Computing 11, 2011.

[13] S.Dehuri, R.Roy, S.Cho and A.Ghosh “An improved swarm optimized functional link artificial neural network (ISO-FLANN) for classification” The Journal of Systems and Software 85, 2012.

[14] Z.Beheshti, S. Shamsuddin , E.Beheshti · S.Yuhaniz “Enhancement of artificial neural network learning using centripetal accelerated particle swarm optimization for medical diseases diagnosis” Soft Computing, December 2014.

[15] H.Faris, I.Aljarah, S.Mirjalili “Enhancement of artificial neural network learning using centripetal accelerated particle swarm optimization for medical diseases diagnosis” Applied Intelligence , 2016.

[16] A. Saleh, S. Shamsuddin1and H. Hamed “Multi-Objective Differential Evolution of Evolving Spiking Neural Networks for Classification Problems” IFIP International Federation for Information Processing, 2015.

[17] H. Suk, S. Lee and D. Shen “Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis” NeuroImage, 2014.

[18] S. Liu et al. "Multimodal Neuroimaging Feature Learning for Multiclass Diagnosis of Alzheimer's Disease," in IEEE Transactions on Biomedical Engineering, volume. 62, no. 4, pp. 1132-1140, 2015.

[19] M. A. Mazurowskia , P.A. Habasa, J.M. Zuradaa, J.Y. Lob, J.A. Bakerb and G.D.Tourassib. “Training neural network classifiers for medical decision making The effects of imbalanced datasets on classification performance” Neural Networks 427–436,2008.

[20] M. Alweshah and S. Abdullah “Hybridizing firefly algorithms with a probabilistic neural network for solving classification problems” Applied Soft Computing 35, 2015.

[21] "UCI Machine Learning Repository: Data Sets", Archive.ics.uci.edu, 2018. [Online]. Available: https://archive.ics.uci.edu/ml/datasets.html. [Accessed: 27- Mar- 2018].

[22] "OASIS Brains - Open Access Series of Imaging Studies", Oasis-brains.org, 2018. [Online]. Available: http://www.oasis-brains.org. [Accessed: 27- Mar- 2018].

[23] A. Azzini “A new genetic approach for neural network design and optimization”, Ph.D. dissertation, The University of Milan 2005.

[24] J. Kennedy, R.C. Eberhart, Particle Swarm Optimization, in: Proceedings of IEEE International Conference on Neural Networks, IV, Piscataway, 1995, pp. 1942–1948.

[25] J.H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, 1975.

[26] G. Box, J. Hunter and W. Hunter, Statistics for experimenters. Hoboken, N.J.: Wiley-Interscience, 2005.

[27] F. Wilcoxon “Individual comparisons by ranking methods. Biometrics Bull”,1945.

Unsupervised Ads Detection in TV Transmissions

Waseemullah[1], Najeed Ahmed Khan[2]
Computer Science & IT Department
NED University of Engineering & Technology
Karachi, Pakistan

Umair Amin[3]
We4do Solutions Karachi
Pakistan

***Abstract*—A novel framework is presented that can segment semantic videos and detect commercials (ads) in a broadcasted TV transmission. The proposed technique combines SURF features and color Histogram in a weighted combination framework resulting in detecting individual TV ads from the transmission after segmenting semantic videos. Thus, better results are achieved. The proposed framework is designed for TV transmissions those who do not use black frame technique between the ad and non-ad part of the transmission and is commonly used in Pakistani TV channels transmission. The television transmission standards in Pakistan are different from those that are used in other countries of the world. The framework used unsupervised technique to segment the semantic videos.**

***Keywords*—TV ads; video segmentation; semantic analysis; ad segmentation; unsupervised segmentation**

I. Introduction

Semantic analyses involve the extraction or understand the meaning of the subject that is being analyzed and finding relations among unstructured data. It is most commonly used in analyzing text data and Natural Language processing system.

The amount of video information that is being generated now is more than it has been ever before. From TV transmission to CCTV footage and videos made by cell phones everything is adding to the amount of video data present. Therefore, it is important to have tools that can automatically extract meaningful information from these videos.

Advertisements displayed in broadcast TV transmissions are a very important part of transmission as major revenue for a broadcaster is generated by advertising. However, the advertiser is keen in statistics of the broadcasted ad in the transmission. Knowing who what and where advertising can be useful information to measure the market trends and forming business strategy.

If this statistics computed automatically and reliably then things will easy for both the broadcaster and advertiser as well for the market trends analyzer.

The term semantics is a broad term and can be used to measure similarity in different domains such as sports, scenery, image etc. The proposed framework chose for the automatic ads detection that appears in a broadcasted television transmission. The developed framework can differentiate between ads and non-ad part in video of transmission as well as compute the statistics of broadcasted an ad in the transmission. Several research works have been attempted commercials detection and identification however; follow the USA and European television transmission standard which follow black frame transmission technique [1]. There are TV channels transmission e.g. in Pakistan that do not follow appearance of a black frame between the ad and a non-ad part of the transmission.

The developed framework useable for TV transmissions those do not use black frames technique in between ad and non-ad video part in the transmission like in Pakistan TV transmission. Techniques are common such as in [1] use of a black frame between ad and non-ad part in video to detect boundary of an ad. Another technique to detect boundary of an ad in video part is the nonappearance of TV channel logo. During the commercial breaks the TV channel logo is removed that to commercials video part and non-commercial video part in TV transmission [2]. This technique also does not use in Pakistani TV transmissions. Fig. 1 shows the some examples labeled in the dataset of ads that contains different products images.

This paper contains 6 sections including this section. Section I described Introduction of framework. Section II described related work that has been done in relation to video segmentation and semantic analysis for ad detection. Section III described implementation of framework. Section IV presented results of experiments that were conducted. Section V presented conclusion and future work.

II. Literature Review

Most American and European television broadcasters often use the black frame as standard for the transmission for separation of TV advertisements from the other programs. Many researchers have worked on TV adds detection using the black frame in TV transmission (Fig. 1).

Fig. 1. Examples (a-l) from the dataset of ads contain different products images/information.

In [2] L. Y. Duan et al. worked on hassle of recognizing and categorizing TV ads from TV video streams. For boundary detection in each commercial a multimodal approach is followed. Absence of TV channel logo is used for separating a normal transmission program from a TV ad and also used for segmentation of ads. In addition to that Hidden Markov Model was trained using audio and visual feature for TV ad segmentation. Results showed the precision and recall were 90% and 80% respectively.

In research findings [3] Covell M. et al. investigated problem of extracting video fragments that are appeared frequently in video streams. Their real goal was the identification of those TV ads those are identified with telecasters possess programs, for example, programs that will be played one week from now in a single day at an alternate time. Their suggested approach has three noteworthy advances sound redundancy identification, endpoint discovery and discovery of endpoint. Experiment ran video recording of four days on video taken of different TV channels. An precision and recall was achieved of 85% and 94% respectively for sound coordinating part. Precision and recall was 92% and 93% achieved after matching video streams announced and finally outcome values for precision and recall were 99% and 95% respectively.

In [4] the issue of recognizing TV ads in MPEG compressed video footage is investigated. It uses features extracted using MPEG parameters. TV ads detection is achieved using black frame, unicolor frame and in addition to that changed in aspect ratio. It was additionally watched that one minute was length for each TV ads break. For evaluation eight hour length video stream was recorded and appearance of black frame in TV ads break was observed to be the most appropriate advertisement recognizing parameter.

In [5] issue of accessing TV Ads by considering salient semantics is addressed through semiotic point of view. Critical, playful, utopic and practical were recognized as semiotic classes. 150 ads recorded from Italian TV networks and examined to check that groupings made by framework are same as obtained via human specialists. For playful ad highest accurate results were obtained for practical least accurate results were obtained due to wrongly recognizing product in foreground.

In [6] the video contents were analyzed for real time indexing. The proposed achieves video indexing by taken into account the video contents using Hidden Markov Model (HMM). The HMM accepts input feature that is based on image difference that indicates location target object in shot. In addition to that grey level histogram, intensity of motion and average motion deviation features are used in this approach. For assessment videos from 12 news programs obtained from various German TV networks were examined. Half of them were used in training and half used in testing. Nine different classes were recognized. Seven from total of nine had detection rate more noteworthy than 80%. The experimental showed that detection value for short news is higher than long news programs.

El-Khoury et al. [7] suggested a strategy for automatic extraction from TV content based on signals. His technique can be applied on sound, visual contents or a mix of sound and visual contents by making utilization of Bayesian Information Criterion and general likelihood ration. This framework was assessed on recorded videos from French TV and furthermore using TRECVid dataset. Recall value of 93% and 89% was observed for ARGO and TRCVid respectively whereas precision value of 93% and 91% was examined for ARGO and TRECVid respectively.

In [8] X. S. Hua et al. suggested a learning method for recognition of TV advertisements. His proposed method performs classification through Support Vector Machine on the basis of audio and video features. For visual features variance and average of frame difference and variance of edge change ratio were used. For audio Mel frequency Cepstral Coefficient and shot time energy were taken into account. For the purpose of evacuating of shots that have little duration, checking of long ads and refining TV ads limits. For assessment TV transmission recordings of duration10.75 hours used gathered from various TV channels including: NBC, CNN and ESPN2. Recall value 88.21% and precision value 89.39% seen without post preparing and after post handling, it expanded to 91.65%.

In [9] S. C. Chen discussed the issue of identification of scene transformation in video streams by utilization of sound and pictorial data that is accessible in video. The suggested strategy comprises of first deciding shot limits and then unsupervised segmentation using object tracking. During experiment many video recordings were made from different TV stations. The results reported after experiments were 89% and 92% for the average recall and average precision respectively.

In [11] the issue of identifying feature films in different types in light of their preview is investigated. They have grouped movies into four types called Action, Drama, Horror and Comedy depending on visual cues. The proposed approach comprised on group of input features utilized to lessen variance among points of same class and maximize difference among points of other class. They computed average shot length feature in Hue, Saturation, and Value of brightness space. For experiment more than hundred previews of movies were downloaded from Mac website and grouped in four classes. There experiments on 101 videos showed 17 video incorrectly categorized.

In [12] K. Schöffmann et al. focused on the issue of detecting TV ads in recordings coded in H.264/AVC. This approach is different from other existing approaches as it works straight on compressed streams as opposed to uncompressing video. The proposed approach makes utilization unavailability of TV station at the time of TV ads which is valid for European and especially German TV. For assessment of results recording from 19 TV stations was utilized that were prevalently related German watchers. Results showed that recall value and precision value were reported as 97.33% and 99.31% respectively.

In [13] J.R. Smith et al. the issue of automatically marking on TV transmission for later pursuit and indexing is addressed because annotating by human is costly and tedious job. For audio, text and video analysis a multimodal approach is applied. Their framework combines the text, audio and video data for video stream annotation. Using machine learning techniques semantic models library is developed using training dataset. Little human intervention involved for small portion of dataset in training stage. Framework enabled clients to search videos either by feature, choosing key frames, semantically and via text. With the end goal of assessment TRECVID benchmark is used. 90% precision was obtained.

N. Venkatesh et al. [14] suggested a novel technique for TV commercial detection based on cookery programs. TV ads are identified using audiovisual features appeared in transmission. At first sound features are utilized for recognizing begin and end of TV ad break. At that point program name logo is compared with begin and end of commercial break. The audio features i.e. short time energy as well as Zero crossing rate are considered. Visual analysis is carried through edge identification and corner.

In [15], Bingqing Qu et al. investigated the content based segmentation of TV programs through grammatical inference for the sake of video indexing. They proposed symbolic approach that finds a common structure model for repeating programs. Each structural element is having distinct syntactic meaning from other video segments. This distinct structure is represented by a symbol and used in grammatical inference. Their defined approach needs minimal information for segmentation of TV video streams which is based on two steps. In first step symbols are assigned to similar structures of programs. In second step, for searching particular TV program grammatical inference is taken into account using symbolic representation of each structural element. The experiment was performed on four distinct types of repeating programs like talk show, news, magazine and game. The results were obtained using three methods called base line, multiple sequence alignment (MSA) and uniform resampling (UNR) methods. UNR was better than baseline in results whereas MSA out performed UNR in situation where program duration is much lengthy. Precision and recall value for news, talk show and games and magazine were 69% and 54%, 53% and 54%, 69% and 50%, 42% and 64% respectively for MSA method.

In [16] Xiaomeng Wu, et al. suggested a system based on detecting and locating a TV commercial in video. They used Hashing algorithm for detection and localization of commercial. The system performance claimed consistent for both video and audio, high speed and clustering is completely unsupervised. Extraction is improved by algorithm that is responsible for integrating audio and video footage. Accuracies were 97.4% and 98.1% for frame level and sequence level during testing. This system is based on new BOFs model as it smart enough to handle decoding errors. The system is tested on ten hours, a month and five year video effectively and efficiently.

III. Framework

The proposed framework starts with the add detection in TV recorded transmission that contains recording of different programs including ads (commercials). This section address the approach used for detection of ads in video if present.

A. Ad Break Detection

To segment the videos automatically into ad and non-ad, we devised an algorithm. It calculates variance of a window before and after a point. The difference in variance between left window and right window is greater at the beginning and end of ad-break. So it can be used to segment video between ad and non-ad parts.

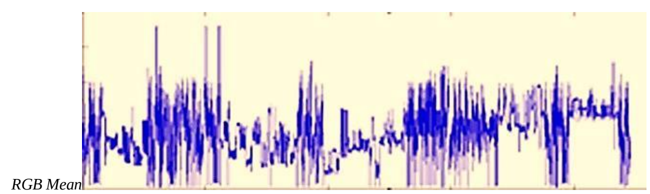


Fig. 2. Plot of RGB for each frame Vs RGB mean where ad breaks had higher distortion in graph and the portion that corresponds to normal transmission had fewer peaks per second.

In the first step for the frames in video RGB mean is calculated and plot them along with the corresponding frame numbers. Fig. 2, 3, and 4 show the RGB mean and the frame numbers. Based on the variance shown in the graph, it is possible to see that there was a pattern being followed by ad breaks and normal transmission where ad breaks had higher distortion in graph and the portion that corresponds to normal transmission had fewer peaks per second.

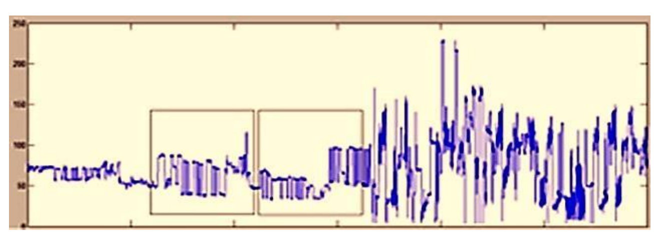


Fig. 3. Left and right window in non_ad region where difference in variance between left window and right window is almost same.

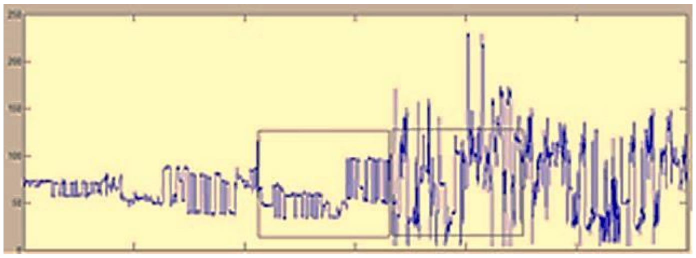


Fig. 4. Left and right window at split region where difference in variance between left window and right window is greater at the beginning and end of ad-break.

In the second step a window size is defined that calculate variance of RGB mean. For each point the variances of that window before that point and after that point is calculated and stores the difference of that variance.

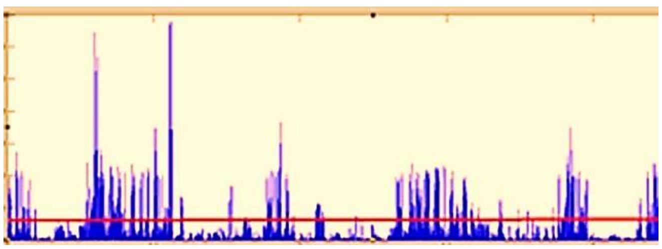


Fig. 5. Plot after step 2 majority of values above red line correspond to ad region using window size 100 frames.

In the next step video is split into ads and non-ad region based on calculating variance, method described above. Main consideration here is that when a value found greater than a threshold value, some other frames seek ahead to make sure that the peak observed is not a false peak. There are several values greater than threshold. Fig. 5 shows add region above the red line after setting a threshold value. Using parameters defined in Table I, we classify values as a sequence of low and high values.

TABLE I. PARAMETERS AND THEIR DESCRIPTION

Parameter	***Description***
Thresh	Threshold to classify a value as low or high
fLowSeek	Number of frames to look ahead when looking from low to high
nLowSeek	Minimum number of high frames that should be present in search space for classification from low to high
fHighSeek	Number of frames to look ahead when going from high to low
nHighSeek	Maximum number of high frames that can be present in search space for it to be classified as change from high to low
fHighSeek2	Minimum number of consecutive low frames for classification from high to low

Parameters and their description were used for classifying values as a sequence of low and high values.

Based on the value of parameters following is established:

- If Value is greater than the threshold mark it as high value if it is lower than threshold mark it as low value.
- If current value is low and previous value was also low mark it as part of same sequence.
- If current value is high and previous was also high mark it as current high sequence.
- If current value is high and previous value was low seek for next “fLowSeek” frames and count number of frames that have value higher than threshold. If that number is greater than “nlowSeek” we mark as a change point from low to high.
- If current value is low and previous value was high we check next “fHighSeek2” frames. If anyone frame in this range is found to be high we consider this value as part of existing high sequence. Otherwise we seek next fHighSeek2 frames, if number of high frames in these is less than nHighSeek we consider this as point of change from high to low sequence otherwise we make this frame as part of existing high sequence and continue.

The proposed Ad Break Detection algorithm was evaluated using the following parameters on a 2.5 hours video recorded from a local TV channel (ARY Digital-Pakistan) at 25 fps.

TABLE II. PARAMETERS AND THEIR CHOSEN WEIGHTS

Parameter	***Value***
Window Size	100
Thresh	800
fLowSeek	2500
nLowSeek	300
fHighSeek	3500
nHighSeek	55
fHighSeek2	800

Parameters and their chosen weights were used for evaluating Ad Break Detection algorithm.

Using these values shown in Table I, the video was split into 15 segments, 7 of them were ad segments and 8 were marked as non-ad. Total length of ad segment was 1 hour 2 minutes. Out of them 4 minute 30 seconds were non-ads that were classified as ads.

Remaining duration was classified as non-ads in them total 4 minutes of content were composed of ads that were wrongly classified as ads.

The (Ad Break Detection algorithm) was again evaluated on altered the values of the parameters in Table II. The altered values are shown in Table III.

TABLE III. PARAMETERS AND THEIR ALTERED WEIGHTS

Parameter	***Value***
Window Size	100
Thresh	700
fLowSeek	2500
nLowSeek	300
fHighSeek	3500
nHighSeek	55
fHighSeek2	2000

Parameters and their altered weights were used for evaluating Ad Break Detection algorithm

Using these values the video was again split into 15 segments, 7 of them were ad segments and 8 were marked as non-ad. Total length of ad segment was 1 hour 4 minutes. Out of them 5 minute 30 seconds were non-ads that were classified as ads.

Remaining duration was classified as non-ads in them total 2 minutes of content were composed of ads that were wrongly classified as ads. This shows choice of parameters values shown in Table II is good as it gives less wrong classification of ads.

B. Ad Boundary Detection

After video segmentation into ad and non-ad parts, it is aim to find automatically the boundaries of individual ads. This task is easy when a black frame exist at the beginning and at the end of each ad part in the video segment. However new techniques are required if black frame not exist a TV transmission, like in Pakistan. Therefore, a novel technique for the ad and non-ad detection automatically in such kind of TV transmission is proposed.

Since the audio was not a good feature to separate ads, as there was no real silence present in between ads. Also several times inside a same ad, block of silence are found that cause wrong attribution of ad boundary. Therefore, instead of audio a new approach based on repeating patterns of scenes was used that assigned ID to each scene and determine ad boundary. The frame work of the proposed technique is shown in Fig. 7.

To detect an ad boundary we employ a method that combines histogram with detecting repeating patterns of scenes in video, as the commercial ads are usually repeated several times during transmission. In the first part of ad boundary detection histogram is calculated for all frame having 30 bins (set threshold value after initial experiments) for each color and scene detection are performed based on histogram. To verify or determine a frame belong to a same scene or not, histogram difference of change point for last 05 scenes is calculated, if appropriate match was found then the frame considered belongs to an existing scene. Fig. 6 shows an output example of different frames belongs to a same scene.

Fig. 6. Frames that marked as belonging to a same scene based on color histogram.

The working of algorithm is divided into two steps.

In first steps scenes are generated from whole video by making use of color histogram. After a scene is identified, it is compared with other scenes that have been already generated from the video. The aim is to assign same scene ID for those scenes which are similar but having not exactly same scene ID. If a new scene appears that was not presented earlier, a new scene ID is assigned. Whereas if that scene already observed at earlier location in video it is added to the scenes list belonging to existing scene ID.

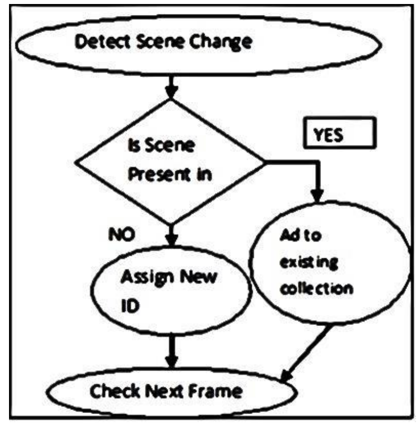


Fig. 7. Proposed framework for boundary detection of an ad.

In second step repeating sequences of scenes are detected. This step allows detecting of unique ads because normally an ad is composed of several scenes and an ad appear several time in a video stream of TV transmission as compared to non-ad transmission. Therefore, if an ad is composed of for example scene number “4, 5, 6” then we find this pattern multiple times in video stream. Thus, if multiple times in video we find that scene 5 comes after scene 4 and scene 6 come from scene 5, it is concluded that these three scenes are related. And because this is a repeating sequence, therefore it is a strong candidate for being an ad.

The details for second step are as follows:

Scenes those have length less than 25 frames are eliminated. Starting from scene that has highest number of occurrences in videos, a repetition threshold is defined as follows:

$$\Theta = (\lambda / 10) \times 3$$

Where, Θ - represents Count Thresh.

λ- Represents Occurrences of pattern.

If Count Thresh is less than three (03) it is increased to three (03). Longest such pattern where number of occurrences of pattern is greater than Count Thresh is searched.

If search is successful then all occurrences of that pattern are marked as checked and added to collection of discovered ads. This process is repeated for remaining scenes by taking unchecked scenes in order of their frequency.

Each discovered pattern generally represents an ad that was repeated several time during the video. First scene of pattern is generally first frame of ad and last scene of ad.

C. Ads segmentation

For Ad segmentation RGB histogram is calculated for all the frames in the video segment and then split ad into scenes by using histogram difference. If histogram difference of two consecutive frames is greater than threshold the point is considered as a change point for scene.

Inside each scene we define key frames. Key frames are those frames whose histogram difference is greater than the threshold value compared to the previous key frame. Using key frame, each ad is divided into scene with each scene having histograms of all frames and key frames marked inside them. To improve the detection and to eliminate false positives, user asks to select a key object that is used for computation of SURF (Speed Up Robust Transform) feature. The key object should be the main object in the ad, which is significant with respect to the product ad. Example of a selected key object is shown in Fig. 8.

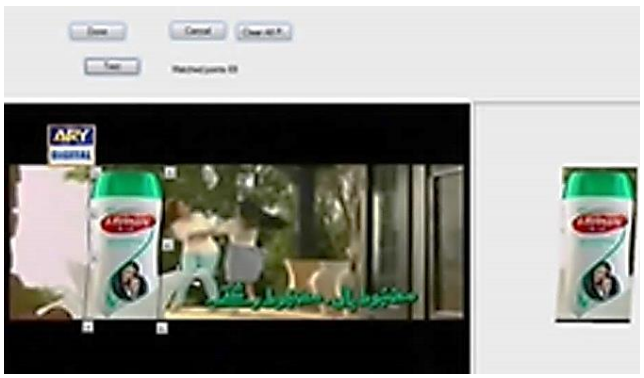


Fig. 8. Key object in an ad for SURF feature computation.

Multiple objects in a same ad can be selected for computation of SURF feature but one object is enough to eliminate false positive ad detections. After this the computed training data can be saved inside a training file that will be used in detection stage.

IV. RESULTS

This section will describe results and outcome statistics of the proposed algorithm evaluated on test dataset. Test dataset composed of 2.5 hours video transmission segment that was recorded from ARY Digital-Pakistan TV channel.

For evaluation of the proposed algorithm, first the ads were marked manually to collect ground truth and then ad detection algorithm was executed on the test dataset segment. To evaluate the performance of the proposed algorithm, automatic labeled outcome results are compared with the ground truth. The target was to find those ads that appear at least 3 times in the provided test video segment. Standard quantitative measures Precision, Recall and F1 value [10] are used to measure the performance of proposed algorithm. Table IV shows the quantitative output values of Precision, Recall and F1 values.

TABLE IV. PRECISION, RECALL AND F1 VALUES

Test Video	*True +ve*	*False + ve*	*False -ve*	*Precision*	*Recall*	*F1-Value*
2.5 hours video transmission ARY Digital-Pakistan	1325	50	116	96%	92%	93%

In Table IV, Columns, True +ve shows number of correct detections duration in seconds, False +ve shows number of incorrect detections duration in seconds, False –ve shows number of missed detections duration in seconds for ads.

Table IV shows Precision; Recall and F-1 Value results evaluated on a Pakistani TV channel (**ARY Digital-Pakistan**) recorded video of duration 2.5 hours. Average precision 96% and average recall rate is 92% and F1 value is 93%.

V. CONCLUSION AND FUTURE WORK

The proposed methods incorporate video segmentation techniques based on color Histogram and detect commercials by computing SURF descriptors feature from TV transmission. The experimental results show that a SURF feature provides better segmentation results for the commercials in video.

The framework can be extended for multiple ads detection and identification from which any particular TV commercial can be identified. In addition to commercial identification, statistical analysis can be performed to know about the repeated occurrence of a TV commercial in given video stream. It can also be converted into a complete media monitoring and commercials verification package. This will be suitable for the analysis of commercials in a TV transmission where other solutions are not available.

REFERENCES

[1] Chen, Shu-Ching, et al. "Scene change detection by audio and video clues." Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on. Vol. 2. IEEE, 2002.

[2] Colombo, Carlo, Alberto Del Bimbo, and Pietro Pala. "Retrieval of commercials by semantic content: the semiotic perspective." Multimedia Tools and Applications 13.1 (2001): 93-118.

[3] Covell, Michele, Shumeet Baluja, and Michael Fink. "Advertisement detection and replacement using acoustic and visual repetition." Multimedia Signal Processing, 2006 IEEE 8th workshop on. IEEE, 2006.

[4] Dimitrova, N., et al. "Real time commercial detection using MPEG features." Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowlwdge-based Systems (IPMU2002). 2002.

[5] Duan, Ling-Yu, et al. "Segmentation, categorization, and identification of commercial clips from TV streams using multimodal analysis." Proceedings of the 14th ACM international conference on Multimedia. ACM, 2006.

[6] Eickeler, Stefan, Andreas Kosmala, and Gerhard Rigoll. "A new approach to content-based video indexing using hidden markov models." Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS). 1997..

[7] El-Khoury, Elie, Christine Sénac, and Philippe Joly. "Unsupervised segmentation methods of TV contents." International Journal of Digital Multimedia Broadcasting 2010 (2010).

[8] Hua, Xian-Sheng, Lie Lu, and Hong-Jiang Zhang. "Robust learning-based TV commercial detection." Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on. IEEE, 2005.

[9] Lienhart, Rainer, Christoph Kuhmunch, and Wolfgang Effelsberg. "On the detection and recognition of television commercials." Multimedia Computing and Systems' 97. Proceedings., IEEE International Conference on. IEEE, 1997.

[10] Khan, Najeed Ahmed, and David C. Hogg. "Unsupervised Learning of Object Detectors for Everyday Scenes." Training 10 (2014): 11.

[11] Rasheed, Zeeshan, Yaser Sheikh, and Mubarak Shah. "On the use of computable features for film classification." IEEE Transactions on Circuits and Systems for Video Technology15.1 (2005): 52-64.

[12] Schöffmann, Klaus, Mathias Lux, and Laszlo Böszörmenyi. "A novel approach for fast and accurate commercial detection in H. 264/AVC bit streams based on logo identification." International Conference on Multimedia Modeling. Springer, Berlin, Heidelberg, 2009.

[13] Smith, John R., et al. "Learning and classification of semantic concepts in broadcast video." Proceedings of the International Conference of Intelligence Analysis. 2005.

[14] Venkatesh, N., B. Rajeev, and M. Girish Chandra. "Novel TV commercial detection in cookery program videos." Proceedings of the World Congress on Engineering and Computer Science 2009 Vol II, WCECS 2009. 2009.

[15] Qu, Bingqing, et al. "Content-based unsupervised segmentation of recurrent TV programs using grammatical inference." Multimedia Tools and Applications 76.21 (2017): 22569-22597.

[16] Wu, Xiaomeng, and Shin'ichi Satoh. "Ultrahigh-speed tv commercial detection, extraction, and matching." IEEE Transactions on Circuits and Systems for Video Technology23.6 (2013): 1054-1069.

Performance Evaluation of IPv4/IPv6 Transition Mechanisms for Real-Time Applications using OPNET Modeler

Khalid EL KHADIRI[1], Ouidad LABOUIDYA[2], Najib ELKAMOUN[3], Rachid HILAL[4]
STIC Laboratory
Chouaib Doukkali University
El Jadida, Morocco

***Abstract*—The problem of the potential depletion of IPv4 addresses has given rise to the development of a new version of the Internet Protocol named IPv6. This version of the protocol offers many improvements, including an increase in the address space from 2^{32} to 2^{128} and improvements in security, mobility, and quality of service. However, the transition from the current version to the new version (IPv4 to IPv6) is complicated and cannot be performed in a short time. The size and complexity of Internet make this migration task extremely difficult and time-consuming. The Internet Engineering Task Force (IETF) took into account this migration problem and proposed transition mechanisms as temporary solutions allowing IPv4 to coexist and operate in parallel with IPv6 networks. The dual stack, manual tunnel, and 6to4 automatic tunnel appear to be promising solutions depending on their characteristics and benefits. In this paper, we study the performance of these transition mechanisms on real-time applications (VoIP and Video Conferencing) using the network simulator OPNET Modeler. Performance parameters such as delay, delay variation, jitter, MOS, and packet loss are measured for these transition mechanisms. The obtained results showed that the dual stack transition mechanism gave better network performance than the tunneling mechanisms.**

***Keywords*— Dual stack; manual tunnel; 6to4; OPNET; VoIP; video conferencing**

I. Introduction

IPv4, the current version of the Internet Protocol (IP), is the version on which the Internet cloud is currently based [1]. IPv6 is the new version of the protocol, developed by the IETF, which has many advantages, including addressing, auto-configuration, mobility, quality of service, and security [2]. Furthermore, IPv4 and IPv6 are two protocols that are not compatible. Therefore the transition from IPv4 to IPv6 is complicated and is not a project that will succeed overnight. The deployment of IPv6 can only be done gradually (step by step) and can be divided into three phases as follows. Phase I: IPv6 is an island in the IPv4 ocean, where IPv4 dominates the global network. Phase II: IPv4 will become an island while IPv6 will be oceanic a few years later. That means that at this point IPv6 will be much wider-ranging than IPv4. The final phase, Phase III: most networks will be in native IPv6. During the transition phase from IPv4 to IPv6, a number of transition and migration mechanisms, proposed and implemented by the IETF, can be used. Most of them have advantages and disadvantages depending on their deployments.

In this research paper, three IPv4/IPv6 transition mechanisms were examined. Namely: the dual stack, the manual tunnel, and the 6to4 automatic tunnel. These mechanisms were evaluated on a simulation network infrastructure under OPNET Modeler using two real-time applications (VoIP and Video Conferencing). The obtained results were compared with those of native IPv4 and IPv6 networks. A comparative analysis of the simulation results involves various parameters such as delay, delay variation, jigger, MOS, and packet loss. The remainder of the document is organized as follows. Section II will present a background of some IPv4/IPv6 transition mechanisms and their classification. Section III will discuss a non-exhaustive state of the art of the research work carried out in this field of research. The simulation scenarios for the chosen IPv4/IPv6 transition mechanisms will be described in Section IV. The results of the simulation and the comparative analysis will be discussed in Section V. The conclusions and perspectives will be presented in the final section of this paper.

II. Background

During the transition phase to IPv6 and because of the incompatibility between IPv4 and IPv6 (the number of fields, the format, and the size of the addresses are different), several transition mechanisms to IPv6 have been developed. Each mechanism has specificities that make it suitable for a particular use. These mechanisms can be classified into three main families as shown in Fig. 1 below:

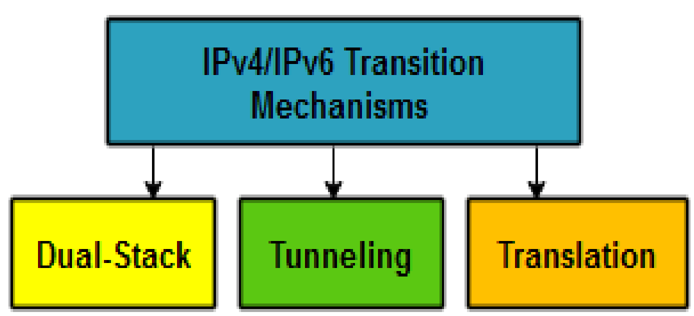


Fig. 1. IPv4/IPv6 transition mechanisms classification.

A. Dual Stack

The dual stack [3] is the simplest transition mechanism to deploy. It requires that all network peripherals (computers, routers, servers, etc.) support both IPv4 and IPv6 protocols working in parallel and side by side. The applications communicate with IPv4 and IPv6. That means that we are on an IPv4/IPv6 network, and therefore we do not need additional mechanisms to access both IPv4 machines and IPv6 machines. In this case, the communications are transmitted by the IP layers corresponding to used addresses, and there is no conversion problem. The choice of the IP version is based on the result of the Domain Name System (DNS) query or application preference.

B. Tunneling

As IPv4 and IPv6 headers are different from each other, the IPv6 packet must be tunneled to head for its destination point across the incompatible IPv4 network. Tunneling [4] is a technique in which one protocol is encapsulated in another protocol depending on the network where the packet must be routed. In the case of an IPv6 tunnel, if an IPv6 source communicating with an IPv6 destination and an IPv4 network is between them, the IPv6 packets must be encapsulated in IPv4 headers in order to be routed for over the IPv4 network and reach their IPv6 destination.

1) Manual Tunnel

Here, IPv6 data are encapsulated in IPv4 packets and then transferred through the tunnel. At the endpoint of the tunnel, the packets will be decapsulated and transmitted to their destination [5]. The endpoint address of the tunnel is determined from the configuration information stored on the encapsulation/decapsulation points of the tunnel. These tunnels can be used from router to router, from host to router, from host to host or from router to host.

2) 6to4 Automatic Tunnel

This one allows communication between two IPv6 sites over an IPv4 network without needing for an explicitly configured tunnel or an IPv6-IPv4 compatible address. Thus, the tunnel end-node configuration is minimal [6]. Other tunneling mechanisms can be used such as: an IPv4-compatible automatic IPv6 tunnel [7], 6over4 [8], ISATAP [9], 6rd [9], etc.

C. Translation

This mechanism works by allowing a native IPv6 network to communicate with nodes on an IPv4 network and vice versa. That is an intermediate peripheral or service that converts packets headers to the border network.

1) Network Address Translation-Protocol Translation (NAT-PT)

It is a device residing at the edge of an IPv4/IPv6 network allowing communication between IPv4 nodes residing in an IPv4 network and IPv6 nodes residing in an IPv6 network and vice versa by translating IP addresses [10]. NAT-PT maintains a globally routable range of IPv4 addresses and assigns IPv4 addresses to IPv6 nodes and vice versa.

2) Dual Stack Application Level Gateway (DS-ALG)

Here, a dual stack IP peripheral is used and it can access IPv4 and IPv6 services in native mode [11]. It is a simple and robust solution, but it requires that all customers are configured to use ALG, and it only works for specific applications supported by ALG. Other translation mechanisms can be used such as: BIH [12], NAT64 [13]/DNS64 [14], SIIT [15], etc.

III. Related Works

Performance evaluation of IPv4/IPv6 transition mechanisms is a very active field of research. Several research works have been carried out in this area. Here, we present some of the most relevant ones.

The author Quintero and his colleagues compared the performance of three transition mechanisms, namely ISATAP, 6to4, and NAT64 with IPv4 and IPv6 networks [16]. This comparison in terms of delay and throughput was performed on different operating systems (Debian, Windows 7, Windows 8 and Windows 10) for two types of traffic: TCP and UDP. The obtained results showed that native IPv4 gave better performance, closely followed by native IPv6. The difference is mainly due to the length of the IP header (20 bytes in IPv4 compared with 40 bytes in IPv6). The selected tunneling solutions (ISATAP and 6to4) have similar performance in terms of delay and throughput. They represent the low level of the technologies studied on account of the additional IPv4 header added by the tunnel.

In [17], the author Lu and his colleagues carried out a performance evaluation in terms of a routing path between two mechanisms that are the 4over6 and dual stack. Consequently, the authors found that the experimental routing performance of a dual stack mechanism gave better performance than 4over6. The 7-node routing path with dual stack was better than 4over6 by 17,329%.

In a detailed manner, the authors El Khadiri and his colleagues [18] and the author Govil [19] realized comparative studies of the mechanisms of the transition from IPv4 to IPv6 in order to help, among others, the customers, who want to use or connect using IPv6, to choose the suitable mechanism according to their needs. The study showed that the right solution for a particular customer or organization depended on its existing infrastructure and the time it takes to migrate to IPv6.

Other performance comparisons between IPv4, IPv6 networks, and the 6to4 tunnel have been discussed in [20], [21]. VoIP has been used as test traffic according to certain parameters such as delay, queuing delay, response time, jitter and throughput. Consequently, IPv4 gave better performance than IPv6. The difference is due to the IPv6 header length, which is higher than the one of IPv4. Similarly, these networks (IPv4 and IPv6 networks) gave better results compared with the 6to4 tunnel. This difference is due to the additional header added by the tunnel during the encapsulation/decapsulation operations of IPv6 packets in IPv4. However, these works did not take into account the impact of packet loss, which is considered as a very important parameter in the measure of performance of a network, on the one hand between IPv4 and IPv6 networks, and on the other hand related to the tunnel.

In view of this reflexive synthesis, it turned out to be indispensable to complete the previous works. On one hand, by adding other IPv4/IPv6 transition mechanisms and on the other hand, we intend to evaluate the behavior of real-time applications, such as VoIP and Video Conferencing, taking into account the impact of the packet loss rate on the performance of the mechanisms that we are going to evaluate.

IV. Scenarios of the Simulation

A. Simulation Environment

The proposed simulation network topology has been configured for five different scenarios. Three of them concern the dual stack transition mechanisms, manual tunnel, and 6to4 automatic tunnel. The remaining two scenarios target native IPv4 and IPv6 networks if all communications are respectively only IPv4 or IPv6. Our simulation was implemented using the simulation tool OPNET (Optimized Network Engineering Tool) as shown by the simulation network topology represented in Fig. 2 below. Two IPv6 sites (the first is set in Rabat and the other in Tangier) want to communicate with each other through an IPv4 backbone.

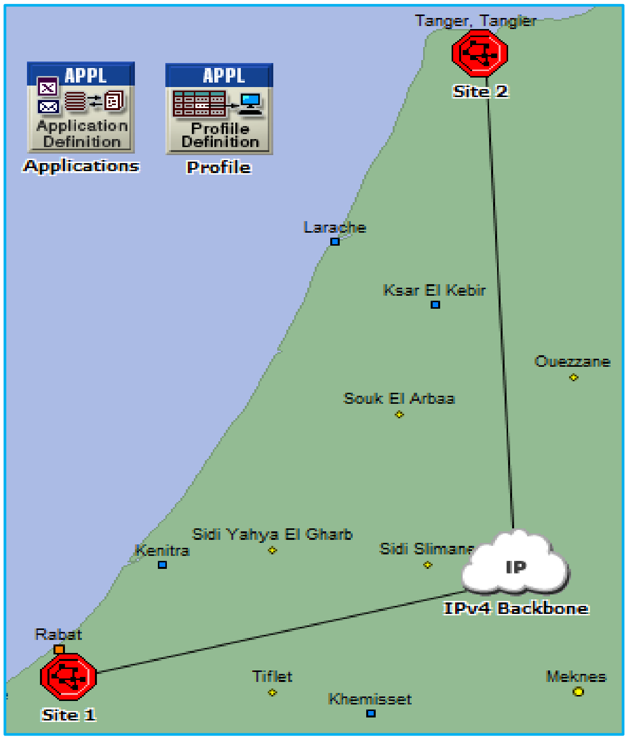


Fig. 2. The typology of simulation network.

To that end, we used the version 17.5 of OPNET (Riverbed Modeler Academic Edition) [22]. In order to simulate the different chosen IPv4/IPv6 transition techniques, we used real-time applications (VoIP and Video Conferencing) as test traffic. Regarding the routing, we have configured RIPv2 routing on the IPv4 backbone and RIPng routing on the IPv6 sites.

B. Simulation Parameters

Five simulation parameters were used during this simulation as follows:

1) End-to-End Delay: This parameter represents the end-to-end delay that is measured between the time a packet is created and sent from a source until it is received at its destination.

2) Delay Variation: This is the variation between the unidirectional delay of selected packets in the same packet stream [23]. It is based on the end-to-end delay difference of these selected packets. This measure has a significant impact assessing the quality of voice/video conferencing applications.

3) Jitter: Jitter is defined as the end-to-end transmission delay difference between selected packets in the same packets stream, without taking into account of eventually lost packets [23], [24]. This parameter is important for a voice application because if the transmission delay varies for a VoIP conversation, the voice quality will be degraded. The best jitter value is the one closest to zero.

4) MOS: That is an important indicator for assessing the quality of a voice application [24] by giving it one of these values (1, 2, 3, 4, 5) where 5 indicates excellent quality and 1 indicates poor quality.

5) Loss Rate: That is the number of lost packets in percent compared with sent packets.

V. Simulation Results and Analysis

In this section, we present the results of the simulation by analyzing and comparing the five scenarios. For each specific case, the simulation process lasted 300 seconds using DES (Discrete Event Simulation). OPNET calculates the desired value at all times for the simulation time. In this simulation, the average values were monitored from the OPNET results viewer and exported to Excel in order to draw bar graphs for comparison purposes.

A. End-to-End Delay

The obtained results in Fig. 3 below represent the Video Conferencing Packet End-to-End Delay in milliseconds for the five proposed scenarios.

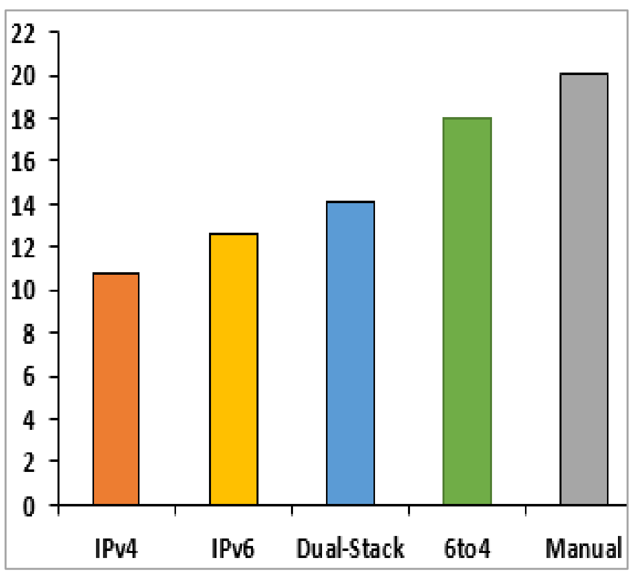


Fig. 3. Video conferencing packet end-to-end delay (ms).

These results show that the dual stack is better than the other two IPv4/IPv6 transition mechanisms: with an average delay value of about 14ms for the dual stack compared with 18ms and 20ms for the 6to4 tunnel and manual tunnel. That is due to the delay caused by encapsulation and decapsulation processes in the tunneling mechanisms whereas, in the dual stack, the two protocols simultaneously work without involving encapsulation or decapsulation. The comparison between the two protocols in relation to the same criterion indicates that IPv4 has better performance than IPv6. Indeed, IPv4 presents an average delay value of about 10.8ms compared with 12.6ms for IPv6. That is due to the IPv6 header length, which is higher than the one of IPv4.

Fig. 4 below shows the results of the Voice Packet End-to-End Delay in milliseconds for the five scenarios.

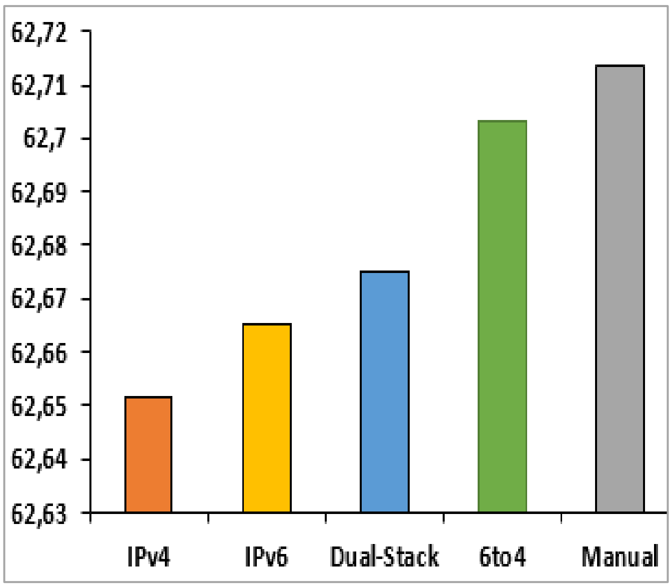


Fig. 4. Voice packet end-to-end delay (ms).

These results indicate that tunneling mechanisms recommend average delay values higher than those of dual stack. The difference is due to the additional IPv4 header that is added by tunneling for encapsulation/decapsulation operations. The comparison between the two protocols shows that IPv6 provides an average delay value higher than the one of IPv4. That is due to the IPv6 header length that is higher than the one of IPv4.

B. Delay Variation

Fig. 5 and 6 below represent the results of the Packet Delay Variation in milliseconds for the two real-time applications (Video conferencing and VoIP) for the five scenarios.

According to a first reading, it is clear that the dual stack is more performing than manual tunneling mechanisms and 6to4. Indeed, it presents a lower value in terms of Packet Delay Variation compared with the tunneling mechanisms. The comparison between the two protocols showed that IPv4 provides a lower delay variation than IPv6. That indicates that IPv4 provides a better quality regarding VoIP and Video Conferencing applications compared with IPv6.

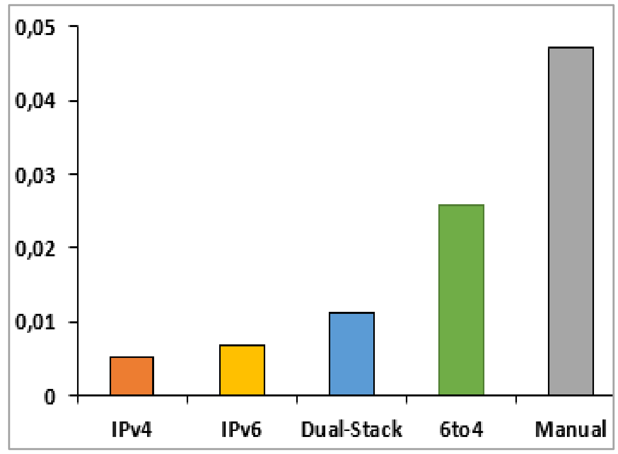


Fig. 5. Video conferencing packet delay variation (ms).

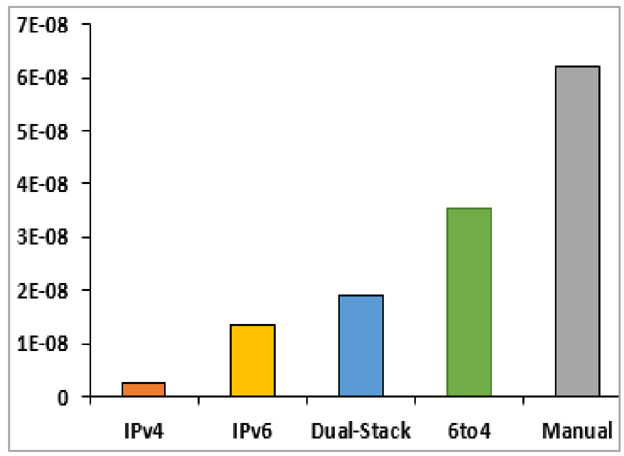


Fig. 6. Voice packet delay variation (ms).

C. Jitter

Fig. 7 below shows the average values of the Voice Jitter in milliseconds for the five scenarios.

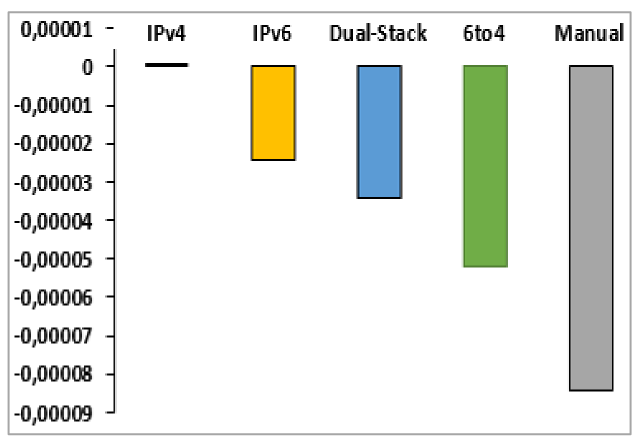


Fig. 7. Voice Jitter (ms).

These results indicate that the dual stack is better than the other two IPv4/IPv6 transition mechanisms. Indeed, it presents a lower average jitter value than those of the manual tunneling mechanisms and 6to4. The comparison between the two protocols in relation to the same criterion indicates that IPv4 is better than IPv6. That indicates that IPv4 offers better quality regarding VoIP applications compared with IPv6.

D. MOS

The obtained results in Fig. 8 below represent the average MOS values for the five scenarios. The highest MOS value indicates one better performance. As it can be seen, the results indicate that dual stack provides better voice quality than the other two IPv4/IPv6 transition mechanisms that appeared similar. The comparison between the two protocols in relation to the same criterion indicates that IPv4 has better voice quality than IPv6.

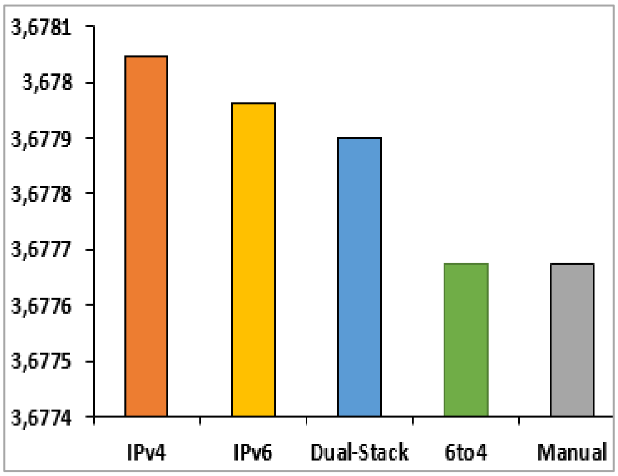


Fig. 8. Voice MOS.

E. Loss Rate

Fig. 9 below shows the results of the Video Conferencing Packet Loss Rate for the five scenarios. As it can be seen, the results show that the packet loss rate of manual and 6to4 tunneling mechanisms is higher than the one of the dual stack: a loss rate of about 3,6% and 4,1% for respectively 6to4 and manual tunneling mechanisms compared with 2,7% for the dual stack. That is due to the encapsulation/decapsulation operations of IPv6 packets encapsulated in IPv4 packets by the tunneling mechanisms. IPv4 presents a lower loss rate than IPv6.

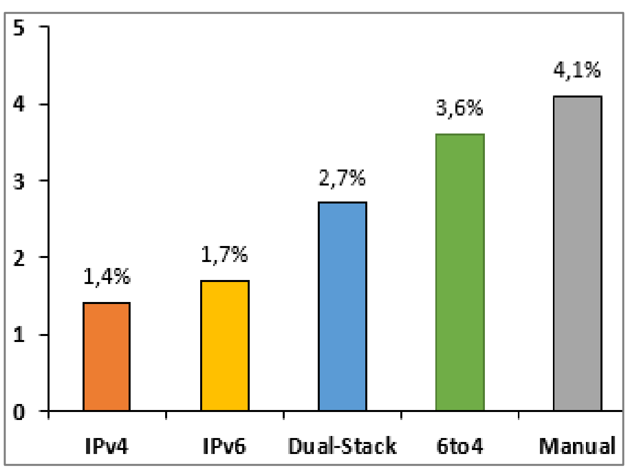


Fig. 9. Video conferencing packets loss rate (%).

Fig. 10 below shows the results of the Voice Packets Loss Rate for the five scenarios. As it can be seen, the results reveal that the loss rate of the dual stack is lower than the one of the tunneling mechanisms: a loss rate of about 2,3% for the dual stack against respectively 3,2% and 3,7% for the 6to4 and manual tunneling mechanisms. The difference is due to encapsulation/decapsulation operations of IPv6 packets encapsulated in IPv4 packets by tunneling mechanisms. The comparison between the two protocols indicates that IPv4 presents a lower loss rate than IPv6.

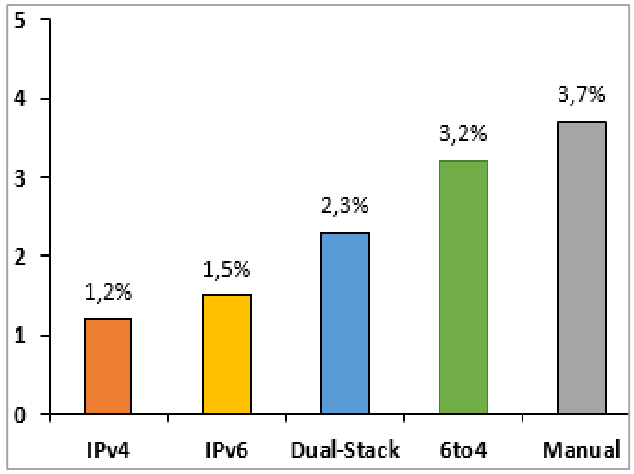


Fig. 10. Voice packets loss rate (%).

VI. Conclusions and Perspectives

In this article, we studied and evaluated three IPv4/IPv6 transition mechanisms: the dual stack, the manual tunnel, and the 6to4 automatic tunnel. These mechanisms were evaluated on a simulation network infrastructure under OPNET Modeler using two real-time applications (VoIP and Video Conferencing). The obtained results were compared with those of native IPv4 and IPv6 networks in terms of delay, delay variation, jitter, MOS, and packets loss.

According to our results, the dual stack mechanism presents better performance compared with the two studied tunneling mechanisms: the manual tunnel and the 6to4 automatic tunnel. The performance degradation of these lasts is due to the additional IPv4 header added by tunneling during encapsulation/decapsulation operations of IPv6 packets in IPv4 packets. In contrast, in the dual stack, the two IP stacks work in parallel and side by side without involving encapsulation or decapsulation.

In terms of deployment, the dual stack mechanism requires that all network peripherals take over both protocols (IPv4 and IPv6), which is difficult in the case of a large network that will require many configurations, including problem-solving in case of a crash. Consequently, this transition mechanism can be deployed in a small network. In contrast, tunneling mechanisms represent a good choice for large networks where only both ends of the network have to be configured with IPv4 and IPv6 protocols.

Regarding both protocols (IPv4 and IPv6), the results showed that IPv4 presents better performance than IPv6. Indeed, that is the price to pay for resolving the problem of the potential depletion of the IPv4 addresses. The difference is actually due to the IPv6 header length, which is higher than the one of IPv4 (40 bytes in IPv6 compared with 20 bytes in IPv4).

Our study was limited by the OPNET Modeler simulator, which only supports the study of integrated IPv4/IPv6 transition mechanisms (the ones we studied). Our future research work focuses on an experimental evaluation of IPv4/IPv6 transition mechanisms by adding other mechanisms and by studying their performance for various types of applications.

References

[1] J. Postel, "Internet protocol," RFC 791, Sep. 1981.

[2] R. Hinden, "Internet protocol, version 6 (IPv6) specification," RFC 8200, 2017.

[3] S. Aravind and G. Padmavathi, "Migration to Ipv6 from IPV4 by dual stack and tunneling techniques," in Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2015 International Conference on, 2015, pp. 107–111.

[4] J. L. Shah and J. Parvez, "An examination of next generation IP migration techniques: Constraints and evaluation," in Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on, 2014, pp. 776–781.

[5] S. Steffann, I. van Beijnum, and R. van Rein, "A Comparison of IPv6-over-IPv4 tunnel mechanisms," RFC 7059, 2013.

[6] A. Misra and H. Chawla, "Performance Analysis of IPv6 Dual-Protocol Stack and Tunnel Transition," Int. J. Sci. Eng. Technol. Res. IJSETR Vol., vol. 5, 2016.

[7] B. Chhetri, "Transition from IPV4 to IPV6," 2015.

[8] R. Kaur, S. Kumar, and V. K. Patle, "Analysis of Tunneling Transition Mechanism in IPv6," Res. J. Sci. Technol., vol. 5, no. 3, p. 3, 2013.

[9] G. LENCSE and Y. KADOBAYASHI, "Survey of IPv6 transition technologies for security analysis," in IEICE Technical Committee on Internet Architecture (IA) Workshop, Tokyo Japan, 2017, pp. 19–24.

[10] A. H. A. Mohamed, A. B. Abdelnabi, and H. E. Edris, "Performance Analysis of IPv4, IPv6 Transition Technologies," 2016.

[11] A. Singh, A. Maheshawari, and D. Shukla, "Study for IPv4 and IPv6 Coexistence," 2013.

[12] B. Huang, H. Deng, and T. Savolainen, "Dual-stack hosts using" bump-in-the-host"(BIH)," 2012.

[13] M. Bagnulo, P. Matthews, and I. van Beijnum, "Stateful NAT64: Network address and protocol translation from IPv6 clients to IPv4 servers," RFC 6146, 2011.

[14] M. Bagnulo, P. Matthews, A. Sullivan, and I. van Beijnum, "DNS64: DNS extensions for network address translation from IPv6 clients to IPv4 servers," RFC 6147, 2011.

[15] J. Hanumanthappa and H. Annaiah, "DW&C: Dollops Wise Curtail IPv4/IPv6 Transition Mechanism using NS2," ArXiv Prepr. ArXiv14093767, 2014.

[16] A. Quintero, F. Sans, and E. Gamess, "Performance evaluation of IPv4/IPv6 transition mechanisms," Int. J. Comput. Netw. Inf. Secur., vol. 8, no. 2, p. 1, 2016.

[17] T. Te Lu, C. Y. Wu, W. Y. Lin, H. P. Chen, and K. P. Hsueh, "Comparison of IPv4-over-IPv6 (4over6) and Dual Stack Technologies in Dynamic Configuration for IPv4/IPv6 Address," in Advances in Intelligent Information Hiding and Multimedia Signal Processing, Springer, 2017, pp. 259–269.

[18] K. EL KHADIRI, O. LABOUIDYA, N. ELKAMOUN, and R. HILAL, "Etude comparative des mécanismes de transition de l'IPv4 à l'IPv6," Mediterranean Telecommunications Journal, vol. 7, no. 1, Jan. 2017.

[19] J. Govil, J. Govil, N. Kaur, and H. Kaur, "An examination of IPv4 and IPv6 networks: Constraints and various transition mechanisms," in Southeastcon, 2008. IEEE, 2008, pp. 178–185.

[20] A. Ahmed, A. Mustafa, and G. Ibrahim, "Performance Evaluation of IPv4 Vs Ipv6 and Tunnelling Techniques Using Optimized Network Engineering Tools (OPNET)," IOSR Journal of Computer Engineering (IOSR-JCE), vol. 17, no. 1, 2015.

[21] A. Salam and M. A. Khan, "Performance Analysis of VoIP over IPV4, IPv6 and 6-to-4 Tunneling Networks," (IJCSIS) International Journal of Computer Science and Information Security, vol. 14, no. 6, 2016.

[22] "Modeler OP. Riverbed Technology. Inc. https://www.riverbed.com," 2018.

[23] K. Neupane, V. Kulgachev, A. Elam, S. H. Vasireddy, and H. Jasani, "Measuring the performance of VoIP over Wireless LAN," in Proceedings of the 2011 conference on Information technology education, 2011, pp. 269–274.

[24] O. Slavata, J. Holub, and P. Hübner, "Impact of jitter and jitter buffer on the final quality of the transferred voice," in Wireless Systems (IDAACS-SWS), 2012 IEEE 1st International Symposium on, 2012, pp. 120–123.

SVM Optimization for Sentiment Analysis

Munir Ahmad, Shabib Aftab, Muhammad Salman Bashir, Noureen Hameed, Iftikhar Ali, Zahid Nawaz
Department of Computer Science
Virtual University of Pakistan
Lahore, Pakistan

***Abstract*—Exponential growth in mobile technology and mini computing devices has led to a massive increment in social media users, who are continuously posting their views and comments about certain products and services, which are in their use. These views and comments can be extremely beneficial for the companies which are interested to know about the public opinion regarding their offered products or services. This type of public opinion otherwise can be obtained via questionnaires and surveys, which is no doubt a difficult and complex task. So, the valuable information in the form of comments and posts from micro-blogging sites can be used by the companies to eliminate the flaws and to improve the products or services according to customer needs. However, extracting a general opinion out of a staggering number of users' comments manually cannot be feasible. A solution to this is to use an automatic method for sentiment mining. Support Vector Machine (SVM) is one of the widely used classification techniques for polarity detection from textual data. This study proposes a technique to tune the SVM performance by using grid search method for sentiment analysis. In this paper, three datasets are used for the experiment and performance of proposed technique is evaluated using three information retrieval metrics: precision, recall and f-measure.**

Keywords*—*Sentiment analysis; polarity detection; machine learning technique; support vector machine (SVM); optimized SVM; grid search technique

I. INTRODUCTION

In order to compete in market, it is essential for the organizations and companies to get aware of the consumer's opinion regarding the products and services, they provide. Now days micro-blogging websites are the rich sources of textual data including opinions and reviews about products, services, brands, movies, music, news and politics, etc. These sources can be used to extract public opinion about anything i.e. from comments regarding particular product or brand to the views about next election etc. In past, this type of information was obtained through telephonic and door to door surveys which were really time consuming and complex task. So the need of automatic technique for opinion extraction from textual data brought the researchers to this domain. Today, there are fundamentally three approaches available to extract the opinion from text: Lexicon driven [1], machine learning based [2] and finally the hybrid of both [3]. The lexicon-based approach depends upon its dictionary of weighted words to generalize the polarity of any given text. It does not require any form of prior training as its every aspect is pre-defined and pre-programmed. Machine learning approach can be further categorized as supervised, un-supervised and semi supervised technique. Supervised technique needs to get trained in order to perform classification. For training purpose pre-classified/pre-labeled data (training data) is used and then it can be capable to classify the real input data (test data) [2], [4], [5]. The hybrid approach integrated both approaches (lexicon driven and machine learning) and usually brings more accurate results. SVM is one of the widely used supervised machine learning algorithms for sentiment analysis. It has proved to be highly effective in categorization of traditional texts. This research intends to tune the performance of SVM for sentiment analysis using grid search technique. Optimized Sentiment Analysis Framework (OSAF) is used in this research which is an extension of Sentiment Analysis Framework (SAF) [5]. OSAF consist of four phases: Dataset, Pre-processing, Classification and Results. Tuning is performed in the classification phase. Three datasets were used in this research, two from twitter and one from IMDB. In proposed technique, Grid-Search keeps tweaking the SVM's parameters and compares the accuracy of the obtained results with the pre-labeled data and chooses the parameters with best result. On the other hand cross-validation takes the accuracy further up to a notch by changing the testing data ratio to the training data in accordance with a pre-defined range until it obtains the highest possible accuracy. To evaluate the performance of proposed technique three information retrieval metrics are used precision, recall and f-measure.

Remaining paper is organized as follows. Section II is about related work in this domain. Section III elaborates the materials & methods used for this research. Section IV presents results & discussion. Section V finally concludes the paper.

II. RELATED WORK

Many researchers have been working to extract the sentiment from textual data by using data mining techniques. Some of the selected studies are discussed here. In [6], J48 and MLP were used to classify five datasets. Performance was evaluated in terms of TP rate, FP rate, Precision, Recall, F-measure and ROC Area. According to results, MLP performed better on each dataset. Neural Network also showed better learning capability and was suggested by the authors as good alternate option for classification. Authors in [7] presented an application for Arabic sentiment analysis of twitter data. Machine learning techniques: NB and SVM were used to analyze 1000 tweets for polarity detection. Feature vector approach was used with machine learning classifiers to improve the accuracy. Some problem areas were also identified by the researchers regarding training data such as multiple occurrences of tweets, opinion spamming and dual opinion tweets. It was mentioned that issues like these can compromise the level of accuracy. Authors in [4] performed sentiment classification of Arabic tweets by using three machine learning algorithms: Naïve Bayes, Decision Tree and Support Vector Machine. Authors in this research used a framework for classification which includes: Term Frequency-Inverse Document Frequency (TF-IDF) and Arabic stemming as sub tasks. One dataset was used for all three techniques to evaluate the performance. Accuracy was measured in terms of three information retrieval metrics: precision, recall, and f-measure. Authors in [8] presented a feature vector technique by dividing the feature extraction process in two steps, where process starts with the extraction of twitter specific features and then integrated with feature vector. After that these features are removed from the tweets and then again the feature extraction process is completed, like the case with normal text. These extracted features are also added to the feature vector. The accuracy of the proposed feature vector technique is same for Nave Bayes, SVM, Maximum Entropy and Ensemble classifiers. In [9], students' academic performance was analyzed and predicted by using three data mining techniques: Decision tree (C4.5), Multilayer Perception and Naïve Bayes. These techniques were applied on the student's data of 2 undergraduate courses from two semesters. According to the results, Naïve Bayes showed the prediction accuracy of 86% which was higher than MLP and Decision tree. This type of prediction can help the teachers to detect those students in early stage, who are expected to get F grade. So ultimately, with the teacher's special care to those students, the academic performance can be improved. In [10], authors predicted the rainfall in Malaysia by using five classification algorithms: Naïve Bayes, Decision Tree, Support Vector Machine, Neural Network and Random Forest. A comparative analysis was performed to identify the technique (s), which can give good result with low training data. Results showed that Decision Tree and Random Forest both have potential to bring higher F-measure after getting trained with lower amount of training data. Support Vector Machine and Naive Bayes brought lower F-measure, when got trained with little amount of data. Neural Network showed less performance in prediction even after getting trained with large amount of training data. In [11], authors performed a comparative analysis of various data mining techniques upon the practitioner's decision of potential therapy change for 40 patients of posterior cancer. J48, MLP, Naïve Bayes, Radial Basis Function and K-Nearest Neighbor were used for classification. According to results, Radial Basis Function performed well and brought high accuracy and Kappa score with low error rate. In [5], performance of SVM was analyzed for polarity classification of textual data. The authors have proposed Sentiment Analysis Framework which consists of 4 phases: Dataset, Pre-processing, Classification and Results. Three datasets were used for the experiment, two from twitter and one from IMDB reviews. Performance of SVM was evaluated for each dataset by keeping in view three different ratios of training data and test data: 70:30, 50:50 and 30:70. Precision, recall and f-measure were used for performance evaluation. In [12], the authors have used SVM for sentiment analysis of twitter data. In this experiment, default parameters were selected in Weka along with 10 fold cross validation technique. Performance of SVM was analyzed on two datasets of pre-labeled tweets. Accuracy was evaluated In terms of Precision, Recall and F-Measure. In [13], authors conducted a systematic literature review on sentiment analysis. Latest research regarding use of SVM for sentiment analysis was the focus of this SLR and published research papers from year 2012 to 2017 were considered. Total of 901 articles were collected and by following a thorough systematic approach, 8 research articles were selected for critical review. Research objectives were identified in the form of research question and during the critical review the answers to those questions were provided.

III. MATERIALS AND METHODS

This research aims to optimize the performance of SVM for sentiment analysis using grid search technique. To analyze the performance of proposed technique, results are compared with three pre-labeled datasets, two from twitter [14], [15] and one from IMDB reviews [16]. An Optimized Sentiment Analysis Framework (OSAF) is used in this research (Fig. 1), which is an extension of Sentiment Analysis Framework (SAF) [5]. The proposed framework consists of four phases: Dataset, Pre-processing, Classification and Results. Dataset phase deals with the insertion of data into the WEKA (Waikato Environment for Knowledge Analysis) environment on which the classification is going to be performed. Pre-processing Phase deals with the conversion of strings into vectors, which is a pre requisite process for the processing of SVM. This phase has further five steps: 1) Term Frequency-Inverse Document Frequency (TF-IDF), 2) Stemming, 3) Stop Words, 4) Tokenizing and 5) Words to Keep. Classification phase deals with the working of SVM in WEKA using K-fold cross validation grid search technique. Result phase produces the results in the form of tables and graphs. For performance evaluation, three information retrieval metrics are used: precision, recall and f- measure.

A. WEKA

This study used WEKA for the implementation of SVM grid search technique. WEKA is one of the widely used data mining software, developed in Java language at the University of Waikato, New Zealand [17]. The reason behind the widely acceptance is it's easy to use GUI interface for different functionalities such as data analysis, classification, predictive modeling and visualizations. Further advantages of this software include its general public license and its portability.

B. Data Sets

Three datasets are used in this research (Table I). First dataset [14] consisted of tweets related to following four topics: 'Apple', 'Google', 'Microsoft' and 'Twitter'. It contains 571 positive, 519 negative, 2331 neutral and 1689 irrelevant tweets. In the second dataset [15], tweets are related to U.S. airlines and categorized as 2362 'positive', 9178 'negative' and 3099 'neutral'. Third dataset [16] was taken from the Internet Movie Database (IMDB) reviews and contains 1000 positive and 1000 negative comments.

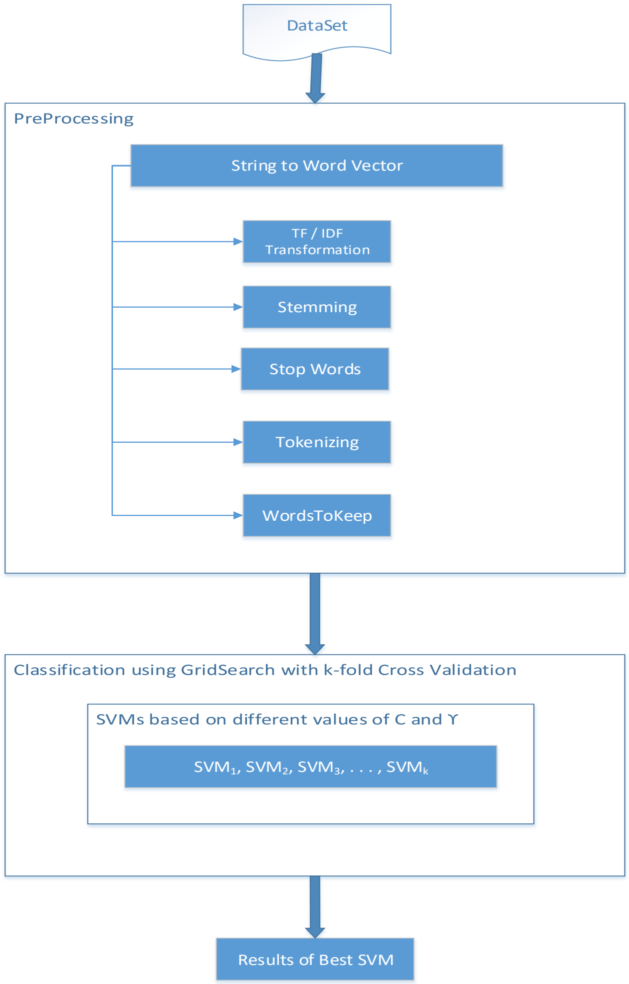


Fig. 1. Optimized Sentiment Analysis Framework (OSAF).

TABLE I. DATASETS DETAIL

Source	Positive	Negative	Neutral	Irrelevant	Total
Twitter [14]	571	519	2331	1689	5110
Airline [15]	2362	9178	3099	-	14639
IMDB [16]	1000	1000	-	-	2000

Dataset phase deals with the arrangement of relevant data and its transformation to CSV/ARFF format to use in WEKA Workbench [17]. Simple CLI can be used to convert text files into ARFF format using "WEKA.core.converters.TextDirectoryLoader" function.

C. Pre-Processing

It is the most important phase of the framework. Purpose of this phase is to normalize the data by converting the strings into vectors for the classification process. Following sub tasks are performed in this phase.

1) Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF provides important information in pre-processing phase by evaluating the frequency of useful words, which essentially makes the sentiment detection process easy. Frequency of these terms plays an important role in identification of important information as explained by [18]. For example, frequently appearing words in a text document can be 'Good', 'Bad', 'Happy' or 'Sad' etc. Identification and frequency of these words can play a vital role in the process of opinion mining. Term Frequency (TF) is the number of occurrences of a term in a given document and can be calculated with following equation:

$$\mathrm{W_d(t) = TD(t, d)} \tag{1}$$

Where TD corresponds to frequency of term t in a given document d. TF-IDF contains the inverse document frequency (IDF) that reverts higher weight-age for rare conditions while maintaining lower weight-age for common conditions as explained by [19]. IDF can be calculated with following equation:

$$\mathrm{IDF_t} = \log\left(\frac{\mathrm{N}}{\mathrm{N_t}}\right) \tag{2}$$

Where N represents number of documents and N_t represents the number of terms. When both TF and IDF parameters are set to true, the results are calculated using the following equation:

$$W_t = TF(t,d).IDF_t \tag{3}$$

In WEKA, TF & IDF transformations are also available along with other filters.

2) Stemming

The process of stemming is immensely useful in many areas of computational linguistics and information retrieval as it reduces all words with the same stem/ base to a common form [20], for example, the word 'working' will be stemmed in to 'work' and so on. Word stemming is one of the essential features of pre-processing in text mining [21]-[25]. In this study, "IterativeLovinsStemmer" is selected in WEKA as the word stemmer in the pre-processing phase. It is based on the LovinsAlgortihm which was the first Stemming algorithm by Lovins JB in 1968 [26].

3) MultiStopWords

The Concept of stop words was originally introduced by [27]. These are common high frequency words like "A", "the", "of", "and", "an". This data is not use full and also does not affect the performance of classification; thus, it has to be removed. There are several methods available for stop word removal as explained by [20], [21], [23], [28], [29]. In this research, "MultiStopwords" was selected for stop words criterion for the pre-processing phase in WEKA.

4) N-GramTokenizer

"N-GramTokenizer" was selected as the Tokenizer in WEKA for pre-processing of data. It first breaks the text into words whenever one of the listed specified characters is detected in it. Afterwards it emits N-Grams of each word of the specified length.

5) WordstoKeep

1000 words were short-listed in "wordstokeep" parameter to narrow down the results within a limited amount of time.

After applying these parameters, pre-processing was performed on all three datasets. Then the processed datasets were forwarded to the classifier.

D. Classification

In supervised machine learning approach, first the algorithm has to get trained with pre classified data (training data) with which it makes rules for classification and then on the basis of these rules it classifies the input data (test data). For performance analysis of any supervised machine learning algorithm, pre classified data is provided as test data and then the results of the algorithm can be compared with this pre labeled data. Same strategy is used is this study to analyze the performance of proposed grid search technique. Pre labeled datasets are obtained from social forums: Twitter and IMDB. For classification, Support Vector Machine (SVM) with grid search and K-fold cross validation technique is used. Grid-Search is basically a model for hyper parameter optimization. Hyper parameter tuning is an important task in SVM to extract more accurate results [30]-[33]. In Grid-Search, different models having different parameter values are trained and then evaluated using cross validation. For an RBF kernel, there are two parameters: C and ϒ. It cannot be ascertained in advance that which values of C and ϒ are best suited for a given problem, so an optimized model is required which can identify the ideal pair of values for these parameters to achieve maximum accuracy. The process of 10-k cross validation is performed on each model of C and ϒ and the pair with optimum results is selected. Cross validation is a method used to test multiple models under a particular classifier with the subset of input data as explained by [35]. For K-fold cross validation, the training data is first divided into k subsets of same size. One subset is tested using the classifier on the remaining k-1 subsets. The cross validation procedure can prevent the overfitting problem [34], [36], a binary classification problem is shown in Fig. 2 to illustrate this issue. Filled circles and triangles are the training data while hollow circles and triangles are the testing data. The testing accuracy of the classifier in Fig. 2(a) and (b) is not good as it overfits the training data. On the other hand, the classifier in Fig. 2(c) and (d) does not overfit the training data and gives better accuracy with cross validation.

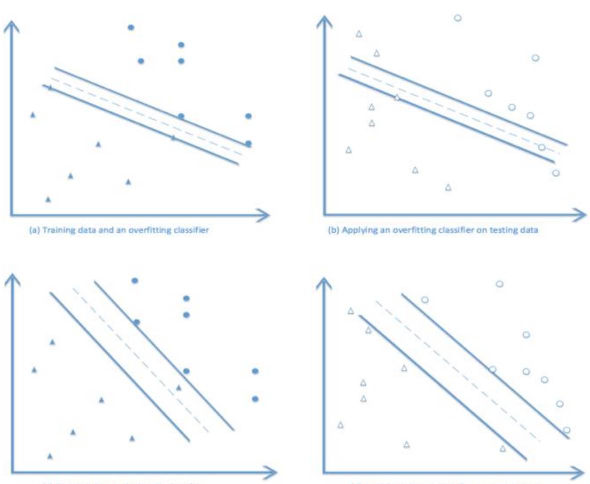


Fig. 2. Over-fitting with cross validation [36].

IV. Results and Discussion

Performance of proposed technique is evaluated for all of three selected datasets and for this purpose three information retrieval parameters are used: Precision, Recall and F-Measure.

The precision can be calculated using TP and FP rate as shown below:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (4)$$

TP is used for the sentences which are correctly classified whereas FP is used for sentences, which are wrongly classified.

Recall can be calculated as shown below:

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (5)$$

FN is used for non-classified sentences and TP are the sentences, which are correctly classified (as explained above).

F-measure can be computed as below:

$$\text{F-measure} = \frac{\text{Precision} * \text{Recall} * 2}{(\text{Precision} + \text{Recall})} \quad (6)$$

A. Results with the First Dataset

First dataset is taken from [14] and consisted of tweets about following topics: ‘Apple’, ‘Google’, ‘Microsoft’ and ‘Twitter’. According to results the average Precision, Recall and F-Measure scores are 0.745, 0.752 and 0.747, respectively.

By keeping in view the f-measure score for each class, it can be analyzed that the proposed technique performed well for irrelevant class with the score of 0.87. Detailed results are available in Table II whereas shown graphically in Fig. 3.

TABLE II. Results of First Dataset

Class	Precision	Recall	F-Measure
Negative	0.583	0.539	0.561
Positive	0.504	0.378	0.432
Neutral	0.745	0.804	0.773
Irrelevant	0.872	0.867	0.87
Average	0.745	0.752	0.747

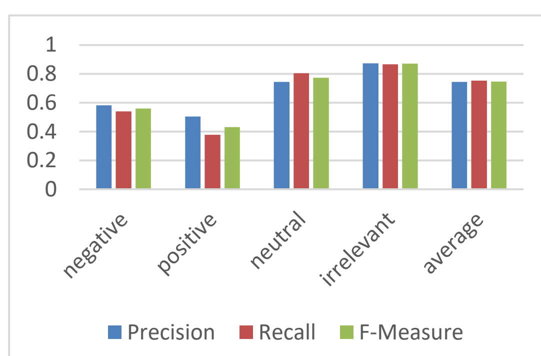


Fig. 3. Results of first dataset.

B. Results with the Second Dataset

Second dataset is obtained from [15] and contained the tweets regarding US airlines. The results show that the average Precision, Recall and F-Measure values are 0.786, 0.79 and 0.787 respectively (Table III and Fig. 4). Highest f-measure is reported in negative class which is 0.868.

TABLE III. RESULTS OF SECOND DATASET

Class	Precision	Recall	F-Measure
Negative	0.85	0.887	0.868
Neutral	0.632	0.62	0.626
Positive	0.74	0.634	0.683
Average	0.786	0.79	0.787

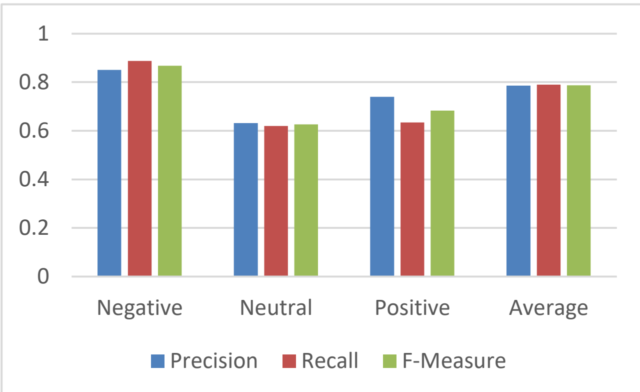


Fig. 4. Results of second dataset.

C. Results with the Third Dataset

Third dataset is taken from [16] and consisted of IMDB reviews. The results show that Precision, Recall and F-Measure values on average are 0.841, 0.841 and 0.841 respectively. Highest f-measure is reported in negative class with the score of 0.843 (Table IV, Fig. 5).

TABLE IV. RESULTS OF THIRD DATASET

Class	Precision	Recall	F-Measure
Negative	0.832	0.855	0.843
Positive	0.851	0.827	0.839
Average	0.841	0.841	0.841

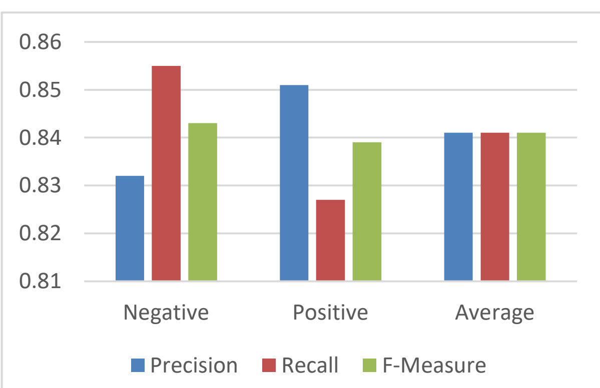


Fig. 5. Results of third dataset.

D. Performance Evaluation of Grid Search SVM

To evaluate the performance of grid search technique, f-measure scores of this research are compared with f-measure scores achieved with 70:30 in [5]. In that study, Sentiment Analysis Framework (SAF) was proposed and SVM was used along with various proportions of training and test data. The selected study is an ideal benchmark for comparison as the datasets used in both the studies are same. For comparison, f-measure score is considered as it is the average of precision and recall. OSAF used in this research is the extension of SAF. F-measure scores of both the studies are compared in Table V and graphically represented in Fig. 6. It can be seen that the grid search technique explored in this research has improved the performance in each class of all three data sets.

TABLE V. COMPARISON OF SAF AND OSAF

Dataset	Class	F-Measure (SAF)	F-Measure (OSAF)
Twitter [14]	Negative	0.532	0.561
	Positive	0.394	0.432
	Neutral	0.721	0.773
	Irrelevant	0.835	0.870
	Average	0.563	0.747
Twitter [15]	Negative	0.862	0.868
	Neutral	0.604	0.626
	Positive	0.671	0.683
	Average	0.776	0.787
IMDB [16]	Negative	0.791	0.843
	Positive	0.785	0.839
	Average	0.788	0.841

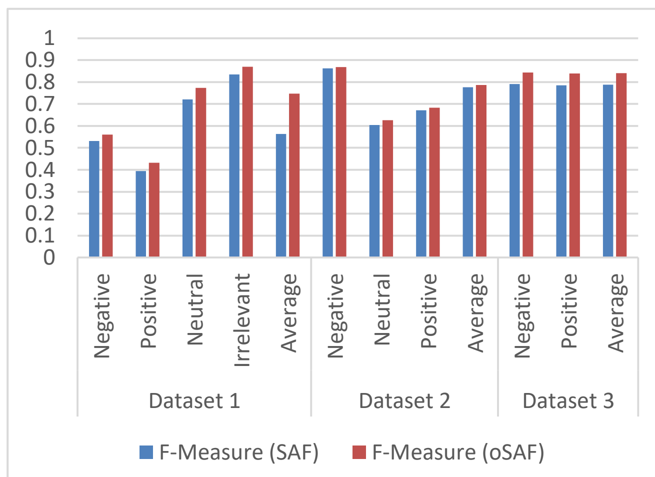


Fig. 6. Comparison of SAF and OSAF.

V. CONCLUSION AND FUTURE WORK

In this research, Grid- search technique for SVM is explored in Weka. SVM is one of the widely used supervised machine learning algorithms for sentiment analysis. Optimized Sentiment Analysis Framework OSAF is proposed in this study for text classification which used the SVM grid search technique along with 10k cross validation. Grid search method continually changes parametric values of SVM: gamma and cost, until the highest polarity detection accuracy is achieved for the given dataset. This characteristic helps to tune the performance of SVM in classification. Cross validation on the other hand keeps changing the ratio of training and test data until finds the particular proportion which further enhance the accuracy to the maximum point. Finally performance of the explored technique was evaluated by comparing the f-measure scores with the published research, which have used the SVM with same datasets. It has been observed that grid search technique performed well and it is further suggested to use this technique on further datasets to explore its accuracy.

REFERENCES

[1] M. Ahmad, S. Aftab, S. S. Muhammad, and U. Waheed, "Tools and Techniques for Lexicon Driven Sentiment Analysis : A Review," Int. J. Multidiscip. Sci. Eng., vol. 8, no. 1, pp. 17–23, 2017.

[2] M. Ahmad, S. Aftab, and S. S. Muhammad, "Machine Learning Techniques for Sentiment Analysis: A Review," Int. J. Multidiscip. Sci. Eng., vol. 8, no. 3, p. 27, 2017.

[3] M. Ahmad, S. Aftab, I. Ali, and N. Hameed, "Hybrid Tools and Techniques for Sentiment Analysis: A Review," Int. J. Multidiscip. Sci. Eng., vol. 8, no. 3, 2017.

[4] M. M. Altawaier and S. Tiun, "Comparison of Machine Learning Approaches on Arabic Twitter Sentiment Analysis," vol. 6, no. 6, pp. 1067–1073, 2016.

[5] M. Ahmad and S. Aftab, "Analyzing the Performance of SVM for Polarity Detection with Different Datasets," Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 10, pp. 29–36, 2017.

[6] R. Arora and Suman, "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA," Int. J. Comput. Appl., vol. 54, no. 13, pp. 21–25, 2012.

[7] A. Shoukry and A. Rafea, "Preprocessing Egyptian Dialect Tweets for Sentiment Mining," Fourth Work. Comput. …, no. November, pp. 47–56, 2012.

[8] M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," 2013 Fourth Int. Conf. Comput. Commun. Netw. Technol., pp. 1–5, 2013.

[9] A. Mueen, "Modeling and Predicting Students ' Academic Performance Using Data Mining Techniques," no. November, pp. 36–42, 2016.

[10] S. Zainudin, D. S. Jasim, and A. A. Bakar, "Comparative Analysis of Data Mining Techniques for Malaysian Rainfall Prediction," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 6, no. 6, pp. 1148–1153, 2016.

[11] N. Mallios, E. Papageorgiou, M. Samarinas, and K. Skriapas, "Comparison of machine learning techniques using the WEKA environment for prostate cancer therapy plan," in Proceedings of the 2011 20th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE 2011, 2011, pp. 151–155.

[12] M. Ahmad, S. Aftab, and I. Ali, "Sentiment Analysis of Tweets using SVM," Int. J. Comput. Appl., vol. 177, no. 5, pp. 25–29, 2017.

[13] M. Ahmad, S. Aftab, M. S. Bashir, and N. Hameed, "Sentiment Analysis using SVM : A Systematic Literature Review," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 2, pp. 182–188, 2018.

[14] N. J. Sanders, "Sanders-twitter sentiment corpus," Sanders Anal. LLC., 2011.

[15] E. Ikonomovska, "Airline dataset." .

[16] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," 2004.

[17] E. Frank, M. A. Hall, and I. H. Witten, "The WEKA Workbench," in Morgan Kaufmann, Fourth Edition, 2016, pp. 553–571.

[18] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Found. Trends Inf. Retr., vol. 2, no. 1–2, pp. 1–135, 2008.

[19] T. Tokunaga and I. Makoto, "Text categorization based on weighted inverse document frequency," Spec. Interes. Groups Inf. Process Soc. Japan (SIG-IPSJ, 1994.

[20] C. C. Aggarwal and C. X. Zhai, Mining text data. 2013.

[21] M. Hu and B. Liu, "Mining and summarizing customer reviews," Proc. 2004 ACM SIGKDD Int. Conf. Knowl. Discov. data Min. KDD 04, vol. 4, p. 168, 2004.

[22] N. Kobayashi, K. Inui, and Y. Matsumoto, "Opinion Mining from Web Documents: Extraction and Structurization," Trans. Japanese Soc. Artif. Intell., vol. 22, pp. 227–238, 2007.

[23] H. Jeong, D. Shin, and J. Choi, "FEROM: Feature extraction and refinement for opinion mining," ETRI J., vol. 33, no. 5, pp. 720–730, 2011.

[24] G. Mishne, "Experiments with mood classification in blog posts," Proc. ACM SIGIR 2005 Work. Stylist. Anal. Text Inf. Access, p. 19, 2005.

[25] S. Stymne, "Pre- and Postprocessing for Statistical Machine Translation into Germanic Languages," Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Student Sess., no. June, pp. 12–17, 2011.

[26] J. B. Lovins, "Development of a stemming algorithm," Mech. Transl. Comput. Linguist., vol. 11, no. June, pp. 22–31, 1968.

[27] H. P. Luhn, "The Automatic Creation of Literature Abstracts," IBM J. Res. Dev., vol. 2, no. 2, pp. 159–165, 1958.

[28] D. K. Ly, K. Sugiyama, Z. Lin, and M.-Y. Kan, "+Product Review Summarization from a Deeper Perspective," Proc. 11th ACM/IEEE-CS Jt. Conf. Digit. Libr., no. July, pp. 311–314, 2011.

[29] N. Archak, A. Ghose, and P. G. Ipeirotis, "Show me the Money ! Deriving the Pricing Power of Product," Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '07, pp. 56–65, 2007.

[30] D. Anguita, S. Ridella, F. Rivieccio, and R. Zunino, "Hyperparameter design criteria for support vector classifiers," Neurocomputing, vol. 55, no. 1–2, pp. 109–134, 2003.

[31] K. Duan, S. Keerthi, and A. Poo, "Evaluation of simple performance measures for tuning SVM hyper parameters. Technical report," Neurocomputing, vol. 51, pp. 41–59, 2001.

[32] J. Wainer and G. C. Cawley, "Empirical Evaluation of Resampling Procedures for Optimising SVM Hyperparameters," J. Mach. Learn. Res., vol. 18, no. 15, pp. 475–509, 2017.

[33] E. Duarte and J. Wainer, "Empirical comparison of cross-validation and internal metrics for tuning SVM hyperparameters," Pattern Recognit. Lett., vol. 88, pp. 6–11, 2017.

[34] D. M. Hawkins, "The Problem of Overfitting," J. Chem. Inf. Comput. Sci., vol. 44, no. 1, pp. 1–12, 2004.

[35] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," Encycl. Database Syst., pp. 532–538, 2009.

[36] C. W. Hsu, C. C. Chang and C. J. Lin, "A Practical Guide to Support Vector Classification," BJU Int., vol. 101, no. 1, pp. 1396–400, 2008.

Defining Network Exposure Metrics in Security Risk Scoring Models

Eli Weintraub[1], Yuval Cohen[2]
Afeka Academic College of Engineering
School of Industrial Engineering
Tel Aviv, Israel

***Abstract*—Organizations are exposed to cyber-attacks on a regular basis. Managers in these organizations are using scoring systems to evaluate the risks of the attacks they are exposed to. Information security methodologies define three major security objectives: confidentiality, integrity and availability. This work is focused on defining new network exposure measures affecting the availability. According to existing security scoring models network exposure risks are assessed by assigning availability measures on an ordinal scale using users' subjective assessment. In this work quantitative objective measures are defined and presented, based on the specific organizational network, thus improving accuracy of the scores computed by the current security risk scoring models.**

Keywords*—*Security; cyber-attack; risk scoring; vulnerability; exposure

I. Introduction

Various kinds of damages are caused to organizational computerized systems by anonymous hostile entities. Damage can range from stealing data, to changing software, or paralyzing websites [1]. Organizations' computers are exposed to attacks for long periods of time, sometimes for weeks, from the moment a vulnerability has been detected until the time a patch is prepared. According to [2] there is a need for a solution that can rapidly evaluate system damages after cyber-attacks for recovery purposes of their information system. Evaluation of potential damages is important for configuration management planning decisions. Information systems contain large amounts of software components which might contain vulnerabilities stemming from logical planning or programming bugs. Attackers plan their attacks on components having specific vulnerabilities using exploits. Organizations are exposed to damages of three kinds named 'CIA triad": Loss of Confidentiality, Integrity and Availability. Organizations wishing to defend their network should have accurate knowledge of their network, focusing on systems' vulnerabilities. This article focuses on using accurate knowledge of computers' components' characteristics and networks' configuration. Defense strategy and execution is effective only if it considers the amount of potential damage and the vulnerability characteristics [3]. Risk is defined in literature as "An event where the outcome is uncertain" [4]. The approach leading this research is lessening the uncertainty by proposing quantitative objective metrics instead of qualitative subjective assessment of organizational risk managers.

This work focuses on risks to systems' availability. There are several definitions for availability. We use the definition of [5]: Availability is the capability of an information system to make information available including all the logical and physical resources and accessible whenever they are needed. Availability is usually evaluated using the MTBF and MTTF measures. Unavailability is used to measure the percentage of components that could be impacted by an attack on systems' components. Availability is the complement to 1 of that percentage. Current security risk models use an ordinal scale of three availability measures: High, Low and No impacts on availability, assigned by organizations' users. In this work availability measures are assigned real numbers in the range [0..1]. The greater the proportion of vulnerable components is the higher is the risk.

This work proposes a new measure called 'network availability exposure', which has not been considered, so far, in current security risk scoring models. Network Availability Exposure reflects the structure and characteristics of the software/hardware components of the network and the interrelationships between the components, which contribute to achieving good / bad network availability. For example, a network containing many vulnerable software components is exposed to external attackers exploiting the vulnerable components. Literature does not define any specific measure, nor scale for calculating the exposure of systems' network configuration to attacks. The proposed measure presented in this paper is based on the real-time information of systems' configuration, as proposed by [6].

This work focuses on measuring networks' availability exposure by quantifying the impacts of cyber-attacks on network components. The quantification is based on formulas developed for this purpose. According to the proposed model, each time a new vulnerability is published or when its status is updated, the metric is calculated and risk scores are re-evaluated. Developing new accurate assessments of risk scores is critical for planning organizations' risk management activities. Risk scores based on qualitative (subjective) estimates (which are currently been used) are prone to errors, and may cause organizations to under-estimate or over-estimate the risk. This in turn may lead to under-mitigation in situations of major risks, or over-mitigation investments in cases of minor risks. By using quantitative accurate risk measures, organizations will be able to build IT configurations in proportion to the risk. According to [7] the secure management of information under the conditions of frequent changes is a complex recognized problem, but the common solution is still absent. This work defines new metrics based on the real-time network configuration and on the updated vulnerabilities of networks' components. The proposed metrics are based on three grounds: First, metrics are based on the knowledge concerning the characteristics of the attacked component within the actual configuration of the system. Second, risk calculations take into consideration the published history of actual vulnerabilities of the specific component. Third, the metrics are defined on a standard scale assigning quantitative, enabling comparisons to other networks' or other internal or external configurations.

The rest of the paper is organized as follows: Section II describes current known existing solutions. Section III presents network exposure metrics and computations. Section IV presents an illustration example. Section V concludes and suggests future research directions.

II. Existing Solutions

External vulnerabilities databases are used by security risk scoring systems for evaluating the risks organizations are facing. There are several owners of vulnerability databases [1]. Two popular systems are the Sans Internet Storm Center services and The National Vulnerability Database (NVD). Risk scoring systems make use of various parameters for estimating vulnerabilities' impacts on the target organization. Risk scores are evaluated through running a scoring algorithm while using the parameters for predicting potential attacks' damages. The Common Vulnerability Scoring System (CVSS) enables characterizing vulnerabilities and predicting risks by IT risk management professionals and researchers [1].

CVSS uses three groups of parameters: basic, temporal and environmental. Each group is represented by score compound parameters used for scoring computations. Basic parameters represent the intrinsic characteristics of the vulnerability, temporal parameters represent the vulnerabilities' specifications that might change over time due to defense activities taken. Environmental parameters represent the characteristics of vulnerabilities as configured by the specific organization, considering potential damages to that organization when exploits are being used by attackers. Basic and temporal parameters are specified by products' vendors who have the best knowledge of their product. Environmental parameters are specified by the users who have the best knowledge of their environments and attacks' impacts on their organization.

For Availability Impact (AI) evaluation CVSS uses three parameters: two base parameters 'Scope' and 'Availability', and one environmental parameter 'Availability Requirements'. Scope refers to the ability for a vulnerability in one component to impact resources beyond its privileges, assigned values 'unchanged' or 'changed'. Availability parameter measures the impact to the availability of the impacted component resulting from a successfully exploited vulnerability. Availability Requirement environmental parameter is assigned values 'high', 'medium' or 'low'. Environmental parameters include three groups of parameters indicating security importance measures in the organization: 'Confidentiality Requirement', 'Integrity Requirements' and 'Availability Requirements'. 'Availability Requirement' represents the damage to the availability of the system in case of a successful attack on a component. Thus, no environmental parameter exists for measuring networks' exposure – which is the focus of this work. The environmental group of parameters enables to customize the CVSS score depending on the importance of the impacted IT asset to a user's organization. The full effect on the overall risk score is determined by the scoring algorithm by incorporating the base impact metrics into the environmental metrics producing the overall security risk score. This work suggests adding the new environmental exposure impact measures into the computations of the availability measure. The availability measure is used for overall risk scoring computations.

All CVSS parameters are assigned ordinal qualitative values which are based on the knowledge of human experts. For example, the 'Availability Requirement' parameter is assigned values High, Medium, Low which do not differentiate between 0.99 availability and 0.999 availability, both are considered high. Also, organizations might assign a specific availability measure as high while other organizations might assign the same availability to medium. Parameter values are not based on the specific characteristics of the network. The new suggested exposure measure will be quantitative, in contrast to current metrics.

According to [8] unavailability is not an option in todays' echo systems, given the heavy dependence of modern organizations on information resources. Availability is the least discussed and researched security attribute, although it is not the least important attribute. In fact, it plays an important role in determining the other attributes of security (confidentiality and integrity). According to [5] availability measurements should take into consideration the logical and physical resources, to enable accessibility whenever the information is needed. In [8] describe the factors availability is dependent upon as software, hardware and network. A system might be exploited step by step by gaining access from the upper layers. Current models do not take into consideration those issues. The proposed model measures network exposure to the actual known vulnerabilities. The model considers networks' components exposure specifically and evaluates network exposure by considering components' interrelationships.

According to Federal Information Processing Standards (FIPS) 199.5 [9] organizations assign their IT resources importance measures based on component location, business function using it, and potential losses in case the component is damaged. For example, U.S. government assigns every IT asset to a group of assets called a system. Every system must be assigned three "potential impact" ratings according to three security objectives to represent the potential impact on the organization in case the system is compromised. Thus, every IT asset in the U.S. government has a potential impact rating with respect to security objectives. CVSS follows this general model but does not require organizations to use specific tools for assigning the impact ratings. In [10], author states that organizations should define the specifications of security risks of their specific environment. The Department of State has implemented a scoring program called iPost that is intended to provide continuous monitoring capabilities of information security risks for IT infrastructure. According to [11], the iPOST scoring model does not define the base scores of CVSS to reflect the characteristics of its specific environment. This work presents a model aimed to close this gap.

Quantification of the environmental parameters in CVSS algorithm has been recently presented in a research demonstrating improvements in accuracy of risk scores by using the actual IT configuration [12]; Incorporating the information relating to the actual IT components into the scoring algorithm metrics changes the risk scores to be objective rather than subjective measures [13]. Moreover, components' specifications are expressed in high resolution of the smallest IT elements rather than an overall configurational metric which shades smaller components' characteristics. Risks quantification is based on a configuration management database system (CMDB) which makes use of systems' metadata on the elementary components and their interrelationships according to security specifications (ibid). This paper continues the same line of research, aimed at improving risk scoring accuracy in relation to existing risk scoring models such as CVSS, by adding a quantitative metric rather than an ordinal subjective assessment and basing evaluations on the specific organizations' configuration.

III. Network Exposure Metrics and Computations

The proposed approach produces estimates to the risk of losing availability, meaning the risk of system malfunction or system failure. It gives both overall network measures, as well as risk measures for single components. An information system is consisted of one computer operating many hardware /software components, or a communication network consisting of many computers communicating between each other. The network is represented as a graph, components are represented by nodes. Links between nodes represent information passing between the connected nodes. Components represent hardware or software entities. Some definitions are now in order to be used for further developments.

A. Definitions

Diameter – Minimal number of links between two points defined on the edge of the network or subnetwork.

Working nodes – Uncompromised nodes

Impacted nodes – Compromised nodes

Impacted link – A link having at least one compromised node.

M – Total number of impacted nodes

m – Number of impacted nodes having at least 2 neighboring impacted nodes

d - Maximal diameter of the network generated by the impacted (compromised) nodes

D' – *M*aximal total network diameter

D – Degree = Number of links emanating from a node

Dir – Number of impacted links emanating from a node

Sec - Number of links to risky nodes, which are still working but have direct links to impacted nodes.

W – The number of arcs connected to working nodes

For the reader's convenience we shortly repeat the definitions in the proposed availability measures. In case, where the organizational network is attacked the following three network measures are suggested.

1) **Damage%:** Percentage of nodes impacted: range of values is [0 to 1]. This measure represents the damage to the whole network, calculated by computing the percentage of impacted nodes out of all nodes.

2) **Dispersion:** The ratio *d/*D' between *d*=maximal diameter of the impacted nodes, and *D'*=maximal total network diameter: range of values is [0 to 1]. *D'* the number of links in the maximal shortest path that connects between each pair of nodes; and *d* is the number of links in the maximal shortest path that connects between each pair of un-impacted neighbors of the impacted nodes.

3) **Concentration:** The ratio of *m/M* between *m*= number of impacted nodes with at least 2 neighboring impacted nodes, and *M*=total number of impacted nodes. Range of values is [0 to 1]. The measure represents the proportion of the seriously impacted nodes. A concentration of disconnected nodes signifies the severity of the impacts on systems' availability. possibly leading to situations of paralyzing the impacted area. Nodes connected to 2 or more impacted nodes might be disconnected more easily.

In case where the organization is a node in the larger network, and in cases of measuring the exposure of a specific node in an organizational network, the following three node risk measures are proposed:

1) **Directs**: defined as *Dir/D*; where *D* is the node degree (number of arcs emanating from it), and *Dir* is the number of impacted direct links: range of values is [0 to 1]. This measure represents the proportion of the directly impacted links between the node and its' neighboring nodes.

2) **Seconds:** defined as *Sec/D* where *D* is the node degree (number of arcs emanating from it), and ***Sec*** is the number of links to working nodes having direct links to impacted nodes. The Seconds range of values is [0 to 1]. The measure represents the proportion of links to working nodes which have direct links to other impacted nodes. Those working nodes are exposed now to risks in cases of a second forthcoming attack coming from an already attacked component, thus, leading to a direct attack on the subject component.

3) **Disconnection risk**: defined as (1/*W*), where *W* is the number of arcs connected to working nodes (with the exception that when this number is zero (disconnection), *W*=1). Thus, when only one arc is connected to a working node *Disconnection risk* =1; and when all arcs are connected to working nodes, the *Disconnection risk* =1/*D* (where *D* is the node degree). The measure represents the efforts needed by an attacker who wishes to disconnect a node. The proportion means he needs to disconnect W links. As W is higher the effort are higher and the organizations' disconnection risk is smaller.

IV. Illustration Example

The following case study illustrates the suggested measures, their computations, their meaning and effectiveness and their importance. Fig. 1 describes an example network with 20 computerized nodes after a first wave of attacks which culminated with failures of nodes 14, 16, and 17.

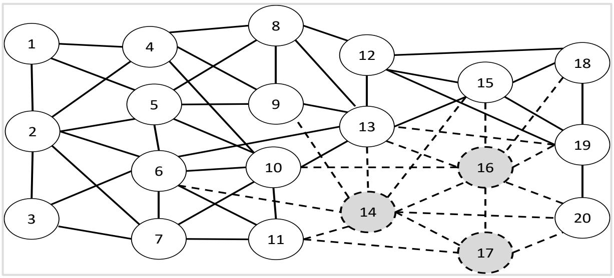


Fig. 1. Computer Nodes Network after attack wave – 1.

The 20 computer nodes network example with 3 impacted nodes (denoted by grey and dashed lines) and dashed lines are disconnected communication lines due to the hackers' attack.

In Fig. 1 the network exposure measures are:

1) **Damage %:** Percentage of nodes impacted: 3/20= 0.15.

2) **Dispersion**: The ratio *d/D'* =3/5 = 0.6. Where *d* the diameter of the impacted nodes: **3**, and ***D'*** the network diameter is *5*.

3) **Concentration**: The ratio of *m/M* = 3/3 = 1 the number of impacted nodes: **3** each is connected to the other two.

To illustrate the node measures for Fig. 1, we chose to compute measures for nodes: 20, 15, 9, 6, 3. For brevity purpose we shall use: "Directs" for directly impacted neighbors, "Seconds" for second degree impacted neighbors, and "Arcs" for the degree of the non-impacted node.

For node 20: *Directs* = 3/4, *Seconds* = 1/4 (node 19 "Directs">0), *Disconnection risk* =1/1=1

For node **15**: *Directs* = 2/6, *Seconds* = 3/6 (13, 18, 19 have "Directs">0), *Disconnection risk* =1/4

For node **9**: *Directs* = 1/5, *Seconds* = 1/5 (node 13 have "Directs">0), *Disconnection risk* =1/4

For node **6**: *Directs* = 1/8, *Seconds* = 3/8 (10, 11, 13 have "Directs">0), *Disconnection risk* =1/7

For node **3**: *Directs* =0, *Seconds* = 0, *Disconnection risk* =1/3

It follows that immediate risk is highest for node 20 with *Directs*=3/4 (followed by node 15), while risk evolution potential is highest for nodes 15 and 6 with *Seconds*=3/6, and 3/8, respectively, and disconnection risk is highest for node 20 (*Disconnection risk* =1).

The example follows with hacker attack evolution as depicted in Fig. 2.

Fig. 2 illustrates the example network with 20 computerized nodes after a second wave of attacks which culminated with failures of nodes 9, 13, 19 (in addition to previously failed nodes: 14, 16, and 17).

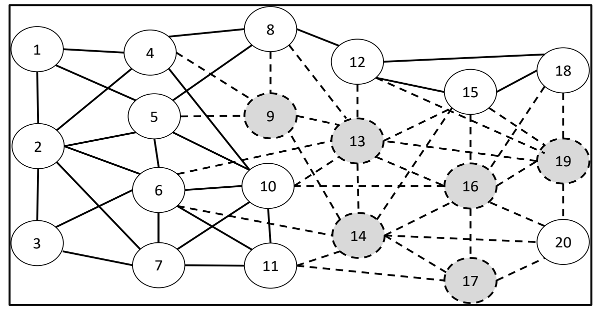


Fig. 2. Computer Nodes Network after attack wave – 2.

The 20 computer nodes network example with 6 impacted nodes (denoted by grey and dashed lines - dashed lines are disconnected communication lines due to the hackers' attack).

In Fig. 2, the network exposure measures are:

1) **Damage %**: 6/20= 0.3 twice the number in Fig. 1.

2) **Dispersion**: The diameter of the impacted nodes: 4/5 (the number of links in the maximal path that connects minimally between each pair of un-impacted neighbors of the impacted nodes) for example: 20-16-13-9-4. Diameter of 4 is an increase from 3 in Fig. 1.

3) **Concentration:** The largest connected group of at least two neighboring impacted nodes = **6** which is as large as the number of impacted nodes=6. So: 6/6=1.

To illustrate the node measures for Fig. 2, we compute measures for nodes: 20, 15, 9, 6, 3. These could be easily compared to the measures for Fig. 1.

For node 20: *Directs* = 4/4, *Seconds* = 0, *Disconnection risk* =1 (disconnected)

For node **15**: *Directs* = 4/6, *Seconds* = 2/6 (12, 18 have "Directs">0), *Disconnection risk* =1/2

For node **9**: Impacted.

For node **6**: *Directs* = 2/8, *Seconds* = 3/8 (5, 10, 11, have "Directs">0), *Disconnection risk* =1/6

For node **3**: *Directs* =0, *Seconds* = 1/3 (node 6 have "Directs">0), *Disconnection risk* =1/3

Thus, node 9 failed, node 20 is disconnected (*Disconnection risk* =1), node 15 faces high immediate and intermediate risk along with disconnection risk. Node 6 has immediate and intermediate risk exposure, but its disconnection is unlikely (*Disconnection risk* =1/6). Finally, node 3 has small intermediate risk exposure (Seconds=1/3).

- Attack Evolution Analysis

Table I summarizes the node exposure metrics evolution along two waves attack. Analyzing node exposure evolution might help decision makers in planning their network mitigation activities. Looking at to Directs column might lead to decision of defending node 15 which has the highest direct-damage measure. Looking at the secondary-damage probability might lead to defending node 15 with highest measure, but in case we have already decided to defend it, we might decide now to defend node 6. Looking at the disconnection column leads to the understanding that it would be reasonable to defend node 15, but if we have already defended it then we will defend node 3.

TABLE I. ATTACK EVOLUTION: 2-WAVES EXAMPLE

Node/wave	Direct impact	Secondary impact	Disconnection
20 1 2	**0.75** **1**	**0.25** **0**	1 1
15 1 2	**0.33** **0.67**	0.50 0.33	**0.25** **0.50**
9 1 2	0.20 **impacted**	0.20	0.25
6 1 2	**0.125** **0.25**	0.375 0.125	**0.14** **0.17**
3 1 2	0 0	**0** **0.33**	0.33 0.33

These measures help decision makers in prevention planning activities. For example, node 20 would be very interested in adding a link to node 15, or 12.

Moreover, the new measures could help in restoration priorities. The impacts of restoring alternative nodes could now be simulated. For example, suppose that the resources for restoration are limited to one node at a time. Comparing alternatives would be an efficient decision support tool. For example, in Fig. 2: comparing restoration of node 16 to node 17 yields the following measures, referring to Fig. 2. A single restoration would change the number of impacted nodes from 6 to 5 and the impacts on network measures are relatively small:

For restoring node 16 – network measures: Damage % =5/20; Dispersion=4/5; Concentration=4/6

For restoring node 17– network measures: Damage %=5/20; Dispersion=4/5; Concentration=5/6.

Thus, analyzing network exposure measures leads to a decision to restore node 16.

Let's see now the impacts on the neighboring nodes.

The impact of restoration is always on specific neighboring nodes. So, for restoring node 16 vs. 17: the neighboring node of 16 and 17 are nodes 20 and 15. The measures for node 20 are the same, but for node 15 the measures are in favor of 16.

For node 20: *Directs* = 3/4, vs. 3/4, *Seconds* = 1 vs. 1, *Disconnection risk* =1 vs. 1

For node 15: *Directs* = 3/6, vs. 4/6, *Seconds* = 3/6 vs 2/6 (12, 18 have "Directs">0), *Disconnection risk* =1/3 vs. 1/2

Thus, restoring node 16 (vs. restoring node 17) has smaller direct risk, higher secondary risk, but more arcs meaning less disconnection risk.

Thus, after analyzing both, network exposure and node measures, it may be concluded that restoring node 16 has a priority over node 17.

The standpoint of the authors of this paper stress that current risk scoring models should be enhanced regarding all security objectives. The enhancements should be implemented taking several phases: First, incorporating the real configurations' characteristics into the scoring model. Secondly, transforming all the parameters defined as inputs to the scoring model as quantitative measures reflecting the actual organization environment rather than ordinal subjective users' assessments. Thirdly, scoring models should define new metrics expressing all risk objectives, relating to a whole environment rather than a specific component. The models should define the characteristics of a whole network and all interrelationships between the network and each of its internal components. Current scoring models do not support modeling the characteristics of the whole network, as has been shown in this research, dealing with the availability security measure. This is just one first example. The fourth phase should include metrics considering the time dimension such as this paper suggests: measuring the development of damages caused to a network as a continuous attack spreading in the different zones of the network. Such time-related measures should incorporate more complex graph-theory models and prediction algorithms looking at the historical network changes.

V. CONCLUSIONS

This work presents a risk assessment model with a focus on new availability risk measures, measuring the network security and node security, as part of an overall risk assessment. According to the proposed model, CVSS will use the new network exposure environmental parameters which are evaluated using a new formula based on the configuration of the system. The new measures are quantitative, normalized to [0..1] and based on the actual networks' configuration. This is contrary to the existing models which do not measure the impacts of network exposure on the overall risk, but consider networks' configuration implicitly relying on intuitive users' subjective assessment.

Regarding the practical use of the theoretic results, the model helps risk managers in assessing the damages caused to firms' components in occasions of cyber-attacks. Using the proposed model will enable predicting accurate measures of organizational damages taking in consideration components' actual characteristics and the availability exposure of the network as an integrated entity. The suggested model enables efficient risk mitigation planning and improved defense to organizations. Risk managers will be able to assign risk budgets according to accurate and actual risk measures, thus focusing on the high-priority risks, and preventing unnecessary budgets to low-order risks.

The new network exposure metrics enable customizing the CVSS score to the characteristics of the attacked IT asset, its interrelationships to other assets, and the exposure characteristics of the network on the damages to user's organization. According to the proposed model, a formula which assigns quantitative measures to exposures' impacts based on the actual updated vulnerabilities of the specific component. The proposed model outlines the structure of a CMS which uses the real organizational environment and components, and the processes which update the network exposure parameters with the planned and actual values. The framework enables getting accurate risk measures, thus enabling the organization making better risk management decisions, allocating risk management budgets in proportion to the risks.

Limitations of the study are related to two issues: First is the feasibility of a managing a graph describing the real-time status of each component and its interrelationships to other components, including all security characteristics. Such a graph might be difficult to update at every attack or configuration-change. Second limitation refers to the various connections possible between two nodes. It is reasonable to assume that not all connections enable transferring the cyber-attack to other nodes. It is possible to assume that there are connection-types that do not transfer the attacker to other connected nodes. This issue is a limitation of the current model and also a research issue.

Further research directions are:

- Designing ways of incorporating the new availability metrics in CVSS framework.
- Algorithm formalization including complexity measures calculations.
- Using more expressive graph models to represent varying node-types and varying connections among network nodes. It is reasonable to assume that certain connections may have more impact on the damages caused to neighboring nodes.
- Further development of the algorithm to calculate N-wave metrics and measures predictions of the N value at which the whole network will be paralyzed.
- Research aimed at predicting attack evolution for decisions concerning mitigation activities for the long run of future attacks.

Future improvements may focus on building a full dash-board of system exposure metrics from its component measures.

More research is needed in studying the impacts of the various proposed network exposure measures on the overall security risk score.

References

[1] P. Mell, & K. Scarfone, & S. Romanosky, "CVSS – Common Vulnerability Scoring System v3.0: Specification Document", FIRST Org, 2015.

[2] Y.F. Nũez, "Maximizing an organizations' security posture by distributedly assessing and remeding system vulnerabilities", IEEE – International Conference on Networking, Sensing and Control, China, April 6-8, 2008.

[3] S. Tom, & D. Berrett, "Recommended practice for patch management of control systems", DHS National Cyber Security Division Control Systems Security Program, 2008.

[4] A. Terje, & R. Ortwin., "On risk defined as an event where the outcome is uncertain", Journal of Risk Research Vol. 12, 2009.

[5] D. Khazanchi, & A.P. Martin, Information Availability: Handbook of Research on Information Security and Assurance, 2008.

[6] E. Weintraub., "Security Risk Scoring Incorporating Computers' Environment", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7(4), April 2016.

[7] B. Carpenter, "The Internet Engineering Task Force. Overview, Activities, Priorities". IETF Report to ISOC BoT, Oct. 2006.

[8] S. Qadir, & S.M.K. Quadri, "Information Availability: An insight into the most important Attributes of Information security", Journal of Information Security, pp. 185-195, 2016.

[9] K. Dempsey, & N.S. Chawia, & A. Johnson, & R. Johnson, & A.C. Jones, & A. Orebaugh, & M. Scholl, & K. Stine, "Information security continuous monitoring (ISCM) for federal information systems and organizations", NIST, 2011.

[10] R. Sandhu, & D. Ferraiolo, & R. Kuhn, "The NIST Model for Role-Based Access Control: Towards A Unified Standard", George Mason Univ., 1999.

[11] A. Keller., & S. Subramanianm., "Best practices for deploying a CMDB in large-scale environments", Proceedings of the IFIP/IEEE International conference and Symposium on Integrated Network Management, pages 732-745, NJ, IEEE Press Piscataway, 2009.

[12] E. Weintraub, "Evaluating Damage Potential in Security Risk Scoring Models", International Journal of Advanced Computer Science and Applications (IJACSA), 7(5), 2016.

[13] E. Weintraub, Y.Cohen, "Security Risk Assessment of Cloud Computing Services in a Ne tworked Environment" International Journal of Advanced Computer Science and Applications (IJACSA), 7(11), 2016, 79-90.

Recovery of User Interface Web Design Patterns using Regular Expressions

Khalid Mahmood*1
*Department of Computer Science
COMSATS Institute of Information Technology, Lahore, Pakistan
*Division of Science and Technology University of Education,, Lahore, Pakistan

Faiza Tariq[2]
Division of Science and Technology
University of Education,
Lahore, Pakistan

Dr. Ghulam Rasool[3]
Department of Computer Science
COMSATS Institute of Information Technology
Lahore, Pakistan

***Abstract*—User Interface Web Design Patterns are standard solutions for the development of web applications. The recovery of these patterns from web applications supports program comprehension, reusability, reverse engineering, re-engineering, and maintenance of legacy web applications. The recovery of patterns from web applications becomes arduous due to the heterogeneous nature of web applications. Authors presented different catalogs and recovery approaches for extracting User Interface Web Design Patterns from source code in last one and half decade. There is still a lack of formal specifications for web design patterns, which are important for their recovery from source code. The objective of this paper is to specify User Interface Web Design Patterns (UIWDP) using semiformal specification technique and use these specifications for the recovery of patterns from the source code of web applications using regular expressions. 55 feature types are identified for the specification of 15 UIWDPs. We evaluated our approach on 75 randomly selected web applications and recovered 15 UIWDPs. The standard deviation, precision, recall and F-score measures are used to evaluate the accuracy of our approach.**

Keywords*—Design patterns; user interface patterns; web applications; web reverse engineering; regular expressions

I. Introduction

The World Wide Web has surprisingly affected many aspects of our life and it will continuously influence society. Due to increasing popularity and use of web applications (WAs) in all fields of life, the WAs are subject to continuous evolution, maintenance and re-engineering [1], [10]. WAs demand continuous evolution due to different reasons such as improvement in usability, quality, efficiency, correction of bugs, introducing the new functionality and modifications in legacy WAs [22], [27]. User Interface Web Design Patterns are standard solutions that are frequently used for developing web applications. The appropriate use of UIWEDPs not only creates consistency among different web pages of web applications, but it also establishes efficient layout of web pages [3]-[6]. The automatic recovery of UIWDPs supports comprehension, maintenance, evolution and reengineering activities.

Web application development too often had no design and modeling principles in early days of web application development processes. Mostly, developers apply ad hoc approaches which make the process of extracting information from web applications difficult. The recovery of information from legacy WAs is a daunting task due to the heterogeneous nature of web applications. Web applications are developed using different technologies such as Java Scripts, HTML, DHTML, XHTML, XML, CSS, ASP, PHP, DOM, database, images etc. Single language based static analysis tools are not capable to extract information from multilingual aspects of WAs.

A major motivation for reverse engineering of web applications is to reuse the tested, reliable and valid artifacts of legacy applications in the development of modern applications. The recovery of UIWDPs helps to comprehend architecture of web applications. The UIWDPs are not just about buttons and menus; they are about the interaction between the users and applications or devices. The UI WDPs are used to create consistency throughout the web development process to give the best, attractive, user-friendly, most usable and effective layouts to the WAs. The different authors have presented various catalogs of UIWDPs in the last one and half decade such as UI pattern library [4], Welie patterns library for interaction design [5], 10 UI design patterns by Jovanovic [9] and 15 UI design patterns by Kayla Night [6]. All the authors of these lists focused only on the organization, usage, naming, problem summary and examples of UIWDPs.

Most previous contributions [11]-[15] in the field of UIWDPs focused on the collection, organization and applications of UIWDPs. The recovery of these patterns from WAs requires standard and formal definitions which are not yet available to the best of our knowledge. Some approaches recover different components and models from web applications, but they did not focus on recovery of specific patterns with complete information. The concept of recovering UIWDPs is applied in [31]. The applied approach only extract whether a pattern is present in a web application or not. The approach is also limited to recognize patterns with fixed tags and fails to recognize the same pattern when it is implemented by using different constructs. Our approach recognizes patterns from WAs along with complete information about the location of patterns which is important for the analysis of source code. We propose a lightweight approach based on lexical analysis to search patterns by file name, file path, line number and its occurrences in the WAs. The regular expression patterns are used in this paper and are easily customizable while handling variations in the detection of UIWDPs.

Our work focuses on following contributions:

- Summarization of up-to-date state of the art work on web design patterns recovery;
- Semiformal specifications of UIWDPs;
- Automatic recovery of UIWDPs from web applications;
- Evaluation of approach.

The rest of this paper is organized as: In Section II, related work is discussed. Section III presents semiformal specifications of the web user interface design patterns used by our approach. This paper presents concept and architecture of our approach in Section IV. Section V presents the evaluation of approach and Section VI concludes the whole approach.

II. Related Work

A number of approaches and tools are presented for recovery of information from web applications. The comparative overview on features of different approaches presented by different authors is given in Table I. The key factors for comparison are tool support, model/framework, source code languages, Analysis type, technique applied and experimental case studies. The important approaches are listed so far in Table I and discuss selected approaches in the following paragraphs.

According to the experiment of the Carlos et al. [28], the proper use of UIWDPs [7] in the development of web applications had positive impact on the quantitative/qualitative performance and usability of WAs [8]. Different libraries of UIWDPs such as Danish web developer [4], Welie (a pattern library for interaction design) [5], 10 UI design patterns by Jovanovic [9] and 15 UIWDPs by Kayla Night [6] are presented. The information about patterns in these libraries is available in informal languages.

Yingtao et al. [37] presented a reverse engineering approach to extract presentation layer from web applications. They specify recovered features in the form of WSDL and these features can be deployed through proxies accessing the original web server and parsing its responses. Authors recover functionalities of websites from the presentation layer instead of focusing on all the source code. The applied recovery process consists of five Components (page collector, pattern miner, pattern visualize, translator and service interface editor).

Amalfitano et al. [34] presented a tool for automatic reverse engineering of dynamic web applications using source transformation technology. Authors extract UML sequence diagrams from the execution traces generated by the resulting instrumentation. The result can be directly imported and visualized in a UML toolset such as Rational Software Architect. The extracted results can be imported and visualized in any UML 2.1 toolset. Authors apply filter execution traces directly on information stored in a database that automatically eliminates redundant information which complicates the understanding process. The approach was limited only to reverse engineering of PHP-based web applications.

Bochiha et al. [15] proposed a semi-automatic approach for re-engineering of multi-language based WAs and recovered the conceptual model by using syntactic and semantic information hidden in the LWAs. The presented framework is based on two steps: web application reverse engineering and SWS forwarding engineering. Reverse engineering step extract class and activity diagrams and forward engineering process generates WSDL and WSMO semantic descriptions. As compared to our approach, authors did not focus on the features of UI patterns that are necessary for pattern's recovery, maintenance and re-engineering from LWAs [23].

Staiger [55] presented an approach for reverse engineering of GUI components from different applications using static analysis. He extracted control flow graph for the examined applications. Author maps source code constructs with GUI components and detects relationship between GUI components through event handlers and their callers. Bauhaus tool is applied by the author to extract GUI components from the source code of C/C++.

Norizan et al. [16] performed the survey to extract the list of user interface design patterns and their impact on the usability of WAs. In their approach, they tried to recover groups of UIWDPs that were used collectively and their impact on the quantitative/qualitative performance and usability of WAs with help of check list based survey. But in our approach, we recover a single or group of UI patterns from LWAs on the basis of pattern's features.

A number of approaches [10]-[13], [35], [36] recovered UML models for the comprehension of behavioral, structural and relational aspects of WAs but these approaches did not focus on the recovery of UIWDPs and their occurrences in WAs. The Marchetteo et al. [14] proposed ReAjax tool to reverse engineer only Ajax based applications. Echeverrıa et al. [17] used MoDisco to reverse engineer only the Strut based WAs. Martin et al. [2] proposed reusing of software engineering tool Rigi as a means of analyzing and visualizing the structure of web applications. Chung and Lee [8] proposed an approach for reverse engineering of Websites and adopt Conallen's UML extensions [53] to describe their architecture. Rasool et al. [32] presented an approach to recover general artifacts from legacy software applications based on abstract regular expressions. Draheim et al. [56] presented a tool that constructs analysis from website based on the concept of from-oriented analysis. Bouchiha et al. [57] presented an ontology based approach for reverse engineering of web applications. Boldyreff et al. [18] proposed a system that exploits traditional reverse engineering techniques to extract duplicated contents and styles from websites in order to restructure them and improve their maintainability.

Authors in approaches [17]-[20], [24], [25] used Eclipse and WARE [21] tools to create the intermediate representations of LWAs. They integrate reverse engineering and model driven engineering to extract the conceptual models of LWAs and user interaction artifacts from web applications. All applied approaches did not address the locations and occurrences of UI patterns in the code that are necessary for maintenance, correction of bugs and abstraction. Key features of different approaches are summarized in Table I.

We see from Table I that most reverse engineering approaches presented for recovery of information from web applications start their process from the selection of a model or a modeling language that will generate an intermediate format of the legacy web application. These intermediate formats may be used by one or more tools for recovering different artifacts from the source code. Some approaches use transform techniques, queries and algorithms to recover different types of artifacts and their relationships from the web applications. Many recovery tools accept models as input and recover different types of information from examined applications. The web models act as main source for the comprehensive reverse engineering of legacy web applications.

III. Semiformal Specification of UIWDPs

The list of UIWDPs proposed by Kayla Night [6] has been used by our approach. We realize that patterns presented in this list are more generic which include maximum UIWDPs that are included in other lists as well. We take user interface graphical items, lexical items and feature types for the specification of patterns. Graphical items are specified with the help of different tags such as HTML, Form, DIV etc. Lexical items are captions/ labels that describe graphical items. For example, user login user name and password are lexical items that are used to describe graphical items. We use the concept of feature types presented in our previous work to specify all 15 UIWDPs [50]. The feature types presented in this paper are different from feature types used to specify GoF (Gang of Four) design patterns in previous work. Feature types in this approach are based on graphical and lexical properties of patterns. They refer to different characteristics of patterns implemented by developers using different tags. The worth and quality of our specifications are based on the appropriate selection of feature types that are used by web designers in the specific area with specific sequence to implement the UIWDPs in the web applications. We identified 55 feature types presented in Table II to specify all 15UIDPs and their variants. We don’t claim that these 55 features can be used to specify User Interface design patterns of different other lists. These features are presented in Table II. We select four UIWDPs (Login, Navigation Bars, Bread Scrum, Lazy Registration) to demonstrate our approach with their intent, visual diagrams and specifications as given in the following subsections. The specifications of rest of patterns are available on our web source [29].

A. Login Pattern

Login Pattern is very common in web applications. The intent and specification of login pattern is given below:

A login pattern [26] is required when users need to identify themselves to either gain access to a restricted area or experience a more personalized user interface based on information provided previously. Fig. 1 below presents the screenshot of a Login pattern.

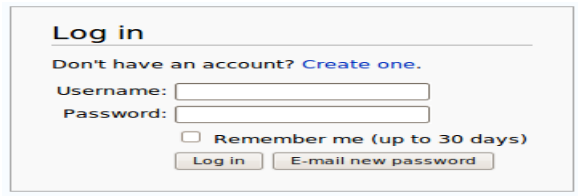


Fig. 1. Login pattern.

- Specification

Specification of login pattern described as feature types and diagrammatic notation is presented in Table III and Fig. 2. We consider variations used by developers while implementing these user interface web design patterns. Login pattern may have features (F11, F12, F13, F14, F15) or (F11, F12, F13, F14, F15, F16).

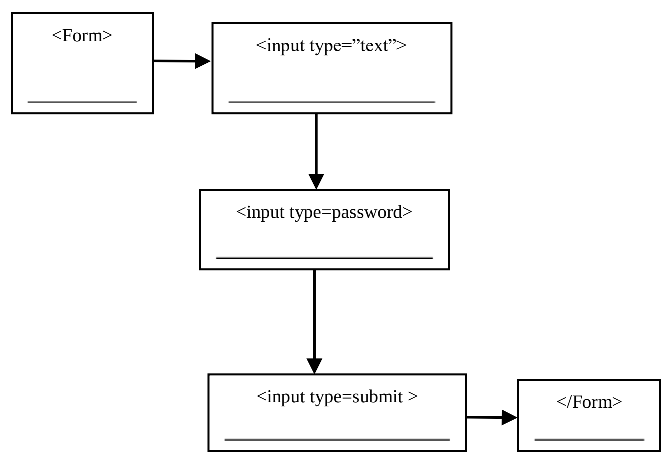


Fig. 2. Login patter.

B. Navigation Bar Patterns

Navigation Bar pattern has become a key feature of all web applications. The intent and specification of Navigation Bar pattern is given below:

Navigation Bar pattern is used when the user needs to locate contents and features necessary to accomplish a task. A vertical/horizontal navigation is a quite common layout that gives much emphasis to the vertical/horizontal orientation. Fig. 3 presents a screenshot of Navigation Bar pattern.

TABLE I. COMPARATIVE OVERVIEW OF FEATURES OF WEB RECOVERY APPROACHES

Reference/Tool	Model /Framework	Technique	Source Code Languages	Analysis Type	Experimental Case Studies
[38] (2001) OOHDM-Web (Framework)	OOHDMFrame. OOHDM-Web(UML)	Manual observation	OO code, ASP ISAPI, COM, Corba	Behavioral &,Object oriented analysis	Web application framework for conference paper review system.
[18] (2001) JavaCC parser	Rational Database	Perl, SQL	HTML, Java script	Code analysis Parse tree analyzer	Palatinate'99, SEG'99 Palatinate'00, Personal Web
[39] (2001) VAQUISTA	Presentation model XIML	Mapping tables matching techniques	HTML4.0	Static Analysis	http://www.acm.org/sigchi/chi2001/registration.html).
[40] (2004) Saxon TidyR	(MultiLingual XHTML) UML	Page matching algorithm	HTML, PHP	Static analysis Syntax tree analysis	www.ien.it.
[41] (2002) WebRemUSINE	Task models	WebRemUSINE	HTML	Static analysis	http://giove.cnuce.cnr.it/sigchi/index.html
[10] (2001) ReWeb TeslWeb	UML Model Meta model	Test cases Node-Reduction algorithm	HTML	Static analysis Dynamic analysis	www.Wordnet.com www.Amazon.com
[13] (2012) WebLabUX	DOM, Three experiment patterns	Survey responses t-test.	N M	users' navigational behavior on a website	NM
[14] (2012) ReAjax	DOM GUI-based state models	Regular Expression FSM(finite state machine)	HTML, CSS Java script, PHP XUL,XML	Static analysis. Dynamic analysis Execution traces	[http://pafm.sourceforge.net], [http://tudu.sourceforge.net]
[15] (2010) OntoWeR	WSMO, SWS class & activity diagrams,UML Model	Transform Algorithms OWL	HTML, PHP JSP,	Syntactic and Semantic Analysis	Self made example
[42] (2008) SEASAT	MDE, Class diagram link-based models	Explanation based on 11 rules	HTML, Java script	Lexical analysis: Syntactic analysis	NM
[43] (2009) Kenyonweb,	Mining Software Repositories (MSR) DBMS	Classification algorithm & Regular expression	Java code	NM	Selected java code
[31] (2008) Google search engine, SPSS	Web mining. User interface Web patterns.	Query with degree scale (+,-,?) Descriptive statistics	NM	Semantic analysis	Nokia mobile websites
[44] (2012) MoDisCo	Navigational models MVC WebFrameworks	ATL Algo Data/Control Flows	HTML, JSP XML, Java	Static analysis.	http://www.unex.es/eweb/migraria/cs/agenda
[45] (2013) RE-UWA UWA-MDD	Model Driven Reverse Engineering(MDRE) UWA models	Query View Transformation - QVT, ATLrules	Java beans, JSP, HTML	Conceptual Analysis Object oriented analysis	NM
[46] (2013) WARE	Re MDWE Meta-Model	Clustering Algorithm, SQL Queries	HTML,JavaScript,XML	Static analysis Dynamic analysis	General Example
[20] (2013) MIGRARIA WARE	MDWE MVC-based web frameworks.	Atlas transformations ATL rules Querying	JSP, Java, XML HTML	Static analysis Dynamic analysis	http://www.eweb.unex.es/eweb/migraria/cs/agenda/
[47](2010) JWebTracer JBPRecovery.	GUI-based Reverse Engineering, UML	Clustering techniques, BPMN, BPEL, Petri Nets	FSS, CSS, APTS	Dynamic analysis	www.softslate.com/
[34] (2009) WARE, PHP2XMI	Database model	ER model, SQL Queries	PHP,SQL HTML,	Static and dynamic analysis	NM
[21] (2007) WSDL2OWLS	MDRE, Model Driven Architecture (MDA)	(QVT) Algo [10] Owl ATL algo	WSDL files	Static, Dynamic and Semantic analysis	NM
[48] (2001) MySQL ,PHP interpreter, JSQLParser,	Application Model, System Overview, Database	SQL queries	PHP, Java	Dynamic analysis	Wordpress, Drupal and Gallery2.
[49](1996) PhpModeler, ReWeb	Reverse engineering (RE) process	Dependency analyzer Algorithm SQL	HTML,CSS, PHP,ASP.NET, Java,JavaScript,	Static analysis	Simple login page iForestFire
[54](2012)	DOM GUI-based State models	Execution traces on State modes	HTML, CSS Java script, PHP XUL,XML	Dynamic analysis	[http://pafm.sourceforge.net], [http://tudu.sourceforge.net]
[11](2004)	XMI	Mutational Techniques	HTML, ASP	Static and Dynamic analysis	XML node construction application

TABLE II. FEATURE TYPES

Feature No	Feature Tag	Feature Description	Feature No	Feature Tag	Feature Description
F1	<UL>	List start point	F29	<Footer>	Footer start
F2	</UL>	List end point	F30	</Footer >	Footer end
F3	<LI>	Item start point	F31	<Label >	Label Text start
F4	</LI>	Navigation item end point	F32	</Label >	Label Text end
F5	<a>	Anchor start	F33	<Section>	Sections start
F6		Anchor end	F34	</Section>	Section end
F7	<DIV>	Division start	F35	<article>	Sub section
F8	</DIV>	Division end	F36	</article>	Sub section end
F9	<SPAN>	Navigation item area start point	F37	<header >	Sub section title/heading start
F10	</SPAN>	Navigation item area end point	F38	</header>	Sub section title/heading end
F11	<Form>	Start of form	F39	<h>	Heading style start
F12	</Form>	End of form	F40	</h>	Heading style end
F13	<input type=”text”>	Text data field	F41	<p>	Para start
F14	<input type=”password”>	Password data field	F42	</p>	Para end
F15	<input type=”submit”>	Submit button	F43	<Strong>	Text Highlighter start
F16	User name /f-name/-l-name	Label/lexical item L1	F44	</Strong >	Text Highlighter end
F17	“Password”	Label/lexical item L2	F45	Digits	Digits
F18	“Login”/text	Label/text	F46		Order list
F19	<Table>	Start of table	F47		Un order list
F20	</Table>	End of table	F48	<Script>	Script language start
F21	<TR>	Start of row	F49	</Script>	Script language end
F22	</TR>	End of row	F50	<tbody>	Table body start
F23	<TD>	Start of column	F51	</Tbody>	Table body end
F24	</TD>	End of Column	F52	<!DOCTYPE>	Html Version
F25	<input type=”hidden”>	Submit with action	F53	<html><head><title>	Webpage head
F26	<button>	Button for action	F54	<linkrel="stylesheet>	CSS link
F27	<IMg>	Image icon	F55	<body>	Page body
F28	Id=Breadcrumb	Div tag Id which contain the path to this page			

TABLE III. FEATURES OF LOGIN PATTERN

UI graphical items	Lexical item	Feature types	Pattern definitions
<form> <input type="text"> <input type="password"> <input type="submit "> </form>	. Login Username Password	F11:Start of form F12:End of form F13: Text data field F14: Password data field F15: Submit button F16:Label/lexical item L1 F17:Label/lexical item L2	F11, F13,*F14,F15,F12 OR F11,F16, *F13,F17,* F14,F15,F12 *: Means that feature type can repeat

Fig. 3. Navigation pattern.

- Specification

Features of Navigation Bar pattern are given in Table IV and diagrammatic specification is presented in Fig. 4. "*" means that a feature type or group of feature types can repeat. The first variant of Navigation Bar pattern has repeating features (F3, F5, F4) as shown in Fig. 4. Features for other variants of Navigation Bar are given in Table IV.

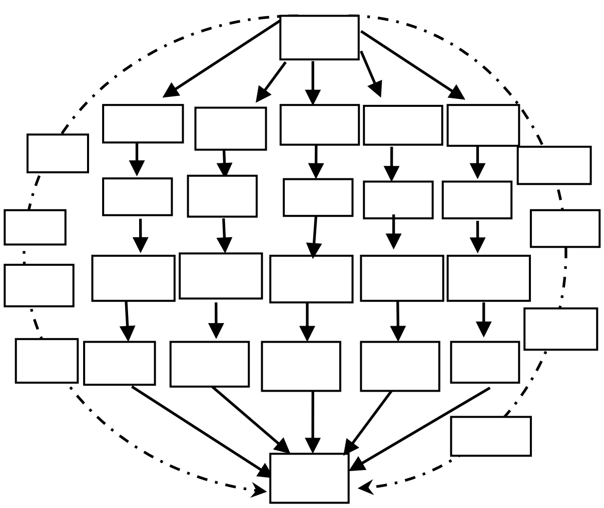


Fig. 4. Navigation bar.

TABLE IV. FEATURES OF NAVIGATION BAR PATTERN

Var.	UI graphical items	Feature types	Pattern definitions
1	<UL> <LI> <a> </LI> </UL>	F1:Navigation list start point F2:Navigation list end point F3:Navigation item start point F4:Navigation item end point F5:Click able link anchor start F6:Click able link anchor end	F1,(F3,F5,F4,)* F2
2	<DIV> <SPAN>> <a><\a> </SPAN> </DIV>	F7:Division or potion start that contain a list F8:Division or potion end that contain a list F9:Navigation item start point F10: Navigation item end point	F7,(F9,F5,F10)*,F8
3	<table> <tr><td> <a><\a></td> </tr> </table>	F19:Start of table F20:End of table F21:Start of row F22:End of row F23:Start of column F24:End of Column	Horizontal F19,F21,(F23,F5,F6, F24)*F22,F20 Vertical F19,(F21,F23,F5,F6, F24,F22)F20 *: Means that feature type can repeat

C. Breadcrumbs Pattern

Breadcrumbs Pattern is mostly used when websites follow a hierarchical structure. The intent and specification of Breadcrumbs pattern is given below.

The web users require the tracking of complete browsing path form home to its current location in order to possibly switch back to a higher level in the hierarchy [3]. Fig. 5 presents a screenshot of Breadcrumbs pattern.

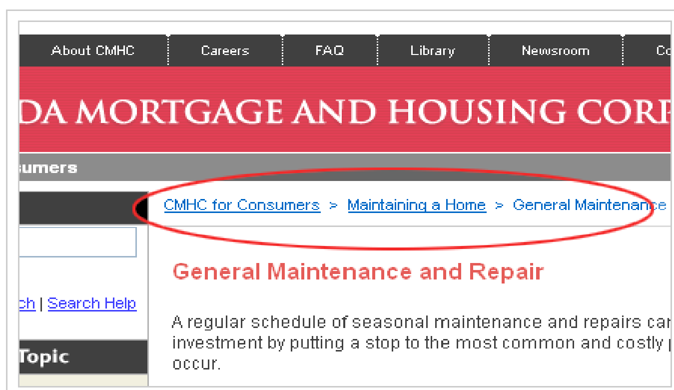


Fig. 5. Breadcrumbs patterns.

- Specification

Features of Breadcrumbs Pattern are given in Table V and diagrammatic specification is presented in Fig. 6. The sequence of features is important for specification and recovery of Breadcrumb. The standard code contains the following tags:

<DIV></DIV>, <UL></UL> ,<LI></Li>, <Span>, <Nav></nav>

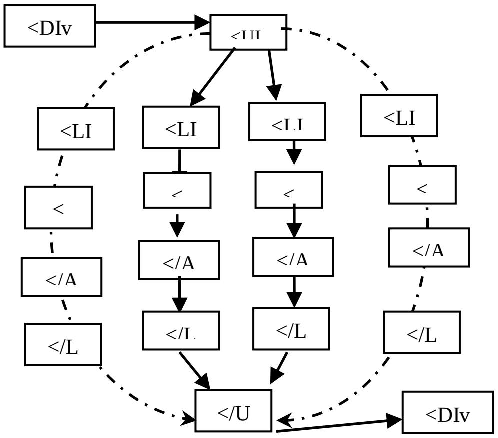


Fig. 6. Breadcrumbs pattern.

D. Lazy Registration Pattern

Lazy Registration Pattern lets the user browse the website without formal registration. The intent and specification of Lazy Registration pattern is given below.

Signup forms have long worked the casual visitor. During the process of discovery, nobody wants to stop and fill out details before they can “unlock” the rest of the site’s potential. As web users become more and more fickle, signup forms are becoming an increasingly large barrier that repels many prospective visitors from great sites. Fortunately, there is a new signup system in web designing that is making it much easier for the visitor to interact with the site and it increases signups. Fig. 7 presents a screenshot of Lazy Registration pattern.

Fig. 7. Lazy registration.

- Semiformal Specification

Features of Lazy Registration Pattern are given in Table VI and diagrammatic specification is presented in Fig. 8.

IV. Detection Approach

A lightweight and customizable approach for recognition of UIWDPs from WAS has been presented. Our approach can handle variations in patterns by customizing specifications and regular expression patterns. A number of regular expression parsing tools are available, but PowerGrep [33] has been selected due to its excellent features and free availability. It can match patterns through large numbers of files/folders in multiple formats. PowerGrep [33] is also capable to map one or more than one patterns at the same time. State of the art approaches [16], [21], [23], [31] discussed in Section II only indicate the presence or absence of patterns by using SQL, checklist survey and descriptive analysis of WAs. The proposed approach takes semiformal specifications of patterns, source code and regular expression patterns as input and recognizes patterns with detailed information. We also plan to automate the process of writing regular expression patterns directly from specifications in the future. The architecture of the approach is given in Fig. 9.

TABLE V. Features of Breadcrumbs Patterns

Var.	UI graphical items	Lexical item	Feature types	Pattern definitions
1	<Div> <UL><Li> <a> *text* </Li> </Div>	<Div> tag contain attribut e breadcr umb > >> /	F7: Division or potion start that contain a list. F1: Navigation list start point F3: Navigation item start point F5: Click able link anchor start F6: Click able link anchor end F4: Navigation item end point F2: Navigation list end point F8: Division or potion end that contain a list	F7,F28,(F1, F3,F5,F6,F 4,F2)*,F8
2	<Div> <Span><a> text </Span></ Div>		F7: Division or potion start that contain a list. F9: Navigation item area start point F5: Click able link anchor start F6: Click able link anchor end F10: Navigation item area end point F8: Division or potion end that contain a list	F7,F28,(F9, F5,F6,F10) *,F8
3	<Div> <a>… <Span > <img> </Span> </Div>		F7: Division or potion start that contain a list. F5: Click able link anchor start F6: Click able link anchor end F9: Navigation item area start point F27: Navigation item area end point F8: Division or potion end that contain a list	F7,F28,(F5, F6,F9,F27) *,F8 *: Means that feature type can repeat

A. Regular Expressions

We develop regular expression patterns for 15 UIWDPs discussed in [6]. There are few false matches due to limitations of regular expressions. We will handle false matches by integrating our present approach with other searching techniques. Table VII below presents the regular expressions for 15 selected UIWDPs. All regular expression patterns are independent of any specific tool and they can be used in any regular expressions based editor. We list variant regular expression patterns for different patterns in column 3.

The above regular expression patterns are based on the compulsory features of UIWDPs. The worth and quality of the recovery process is dependent on identification of correct feature types, their sequence and regular expression patterns which must be free from backtracking and false matches. The presented approach is capable to detect the presence of specified UI pattern in a given WA. It can also recover filename, complete path and line number on which it was implemented. However, we want to clarify that presented approach can recover the UIWDPs according to specified sequence of feature types from the web source codes, but, if any application has implemented a pattern with same feature types according to our semiformal specifications, but in different order, then our approach is unable to recover such types of patterns.

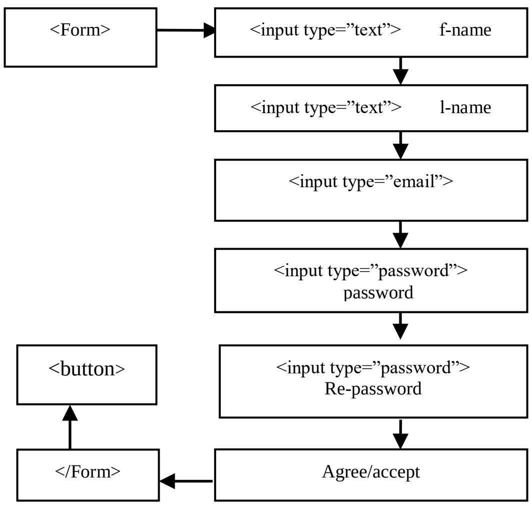


Fig. 8. Lazy registration.

TABLE VI. FEATURES OF LAZY REGISTRATION

Var.a	UI graphical items	Lexical item	Feature types	Pattern definitions
1	<form> <input type=”text” name=”f-name”> <input type=”text” name=”l-name”> <input type=”email/text” name=”email”> <input type=”password” name=”password” > <a>--- <input type=”submit ”>/Button </form>	First name Last name Email Password Agree/accept term and condition	F11:Start of form F12:End of form F13:Text data field F14:Password data field F5: Click able link anchor start F6: Click able link anchor end F15:Submit button F16:Label/lexical item L1 F17:Label/lexical item L2	F11,F12,*F13,*F14,F5,F6 , F16 ,F17 *: Means that feature type can repeat

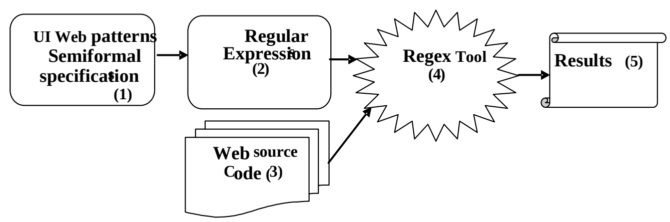


Fig. 9. Architecture of approach

TABLE VII. REGULAR EXPRESSION PATTERNS

<table>
<tr><th>D</th><th></th><th>Vr.</th><th>Regular Expressions</th></tr>
<tr><td></td><td>Login</td><td>1</td><td>((<FORM.*?>(?:(?!ORM>|pe="password"|pe="hidden"|pe="submit").)*?type="text"(?:(?!pe="text").)*?type="password"(?:(?!pe="text"|pe="password").)*?Type=("submit"|"hidden")))|((?><form[^>]*>)((?>(?!ORM>|pe="password"|pe="hidden"|pe="submit").)*?type="text")((?>(?!ORM>|pe="hidden"|pe="submit").)*?type="password")((?>(?!ORM>).)*?type=("submit"|"hidden"))((?>(?<!for).)*?</form>))</td></tr>
<tr><td rowspan="3"></td><td rowspan="3">Navigation</td><td>1</td><td><u>Using <Ul> and <LI> cannot be nested</u>
?><ul[^>]*>)(?>((.(?<!<ul))*?<li[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?))(?>((.(?<!<ul))*?<li[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?))(?>((.(?<!<ul))*?<li[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?))(?>((.(?<!<ul))*?<li[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?))(?>(.(?<!<ul))*?)

<u>Using <Ul> and <LI> can be nested</u>
(?><ul[^>]*>)(?>(?>.*?<li[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?))(?>(?>.*?<li[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?))(?>(?>.*?<li[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?))(?>(?>.*?<li[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?))(?>.*?)</td></tr>
<tr><td>2</td><td>Using <Div> and <Span>

(?><div[^>]*>)(?>(?>.*?<span[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?))(?>(?>.*?<span[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?))(?>(?>.*?<span[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?))(?>(?>.*?<span[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?)) (?>.*?</div>)</td></tr>
<tr><td>3</td><td><u>Using <table> cannot be nested</u>
(?><table[^>]*>)(?>(.(?<!<table))*?<td[^>]*>(.(?<!<table))*?<a[^>]*>(.(?<!<table))*?(.(?<!<table))*?</td>)(?>(.(?<!<table))*?<td[^>]*>(.(?<!<table))*?<a[^>]*>(.(?<!<table))*?(.(?<!<table))*?</td>)(?>(.(?<!<table))*?<td[^>]*>(.(?<!<table))*?<a[^>]*>(.(?<!<table))*?(.(?<!<table))*?</td>)(?>(.(?<!<table))*?<td[^>]*>(.(?<!<table))*?<a[^>]*>(.(?<!<table))*?(.(?<!<table))*?</td>)(?>(.(?<!<table))*?</table>)
<u>Using <table> can be nested</u>
(?><table[^>]*>)(?>(?>.*?<td[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?</td>))(?>(?>.*?<td[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?</td>))(?>(?>.*?<td[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?</td>))(?>(?>.*?<td[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?</td>))(?>.*?</table>)</td></tr>
<tr><td></td><td>d cr u</td><td>1</td><td>(((?><Div[^>]*(?=rumb[a-z]*")[^>]*>)(?>(.(?<!<Div))*?<a[^>]*>(?>.*?))(?>(.(?<!<Div))*?<a[^>]*>(?>.*?))(?>.*?</Div>)))</td></tr>
</table>

<table>
<tr><td rowspan="2"></td><td rowspan="2"></td><td>2</td><td>(((?><Div[^>]*(?=rumb[a-z]*")[^>]*>)(?>(.(?<!<Div))*?<a[^>]*>(?>.*?))(?>(.(?<!<Div))*?<a[^>]*>(?>.*?))(?>.*?</Div>)))</td></tr>
<tr><td>3</td><td>(((?><Div[^>]*(?=rumb[a-z]*")[^>]*>)(?>(.(?<!<Div))*?<a[^>]*>(?>.*?))(?>(.(?<!<Div))*?<a[^>]*>(?>.*?))(?>.*?</Div>)))</td></tr>
<tr><td rowspan="2"></td><td rowspan="2">Lazy Reg.</td><td>1</td><td>((?><form[^>]*>)(?>.(?<!<for))*?type="text"((?>(?!ORM>|pe="password"|pe="submit").)*?<input)(?>(?>.(?<!<for))*?type=)(?>(?>.(?<!<for))*?("email"|"text")(?>.*?)(name="[a-z_]*?email[a-z_]*?")?[^>]*>)(?>(?>.(?<!<for))*?<input)(?>(?>.(?<!<for))*?(?=pe="password"))(?>(?>.(?<!<for))*?(agree|accept)(?>.(?<!<for))*?<a[^>]*>(?>.(?<!<for))*?)?(((?>(?>.(?<!<for))*?a[^>]*>(?>.(?<!<for))*?))|((?>(?>.(?<!<for))*?(<input|<button))(?>(?>.(?<!<for))*?type=)(?>(?>.(?<!<for))*?("submit"|"hiddent")[^>]*>)))(?>(?>.(?<!<for))*?</form>))</td></tr>
<tr><td>2</td><td>((?><form[^>]*>)(?>.(?<!<for))*?type="text"((?>(?!ORM>|pe="password"|pe="submit").)*?<input)(?>(?>.(?<!<for))*?type=)(?>(?>.(?<!<for))*?("email"|"text")[^>]*>)(?>(?>.(?<!<for))*?<input)(?>(?>.(?<!<for))*?(?=pe="password"))(?>(?>.(?<!<for))*?(agree|accept)(?>.(?<!<for))*?<a[^>]*>(?>.(?<!<for))*?)?(((?>(?>.(?<!<for))*?a[^>]*>(?>.(?<!<for))*?))|((?>(?>.(?<!<for))*?(<input|<button))(?>(?>.(?<!<for))*?type=)(?>(?>.(?<!<for))*?("submit"|"hiddent")[^>]*>)))(?>(?>.(?<!<for))*?</form>))</td></tr>
<tr><td></td><td>Pagination</td><td>1</td><td>(((?><Div[^>]*?(?=nation")[^>]*>))|((?><p[^>]*?(?=nation"|rfix")[^>]*>)))(((?>.*?<span[^>]*>)(?>.*?1)(?>.*?</Span>))|((?>.*?<Strong[^>]*>)(?>.*?1)(?>.*?</Strong>))|((?>.*?<a[^>]*>)(?>.*?1)(?>.*?)))((?>.*?<a[^>]*>)(?>.*?\d)(?>.*?)){8}((?>.*?</div>)|(?>.*?</p>))</td></tr>
<tr><td rowspan="2"></td><td rowspan="2">Content footer</td><td>1</td><td>((?><footer[^>]*>)(?>(.(?<!<foot))*?(<div[^>]*>)(.(?<!<foot))*?(<a[^>]*>)(?>(.(?<!<foot))*?</footer>)))</td></tr>
<tr><td>2</td><td>((?><Div[^>]*(?=(footer"|tent"))[^>]*>)((?>.*?<div[^>]*>)(?>.*?<a[^>]*>)(?>.*?</div)){3,}(?>.*</div>))</td></tr>
<tr><td rowspan="2"></td><td rowspan="2">Form wizard</td><td>1</td><td>(?><form[^>]*>)((?>(.(?<!<form))*?((<div[^>]*>)(?>(.(?<!<div))*?<Label[^>]*>)(?>(.(?<!<div))*?type="text")(?>(.(?<!<div))*?</div>)))){2,}((?>(.(?<!<form))*?((<div[^>]*>)(?>(.(?<!<div))*?<Label[^>]*>)(?>(.(?<!<div))*?type="password")(?>(.(?<!<div))*?</div>)))){2,}(?>(.(?<!<form))*?((<input(?>(.(?<!<form))*?type="submit"[^>]*>))|(<button)))(?>.*?</form>)</td></tr>
<tr><td>2</td><td>(?><form[^>]*>)((?>(.(?<!<form))*?<Label[^>]*>)(?>(.(?<!<form))*?</label>)(?>(.(?<!<form))*?<input)(?>(.(?<!<form))*?type="text"[^>]*>)){2,}((?>(.(?<!<form))*?<Label[^>]*>)(?>(.(?<!<form))*?</label>)(?>(.(?<!<form))*?<input)(?>(.(?<!<form))*?type="password"[^>]*>)){1}(?>(.(?<!<form))*?((<input(?>(.(?<!<form))*?type="submit"[^>]*>))|(<button)))(?>.*?</form>)</td></tr>
<tr><td rowspan="2"></td><td rowspan="2">Modular tab</td><td>1</td><td>((<div[^>]*>)|(<article[^>]*>))(((?>.*?<header[^>]*>)|(?>.*?<div[^>]*>))(?>.*?<a[^>]*>)(?>.*?<p[^>]*>)(?>.*?<a[^>]*>)((?>.*?</header>)|(?>.*?</div>))){4,}((?>.*?</div>)|(?>.*?</article>))</td></tr>
<tr><td>2</td><td>(<div[^>]*>)((?>.*?<ul[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?)){5,} (?>.*?)(?>.*?</div>)</td></tr>
<tr><td></td><td>Page Grid</td><td>1</td><td>(<!DOCTYPE[^>]*>)(?>.*?<html)(?>.*?<head)(?>.*?<title>)((?>.*?<link)(?>.*?rel="stylesheet"[^>]*>)){3,}(?>.*?<body)(((?>.*?<div)(?>.*?id=[^>]*>)(?>.*?<div)(?>.*?id=[^>]*>)(?>.*?<div)(?>.*?id=[^>]*>))){35,}</td></tr>
<tr><td rowspan="2">0</td><td rowspan="2">Article List</td><td>1</td><td>(<section[^>]*>)(?>.*?<header[^>]*>)(?>.*?</header>)((?>.*?<article[^>]*>)(?>.*?<h\d[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?</h\d>)(?>.*?<p[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?</p>)(?>.*?<p[^>]*>)(?>.*?</p>)(?>.*?</article>)){2,}?(?>.*?</section>)</td></tr>
<tr><td>2</td><td>(<div[^>]*(wrapper|article|main)[^>]*>)((?>.*?<div[^>]*>)(?>.*?<h\d[^>]*>)(?>.*?</h\d>)(((?>.*?<Div[^>]*>))|((?>.*?<p[^>]*>)))(?>.*?<a[^>]*>)(?>.*?)((?>.*?</div>)|(?>.*?</p>))((?>.*?<p[^>]*>))(?>.*?<a[^>]*>)(?>.*?)(?>.*?</p>)(?>.*?</div>)){3,}?(?>.*?</div>)</td></tr>
<tr><td>1</td><td>Vote to promote</td><td>1</td><td>((<section[^>]*>)|(<div[^>]*>)|(<table[^>]*>))(((?>.*?<div[^>]*>)|(?>.*?<tr[^>]*>))(?>.*?<a[^>]*>)(?>.*?)((?>.*?<div[^>]*>)|(?>.*?<td[^>]*>))(?>.*?<a[^>]*>)(?>.*?)((?>.*?<div[^>]*>)|(?>.*?<td[^>]*>))((?>.*?<span[^>]*>)|(?>.*?<p[^>]*>))((?>.*?)|(?>.*?</p>))){10,}?((?>.*?</section>)|(?>.*?</div>)|(<?>.*?/table>))</td></tr>
<tr><td rowspan="2">2</td><td rowspan="2">Archives</td><td>1</td><td>((?><section[^>]*>)|(?><div[^>]*>)|(?><table[^>]*>))((?>.*?<h\d[^>]*>)(?>.*?</h\d>)((?>.*?<ul[^>]*>)|(?>.*?<ol[^>]*>)|(?>.*?<article[^>]*>))((?>.*?<li[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?)){2,}?((?>.*?)|(?>.*?)|(?>.*?</article>))){4,}?((?></section[^>]*>)|(?>.*?</div>)|(<?.*?</table>))</td></tr>
<tr><td>2</td><td>((?><section[^>]*>)|(?><div[^>]*>)|(?><table[^>]*>))(((?>.*?<tr[^>]*>)|(?>.*?<div[^>]*>)|(?>.*?<article[^>]*>))(?>.*?<h\d[^>]*>)(?>.*?</h\d>)((?>.*?<ul[^>]*>)|(?>.*?<ol[^>]*>)|(?>.*?<article[^>]*>))((?>.*?<li[^>]*>)(?>.*?<a[^>]*>)(?>.*?)(?>.*?)){2,}?((?>.*?)|(?>.*?)|(?>.*?</article>))){4,}?((?></section[^>]*>)|(?>.*?</div>)|(<?.*?</table>))</td></tr>
<tr><td>3</td><td>Dashboard</td><td>1</td><td>(?>(<Script[^>]*"text/javascript"[^>]*>)(?>.*?Dashboard){1,}?(?>.*?</script>))|(?><\w+[^>]*Dashboard[^>]*>){1,}?</td></tr>
<tr><td rowspan="2">4</td><td rowspan="2">Hierarchical Actions</td><td>1</td><td>(?><form[^>]*>)(?>(?>.(?<!<for))*?<input)(?>(?>.(?<!<for))*?type="submit"[^>]*>)(?>(?>.(?<!<for))*?<a[^>]*>)(?>(?>.(?<!<for))*?)(?>(?>.(?<!<for))*?</form>)</td></tr>
<tr><td>2</td><td>(?><form[^>]*>)(?>(?>.(?<!<for))*?<a[^>]*>)(?>(?>.(?<!<for))*?)(?>(?>.(?<!<for))*?<input)(?>(?>.(?<!<for))*?type="submit"[^>]*>)(?>(?>.(?<!<for))*?</form>)</td></tr>
<tr><td>5</td><td>Display Collection</td><td>1</td><td>((?><table[^>]*>)|(?><div[^>]*>))(((?>.*?<tbody[^>]*>)|(?>.*?<div[^>]*>)|(?>.*?<UL[^>]*>))(((?>.*?<td[^>]*>)|(?>.*?<div[^>]*>)|(?>.*?<li[^>]*>))(?>.*?<a[^>]*>)(?>.*?)((?>.*?</div>)|(?>.*?))){20,}?((?>.*?</tbody>)|(?>.*?</div>)|(?>.*?</UL>)))((?>.*?</table>)|(?>.*?</div>))</td></tr>
</table>

V. Evaluation

We evaluate our approach on 75 randomly selected websites. The source code for all these websites is available freely on web. We cannot find a standard benchmark system from the literature that can be used for comparison of our results. Due to this reason, we performed manual analysis on the source code which was very time consuming and daunting task. The results of manual analysis are compared with automated results on five selected websites for each pattern. The reason for performing manual and automated on five selected websites is due to the reason that we want to validate patterns definitions manual analysis of source code. Secondly, we performed an experiment on all websites for 15 UIWDPs and extracted results are shown in Table X. We want to clarify that we calculated accuracy of our approach on five selected websites for each pattern. The detail about automated and manual experiments is given below.

A. Automated Experiment

In order to validate our semiformal specifications and regular expression patterns, we performed experiments on 75 different websites using PowerGrep [33] tool and recovered instances of 15 selected patterns. PowerGrep tool has excellent capabilities for parsing source code by using regular expressions. The recovered results are presented in Table VIII.

B. Manual Experiment

Manual analysis of source code is very time consuming and daunting task. It was important for us to validate our semiformal specifications of UIWDPs in the source code manually. In order to recover a pattern, we start searching from folders, subfolders and files, line by line, that was very time consuming and very difficult to mark starting and end point of the patterns. The multiple occurrences of patterns and finding the precise location of patterns in the code are very hard.

There exists slight difference between the manual and automated results due to the several factors. One of the major reasons is that different web designers implement UIWDPs according to their skill, nature of applications and frameworks. The manual detection involves personal experiences and understanding of multi-language source code, but our approach can recover patterns automatically by following the semiformal specifications and regular expression patterns.

The difference between automated and manual experiments with their Standard Deviation (S.T.D.) is shown in the following Fig. 10.

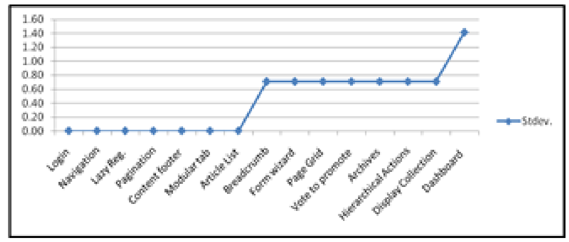


Fig. 10. Comparison between the manual & automatic recovery.

C. Accuracy of Approach

The accuracy in the universal statistical sense indicates the proximity of calculations or estimates to the accurate or exact results. To measure the accuracy of our approach, we compute the Precision, Recall and F-Score for our recovered results.

The data retrieval techniques can be evaluated by using the Precision and Recall metrics. This matrix method is popular for the evaluation of patterns extraction approaches.

TABLE VIII. Evaluation Results

Pattern	Website	Files		Results		Total		SD	TP	FP	FN	Pr	Rc	Fs
		Tot	Imp	M	A	TM	TA							
Login	Olx.com	80	1	1	1	3	3	0.00	3	0	0	1.00	1.00	1.00
	samshine.com	87	0	0	0									
	geomedia.com	103	0	0	0									
	www.fast.edu.pk	10	1	1	1									
	Netvib.com	85	1	1	1									
Navigation	Olx.com	80	1	2	2	9	9	0.00	9	0	0	1.00	1.00	1.00
	aiou.edu	107	2	2	2									
	geomedia.com	103	1	1	1									
	Ngaming.com	80	1	1	1									
	Bath_baby.com	80	3	3	3									
Breadcrumb	hec.gov.pk	17	0	0	1	3	4	0.71	3	0	1	0.75	1.00	0.95
	Ngaming.com	80	1	1	1									
	Bath_baby.com	80	1	1	1									
	Netvib.com	85	1	0	0									
	Vimeo.com	91	1	1	1									
Lazy	hec.gov.pk	17	0	0	0	2	2	0.00	2	0	0	1.00	1.00	1.00

Registration	sis.cuonlineatd.edu	95	0	0	0									
	aiou.edu	107	0	0	0									
	Netvib.com	85	1	1	1									
	Vimeo.com	97	1	1	1									
Pagination	drweb.de/magazin	73	1	1	1	8	8	0.00	8	0	0	1.00	1.00	1.00
	d3.ru/popular	45	2	2	2									
	erweiterungen.de	35	2	2	2									
	subcide.com/articles	95	1	1	1									
	spacecollective.org	88	1	1	1									
Content Footer	www.cssbeauty.com	17	1	1	1	5	5	0.00	5	0	0	1.00	1.00	1.00
	billyhughes.oph.gov	17	1	1	1									
	www.cityofgrace.com	35	1	1	1									
	www. Lastfm.com	84	1	1	1									
	viget.com	28	1	1	1									
Form wizard	club.nokia.com/	116	1	1	1	6	5	0.71	5	1	0	1.00	0.83	0.85
	beanstalkapp.com/	86	1	1	1									
	ui-patterns.com/	87	2	2	2									
	mite.com	20	1	1	1									
	statementstacker.com	36	1	1	0									
Modular Tab	bestwebgallery.com	43	1	1	1	6	6	0.00	6	0	0	1.00	1.00	1.00
	haveamint.com	10	1	1	1									
	mailchimp.com	11	1	1	1									
	www.cbs.com	107	2	2	2									
	viget.com/advance	16	1	1	1									
Page Grid	nps.gov/index	46	1	1	1	5	4	0.71	4	1	0	1.00	0.80	0.83
	nytimes.com/	158	1	1	1									
	yeeaahh.subtraction.com	30	1	1	0									
	theguardian.com	131	1	1	1									
	yui.github.io	37	1	1	1									
Article list	alistapart.com	17	1	1	1	5	5	0.00	5	0	0	1.00	1.00	1.00
	businesscatalyst.com	55	1	1	1									
	visitmix.com	23	1	1	1									
	www.popsci.com	106	1	1	1									
	valetmag.com	82	1	1	1									
Vote to promote	digg.com/	99	1	2	3	9	10	0.71	9	0	1	0.90	1.00	0.98
	news.layervault.com	9	1	1	1									
	news.ycombinator.com	3	1	1	1									
	stackoverflow.com/	65	1	3	3									
	www.squidoo.com/	62	1	2	2									
Archives	astheria.com	9	1	1	1	6	6	0.71	6	0	0	1.00	1.00	1.00
	ismaelburciaga.com	36	1	1	1									
	colly.com	6	1	1	1									
	www.cssmania.com	17	1	1	1									

	www.darrenhoyt.com	31	1	1	1									
Dashboard	www.tripit.com	48	1	1	1	8	7	1.41	6	2	0	1.00	0.75	0.78
	www.mint.com	23	1	1	0									
	www.weatherspark.com	10	1	1	1									
	www.udemy.com	77	2	4	3									
	www.last.fm	120	1	1	1									
Hierarchical Actions	www.fedex.com	40	1	1	1	5	4	0.71	5	1	0	1.00	0.83	0.85
	www.twitter.com	30	1	2	2									
	www.paypal.com	29	1	1	1									
	www.barnesandnoble.com	22	1	1	0									
	www.reg.ebay.com	11	1	1	1									
Display Collection	www.smileycat.com	63	2	3	3	9	8	0.71	8	1	0	1.00	0.89	0.90
	http://patterntap.com	34	1	3	3									
	www.webdesignpractices.com/	26	1	1	0									
	www.quince.infragistics.com	123	1	1	1									
	http://www.patternry.com	55	2	1	1									
Total/ Average		-	79	90	86	90	86	-	84	6	2	0.98	0.93	0.94

SD: Standard Deviation, TP: True Positives, FP: False Positives, FN: False Negatives, Pr: Precision, Rc: Recall, Fs: F-Score, Tot: Total Files, Imp: Files in which pattern is implemented, M: Manual Result, A: Automated Result, TM: Total Manual Instances, TA: Total Automated Instances

The Precision and Recall have been used to evaluate the quality of different systems from the last few decades. This method can evaluate how many patterns retrieved are relevant and how many relevant patterns are retrieved [51]. The accuracy of any approach can be measured by finding the relationship between the Precision and Recall metrics. In the ideal situations, the Precision results of an approach should remain high when the Recall increases [52]. The following parameters are used to calculate the Precision and Recall for pattern recovery techniques as given in Table IX.

Precision: The fraction of retrieved documents those are relevant: P=TP/(TP+FP)

Recall: The fraction of relevant documents that are retrieved: R=TP/(TP+FN)

TABLE IX. PRECISION AND RECALL

TP= True Positives **TN=True Negatives** **FN=False Positives,** **FN=False Negatives**	**Relevant**	**Non Relevant**
Retrieved	TP	FP
Not Retrieved	FN	TN

F-Score: The accuracy of pattern recovery approaches can be effectively measured by using the Precision and Recall, although the combination of both factors yields a combined effect. This common factor for evaluating the Precision and Recall metrics for any recovery technique is addressed by Peterson et al. [30]. They proposed a standard solution by using the Precision and Recall of any approach. They defined F-score (Fw) as:

$$Fw = (1+W2)(PR)/(W2P+R)$$

The value of W is constant (W =2.28). If the Precision and Recall of any approach are high, then F-Score obtained is also high. The Precision and Recall for manual and automated experimental results are given in Table VIII and presented in Fig. 11. The average Precision, Recall and F-Score is 98%, 93% and 94% respectively for all 15 UI WDPs detected from the 75 web applications.

D. Validity Threats

Validity is an important concern for empirical acceptance of results extracted by different approaches. A major threat to results of our approach is a lack of standard definitions for UIWDPs and their variants. These definitions are not available in literate to the best of our knowledge. Internal validity refers to the consistency of measurements across all methods and tools. It is affected by experimental biases. We tried to mitigate this threat by manually performing experiments on source code of 75 selected web applications. All the experimental results of our approach are presented in paper and researchers can validate our results. External validation is affected by generalization of results. We specified all 15 UIWDPs using semiformal specifications and these specifications are available on web for community. The source code of 75 web applications is also freely available. One possible threat to external validity of our results may be application of different constructs for the implementation of UIWDPs. Reliability validity affects the replicability of our results. All selected web applications and their source code are available on web for validation. Reliability validity threat is eliminated because our regular expression patterns can be validated with different regular expression parsing tools which are available on the web.

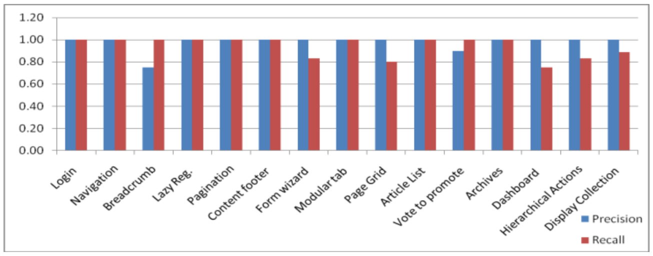


Fig. 11. Precision and recall for all patterns.

VI. CONCLUSION

UIWDPs recovery from WAs is a challenging task due to ever increasing applications of new technologies for the development of web applications. The recovery of information from web applications provides valuable information to maintenance, comprehension, refactoring, reuse and re-engineering disciplines. A number of techniques and tools are presented for recovering information from web applications, but they are not capable enough to deal with the heterogeneous nature of web applications completely. The recovery of information from web applications is difficult due to number of technologies and external dependencies in web applications. State of the art approaches focused on extraction of UML models from WAs. In this paper, we present an approach that can recover UIWDPs from LWAs with their necessary attributes such as filename, line number, numbers of matches per file, etc. The deviation in the automated results extracted by our approach and manual results shows that there is no consensus on the definitions of UIWDPs from community. The implementation variations are another cause in the disperse results extracted by our approach. Our approach can handle multi-language source code partially for recovery of patterns directly from the source code. Moreover, the recovered UI pattern's information can be effectively used in the maintenance, abstraction, comprehension, upgradation, migration of applications from one framework to another and re-engineering of LWAs. In future, we plan to extend the scope of our automatic recognition of UIWDPs approach on other UI pattern libraries such as Yahoo pattern library, Weli pattern library, etc.

TABLE X. EVALUATION OF ALL WEBSITES FOR ALL PATTERNS

Pattern #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Website name															
www.olx.com	1	2	x	x	x	1	x	2	x	x	1	x	x	x	2
www.samshine.com	x	1	x	x	x	1	x	1	x	1	x	1	x	x	1
www.geomedia.com.au	x	1	x	x	x	1	x	2	1	1	x	2	x	x	x
www.fast.edu.pk	1	x	x	x	x	x	x	x	x	x	x	x	x	x	x
www.netvibes.com	1	x	x	1	1	1	1	1	1	x	x	x	x	x	x
www.aiou.edu	x	2	x	x	x	x	x	1	x	x	1	x	x	x	1
www.ngaming.com	x	1	1	x	x	1	x	2	1	x	x	1	x	x	1
www.buybuybaby.com	x	3	1	x	x	1	x	x	x	x	x	x	1	x	x
www.hec.gov.pk	x	1	1	x	x	1	x	2	1	x	1	x	x	x	2
www.buybuybaby.com	x	x	1	1	x	1	1	x	1	1	x	1	x	1	1
www.netvibes.com	1	2	x	x	x	1	x	2	x	x	1	x	x	x	2
www.vimeo.com	x	1	x	x	x	1	x	1	x	1	x	1	x	x	1
www.sis.cuonlineatd.edu.pk	x	x	x	x	x	x	x	1	1	x	1	x	x	1	x
www.drweb.de/magazin	x	3	x	x	1	1	x	1	1	1	1	x	2	x	3
http://d3.ru/popular	x	3	x	1	2	2	x	1	x	x	x	1	x	1	1
www.erweiterungen.de	x	x	x	1	2	x	x	x	x	1	x	x	x	x	1
www.subcide.com/articles	x	x	x	x	1	x	x	1	x	x	1	1	x	x	x
http://spacecollective.org	x	x	x	x	1	1	x	x	x	x	2	x	x	x	x
www.cssbeauty.com	x	x	x	x	2	1	x	1	1	x	x	1	x	x	x
www.billyhughes.oph.gov	x	x	x	x	x	1	x	1	1	1	x	x	x	x	1
www.cityofgrace.com	x	x	x	x	x	1	x	x	1	x	x	x	x	x	x
www.last.fm	x	x	x	x	x	1	x	1	1	x	1	2	x	x	2
www.viget.com	x	x	x	x	x	1	x	1	x	1	x	1	x	x	1
www.club.nokia.com	x	x	x	1	x	1	1	1	x	1	1	1	x	x	5
www.beanstalkapp.com	1	x	x	1	x	1	1	1	2	1	1	1	x	2	1

http://ui-patterns.com	2	x	x	2	x	2	2	2	2	x	x	x	x	2	2
www.mite.com	1	x	x	1	x	x	1	x	1	x	x	x	x	1	x
www.statementstacker.com	x	1	1	x	1	1	x	x	1	1	x	1	x	1	x
www.bestwebgallery.com	x	x	x	x	x	2	x	1	1	x	2	4	x	x	3
www.haveamint.com	x	x	x	x	x	1	x	1	x	x	x	x	x	x	x
www.mailchimp.com	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
www.cbs.com	x	x	x	x	x	1	x	1	x	x	x	1	x	x	1
www.viget.com/advance	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
www.nps.gov/index	x	x	x	x	x	1	x	1	1	x	1	x	x	x	3
www.nytimes.com/	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
ww.yeeaahh.subtraction.com	x	x	x	x	x	1	x	1	x	1	x	1	x	x	1
www.theguardian.com	x	x	x	x	x	1	x	1	1	x	x	2	2	x	2
www.yui.github.io	x	x	x	1	x	1	x	2	1	1	3	1	x	x	4
www. alistapart.com	x	x	x	x	x	1	x	x	1	x	x	1	x	x	1
businesscatalyst.com	x	x	x	x	x	1	x	1	1	1	x	4	1	x	3
http://visitmix.com	x	x	x	x	2	3	x	2	5	2	1	1	3	x	1
http://www.popsci.com	x	x	x	x	x	x	x	1	x	1	x	x	x	x	x
www.valetmag.com	x	x	x	x	x	1	x	1	x	1	x	1	x	x	2
http://digg.com	x	x	x	x	x	1	x	2	x	2	x	x	x	x	2
https://news.layervault.com	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
www.news.ycombinator.com	x	x	x	x	x	2	x	1	1	1	1	1	x	x	1
http://stackoverflow.com	x	x	x	x	x	2	x	2	x	1	1	x	x	x	3
http://www.squidoo.com	x	x	x	x	x	1	x	1	x	1	3	1	x	x	1
http://astheria.com	x	x	x	x	x	x	x	x	x	x	x	1	x	x	1
http://ismaelburciaga.com	x	x	x	x	x	2	x	1	1	1	1	1	x	x	2
http://colly.com	x	x	x	x	x	x	x	x	x	x	x	1	x	x	1
www.cssmania.com	x	x	x	x	x	x	x	1	x	x	x	1	x	x	1
www.darrenhoyt.com	x	x	x	x	1	1	x	1	x	x	x	1	x	x	x
www.tripit.com	x	x	x	x	x	1	x	x	1	x	x	x	1	1	1
www.mint.com	1	x	x	x	x	1	x	1	1	1	x	x	0	x	x
www.weatherspark.com	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
www.udemy.com	x	x	x	x	x	1	x	1	1	1	x	x	5	x	1
www.last.fm	x	x	x	x	x	x	x	1	x	1	2	5	1	x	x
www.fedex.com	x	x	x	x	x	x	x	x	x	x	x	x	1	x	x
www.twitter.com	1	x	x	x	x	1	x	1	x	x	x	4	x	1	1
www.paypal.com	3	x	x	2	x	3	x	2	1	x	x	1	1	2	1
www.barnesandnoble.com	x	x	x	x	x	1	x	x	1	x	x	x	x	1	x
www.reg.ebay.com	x	x	x	x	x	1	x	1	1	x	1	1	x	x	3
www.smileycat.com	1	x	x	x	x	1	1	x	x	x	x	x	x	1	x
http://patterntap.com	x	x	x	x	x	3	x	3	2	2	x	2	x	x	1
www.webdesignpractices.com	x	x	x	x	x	1	x	1	x	x	1	1	x	x	1
www.quince.infragistics.com	1	x	1	x	1	1	x	1	x	1	x	x	1	1	1
www.patternry.com	1	x	x	x	x	x	x	x	x	x	x	x	1	1	1
Total	**15**	**21**	**6**	**10**	**15**	**61**	**8**	**60**	**38**	**30**	**29**	**52**	**19**	**17**	**72**

1: Login, 2: Navigation, 3: Breadcrumb, 4: Lazy Registration, 5: Pagination, 6: Content Footer, 7: Form wizard, 8: Modular Tab, 9: Page Grid, 10: Article list, 11: Vote to promote, 12: Archives, 13: Dashboard, 14: Hierarchical Actions, 15: Display Collection

References

[1] A. Rezazadeh, "Formal Patterns for Web-based Systems Design", PhD Thesis, http://eprints.soton .ac.uk/267101/1.hasCoversheetVersion/Formal Patterns for Web-based_Systems_Design.pdf, 2007.

[2] J. Martin and L. Martin, "Website maintenance with software engineering tools", In Proceedings of 3rd International Workshop on Web Site Evolution, pp. 126-131, 2001.

[3] J. Wedekind, "Web Design Patterns A future approach?", Bachelor Thesis, www.patterns.jana-wedekind.de/downloads/BachelorThesis_Wedekind.pdf, 2008.

[4] A. Toxboe, "UI Patterns," http://ui-patterns.com, 2008.

[5] W. Van, V. Martijn, C. Gerrit, "Pattern Languages in Interaction Design: Structure and Organization", http:// www.welie.com/papers/Welie-Interact2003.pdf, 2003.

[6] K. Night, "15 UI Design Patterns Web Designers Should Keep Handy", 2010.

[7] C. Alexander, "The Timeless Way of Building," Oxford University Press, New York, 1979.

[8] Y. Chung, S. Lee, "Reverse software engineering with UML for website maintenance", In Proceedings of 1st International Conference on Web Information Systems Engineering , IEEE Computer Society Press: Los Alamitos, CA, Vol.2. pp. 157-161, 2001.

[9] J. Jovanovic, "10 UI Design Patterns You Should Be Paying Attention To", http://uxdesign. smashingmagazine.com/2009/06/23/10-ui-design-patterns-you-should-be-paying-attention-to, 2009.

[10] F. Ricca and P. Tonella, "Analysis and Testing of Web Applications", In Proceedings of the 23rd international conference on Software engineering, pp.25-34, 2001.

[11] C. Bellettini, A. Marchetto, "WebUml: Reverse Engineering of Web Applications", ACM Symposium on Applied Computing, Vol. SAC(04), pp.1662-1669,2004.

[12] C. Wei, E. Cuddihy, J. Barrick and J. H. Spyridakis, "Conducting Usability Research through the Internet", In Proceedings of the Usability Professional Association, 2005.

[13] E. Cuddihy, J. H. Spyridakis, "From Web Designer to Researcher: Using Experiment Patterns to Research Web Design Decisions" In Proceedings of Professional Communication Conference, Vol. 7(12), pp. 1-10, 2012.

[14] A. Marchetto, P. Tonella, F. Ricca, "ReAjax: a reverse engineering tool for Ajax Web Applications", IET Software, Vol. 6, pp.33–49, 2012.

[15] D. Bouchiha, M. Malki, "Towards Re-engineering Web Applications into Semantic Web Services", In Proceedings of International Conference on Web and Machine Intelligence, pp.350-353, 2010.

[16] W. W. Norizan, W. Hashim, N. Laila, M. Noor, W. Adilah, W. Adnan, "Web Aesthetic Interaction Design Pattern in Popular Clothing Brands Websites: An Initial Investigation", In Proceedings of International Conference on Systems, Man and Cybernetics, Vol. 978-1, pp. 2323-2327, 2010.

[17] R. R. Echeverrıa, M. J. Conejero, J. Pedro, Clemente, D. M. Villalobos and F. S. Figueroa, "Generation of WebML Hypertext Models from Legacy Web Applications", In Proceedings of 14th IEEE International Symposium on Web Systems Evolution, pp.91-95, 2012.

[18] C. Boldyreff, R. Kewish, "Reverse Engineering to Achieve Maintainable WWW Sites", In Proceedings of 8th Working Conference on Reverse Engineering IEEE Computer Society Press: Los Alamitos, CA, pp. 249-257, 2001.

[19] De Lucia, A., Francese, R., Scanniello, G., and Tortora, G., " Reengineering Web Applications Based on Cloned Pattern Analysis", In Proceedings of the 12th IEEE International Workshop on Program Comprehension (IWPC'04), 2004.

[20] R. R. Echeverrıa, M. J. Conejero, J. Pedro, D. M. Villalobos and F. S. Figueroa, " Model Driven Extraction of the Navigational Concern of Legacy Web Applications," ICEWE Workshop LNCS 7703,pp.56-70, 2012.

[21] R. Patel, F. Coenen, R. Martin, and L. Archer, "Reverse Engineering of Web Applications: A Technical Review," https://intranet.csc.liv.ac.uk/research/techreports/tr2007/ulcs-07-017.pdf, pp.1-12, 2007.

[22] G. Rasool, "Customizable Feature Based Design Pattern Recognition Integrating Multiple Techniques," Ph.D. Thesis, 2010.

[23] J. Conallen. "Modeling Web application Architectures with UML", In Communications of the ACM, Vol. 42, pp. 63–70,1999.

[24] J. Maras, A. Petric, and M. Stula. "Reverse engineering legacy Web applications with php Modeler", Conference Paper at Mälardalen University Software Engineering Workshop (MUSE09), 2009.

[25] S. Tilley and S. Huang, "Evaluating the reverse engineering capabilities of web tools for understanding site content and structure: A case study", In Proceedings of the 23rd International Conference on Software Engineering, pp.514–523, 2001.

[26] P. Tramontana, "Reverse Engineering of Web Applications", Ph.D. Thesis, Chapter 10, 2005.

[27] F. Ricca and P. Tonella."Website understanding and re-structuring with the reweb tool", IEEE MultiMedia, Vol, 8, No. 2, pp. 40-51, April-June, 2001.

[28] A. Carlos, L. Mmarc and Maldonda, "Do Common user Interface Design Patterns improve Navigation" In Proceedings of the Human and Ergonomics Society 46th Annual Meeting, Vol, 46, No 14, pp. 1315-1319, 2002.

[29] http://research.ciitlahore.edu.pk/Groups/SERC/DesignPatterns.aspx

[30] N. Pettersson, W. Löwe and J. Nivre, "Evaluation of Accuracy in Design Pattern Occurrence Detection", IEEE Transactions on Software Engineering, IEEE computer Society Digital Library, IEEE Computer Society, pp. 575-590, Volume 36, No. 4, July/August 2010

[31] Karel, k., Jana, K., Ondrej, L., Milos, K., Vaclav, S., Hana, R., "Web Page Analysis: Experiments Based on Web Patterns", In proceedings of 4th International Conference on Innovations in Information Technology, pp. 16-20, 2007.

[32] G. Rasool and N. Asif, "Software artifacts recovery using abstract regular expressions", In Proceedings of Multitopic Conference INMIC, pp. 1-6, 2007.

[33] PowerGrep, Home page:https:// www.powergrep.com

[34] D. Amalfitano, A. R. Fasolino, and P. Tramontana, "Experimenting a reverse engineering technique for talian the talian of rich internet applications, "In Proceedings of IEEE International Conference on Software Maintenance, pp. 571–574, 2009.

[35] W. Sun, S. Li, D. Zhang, and Y. Yan, "A model-driven reverse engineering approach for semantic web services composition," In Software Engineering, 2009. WCSE'09. WRI World Congress on, 2009, pp. 101-105.

[36] Djelloul, B., Mimoun, M., & El Kader, M. A., "Towards reengineering web applications to web services", The International Arab Journal of Information Technology, 6(4), 359-364, 2009.

[37] Jiang, Y., & Stroulia, E., "Towards reengineering Web sites to Web-services providers", In Proceeding of Eighth European Conference on Software Maintenance and Reengineering, pp. 296-305, 2004.

[38] D. Schwabe, L. Esmeraldo, G. Rossi, F. Lyardet, "Engineering web applications for reuse ",IEEE Multimedia, 8(1), pp. 20–31, 2001.

[39] J. Vanderdonckt, L. Bouillon, N. Souchon, "Flexible reverse engineering of web pages with VAQUISTA", In Proceedings of Eighth Working Conference on ReverseEngineering, pp. 241–248, 2001.

[40] P. Tonella, F. Ricca, E. Pianta, C. Girardi, "Restructuring Multilingual Web Sites", In Proceedings of International Conference on Software Maintenance, pp. 290-299, 2002.

[41] L. Paganelli, F. Paternò, "Automatic Reconstruction of the Underlying Interaction Design of Web Applications", In Proceedings of the 14th International Conference onSoftware Engineering and Knowledge Engineering, pp.439-445, 2002.

[42] J. Pu, H. Yang, B. Xu, L. Xu and W. Cheng-Chung Chu, "Combining MDE and UML To Reverse Engineer Web Based Legacy Systems", In proceedings of IEEE International Computer Software and Applications Conference, pp. 718-725, 2008.

[43] S. Kim, S. Shivaji, E. James Whitehead Jr., Kenyon "Web: Reconfigurable Web-based Feature Extractor", In Proceedings of IEEE ICPC, pp. 287-288, 2009.

[44] R. Rodr´ıguez-Echeverr´ıa, J. Conejero, P. J. Clemente, M. D. Villalobos and F. S´anchez-Figueroa, “Generation of WebML Hypertext Models from Legacy Web Applications”, In Proceedings of 14th Symposium on Web Systems Evolution, pp.91-95, 2012.

[45] M. L. Bernardi, G. A. Di Lucca, D. Distante and M. Cimitile, “Model Driven Evolution of Web Applications”, In Proceedings of IEEE Conference, pp. 45-50, 2013.

[46] P. Tramontana, D. Amalfitano and A. R. Fasolino, “Reverse Engineering Techniques: from Web Applications to Rich Internet Applications”, In Proceedings of 15th Symposium on Web Systems Evolution, pp.83-86, 2013.

[47] D. Amalfitano, A. R. Fasolino, and P. Tramontana, “An Iterative Approach for the Reverse Engineering of Rich Internet Application User Interfaces”, In Proceedings of Internet and Web Applications and Services Conference, pp. 401–410, 2010.

[48] D. Lucca, G.A. Di Penta, G. Antoniol, G. and G Casazza, “An approach for Reverse Engineering of Web Based Applications”, In Proceedings of 8th Working Conference on Reverse Engineering, pp. 208-207, 2001.

[49] M. Moore, “Representation Issues for Reengineering Interactive Systems”, ACM Computing Surveys, 28(4), pp. 199, 1996.

[50] G. Rasool and P. Mäder, “A Customizable Approach to Design Patterns Recognition Based on Feature Types”, Arabian Journal for Science and Engineering, pp. 1-23, 2014.

[51] M. Voorhees, "The philosophy of information retrieval evaluation, “In Evaluation of cross-language information retrieval systems, pp. 355-370, 2002.

[52] C. K. Roy, "Detection and analysis of near-miss software clones," In Proceedings of International Conference on Software Maintenance ICSM 2009. pp. 447-450, 2009.

[53] [53] F. Ricca, M. Di Penta, M. Torchiano, P. Tonella and M. Ceccato, “An empirical study on the usefulness of Conallen's stereotypes in Web application comprehension”, In Proceedings of Eighth IEEE International Symposium on Web Site Evolution, pp. 58-68, 2006.

[54] A. Marchetto A, P. Tonella and F. Ricca ,” ReAjax: a reverse engineering tool for Ajax Web applications”, IET Software Vol 6, No 1, pp. 33-49, 2012.

[55] S. Staiger, ‘Reverse engineering of graphical user interfaces using static analyses’. Working Conf. on Reverse Engineering (WCRE), Washington, DC, USA, 2007.

[56] Draheim, D., Lutteroth, C., and Weber, G., “A Source Code Independent Reverse Engineering Tool for Dynamic Web Sites”, In Proceedings of 9th European Conference on Software Maintenance and Reengineering (CSMR'05), pp. 168-177, 2005.

[57] Bouchiha, D., Malki, M., and Benslimane, S-M., “Ontology Based Web Application Reverse-Engineering Approach”, Journal of Computer Science (INFOCOMP), VOLUME 6-N. 1-. pp: 37-46, 2007.

Internet of Plants Application for Smart Agriculture

Khurshid Aliev
Department of Management and Production Engineering
Politecnico Di Torino
Torino, Italy.

Eros Pasero
Department of Electronics and Telecommunication
Politecnico Di Torino
Torino, Italy.

Mohammad Moazzam Jawaid, Sanam Narejo
Department of Computer Systems Engineering
Mehran University
Jamshoro, Pakistan

Alim Pulatov
Head of EcoGIS centre
Tashkent Institute of Irrigation and Agricultural
Mechanization Engineers, Tashkent, Uzbekistan

***Abstract*—Nowadays, Internet of Things (IoT) is receiving a great attention due to its potential strength and ability to be integrated into any complex system. The IoT provides the acquired data from the environment to the Internet through the service providers. This further helps users to view the numerical or plotted data. In addition, it also allows objects which are located in long distances to be sensed and controlled remotely through embedded devices which are important in agriculture domain. Developing such a system for the IoT is a very complex task due to the diverse variety of devices, link layer technologies, and services. This paper proposes a practical approach to acquiring data of temperature, humidity and soil moisture of plants. In order to accomplish this, we developed a prototype device and an android application which acquires physical data and sends it to cloud. Moreover, in the subsequent part of current research work, we have focused towards a temperature forecasting application. Forecasting metrological parameters have a profound influence on crop growth, development and yields of agriculture. In response to this fact, an application is developed for 10 days ahead maximum and minimum temperatures forecasting using a type of recurrent neural network.**

Keywords*—Internet of Things; wireless sensor networks; smart agriculture; smartphone applications; artificial neural network; nonlinear autoregressive model; temperature forecasting

I. Introduction

Agriculture is a cultivation of products to feed the population. For centuries, it has remained as a key development factor for human civilization. Moreover, today the demand for efficient agriculture products is increasing [1]. In order to improve agriculture processes, we can acquire field data with sensors, make data analytics, perform analysis and take appropriate decisions and actions. Collecting big data from the field gives us a clearer understanding of product variability and quality of products [2].

In agriculture, physical parameters such as temperature, relative humidity and soil moisture are important [3]. There are several applications and well established measuring instruments to collect these data [4]. In addition, sensors to measure soil properties [5], detect and monitor foliar disease [1] or fertilizer management [6] already exist.

Data acquisition systems in agriculture need to cover large areas, collect representative samples and exchange measured information and control commands. In precision agriculture, one of the vital problems is how to distribute sensors and to establish reliable data communications. A robust and reliable data acquisition system enables synchronizing, exchange and storing of measured data. This kind of system is required to efficiently evaluate sensor signals and allow real-time or a-posteriori analysis of the behaviour of single parameters and their mutual impact [7].

Internet of Things (IoT) is receiving a great attention due to its potential strength and ability to be integrated into any complex system. It is becoming a great tool to acquire data from particular environment to the cloud. One of the use case fields of IoT is smart agriculture. However, there are some issues on developing low cost and power efficient WSN using advanced radio technologies for short and long-range applications. To satisfy the need for population, farmers and agriculture companies are turning to the Internet of Things (IoT). The IoT is pushing the future of farming to the next level. Smart agriculture is becoming more commonplace among farmers, and high tech farming is quickly becoming the standard thanks to the agricultural sensors.

WSN are playing important role in IoT technologies since it has been discovered. WSN systems are a strong and effective tool to distribute data among sensor nodes. The use of IoT applications and WSN (Wireless Sensor Networks) in the agriculture domain as proposed by other authors are discussed as follows. In [8], the authors have identified the issues related to reliability, autonomy, cost and accessibility to the application domain. The authors of [9] showed farm management system and its architecture is based on Internet of Things features. This architecture gives easy access to acquired data and advice. The authors of [10] have offered the use cases of Cloud Computing in the agriculture area. Moreover, they further describe it in the context of service providers and supply chain for cost-effective services for farmers. In [11], authors have illustrated the controlled architecture of smart agriculture based on IoT and Cloud Computing. Another issue arose with Wireless Sensor Networks (WSN) which must be connected to all sensor nodes in a smart way. The authors of [12] proposed a WSN for precision agriculture where a real-time data of the climatologically and other environmental properties are sensed and relayed to a central repository. Moreover, they proposed Wireless Mesh Network architecture which is convenient for agriculture applications [13]. The authors in [14] proposed an architecture Bayesian event prediction model which uses historical event data generated by the IoT cloud to predict future events. In [15], researchers reviewed the application of the data mining techniques for solving different challenges in event prediction system on time series. Despite the fact that there are many types of research, proposed architectures and WSN applications in agriculture domain, still, main problems rely on allocation sensors and thereby establishing reliable data communications in long-range areas. A robust and reliable data acquisition system enables synchronizing, exchange and storing of measured data. This kind of system is required to efficiently evaluate sensor signals and allow real-time or a posterior analysis of the behaviour of single parameters and their mutual impact [17].

This paper proposes a novel WSN approach and predictive model for smart agriculture sector which is different from above-mentioned technologies. Section 2 presents the detailed specifications followed and implemented to develop an IoT system and its Android application software. Our proposed IoT is capable to monitor ambient temperature, humidity and soil moisture. In Section 3, we introduce the use of an Artificial Neural Network (ANN) as time series forecaster to predict the temperature. Section 4 demonstrates the obtained results and evaluations for the chosen model. It also presents the discussion and summarized results. Section 5 concludes our work.

II. Research Methodology

In this section, we focused on the requirements and specifications of the system and described the developed hardware and software configurations of the Internet of Plants (IoP) system. In the foremost subsections, we specify the general architecture and components of our IoP based system. We further explain the basic flow of the designed system. Last two subsections are related to the server for storing data and developing PCB layout with Allegro ORCAD. The proposed system is based on monitoring temperature, humidity [16] and soil moisture of plants. We achieve this through developed IoP device and Android application, which provides the data, acquired from agricultural environment to the Internet through the Thingspeak platform. ThingSpeak [18] is an open source IoT platform. It is user-friendly and easy to extract data using API and HTTP protocol over the Internet or via a Local Area Network (LAN). The ThingSpeak IoT platform provides applications that let you analyze and visualize your data in MATLAB [19], and then act on the data. Sensor node data can be sent to ThingSpeak from Arduino [20], Raspberry Pi [21], BeagleBone Black [22], and other hardware.

A. General Architecture and Components of IoT System

The IoT is defined as a network of physical objects and devices embedded with Internet connectivity. It allows objects to be sensed and controlled remotely through embedded devices. The proposed prototype device gets information from plants and makes it available to the user. The design architecture of the IoP system is shown in Fig. 1. Next subsections will describe each stage of IoT based plant monitoring system.

B. Requirements and Specifications of the System

The IoP temperature, relative humidity and soil moisture acquisition system of plants consists of a device including a sensor and a microcontroller which wirelessly transmits plants data to the server using Wi-Fi protocol. DHT11 humidity and temperature sensors are selected for the system. 8-bit resolution digital output was used for reading the data. This sensor is ranked IP65, which means complete protection from dust and protection from water. It requires 2.5mA maximum current. The temperature ranges from 0 to 50 °C and response time from 6sec to 30sec, humidity ranges from 20% to 90% at 25 °C, response time 6 sec to 15sec.

To measure soil moisture, we have selected six Groovec soil moisture sensor. It is based on soil resistivity measurement, so with one device, it's possible to read the ground humidity in six different locations. Current consumption of the sensor is 35mA.

For data elaboration and communication management, the STM32L476RG microcontroller unit (MCU) of ST Microelectronics has been chosen. This component supports a Universal Synchronous/Asynchronous Receiver/Transmitter (USART) communication protocol required to communicate with the ESP-12E Wi-Fi module. It can drive various digital pins, that are required to communicate with the DHT11, drives the light-emitting diode (LED)s and performs other required functions. Maximum current usage of the MCU is about 150 mA. The main feature of this MCU is very low power consumption even in fully operative mode. This microcontroller enables users to serially upload the firmware through an embedded Joint Test Action Group (JTAG) serial wire debug port, which employs only two pins: PA13(the 46th pin) and PA14(the 49th pin).

For uploading the code it's possible to use an ST-Link device that can be connected to provide serial communication between the MCU and a personal computer (PC). In order to implement the connection between the device and Internet, we used ESP-12E Wi-Fi module. The WiFi module runs Lua language based scripts that are provided through USART communication by the MCU. WiFi module creates a client socket connection to the ThingSpeak internet protocol (IP) address through the predefined port and provides to a personal channel, which is created in the website database. In order to send data, we send an access request to the server. The thingspeak platform provides application programming interface (API) key which sends the package so that the data will be correctly processed. The maximum current consumption of the module is 250 mA.

We calibrated all sensors before prototyping. The calibration has been performed under standard working conditions. This means, for the DHT11, with a temperature range between (10 ÷ 35) °C and with a humidity range between (35 ÷ 65) %RH.

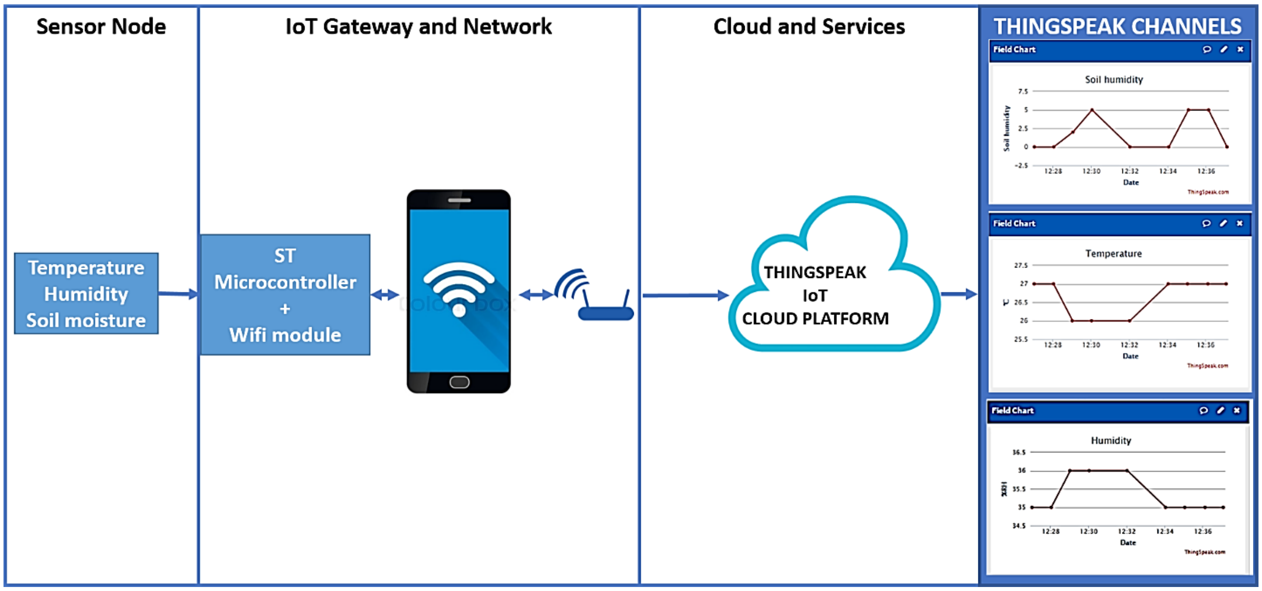


Fig. 1. IoT based plant monitoring system on the Thingspeak platform.

For the soil moisture sensor, another approach has been considered: First of all, the maximum reading value has been considered by measuring the value in water, and afterwards, the values in a dry soil that is gradually damping have been sensed. Considering the readability of the moisture data for the user, the result values have been divided into five ranges, that are represented by integer values as presented in Table I.

TABLE I. SOIL MOISTURE SENSOR CALIBRATION VALUES

Value	Data
0	Very dry soil
1	Dry soil
2	Humid soil
3	Very humid
4	Wet soil

C. The Design flow of the IOP

In order to have a working product, the design has been divided into three steps:

- First step: to create a working prototype sensor node using the NUCLEOL476RG developing board of ST Microelectronics and the NodeMCU WiFi board which is depicted in Fig. 2.
- Second step: to draw the schematics of design using the OrCAD Capture tool.
- Third step: to create the PCB layout using the OrCAD Allegro tool.

For developing prototype board we have been used STM MCU of the NUCLEOL476RG and ESP-12E component of the NodeMCU thus having an easier access to the offered features. Fig. 3 shows the final prototype of the IOP system.

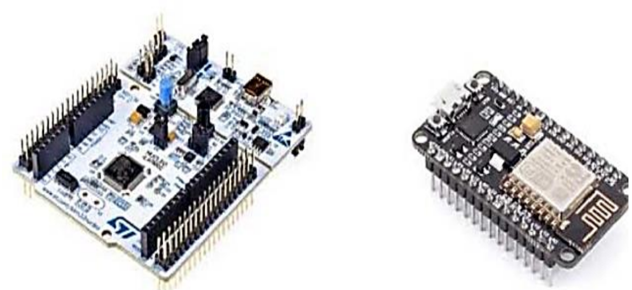


Fig. 2. The used boards NUCLEO-L476RG and NodeMCU.

D. WiFi Configurations

Whenever the device is booted, the WiFi connection needs configuration. Therefore, the first part of the firmware consists in a loop to correctly set the connection. For this phase, an Android application has been developed, so that the required information can be sent. The information required is:

- Service set-identifier (SSID) of the WiFi network in which the connection needs to be established.
- Password to connect to the network.
- API code for the ThingSpeak channel.
- Number of moisture sensors connected to the device.

Once the WiFi connection has been correctly set, the firmware enters into an infinite loop, which restarts every minute. Initially, data was collected from the sensors.

At this point, the values are correctly converted so that they can be transferred to the ESP module. Afterwards, the MCU through Lua scripts makes the ESP module open the socket connection. Data is thus sent to the ThingSpeak channel and plotted into each proper graph.

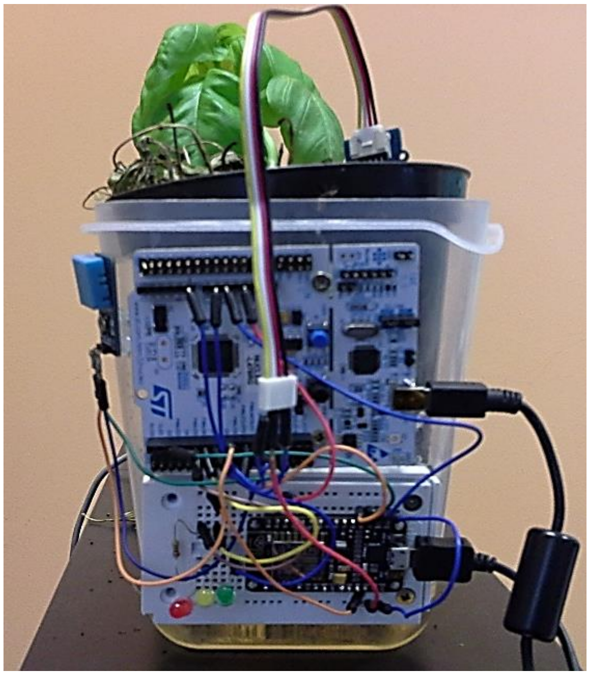


Fig. 3. Final development prototype of the IoP system.

E. IoP Android Application

Whenever the device is turned on or it's rebooted it is necessary to configure the WiFi connection to the access point. Solution before developing the application was to generate each time a telnet connection, with a computer or a general mobile application, by specifying the connection parameters and tokens. In order to make this procedure easier, we developed an application using Android Studio.

The application gives the possibility of configuring the connection by just entering the four required fields SSID, password, API code for the ThingSpeak channel, number of moisture sensors.

The Application is developed with a single activity that gives the possibility of entering the fields and by pressing the send button a Transmission Control Protocol (TCP). TCP/IP connection is established with the device and with a predefined timing the required data is sent. The user interface of the android application is shown in Fig. 4.

F. Web Service Settings

As mentioned earlier, Thingspeak is a free web service that gives the possibility to collect and store sensor data in the cloud for developing Internet of Things applications. The following steps are required to properly use this service.

Account registration

Channel settings:

Name: The name of the monitoring object

Field 1: Temperature

Field 2: Humidity

Field 3: Soil humidity

API generation: Each channel has an API key used to protect the user from undesired connections. When connecting to a channel, the API key specified to the App and the one specified on the portal must coincide. In Fig. 5, 6, 7 we have shown that how ThingSpeak plots data sent by the device and further collected by the sensors. Air temperature (Fig. 5) and humidity (Fig. 6) have been compared to the readings of a commercial thermo-hydrometer. Soil moisture sensor which is in Fig. 7 has been tested first in outside the soil with the absence of water, then, by inserting it into an almost dry soil; thus, by adding some water to the soil and finally into the pure water.

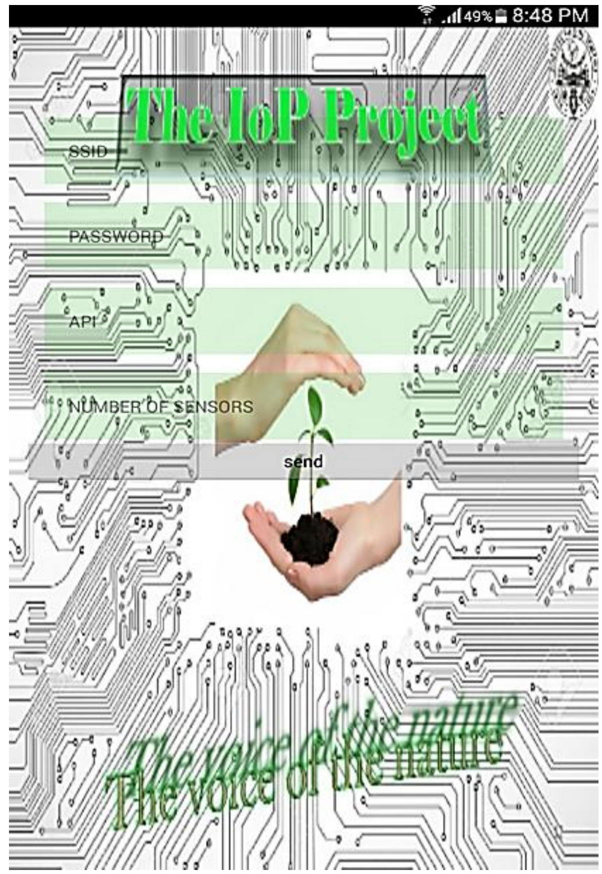


Fig. 4. The user interface of the IoP android application.

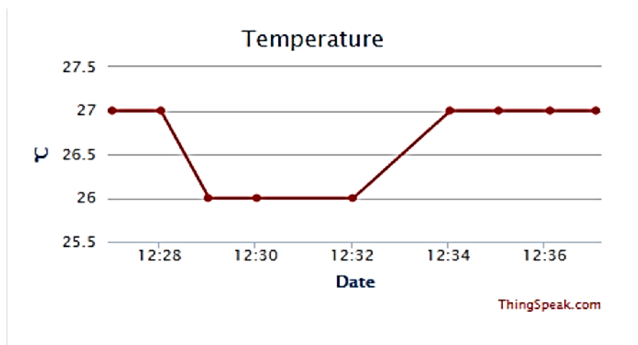


Fig. 5. Air temperature.

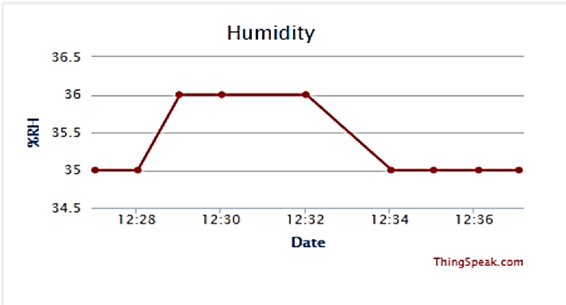


Fig. 6. Relative humidity.

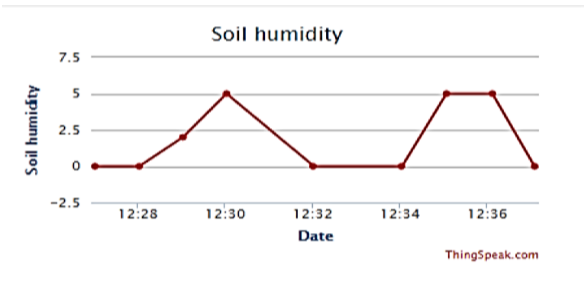


Fig. 7. Soil moisture.

G. PCB Layout with OrCAD Allegro

After creating a footprint of the PCB, the netlist of the circuit has been generated from the Capture tool and a file with the extension .brd is generated. Some constrain on the nets have been defined, in our case the power supply, +3V3 and GND nets need to change the configuration from the default one. After setting the correct environment, the components are placed. The board consists of a double layer PCB. The top layer holds components and provides a ground plane, the bottom one holds coupling capacitors and another ground plane. These two-ground plane are connected through series of vias to avoid big ground loops. In order to design correctly the PCB, several rules were employed for noise rejection and power dissipation. The methodology employed for placing and routing components is elaborated as follows.

PCB traces width: Maximum current consumption through the board occurs when the ESP module communicates over a WiFi connection. The total current, considering also the current required for the MCU module and the other circuitry is approximately 150 + 215 + (35*6) + 2.5 =577.5 mA. Thus, for 577.5 mA maximum current owing to power supply traces, 0.66 mm has been set for power traces.

PCB traces geometry and position: To avoid reactions 90 angles traces have been avoided. Since the board doesn't employ high-frequency signals, cross-talk effects have been neglected.

Decoupling capacitors: Near each power supply pins of each device decoupling capacitors have been used, to isolate direct current (DC) power supply pins from alternating current (AC) radiated noise.

Ground: Two ground planes, one for each side of the PCB have been used. These improve electromagnetic noise immunity but must be properly designed, to avoid undesired ground loops, by employing series of via between the two planes.

After defining the restrictions, it was possible to route the components. For this task, the automatic router of Allegro has been used and some more critical connections have been re-positioned.

The final PCB with highlighted measurements can be viewed in Fig. 8. It's also possible to see a 3D view of the PCB in Fig. 9.

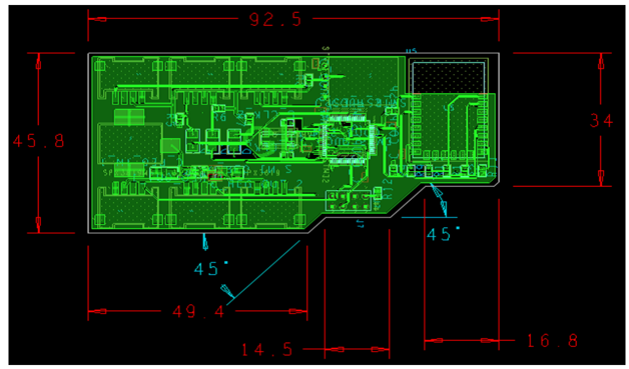


Fig. 8. PCB design with OrCAD Capture.

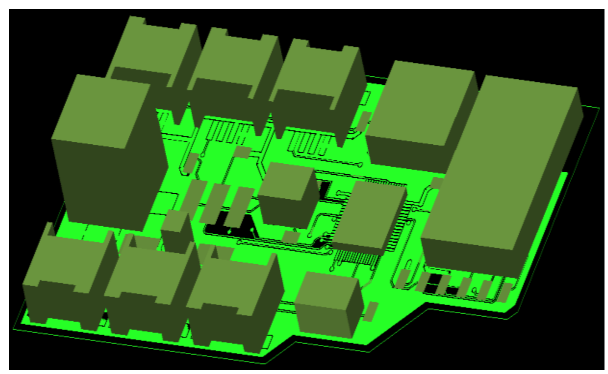


Fig. 9. 3D view of the PCB design.

The final cost of the product is around 25 Euro, where the main costs of production are provided from the DHT11 sensor, the MCU and the WiFi module, that cover about the 72% of the costs

III. Artificial Neural Network as Forecaster

Artificial Neural networks are becoming a significant method for prediction and forecasting in almost every field. Utilizing the time series approach to forecasting problem, forecasters actually collect and analyze the historical observations to create a model to capture the underlying behaviour of the data.

In Scientific research of meteorology, the weather forecasting is actually a typical unbiased time series forecasting problem. The real-life process is mostly nonlinear which brings deficiency in the traditional mathematical model. The metrological data is very irregular and it follows nonlinear Trend. Application of neural network in time series forecasting is based on the ability of the neural network to approximate nonlinear functions.

ANN is a mathematical model based on the structure and functions of biological neural networks. Information that flows through the network affects the structure of the ANN and it is generally based on three layers: an input layer, hidden and an output layer. Each layer contains nodes or neurons. Input layer consists of input nodes and output layer has as many nodes as the output. The middle layer or hidden layer contains an arbitrary number of nodes which is chosen after some trial and error initially. Fig. 10 shows a schematic view of feed forward Multilayer Perceptron (MLP). The mathematical representation is demonstrated in equation 1 where θ is the vector parameter which contains all the adjustable parameters of the network such as weights and biases$\{\boldsymbol{w_{j,l}}, \boldsymbol{W_{i,j}}\}$ and they are determined from a set of examples through the process called training. The examples, or the training data as they are usually called, are a set of inputs, u(t), and corresponding desired outputs, y(t).

$$\hat{y}(t) = g_i[z,\theta] = F_i\left[\sum_{j=1}^{n_h} W_{i,j} f_j\left(\sum_{l=1}^{n_z} w_{j,lzl} + w_{j,0}\right) + W_{i,0}\right]$$

In general, the aim of the experiment is to acquire a set of data that describes how the system behaves over its entire range of operation. Different input(s), u, is given on input layer and the impact on the output(s), y is observed. If the system to be identified is unstable or contains lightly damped dynamics it may be necessary to conduct the experiment again.

Afterwards, the network must be trained and when it starts to give reasonably accurate results, we can use the same network to predict unseen data. The flow chart for the selection and training of a predictive model is presented in Fig. 11.

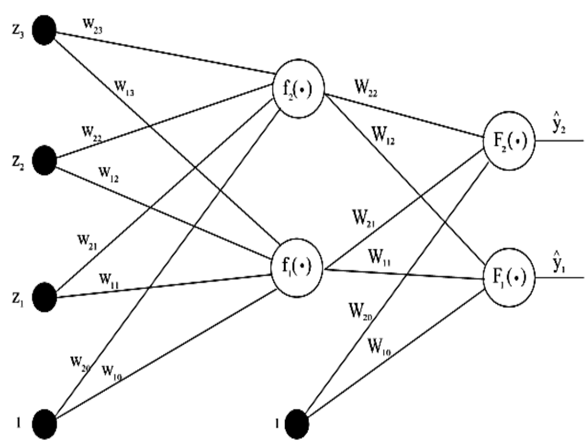


Fig. 10. MLP network with three input nodes and one bias, 2 hidden layers and an output layer with 2 nodes.

IV. Experimental Results

The primary goal of our research was to develop an IoT system which can acquire the remotely sensed weather data through an android application. Additionally, it should be capable to foresee the future data by implementing the neural networks for forecasting. Consequently, this aims to present an application for smart agriculture. In this part, we present the time series prediction of meteorological parameters using ANN.

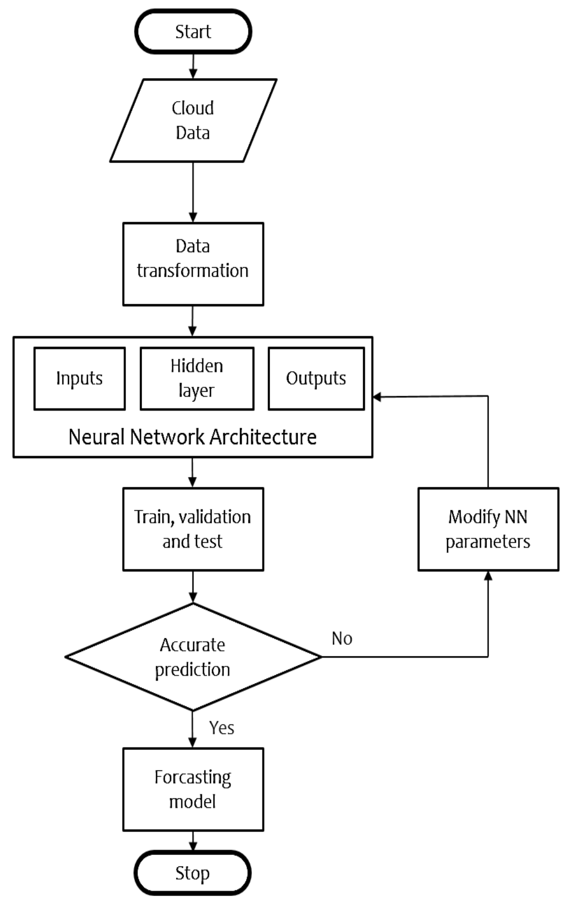


Fig. 11. The flow chart for the predictive model.

In order to select the appropriate predictive model, we took three different types of neural architectures into consideration which are further discussed in next section. Depending on the choice of input or regression vector, different nonlinear model structures emerge.

A. Time Series Temperature Prediction Model

Forecasting of temperature is important for many agriculture applications. Forecasts of the temperature of the soil, water, crop canopies or specific plant organs are also important in some specific cases. Crop species exhibit the phenomenon of thermo-periodicity, which is the differential response of crop species to daytime, nocturnal and mean air temperatures. It is possible to derive mean day and night-time temperatures from maximum and minimum temperatures data.

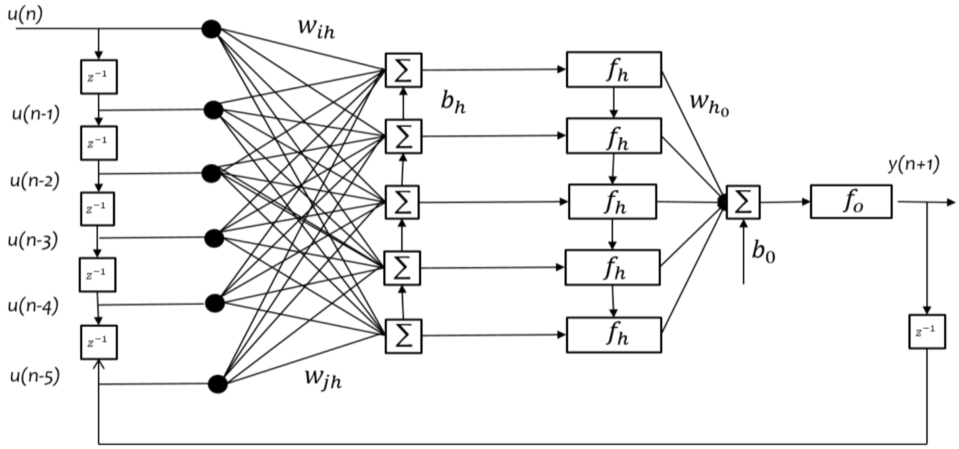


Fig. 12. NARX architecture model of 10-day forecasting maximum and minimum temperatures.

In order to predict maximum and minimum temperatures, we developed a nonlinear autoregressive exogenous model (NARX) ANN time series model. The followed approach is not intended for proposing a new method, but to use the dynamic and more robust method from the family of neural networks for forecasting the atmospheric temperature related to our aim. NARX has shown promising results and an accurate prediction in earlier studies [23]-[25]. NARX belongs to the family of Recurrent Neural Networks. The RNN have been widely used as one of the most popular data-driven dynamic models. In next sections, we present the data preprocessing and the followed steps for experimental setup.

B. Data Preprocessing and Feature Extraction

Since data collection takes a long time in agriculture applications, we took data from Oak Park weather station. In order to compute accurate predictions, we must have some meaningful attributes that provide content contribution and possibly reduced error rates. However, the valuable information can be lost as the choice of filtering and signal processing techniques require the expertise of data.

Feature selection, also known as variable selection or attribute selection. It is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. As the input attributes, we used date (day, month, year), rainfall and season. The season attribute is related to each of the four divisions of the year (winter, spring, summer and autumn). As an input feature, this season attribute is represented by a numerical value consecutively. In total, there are five nodes in the input layer. Our target (output layer) in the model was to predict maximum and minimum temperatures.

In order to remove the noisy fluctuations from the time series data, we filtered the target data with a Savitzky–Golay filter. Savitzky-Golay filtering can be thought of as a generalized moving average. In fact, the moving average filter smoothes the data by replacing each data point with the average of the neighbouring data points defined within the span.

For more accurate predictions, it is necessary to choose statistically the lagged terms, which are highly correlated with the predicted value Autocorrelation and Cross-correlation functions were used to determine the correct input/output patterns for nonlinear time series forecasting [26]. The correct combination and selection of lag terms also place strong impact on proper forecasts. These terms are also known as delay terms.

Moreover, the time series data usually have an explicit dependency on time variable. It can be observed that, given an input x(t) at time t, the model computes y(t), albeit similar input at later in time can be related to different prediction. Due to this reason, the predictive model must be equipped with more data input from past or must have memory for past inputs.

C. Forecasting Model

Prior to network modelling, the next step is the selection of meaning full attributes and proper lag terms. Consequently, the subsequent step is to develop and train ANN model. Since our aim was to predict both, the maximum and minimum temperatures, we created two separate models. The above significant lag terms for both models were determined by the correlation function.

The number of hidden neurons was selected by optimizing the model several times depending on validation set error. The approach followed here is in line with [25]. Finally, two hidden layers were selected with [10 50] number of neurons which fitted well to both time series data better than others. These trained models were further transformed into closed loop models so that multi-step ahead predictions could be computed.

The general NARX architecture of the multi-step forecasting model for maximum and minimum temperatures is shown in Fig. 12.

Where W_{ih}, W_{jh} and who represents weights of network and they are multiplied by inputs and added together in bias $b_h$and $b_o$and subjected to an activation function which decides the output of the network. The training was performed with Levenberg-Marquardt training algorithm, which is good at curve fitting for nonlinear regression problems.

D. Discussion

This paper proposed Internet of Plants (IoP) system based on monitoring temperature, humidity and soil moisture of plants. Further, collected data could be used to train a predictive ANN model. In order to train ANN model, we used Oak Park weather station dataset. We selected NARX as a forecasting model due to its dynamic nature and long-term dependency. For both maximum and minimum temperatures data, we trained separate neural network models. Mean Square Error (MSE) is chosen as performance measuring criteria. Prior to training, the data was divided into a training set, validation set and test set.

Trained ANN model was further extrapolated in a closed loop. Afterwards, this closed loop model was able to predict 10-days forecast for maximum and minimum temperatures. As we increased the steps ahead for forecasting, it is obvious that the propagation of error in each sample will be raised.

In the following section, we present results of performance of ANN model. The performance of trained models is shown in Fig. 13 and 14. The model for maximum temperature prediction provided an error of 0.8826 on unseen data for the month of September. Similarly, the performance of the model predicting minimum temperature was tested. It resulted in an error value of 0.944.

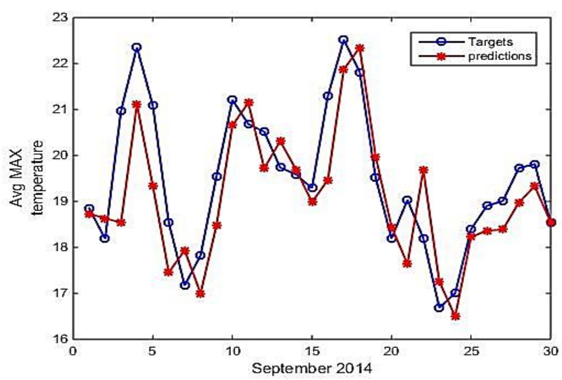


Fig. 13. Performance of the maximum temperature.

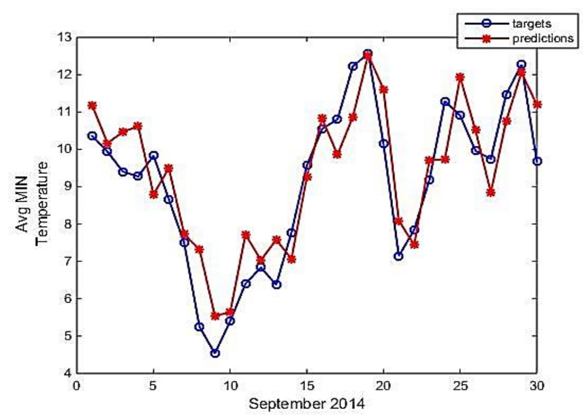


Fig. 14. Performance of the minimum temperature.

As mentioned earlier, the trained models were extrapolated in closed loop form to perform multi-step ahead predictions. The multistep forecasting computed for the next 10 days. The estimation for 10-days forecast maximum and minimum temperatures were obtained by a forecaster for the month of October. The performance of both predictive models for maximum and minimum temperatures for 10-days forecast is shown in Fig. 15 and 16 respectively. The figure shows, that the estimated forecasts for longer horizon can be achieved but with reduced accuracy. Since multistep forecasting is much more challenging than a one-step-ahead prediction. Thus, the accumulation of error increases, which degrades the prediction performance. This application can be considered general: it can be further applied to predict other climate attributes which are necessary for the deployment of smart agriculture.

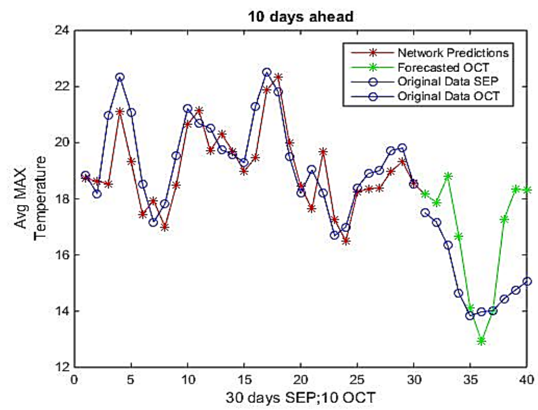


Fig. 15. Ten-days forecast of maximum temperature.

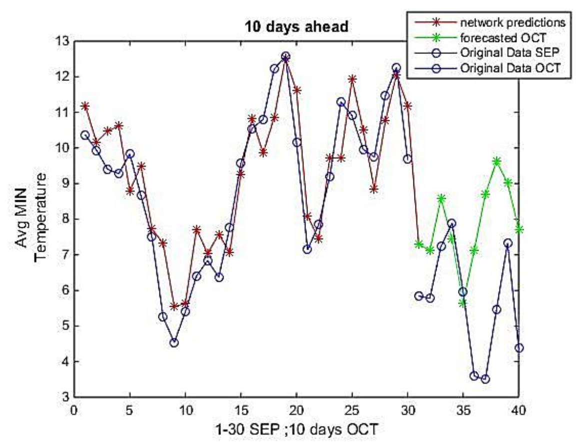


Fig. 16. Ten-days forecast of minimum temperature.

V. Conclusion

The main objective of our research was to develop an IoP based system which contains prototype device, the Android application as the hotspot to acquire physical data of the objects which are in long ranges and further send it to the cloud. We studied the network architecture, standards and protocols of IoT and selected the best communication protocols for our IoT designs. Considering the fact that in agriculture domain we need long distance communication and power efficient technologies, we proposed the WSN technology which is useful for field smart agriculture applications.

Additionally, we did 10 days ahead temperature forecasting using ANNs. Temperature prediction is important for many agriculture applications. However, the weather time series data is non-stationary and complex with unique characteristics that make them more challenging to get analyzed. In order to forecast the maximum and minimum temperatures for ten days ahead our preferred neural model is NARX. The deployed ANN application can be further applied to predict other climatic parameters of a smart agriculture.

ACKNOWLEDGEMENTS

Authors are thankful to Prof. Giansalvo Cirrincione for his support on supervising on the neural networks for modelling dynamic systems.

The work has been done during PhD period supported by TIMUR project, Erasmus Mundus Action 2 program and Higher Education Commission (Pakistan).

REFERENCES

[1] Adamchuk, V. I., Hummel, J. W., Morgan, M. T., & Upadhyaya, S. K. (2004). On-the-go soil sensors for precision agriculture. Computers and Electronics in Agriculture.

[2] Camilli, A., Cugnasca, C. E., Saraiva, A. M., Hirakawa, A. R., & Corrêa, P. L. P. (2007). From wireless sensors to field mapping: Anatomy of an application for precision agriculture. Computers and Electronics in Agriculture, 58(1).

[3] Zhang, N., Wang, M., & Wang, N. (2002). Precision agriculture—a worldwide overview. Computers and Electronics in Agriculture, 36(2–3), 113–132. https://doi.org/10.1016/S0168-1699(02)00096-0

[4] Schellberg, J., Hill, M. J., Gerhards, R., Rothmund, M., & Braun, M. (2008). Precision agriculture on grassland: Applications, perspectives and constraints. European Journal of Agronomy. https://doi.org/10.1016/j.eja.2008.05.005

[5] Cox, S. (2002). Information technology: The global key to precision agriculture and sustainability. In Computers and Electronics in Agriculture (Vol. 36, pp. 93–111).

[6] Corwin, D. L., & Lesch, S. M. (2005). Apparent soil electrical conductivity measurements in agriculture. Computers and Electronics in Agriculture. https://doi.org/10.1016/j.compag.2004.10.005

[7] Lee, W. S., Alchanatis, V., Yang, C., Hirafuji, M., Moshou, D., & Li, C. (2010). Sensing technologies for precision specialty crop production. Computers and Electronics in Agriculture.

[8] Mihai T. Lazarescu. (2017). Wireless Sensor Networks for the Internet of Things: Barriers and Synergies. In Components and Services for IoT Platforms (pp. 155–186). Springer International Publishing.

[9] Xiang YAN, Ji-yun JIN, Ping HE, M. L. (2008). Recent Advances on the Technologies to Increase Fertilizer Use Efficiency. Agricultural Sciences in China, 7(4), 469–479.

[10] Kaloxylos, A., Eigenmann, R., Teye, F., Politopoulou, Z., Wolfert, S., Shrank, C., … Kormentzas, G. (2012). Farm management systems and the Future Internet era. Computers and Electronics in Agriculture, 89, 130–144

[11] F. TongKe. (2013). Smart agriculture based on cloud computing and iot. Journal of Convergence Information Technology, 8(2).

[12] Siuli Roy, A. D., & Bandyopadhyay, S. (2008). Agro-sense: Precision agriculture using sensor-based wireless mesh networks. In International Telecommunication Union - Proceedings of the 1st ITU-T Kaleidoscope Academic Conference, Innovations in NGN, K-INGN.

[13] Yinbiao, Shu, P. Lanctot, and F. J. (2015). Internet of things: wireless sensor networks. White Paper, International Electrotechnical Commission.

[14] Karakostas, B. (2016). Event Prediction in an IoT Environment Using Naïve Bayesian Models. In Procedia Computer Science (Vol. 83, pp. 11–17).

[15] Molaei, S. M., & Keyvanpour, M. R. (2015). An analytical review for event prediction system on time series. In 2015 2nd International Conference on Pattern Recognition and Image Analysis (IPRIA) (pp. 1–6).

[16] Aliev Khurshid, Rugiano Francesco, Pasero Eros. (2016b). The Use of Bluetooth Low Energy Smart Sensor for Mobile Devices Yields an Efficient Level of Power Consumption. In The First International Conference on Advances in Sensors, Actuators, Metering and Sensing (pp. 5–9). Venice.

[17] Aliev Khurshid, Rugiano Francesco, Pasero Eros. (2016a). Smartphone and Bluetooth Smart Sensor Usage in IoT Applications. Sensors & Transducers, 201(6), 27–34.

[18] https://thingspeak.com/

[19] www.mathworks.com

[20] https://www.arduino.cc/

[21] https://www.raspberrypi.org

[22] http://beagleboard.org.

[23] Bektas Oguz, Jeffrey Jones. (2016). NARX Time Series Model for Remaining Useful Life Estimation of Gas Turbine Engines. In Third European Conference of the Prognostics and Health Management Society (pp. 1–10).

[24] Aliev Khurshid, Narejo Sanam, Pasero Eros, I. J. (2016). A Predictive Model of Artificial Neural Network for Fuel Consumption in Engine Control System. In WIRN 2016 26th Italian Workshop on Neural Networks

[25] Sanam Narejo and Eros Pasero. (2016b). Time series forecasting for outdoor temperature using nonlinear autoregressive neural network models. Journal of Theoretical and Applied Information Technology, 94(2), 451–463

[26] Sanam Narejo and Eros Pasero. (2016a). A Hybrid Approach for Time series Forecasting using Deep Learning and Nonlinear Autoregressive Neural Network. In INTELLI 2016 : The Fifth International Conference on Intelligent Systems and Applications (includes InManEnt 2016) (pp. 69–75).

A Penalized-Likelihood Image Reconstruction Algorithm for Positron Emission Tomography Exploiting Root Image Size

Munir Ahmad[1]
Institute of Nuclear Medicine and Oncology,
Lahore, Pakistan
Physic Department, The University of Lahore,
Lahore, Pakistan

H. M. Tanveer[2], Z.A. Shaikh[3]
Department of Electrical Engineering
University of the Punjab Lahore, Pakistan

Furkh Zeshan[4]
Department of Computer Science
COMSATS, Institute of Information and Technology
Lahore, Pakistan

Usman Sharif Bajwa[5]
University of the Punjab Lahore,
Pakistan

***Abstract*—Iterative image reconstruction methods are considered better as compared to the analytical reconstruction methods in terms of their noise characteristics and quantification ability. Penalized-Likelihood Expectation Maximization (PLEM) image reconstruction methods are able to incorporate prior information about the object being imaged and have flexibility to include various prior functions which are based on different image descriptions. Median Root Priors intrinsically take into account the salient image features, such as edges, which becomes smooth owing to quadratic priors. Generally, a 3*3 pixels neighborhood support or root image size is used to evaluate the median. We evaluate different root image sizes to observe their effect on the final reconstructed image. Our results show that at higher parameter values, root image size has pronounced effect on different image quality parameters evaluated such as reconstructed image bias as compared to the phantom image, contrast and resolution in the reconstructed object. Our results show that for the small-sized objects, small root image is better whereas for objects of diameter more than two to three times of the resolution of the reconstructed object, larger root image size is preferable in terms of reconstruction speed and image quality.**

***Keywords*—Penalized-Likelihood expectation maximization; median root priors; maximum-likelihood expectation maximization; full-width-at-half-maximum**

I. Introduction

Positron emission tomography (PET) and Single photon emission tomography (SPECT) are used to image human body functions non-invasively. Data obtained from these scanners, is reconstructed by analytical or iterative image reconstruction methods to estimate emission object's activity distribution. Analytical image reconstruction methods use line integral model and simply ignore any underlying noise distribution [1], [2]. It is also not so simple to incorporate emission or detection physics model into the reconstruction problem and these methods do not take into account the non-negativity condition due to Fourier transforms employed. On the other hand, statistical iterative image reconstruction methods can include modeling for emission and detection processes into the system matrix and automatically fulfill non-negativity constraint [3]. Owing to this ability, statistical iterative image reconstruction methods are claimed to be superior to analytical image reconstruction methods with respect to noise characteristics and quantification ability [3], [4].

A very popular and basic iterative image reconstruction method is known as maximum-likelihood expectation maximization (MLEM) method [5]. Various techniques, such as post reconstruction smoothing or method of Sieves, have been used to reduce this resultant image noise [7]–[9]. We adopt the same definition of the prior derivative in this paper. The equally popular class of priors based on median root priors (MRPs) is used. MRPs are designed on the basic image description that images are locally monotonic. In MRP's algorithm, neighborhood support of the prior function is termed as root image and most frequently a 3*3 pixels neighborhood is used as the root image. Median value evaluated on this neighborhood is used to penalize image pixels. We evaluate the impact of variation of the root image size on the final reconstructed image by proposing a penalized-likelihood image reconstruction algorithm which exploits root prior knowledge.

To the best of our knowledge, there is not much work done to evaluate the impact of root image size on final reconstructed image characteristics [15]. It is observed that the correlation between image pixels decreases as the distance between two pixels is increased. However, the value of the median evaluated on different root image sizes may vary. This work focuses on analyzing the impact of this variation on the properties of the final reconstructed image.

The rest of this paper is organized as follows. In Section II, we describe the methodological background. In Section III, we outline the proposed algorithm. Experimental results are shown in Section IV. In Section V, the discussion on the experimental results is presented. Finally, the paper is concluded in Section VI

II. Methodological Background

MLEM is an iterative optimization method used to reconstruct images from scanner data known as sinogram. Mathematically, this method attempts to maximize an objective function based on the statistical average of the logarithm of the likelihood function as follows:

$$\hat{X} = arg_{X\in\Omega} \max \mathrm{E}(L(X,Y)) \qquad (1)$$

where the objective function L(X,Y) is defined as the logarithm of the likelihood function of the emission data Y given object X and the statistical expectation of Y is considered to be the mean of the independent Poisson emission model as given below. A_{ij} is the system matrix element and characterizes the probability of an event being detected at the i-th bin and emitted from the j-th pixel point.

$$\overline{Y} = E[Y(X)] = \sum_{j=1}^{N} A_{ij}\, X_j \qquad (2)$$

However, with increasing iteration number this method is known to produce noisier images as image reconstruction is an ill-posed inverse problem and this technique attempts to fit the solution image to the data and does not consider any priori information about the object being imaged [6].

Various techniques such as post reconstruction smoothing or method of Sieves, have been used to reduce this resultant image noise [7]–[9]. However, penalized-likelihood image reconstruction methods are more favorable due to their flexibility to incorporate various penalizing schemes. A penalty function is added to the likelihood function based on some priori image description or knowledge and logarithm of this modified objective function is maximized, instead, as described in Eq. (II). These functions induce additional constraints to reduce the solution image set [10], [11].

$$\hat{X} = arg_{X\in\Omega} \max[E[L(X,Y)] - \beta R(X)] \qquad (3)$$

In relationship given below, R(X) is the prior term and depends on some priori image information or image description and is the hyper parameter which controls prior influence on the final image. Generally, quadratic prior functions are used due to their implementation simplicity [3], [12]. However, these priors produce overly smoothed salient image features due to their penalizing scheme to modify pixel values on the basis of the differences in their values [3], [7], [11], [13]. These priors are designed with the basic concept of image description that images are locally smooth. Hence, quadratic priors attract all image pixels in a local neighborhood which are close to each other in values [1]. Following relationship describes mathematical form of the quadratic priors:

$$R(X) = \sum_{j=1}^{N} \sum_{k\epsilon Nb} w_{ij}\, V\left(x_j - x_k\right)$$
$$= \sum_{j=1}^{N} \sum_{k\epsilon Nb} w_{jk}\, \frac{1}{2}(x_j - x_k)^2 \qquad (4)$$

In the above relationship, w_{jk} are the weights assigned to pixels in a smaller neighborhood and V(x) is energy function as defined in Gibbs distribution [14]. Another equally popular class of priors based on median root priors (MRPs). MRPs are designed on the basic image description that images are locally monotonic. Interestingly, this includes the definition of local smoothness. Hence, we may observe some level of smoothing in the resultant image. MRPs also have their ability to automatically preserve edges inside the image without any need of an additional tuning parameter which is a disadvantage of other non-quadratic edge-preserving priors [1], [15], [16]. MRPs can be described as follows:

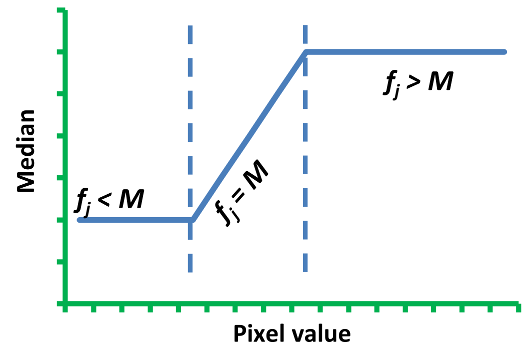


Fig. 1. An empirical dependence of median in a local neighborhood against those pixel values in the same neighborhood.

$$R(X) = \sum_{j=1}^{N} \sum_{k\epsilon Nb} V\left(x_j - x_k\right)$$
$$= \sum_{j=1}^{N} \frac{(X_j - M_j)^2}{2M_j} \qquad (5)$$

Where M_j is the median in local neighborhood of jth pixel and N_b is the number of pixels in the neighborhood of the relevant disk. Pixel neighborhoods are generally considered as Markov Random Fields (MRFs) and described according to Gibbs distribution [14], [17]. These prior functions are incorporated into the reconstruction algorithm according to Greens very popular One-Step-Late (OSL) algorithm. According to this algorithm, we need first derivative of the objective function, including log-likelihood and logarithm of the prior function, to obtain the current image estimate. However, while the current image is being estimated, it is not possible to evaluate first derivative at the current image estimate. Derivative of the prior function is evaluated for image estimated at the previous iteration, hence the name OSL algorithm [18].

OSL algorithm needs first derivative of the prior function and, unfortunately, it is not possible to directly evaluate the gradient of the median function as dependence of median on local neighborhood is nonlinear. We have to resort to some empirical derivation of the gradient of median based priors. MRP has been defined as Gaussian distribution of the prior function, in the form of Gibbs priors [15], [17]. For the sake of the evaluation of the derivative of MRP, authors have assumed an empirical dependence of the local median on the image values as presented in Fig. 1. They have assumed local median to be constant within the functioning limits of the MRP function. Under this assumption, it becomes much easier to find an empirical derivation of the prior function. We adopt the same definition of the prior derivative in this paper. In MRP algorithm, neighborhood support of the prior function is termed as root image and most frequently, a 3*3 pixels neighborhood is used as the root image.

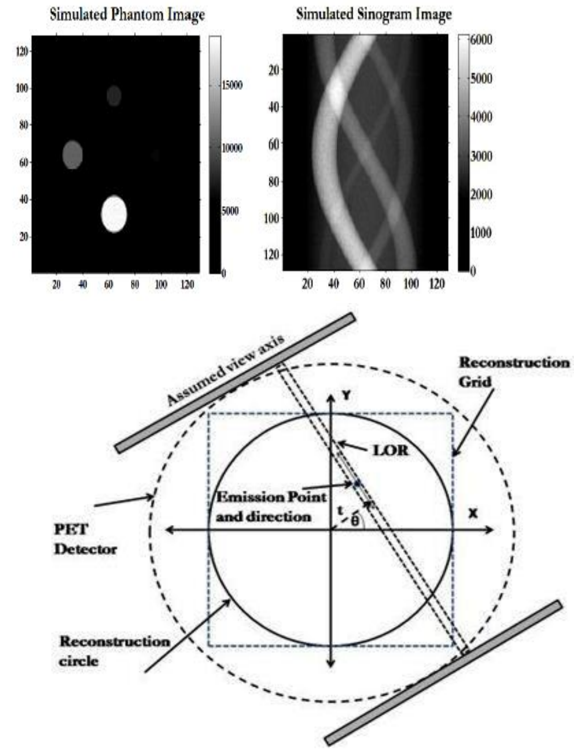


Fig. 2. Data simulation scheme (bottom) simulated Phantom image (left) and simulated sinogram data (right) which was used to reconstruct images for our analysis evaluated on this neighborhood is used to penalize image pixels.

III. Proposed Algorithm

We use a 128*128 pixels image grid with a circular emission object comprised of a main background disk, having four small disks inside it placed at the same distance from center of the background disk as shown in Fig. 2 (left). The simulated phantom image was based on the simulation scheme shown in Fig. 2 (bottom). This figure shows a square grid encircling reconstruction circle inside its boundaries. A random point was generated inside the grid along with a randomly generated angle to simulate a random event. This simulated event, based on a random point and angle pair, was assumed to describe path of a randomly emitted photon pair and detected in a particular line-of-response (LOR). Index of this LOR was calculated by the following (6) and describes perpendicular distance of the assumed path of emitted pair of photons.

$$t = x \cos\vartheta + \ y \sin\vartheta \qquad (6)$$

In this expression, t represents perpendicular distance from the origin to the LOR, defined by the line of flight of the pair of photons emitted at point (x; y) and traveling at an angle with the x-axis. To reconstruct images for our analysis, we have used One-Step-Late (OSL) algorithm by P. Green [18]. That algorithm can be given by the following relationship.

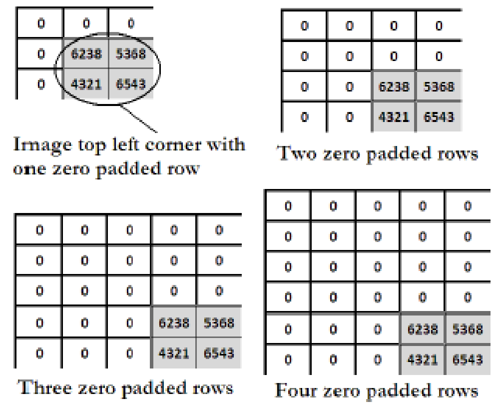


Fig. 3. Zero padding scheme for the application of median.

$$X_j^{k+1} = \ {C_{PL}^k C_{ML}^k X_j^k}\Big/{\textstyle\sum_{i=1}^{M} A_{ij}} \qquad (7)$$

$$X_j^{k+1} = \ {C_{PL}^k C_{ML}^k X_j^k}\Big/{\textstyle\sum_{i=1}^{M} A_{ij}} \qquad (8)$$

$$X_j^{k+1} = \ {C_{PL}^k C_{ML}^k X_j^k}\Big/{\textstyle\sum_{i=1}^{M} A_{ij}} \qquad (9)$$

Where C_{ML}^k is the maximum-likelihood correction factor and C_{PL}^k is the priors correction factor both evaluated at kth iteration. As mentioned by Green [18], prior function needs to be a continuous function having continuous derivatives, unluckily, median priors do not have its derivatives defined. However, Alenius [15] has developed some heuristic approach to visualize dependence of median on the local neighborhood as shown in Fig. 1. According to this approach, median can be considered constant within the operational range of the MRP.

Our work deals with the effect of root image size on the images reconstructed by the PLEM image reconstruction methods including MRPs, hence we have used varying size of root images. Generally, a 3*3 pixels or 5*5 pixels root image is used to reconstruct images with PLEM including MRPs. We used 3*3, 5*5, 7*7 and 9*9 pixels root images to analyze characteristics of the reconstructed images. It is important to note that at the boundaries of the image grid, we need zero padding in order to apply MRPs. We have zero-padded the images as shown in Fig. 3.

Initially, for our analysis, to see if size of the root image does produce any change in the reconstructed image, we evaluate reconstructed image bias (RIB) relative to the true phantom image based on the following equation:

$$RIB = \textstyle\sum_{k=1}^{N} (X_k - Ph_k) \qquad (10)$$

Where X_k and Ph_k are the relative pixels in the reconstructed image and the phantom image respectively and N is the total number of pixels in the image. We analyze the behavior of MRPs relative to the size of their root image in terms of image contrast in a local neighborhood. We select 4 4 pixels ROIs inside each small disk, at higher activity as compared to the background, to obtain positive contrast values. Contrast is evaluated using the following formula:

$$Constrast = \frac{(X_e - X_b)}{X_b} \qquad (11)$$

In this relation, X_e is the estimated mean value in an ROI in the hot disk and X_b is the estimated mean value in the background disk of the same size ROI. Finally, we study the effect of root image size on the resolution of the reconstructed image. We generate two realizations of the sinogram to evaluate reconstructed resolution, one with an impulse added to the centre pixel of each disk and the other one without an added impulse. These sinograms are then reconstructed to obtain two images f(X +ΔX) and f(X). Resolution in Full-Width-at-Half-Maximum (FWHM) is calculated from the differential image evaluated using the following formula:

$$\frac{dI}{dx} = \lim_{\Delta \to 0} \frac{(f(X+\Delta X) - f(X))}{\Delta} \qquad (12)$$

Where ΔX is a very small quantity and logically tends to zero for the differential image. A Matlab Function fwhm2.m, developed by J. A. Fessler is used to calculate resolution in pixels at the centre of each small disk. The proposed algorithm is outlined in Algorithm 1.

Algorithm 1: Penalized-Likelihood Image Reconstruction Algorithm for PET

1: Input: Take an image X_j^k
2: Output: Resolution
3: For each iteration number k, update image using Eqs. (7-9).
4: Reconstruct image with varying neighbourhood size 3*3, 5*5, 7*7, 9*9 in PLEM.
5: Evaluate image bias using Eq. (10) for each reconstructed image to evaluate hyper parameter values for optimum reconstructed resolution.
6: for each beta do
7:     for each disc do
8:         Evaluate contrast using Eq. (11).
9: Evaluate reconstructed resolution by FWHM of impulse response.
10:         Next
11:     end for
12:     Next
13: end for

IV. Experimental Results

We evaluate reconstructed image bias relative to the true phantom image, to observe any effect of root image size on reconstructed images, and presented our results in Fig. 4. We use (10) to calculate image bias and this figure displays bias values for different beta values and root image sizes. We observe that image bias reduces with decreasing value of beta parameter value, reconstructed image moves more and more towards its MLEM solution, hence reducing effect of the prior and, consequently, of the root image size.

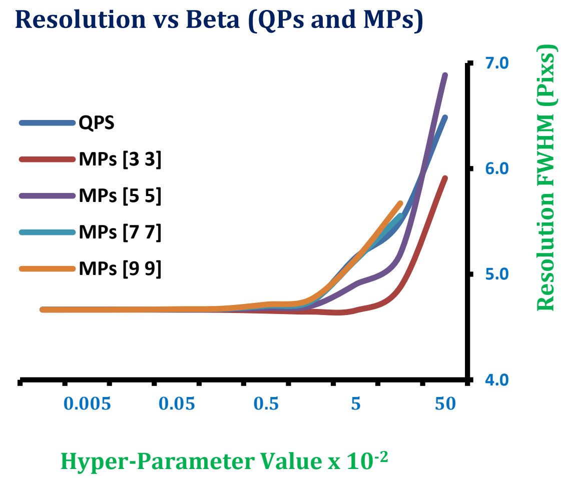


Fig. 4. Reconstructed image bias relative to the true phantom image using MRPs for different root-image sizes and beta parameter values.

It is clear that at low beta values, there is not much difference in bias values for different root image sizes. However, as beta value increases, and gets near 1, image bias increases and clearly is affected by different root image sizes in penalized-likelihood image reconstruction method using MRPs. We also analyze the effect of root-image-size in MRPs on the reconstructed contrast values inside each small disk, using (4), and present our results in Fig. 5.

This image clarifies that the reconstructed contrast varies with different root image size at higher beta values for smaller objects. This contrast values becomes almost same for different root image sizes for the largest disk size having highest activity level. However, for lower beta values root image size does not have much effect on the reconstructed contrast in different objects and for different root image sizes.

Fig. 6 presents results for reconstructed resolution, in FWHM (Pixels), at the centre of each small disk for two different beta parameter values. With higher beta value, reconstructed resolution is much different for different root image sizes. However, for lower beta value where priors influence is reduced, root image size does not seem to have strong influence. This effect is similar to that of different root image sizes over the image bias or reconstructed image contrast values.

Average reconstructed resolution also seems to be higher (lower FWHM values) in case of lower beta value as compared to the one with higher beta values which may be considered as higher smoothing at high beta values. This may not, strictly, be the same kind of smoothing as produced by quadratic priors [2]. It should also be noted that DISK2 and DISK4 are on the vertical axis whereas DISK1 and DISK3 are on the horizontal axis. This indicates an asymmetrical reconstructed resolution response.

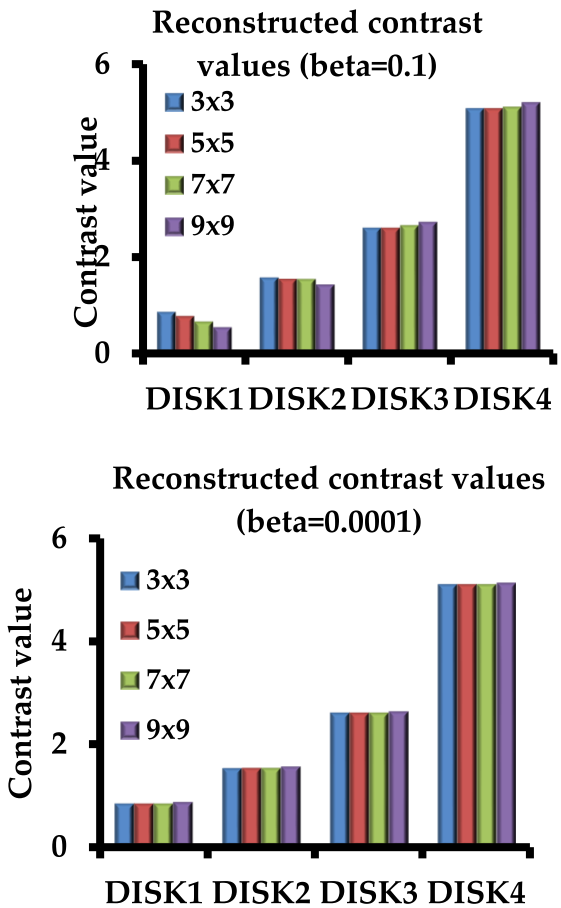
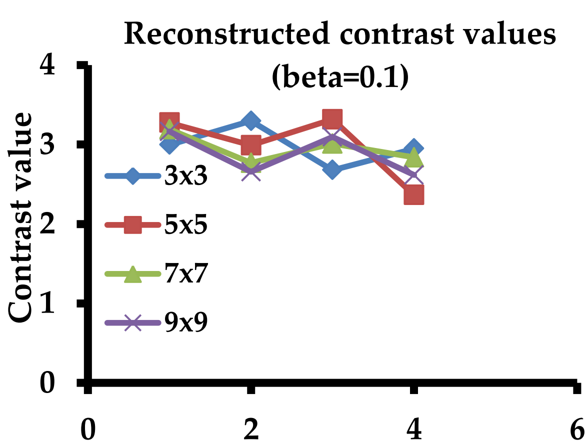


Fig. 5. Contrast recovery results with MRPs for different size of root images. DISK1 to DISK4 present recovered contrast for different activity values and different size of the object for two different values of beta hyper parameter values. At higher beta values, for the smallest DISK1 with least activity, contrast varies for different root-image-sizes, whereas for largest DISK4 with highest activity value, contrast is almost same.



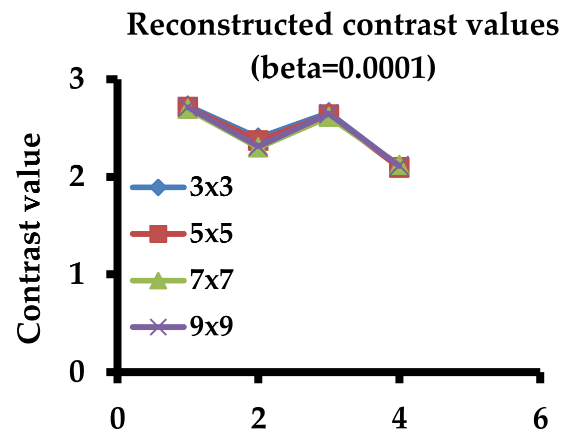


Fig. 6. Resolution in FWHM (Pixels) with MRPs for different size of root images at the center of each small disk. DISK1 to DISK4 present reconstructed resolution for different activity values and different size of the object for two different values of beta hyper parameter values. At higher beta values, for the smallest DISK1 with least activity, contrast varies for different root-image-sizes, whereas for largest DISK4 with highest activity value, contrast is almost same.

V. Discussion

We have analyzed effects of root image size on the final reconstructed image using penalized-likelihood image reconstruction methods including median root based priors. We have evaluated this effect in terms of reconstructed image bias as compared to the true phantom image, reconstructed contrast in high activity regions and resolution in FWHM (in pixels) at certain locations inside the image.

Our results show that root image size has pronounced effect on the bias values for higher beta parameter values. Higher beta value, in case of MRPs, means higher priors influence on the final reconstructed image which means that it is attracted more towards local medians. At higher beta values, shape of the prior is more sharply peaked and locally non-monotonic areas are dragged towards local median with stronger force. Moreover, with higher beta values and bigger root image size, response of MRPs is farthest from that of any local smoothing prior (quadratic priors), hence reduced smoothing will be observed. That could be the reason that we observe higher bias values with bigger root image size at higher beta values. This can also be attributed to less correlation between distant pixels.

Additionally for contrast values, no significant difference is observed at lower beta values with different root image size. However, at higher beta values, contrast is much reduced for bigger root image size for small object size containing low activity value as compared to the smaller neighborhood in the same object. This may also mark presence of partial volume error because this effect is reversed in the bigger size disks containing higher activity values where contrast is higher for bigger root image size.

For higher values of beta parameter, resolution in FWHM (Pixels) varies markedly for different size of root image and even the trend varies in horizontal and vertical directions. DISK1 and DISK 3 are on the horizontal axis where the reconstructed resolution is worse (or higher in FWHM) as compared to the DISK 2 and DISK 4 that are lying on the vertical axis. This effect is totally strange for the smallest root image size and could represent a mixture of root image partial volume error combined due to the disk size smaller than 2 to 3 times of the resolution (FWHM in Pixels).

However, at lower beta values (near to zero) difference between different root image size disappeared, though the effect of object size and activity level still persists in the form of varying resolution at the center of these disks.

VI. Conclusion

We evaluated the effect of varying root image size on the final reconstructed image by Penalized-Likelihood image reconstruction methods with median root based priors. We conclude that higher root image or neighborhood size produces higher bias which means reconstructed image is farther from the true image as compared to the image reconstructed by the smaller root image size. The same is true for the reconstructed contrast and reconstructed resolution. It should be stressed that the effect of the hyper-parameter values be considered along with the effect of root image size.

References

[1] S .Alenius, U. Ruotsalainen,: Generalization of median root prior reconstruction. IEEE Trans.Med.Imaging11, 1413-1420 (2002).

[2] J.W. Stayman, J.A Fessler,: Regularization for uniform spatial resolution properties in penalized-likelihood image reconstruction. IEEE Trans.Med.Imaging6, 601-615 (2000).

[3] J.W. Stayman, J.A. Fessler,: Compensation for nonuniform resolution using penalized-likelihood reconstruction in space-variant imaging systems. IEEE Trans.Med.Imaging3, 269-284 (2004).

[4] H. Wieczorek,: The image quality of FBP and MLEM reconstruction. Phys.Med.Biol.11, 3161-3176 (2010).

[5] L.A. Shepp, Y. Vardi,: Maximum likelihood reconstruction for emission tomography. IEEE Trans.Med.Imaging2, 113-122 (1982).

[6] L. Kaufman,: Implementing and accelerating the EM algorithm for Positron Emission Tomography. IEEE Trans.Nucl.Med37-51 (1987).

[7] J. Nuyts, J.A. Fessler,: A penalized-likelihood image reconstruction method for emission tomography, compared to postsmoothed maximum-likelihood with matched spatial resolution. IEEE Trans.Med.Imaging9, 1042-1052 (2003).

[8] D.G. Politte, D.L. Snyder,: Corrections for accidental coincidences and attenuation in maximum-likelihood image reconstruction for positron-emission tomography. IEEE Trans.Med.Imaging1, 82-89 (1991).

[9] D.L. Snyder, M.I. Miller,: The use of sieves to stabilize images produced with the EM algorithm for emission tomography. IEEE Tr.Nuc.Sci5, 3864-3871 (1985).

[10] J.A. Fessler,: Penalized weighted least-squares image reconstruction for positron emission tomography. IEEE Trans.Med.Imaging2, 290-300 (1994).

[11] J.A. Fessler, A.O. Hero,: Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms. IEEE Trans. Image Process10, 1417-1429 (1995).

[12] J.A. Fessler, W.L. Rogers,: Spatial resolution properties of penalized-likelihood image reconstruction: space-invariant tomographs. IEEE Trans.Image Process9, 1346-1358 (1996).

[13] J.W. Stayman, J.A. Fessler,: Efficient calculation of resolution and covariance for penalized-likelihood reconstruction in fully 3-D SPECT. IEEE Trans.Med.Imaging12, 1543-1556 (2004).

[14] K. Lange,: Convergence of EM image reconstruction algorithms with Gibbs smoothing. IEEE Trans.Med.Img.4, 439-446 (1991).

[15] S. Alenius, U. Ruotsalainen,: Bayesian image reconstruction for emission tomography based on median root prior. Eur.J.Nucl.Med.3, 258-265 (1997).

[16] V. Bettinardi, S. Alenius, P. Numminen, M. Teras, M.C. Gilardi, F. Fazio, U. Ruotsalainen,: Implementation and evaluation of an ordered subsets reconstruction algorithm for transmission PET studies using median root prior and inter-update median filtering. Eur.J.Nucl.Med.Mol.Imaging2, 222-231 (2003).

[17] T.J. Hebert, R.M. Leahy,: Statistic-based MAP image reconstruction from Poisson data using Gibbs priors. IEEE Trans.Sig.Proc.9, 2290-2303 (1992).

[18] P.J. Green,: On the use of the EM algorithm for penalized likelihood estimation. J.Roy.Statist.Soc., B3,443-452 (1990).

Secure and Efficient Routing Mechanism in Mobile Ad-Hoc Networks

Masroor Ali[1], Zahid Ullah[2], Meharban Khan[3], Abdul Hafeez[4]
Department of Electrical Engineering, CECOS University of IT and Emerging Sciences, Peshawar, Pakistan[(1,2,3)]
Department of Computer Science and IT, University of Engineering and Technology, Jalozai Campus, Pakistan[(4)]

***Abstract*—Securing crucial information is always considered as one of the complex, critical, and a time-consuming task. This research investigates a significant threat to the security of a network, i.e., selective forwarding attack. Protection of information is considered as the main stimulating task in the design of an information system for an organization. The research work proposes a framework that detects the selective forwarding attacks and computes the harmful hosts residing an ad-hoc structure. Our solution is further split into two phases: initial phase is the detection of selective forwarding attacks and the second phase performs the identification of malicious nodes. Performance of the proposed model is evaluated based on the network throughput, which is for the enhancement of security. Simulation of the proposed model is performed using NetLogo and the results show an improvement of 20% in throughput of the network.**

Keywords—Ad-hoc on-demand distance vector; control ACK; mobile ad-hoc network; network throughput

I. Introduction

Ad-hoc network is considered as one of the most emerging technologies in today's world. The unique ad-hoc mesh structure of a network is the latest wireless networking model of roving hosts. In an ad-hoc network, the hosts rely upon one another to enable and maintain the entire network communicating and linked together. The ad-hoc networking model has achieved a lot of attention in the industry of information technology as well as in academics. The ad hoc mesh is most developing technologies in the information technology world. Significant research has been conducted in emerging technology of mobile ad-hoc network (MANET).

The use of MANET upon people's day-to-day life style has greatly developed with the passage of time. Thus, in response it has created an immense requirement for a secure ad-hoc structure. Greater amount of investigation and research is focusing upon constructing and modeling efficient and safe ad-hoc network [1], [2]. To provide defense from malicious hosts, different researchers are working specifically to the challenges concern to security of MANETs [11].

The ad hoc structure is built with the help of a number of mobile nodes that are connected wirelessly and they cooperate with each other. These networks transfer the information and data by using the multi-hop forwarding techniques. The deployment for a spectacular MANET structure is very useful into such area where the pre-deployed infrastructure is not supportive. There are several recent works on the security of MANETs but most of them have failed due to different aspects including network throughput, performance, etc. [9]. With the passage of time as the user of the computing technology is dynamically growing, the threat to the computing technology has increased as well, which demands for security requirement. Increasing users of computing technology has also created the problem of finding the loopholes.

To develop a safe and secure routing protocol for an ad-hoc structure is a daunting task owing to its distinctive mesh properties such as missing of central control, quick host mobility, continuously change of topology, safe working environment, and accessibility to resources [10]. A MANET is a group of roving hosts that can create a neighborhood for themselves. It is possible that they leave an area open. Therefore, it is crucial to have a safe and reliable exchange of data within MANET [12].

A. Key Contributions

- The proposed system presents an efficient framework, which detects selective forwarding attacks and identifies malicious nodes in the network.
- Thorough experiments are performed for the verification and validation of the proposed solution. Malicious hosts are created randomly for testing in the simulation environment.
- The harmful hosts are positioned in the forwarding path of simulation environment. The control packet and control ACK is added to the ad hoc on-demand distance vector (AODV) routing protocol for the verification and validation of our proposed framework.

B. Paper Organization

Rest of the paper is organized as follows: Section II provides the prior work. Section III discusses the proposed solution, control packet/ control ACK, and different cases. Section IV gives evaluation of the proposed solution based on the experiments. Finally, Section V concludes the paper and highlights the future work.

II. Previous Routing Mechanisms

The number of information technology users are dynamically increasing. Consequenlty, the issue of the security is increasing. There are many types of protocols, which include SRP [5], SEAD [6], and SAODV [7]. These routing protocols have been addressing the security attacks and have presented a different solution to counter security attacks. Most of these protocols use different authentication techniques including public key management, asymmetric key management, and many other authentication techniques.

The secure key management scheme was given by [2], which was based on the threshold cryptography. The scheme worked efficiently when it had to deploy in large scattered areas. The mobiles nodes contact the servers. A refreshing scheme was used to counter the mobile node adversaries. Scheme in [8] was based on simple architecture where all nodes were considered as servers. Main advantage of the scheme was that it was highly efficient and secure while performing local communication. The scheme reduced security in case when mobile nodes were not physically protected.

Several security routing protocols and framework were introduced in the past years such as SEAD, which was modeled and established for wireless ad-hoc mesh for the distance vector routing. There is a number of researchers who have been extending these techniques on the simulation of different goals. In [1], the design and evaluation of efficient secure AODV routing protocols were mainly focused. A protected ad-hoc structure for routing stands upon frame of "destination-sequenced distance-vector" routing protocol. The work restricted CPU working capacity and protection towards the attacks for the service denial where attacker tries to force other nodes capture excess processing time or bandwidth of the network [3].

In [6] authors presented a complete survey on the solutions and challenges within the security perspective of the wireless network and discussed some improvements to overcome these challenges. The provided solution can be simulated on different techniques. To address a number of challenges as identified in the survey, the survey was simulated around key management techniques, routing protocols, and securing challenges involved in it.

The prior work presented the intrusion detection in order to resolve a number of issues, which have been presented in the survey. The research study in [3] presented a new approach inspired by trust model and clustering algorithm. The solution utilized the certificate authority (CA) for securing the information in the MANET. The solution utilized the self-organization security and PGP (trust model) to secure ad-hoc network. In the solution, all the cluster nodes hold the CA and are registered to the specific authority that authorities issue the certificates to the cluster nodes.

III. Proposed Solution

The solution proposed is divided into two phases; a) to find and b) localize selective forwarding attacks on MANET. In the first phase, selective forwarding attacks are detected. Since a mesh of random numbers is needed, it is generated for the nodes of source and destination. The AODV protocol route requests are flooded in entire MANET. Since a packet counter is required, it is maintained in each node so that a track of all packets sent is kept and received by the node. Second phase of the proposed solution provides a detection scheme, which is based on two different packets. The given method "the control packet and the control ACK" is utilized in the detection scheme.

A. Control Packet and Control ACK

The proposed solution contains the new control packet. This packet is added to the data packet in the protocol of AODV. Main goal is to add these packets to identify the selective forwarding attacks and malicious nodes that are presented in the network. The control packet and control ACK (acknowledge) packet contain the parameter including final hash, hash function, hash field, destination ID and source ID. The control packet contains the parameters given in Table I.

TABLE I. Control Packet and Control Ack

Control Packet & Control ACK				
Final hash	Hash function	Hash field	Destination ID	Source ID

Table I presents the structure of control packet and control ACK. The source ID filed contains the information about source node from where the packet is generated and the destination ID field contains the information about the destination node. Total number of hops is carried from source to destination by the hash. Hash function is used to compute hops between destination and source, which is stored in final hash. In the end, the final hash value is compared with the security value stored in hash field. The hash chain is implemented on each packet to make the communication secure. The control packet is sent with each data packet.

Algorithm 1 Retrieve hash field value algorithm

1: **If** (Final – Hash = F Hop Count (Hash)) **then**
2: Retrieve the packet count value in the Hash field of the control packet
3: **else**
4: Drop the control packet

Algorithm 1 is used by the final node, which accepts the control packet and accomplishes the operation as per steps given in Algorithm 1. There are three cases of control ACK; a) positive control ACK, b) negative control ACK and, c) no control ACK. Table II "Cases of Control ACK" presents the three cases of control ACK. In the first case, hop count is equal to the detection threshold and as a result positive control ACK is sent from destination to source node in order to send data packet for communication. This case has no attacker and thus communication occurs in a safely manner. In the second case, the hop count is less than the detection threshold and as a result destination sends negative control ACK to source host. The negative control ACK mentions the presence of the potential attacker in the forwarding path to the source host. While in the third case, there is no control ACK received by the source node from the destination, demonstrating the existence of attacker in the forwarding path that has dropped the control ACK or has dropped control packet; hence, destination did not receive any packet.

TABLE II. CASES OF CONTROL ACK

Case I	Case II	Case III
Positive Control ACK from Destination	Negative Control ACK from Destination	No Control ACK from Destination
Count (destination packet) = detection threshold	Count (destination packet) < detection threshold	⦿ Condition I Malicious host drops the control ACK, which is send by the destination. ⦿ Condition II Malicious host drops the control packet, as it doesn't reach destination, so destination does not reply with any control ACK.
In case I, the control ACK received by source host.	In case II, control ACK received by source host.	In case III, no control ACK received by the source node.
No presence of attacker	Presence of attacker	Presence of attacker

B. Detection of the Malicious Hosts

When it is detected that a malicious node is present, the source node begins querying all hosts, which are present in the forwarding path and then gets a value of the packet counter. A counter frequency is utilized to choose the intermediate nodes in the way between destination and source nodes. A suitable value of the counter frequency could be computed by using mesh topology experiments [4]. Table III presents two different cases of malicious node detection.

If the malicious node is noticed then the source node forwards the error packet in order to update all the hosts, which are existing in neighborhood about the harmful host. Specific node is dropped from the routing table and in future, during the route discovery, the malicious nodes are not considered again. The removal of harmful host from neighborhood enhances overall network performance of the network as the packets of the malicious nodes are no longer flooding in the network

TABLE III. CASES OF MALACIOUS NODE DETECTION

Case I	Case II
Malicious host drops all packets.	Malicious host drops few packets.
No control ACK received	Negative control ACK to source
Malicious node detected	Malicious node detected

IV. EVALUATION

A. Detection Threshold

In this subsection, analysis of the threshold is performed. Detection of the threshold is designed upon metric of expected transmission count (ETX). The ETX calculates total number of counts, which are needed for the successful delivery of a packet between a source and destination. Total number of the count also includes the retransmitted packets. Table IV describes the threshold parameters.

TABLE IV. THRESHOLD PARAMETER

Symbol	Parameter	Description
D_f	Forward delivery	Probability of the packet deliver successfully.
D_r	Reserve delivery	Probability of the ACK packet received successfully.

The ETX link is computed by using (1).

$$\mathrm{ETX} = \frac{1}{D_f - D_r} \quad (1)$$

B. Simulation

The AODV protocol is simulated on the MANET based in real-time environment. Ad-hoc network environment is deployed. The nodes detect the environment and rain, speed, average humidity, wind speed, temperature and find fire weather index (FWI) based on this parameter. Network throughput in kbps is used as an evaluation metric for the proposed solution. There are 35 nodes deployed in the forest fire simulation model. The AODV protocol is used for communication between these mobile nodes. The data packet of each node contains the information of current temperature (T), rain (R), wind speed (WS), and average humidly (H). Harmful hosts are randomly created in the environment. These harmful hosts are positioned in the forwarding path. The control packet and control ACK are added to the AODV routing protocol for verification and validation of the proposed model. The timeout is set to 2 units before the initialization of the lead-off node. The ETX metric is integrated into the AODV protocol.

C. Performance Results

Performance of the proposed solution was evaluated and analyzed based on the network throughput. The detailed discussion of the experiments performed and their corresponding results has been provided. Two different experiments have been simulated based on different scenarios and parameters. The harmful hosts are also added in the network with the context of the experiment so that the right performance and efficiency can be calculated by the help of the respond time of the simulation environment.

1) *Experiment no. 1:* The forwarding path between source node and destination node has a single attacker. Within this experiment, attacker is randomly deployed in the path of selective forward routing. In the simulation environment, two observations are made. First is the network throughput and the second is detection. Network throughput is decreased as there is a presence of malicious nodes. As the malicious node interrupts with the communication held between the nodes of the network, the throughput of the network drops, which can be seen in Fig. 1

In order to define and measures the results more accurate, the experiment was considered and executed for more than ten times and based on that simulation the following graph has been generated. During simulating environment, bit error rate is also considered while calculating throughput in the context of this research and errors have also been considered that occur in the transmission of digital data and based on the analysis, the overall performance is computed.

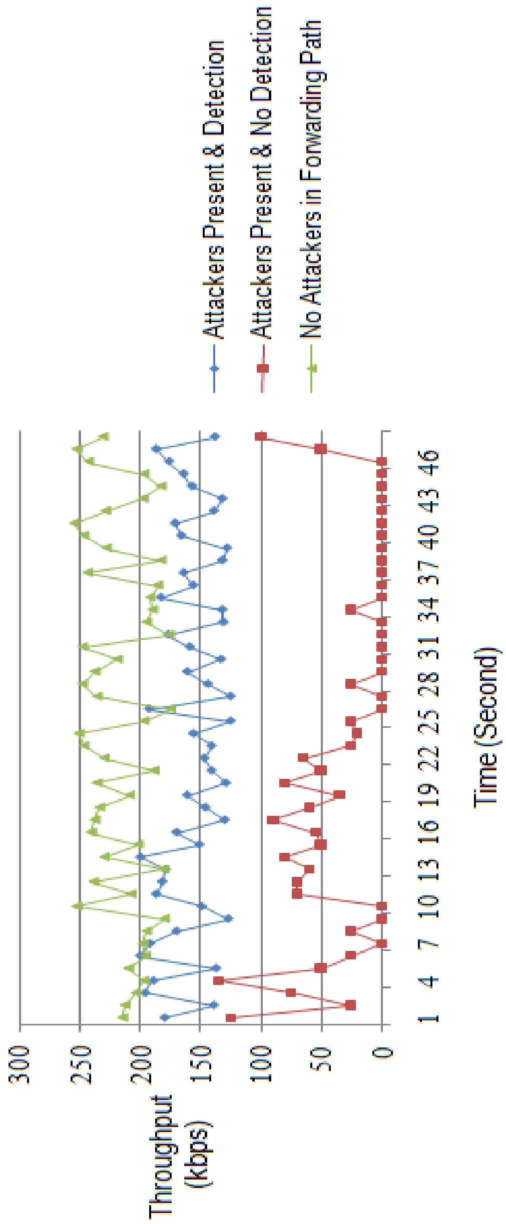


Fig. 1. Single attacher present in the forwarding path.

The time is considered on the X-axis and the network throughput is considered on the Y-axis. It can be observed from the above graph as the time passes the network throughput increases. The graph in Fig. 1 that is mentioned in the experiment is generated from Table V, which contains different values of “attacker present and detection”, “attacker present and no detection” and “no attacker in forwarding path”.

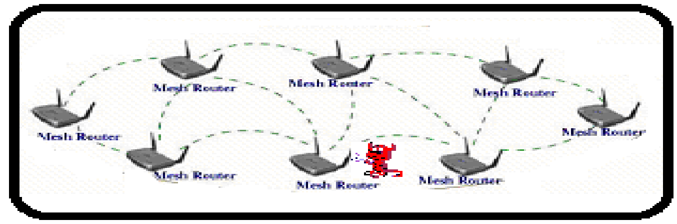


Fig. 2. Single malacious node in the forwarding path.

TABLE V. THROUGHPUT OF THE NETWORK IN CASE OF A SINGLE ATTACKER

Attackers present & detection (kbps)	Attackers present & no detection (kbps)	No attackers in forwarding path (kbps)
182	125	246
142	25	249
194	75	191
129	135	238
151	50	192
162	25	240
176	0	186
125	25	251
168	0	208
161	0	222
182	70	220
148	70	201
141	60	243
142	80	181
194	50	245
127	55	255
155	90	251
154	60	231
172	35	184
150	80	233
178	50	219
186	65	205
200	25	209
200	20	243
164	25	229
153	0	225
150	0	235
177	25	235
186	0	206
180	0	235
156	0	227
127	0	244
192	0	191
187	25	230
166	0	212
145	0	204
133	0	222
145	0	195
143	0	219
192	0	252
148	0	199
154	0	180
137	0	192
159	0	217
132	0	233
137	0	181
187	50	241
144	100	252

Fig. 1 has three different graphs, which are used to show the network performance based on the network throughput. The green graph shows different values of the throughput when there is no attacker present in the forwarding path; the throughput is the highest in this case. The red graph shows the values of throughput when network has the presence of an attacker but not detected; the throughput is the lowest in this case. While the blue graph shows values of the throughput when attacker was spotted and then removed from neighborhood, which results in better network throughput than the case when there is no detection of the attacker. The green and red graphs were obtained by simple AODV while blue graph was obtained by the implementation of the proposed solution. In Fig. 2, forwarding path has a single malicious node, which attempts selective forwarding attack.

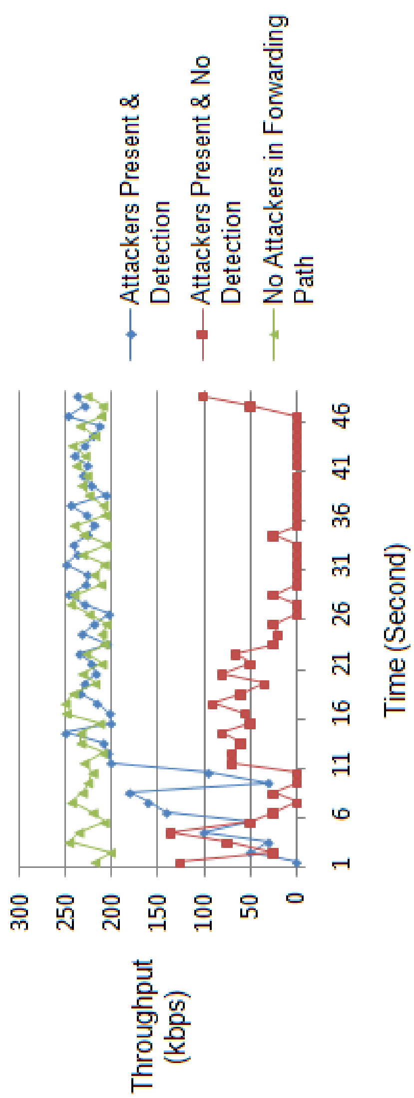


Fig. 3. Multiple forwarding paths between source node and destination node have colluding attackers.

2) *Experiment no. 2:* In mutiple forwarding paths between source and destination, colluding attackers are present. In this experiment, there were more than two attackers randomly.

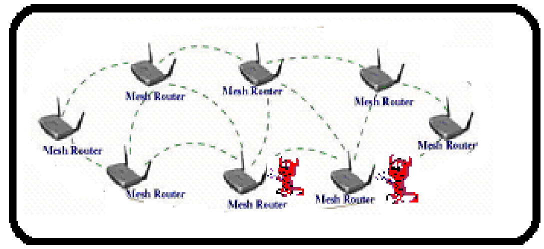


Fig. 1. Colluding malicious nodes in the multiple forwarding paths.

TABLE VI. THROUGHPUT OF THE NETWORK IN CASE OF TWO ATTACKERS

Attackers present & detection (kbps)	Attackers present & no detection (kbps)	No attackers in forwarding path (kbps)
0	125	242
50	25	238
30	75	209
100	135	232
50	50	205
140	25	224
160	0	210
180	25	230
30	0	243
95	0	239
200	70	243
237	70	238
202	60	235
224	80	212
242	50	227
248	55	223
206	90	204
219	60	210
231	35	207
228	80	210
204	50	226
204	65	203
204	25	222
207	20	240
244	25	247
218	0	234
214	0	209
238	25	232
248	0	236
212	0	212
240	0	219
234	0	233
225	0	241
221	25	244
213	0	238
234	0	232
204	0	239
207	0	231
250	0	214
205	0	250
203	0	238
210	0	241
247	0	244
221	0	230
211	0	212
240	0	244
230	50	218
241	100	249

Selected and placed in between multi path of selected forwarding routing. Network throughput is increased as the malicious nodes were detected. In order to obtain accurate results, the experiment was executed for more than ten times and based on that simulation; the graph in Fig. 3 is generated. Bit error rate and errors occur in the transmission have also been considered and based on analysis, the overall performance is calculated.

The graphs in Fig. 3 are generated from the data of Table VI, which contains different values of “attacker present and detection”, “attacker present and no detection” and “no attacker in forwarding path”. Fig. 3 shows that throughput of network increases after the specific time interval. The new route request is generated and the malicious nodes are removed from the routing table. Network performance is improved after specific time interval. In initial phase, performances was low but after detection of attackers within a time interval of 0 -7s new routes were identified in network and the malicious nodes were removed. Time (seconds) is considered on the X- axis and network throughput (kbps) is considered on the Y-axis. In Fig. 4, two malicious nodes are present in the multiple forwarding paths attempting the selective forwarding attack.

D. Network Overhead

The control packet send with each data packet increases the system’s algorithm complexity and processing time, which results in network overhead. Although the control packet design has a slight network overhead but after the detection and elimination of the malicious nodes, their packets are no longer floating in the network plus there is also no need to resend the dropped packet hence increasing the overall performance and reducing the overhead of the network.

TABLE VII. SUMMARIZED OBSERVATIONS

Experiment No.	Scenario	Observation
1.	When there is a single attacker in the forwarding path of the network	Network throughput is decreased as there is a presence of the malicious nodes.
2.	When there are more than one attackers present in multiple forwarding paths between destination and source	Performances of the network throughput increases after the specific time interval.

Table VII summarizes the observations made in the experiments performed for the evaluation of the proposed system.

V. CONCLUSIONS

We developed and simulated a framework for the detection of selective forwarding attacks using MANET technologies. A network environment is deployed and several experiments were performed for the verification and validation of the proposed solution. For the testing purpose, malicious nodes were randomly created in the environment. These malicious nodes were positioned in the forwarding path of the simulation environment. The control packet and control ACK are added to the AODV routing protocol during evaluation of the system. The proposed solution can be applied to the protocols such as DSR and DSDV. There are some parameters such as network overhead and threshold detection, which can be done in future work.

REFERENCES

[1] J. Jubin and D. Tornow, "The DARPA packet radio network protocols," Proceedings of the IEEE, 75(1), pp. 21–32, 1987.

[2] L. Zhou and Z. Haas, "Securing ad hoc networks," IEEE Network Magazine 13(6):, pp. 24–30, 1999.

[3] N. Schacham and J. Westcott, "Future directions in packet radio architectures and protocols ," Proceedings of the IEEE, 75(1), pp. 83–99, 1987.

[4] S. Capkun, L Buttyan, and J. Hubaux , "Self-organized public-key management for mobile ad hoc networks," IEEE Trans Mobile Computer 2(1), 2003.

[5] P Papadimitratos and Z Haa, "Secure routing for mobile ad hoc networks," Proceedings of the SCS communication networks and distributed systems modeling and simulation conference (CNDS 2002), 2002.

[6] D Johnson, A. Perrig, and Y. Hu, "SEAD: secure efficient distance vector routing in mobile wireless ad-hoc networks," Proceedings of the 4th IEEE workshop on mobile computing systems and applications (WMCSA’02), pp. 3–13, 2002.

[7] M. Zapata, "Secure ad hoc on-demand distance vector (SAODV)," Internet draft, draft-guerrero-manet-saodv-01.txt, 2002.

[8] H. Luo and S. Lu, "URSA: ubiquitous and robust access control for mobile ad-hoc networks," IEEE/ACM Trans Networking 12(6), pp. 1049–63, 2004.

[9] M. P. Shelke, A. Malhotra and P. Mahalle, “A packet priority intimation-based data transmission for congestion free traffic management in wireless sensor networks,” Computers & Electrical Engineering, vol. 64, pp. 248-261, 2017.

[10] V. K. Saurabh, R. Sharma, R. Itare and U. Singh, “Cluster-based technique for detection and prevention of black-hole attack in MANETs,” in International conference of Electronics, Communication and Aerospace Technology (ICECA), 2017.

[11] L. Baghel, P. Mishra, M. Samvatsar and U. Singh, “Detection of black hole attack in mobile ad hoc network using adaptive approach,” in Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of, 2017.

[12] D. Gayathri and S. J. Raman, “Pltrust AODV: Physical logical factor estimated trust embedded AODV for optimised routing in Manets,” in Advanced Computing and Communication Systems (ICACCS), 2017 4th International Conference on, 2017.

A New Project Risk Management Model based on Scrum Framework and Prince2 Methodology

Mahdi Mousaei
Department of Computer Engineering
Isfahan (Khorasgan) Branch, Islamic Azad University
Isfahan, Iran

Taghi Javdani Gandomani
Department of Computer Engineering
Boroujen Branch, Islamic Azad University
Boroujen, Iran

***Abstract*—With increasing competition in the software industry, software companies need to effectively manage the risks of software projects with minimal time and cost to deliver high quality products. High frequencies of warning errors and failures in software projects are indicative of human and financial costs in the software projects and teams. One of the reasons for the failure of software projects is the lack of risk management mechanism in the software development process, which can, in case of proper implementation of risk management, increase the success rate of such projects. In most projects, risk management activities are strongly confined to the adopted software methodology. Therefore, is needed a solution or model to overcome this constraint. Scrum is one of the most popular software development methodologies which has recently considered by software teams. This methodology seems not to have paid much attention to risk management. Focusing on this weakness, this research has been trying to provide a model for risk management with the participation of 52 Agile experts from six different countries using the Prince2 project management framework in Scrum methodology. The main goals of this model are to improve the coverage and appropriate risk management mechanism on software projects, increase the project's success rate and to provide a good estimation of the required time, improve product quality and enhance quality parameters, such as the usability, flexibility, efficiency, and reliability.**

Keywords*—Agile software development; risk management; risk management hybrid model; prince2; scrum; agile risk management

I. Introduction

American Project Management Institute (PMI) introduced risk management as one of the twelve principal levels of the general knowledge project [1]. Risk management refers to all processes to identify, analyze, and respond to any uncertainty that includes maximizing the results of desirable events and minimizing adverse events results [2]. Chawan et al. (2013) examined risk management models and the result of adverse events the basic steps such as risk identification, risk planning, risk assessment, risk mitigation and risk monitoring and control and concluded that different models and frameworks for risk management are tools for risk management and control in the critical conditions [3].

Traditional project management (waterfall approach) is suitable for the projects that are well defined areas and it has less complexity and uncertainty [4]. At present, more complex projects and business environments with unique needs and capabilities are changing [5]. Most customers cannot express all their needs clearly [6]. One of the challenges of the traditional models, it needs very time and cost to writing the documentation of the project. The heavyweight methodologies, nature of software development, lead to unrealistic estimates in the design phase and inability to adapt to unforeseen changes in projects. So it is needed an approach to identify and solve these challenges and risks [7].

A review of previous studies shows that there is no a comprehensive model for risk management in the agile software development process [8]-[11]. The evidence for this assertion, by reviewing the prior studies in the literature, is the enhancing the failures in software projects.

Results of proposed model and participation of 52 Agile experts from 6 different countries shows that the proposed model can reduce the risk of projects and increased the usability, flexibility, efficiency, and reliability.

The next section of this research is to investigate the importance of risk management and its challenges. Section III will be explaining the PRINCE2 method. Also, Section IV focuses on research method. In Section VI is explained the designing, analyzing, and results of the performance of the proposed model. Sections VI and VIII discuss the limitations and discussion of the study and Section IX explains the conclusion and future work of the research.

II. Literature Review

After Agile Manifesto was released, the discussion of agile project management was also raised [12]. There are many types of research on the differences between traditional and agile development methods and new approaches to project management are needed [13]-[17]. In agile methodologies, risk management has not been defined clearly. For example, Scrum framework has not formally described project risk management. Therefore, it is necessary, according to project requirements to be included processes for risk management in this framework [8].

Scrum has flaws and defects in project management and risk management in software projects [9]. The absence of effective risk management techniques in agile software development caused many challenges in the software production process [10]. In Scrum Framework, projects are broken into several sprints and can be performed in two to four weeks. However, for large projects, a large number of Sprints in Scrum cause difficult task management for Scrum development teams, which is one of the challenges and weaknesses in the scrum [11], [18], [19].

According to studies conducted the Scrum and Prince 2 methodologies are both process-oriented, which can overlap each other, so can be created a combine framework with them.

Nitin and Ugrasen (2015) introduced a framework called RBSM to risk management processes to improve the Scrum method in order to increase the success of projects. The purpose of this model is to create a general model to a quality and reliable productin in the organizations that use the agile methods [20].

Al-Zoabi (2008) introduced a framework with XP and Prince 2 standard that her aim was the flexible method to project management. The results indicated that using their proposed method in a real project provided high-quality software products at a lower cost and time [21].

III. Prince 2

Prince2 is a well-known standard in project management. Prince2 proposed in 1996 after PMBOK by APMG in the UK and has been expanded by the British state (Government). Prince2 is the result of managers, experts and consultants experience in the field of project management (OGC, 2009). Project management methodology success with Prince 2 that has created consisting of four integrated elements, principles, themes, processes and project environment [22]-[24].

Project management based on Prince2 has two parallel sections called themes and processes; each one divided into seven separate issues, and all moving to forward based on seven principles [25]. Prince2 is a process-based approach in project management. In Prince2 there are seven processes, including a set of activities required for project control, management and successful delivery of the projects [23], [26], [27]. The principles in Prince2 are definitive guidelines that show proximity and compliance in the project management based on Prince2. There are seven principles in Prince2 [23], [27]. Themes, describing the components of project management must be used continuously and in parallel in the project [25], [27]. The Prince2 framework is flexible and can be tailored to any size and type [23], [28]. The proportionating element (of project environment) can be used as a framework for combining it with Scrum framework and to create an integrated model for risk management in agile software development. Prince2 describes how to divide a project into manageable steps.

IV. Research Methodology

The present study is an applied empirical research in the field of software engineering. Fig. 1 shows the steps and processes adopted in the current research [29].

A. Data Collection

The questioner is one of the research tools in which the researcher designs a set of items (question) aiming to collect information on the respondents and statistically analyze the responses. In this research, for measurement of the attitudes, Likert scales were used [30]. In this research the questionnaire items were designed in 5- and 7- option Spectrum.

B. Validity

Validity refers to the ability of the instrument to measure the attribute for the measurement of the test has been designed. In this research for validity measurement, a standard questionnaire was used. After the questionnaire was designed, it was given to a number of researchers and experts in the field of software engineering. After receiving the comments, corrective actions were made.

C. Reliability

Reliability refers to the accuracy of reliance and stability of test results. Cronbach's alpha was used to measure reliability in the present study. After the data collection process, reliability factor (Cronbach's alpha) was calculated using SPSS software. The result was 0.90 indicating the stability and internal consistency of the questionnaire.

The Cronbach's alpha method is one of the most common methods to measure the reliability and validity of the questionnaire. This factor (coefficient) is used to get the respondents' impression of the items (questions). Evidently, the nearer Cronbach's alpha index to 1, the more internal correlation between the items and, as a result, the more It is clear any amount Cronbach's alpha is the index to 1 closer, the internal correlation between questions are more homogeneous. Cronbach's alpha values between 5.0 to 7.0 are average and acceptable, lower than 5.0 lacks reliability, more than 7.0 is good reliability, and higher than 9.0 is considered too high [31].

D. Selection of Research Target Population

The statistical universe and participants in this research include software development teams, senior and junior managers, executive managers, consultants, developers, Scrum Masters, customers, software companies and all stakeholders involved in projects. It should be noted that data were collected by sharing the questionnaire[1] in professional online groups, and sites like Facebook, Google +, LinkedIn, Researchgate, Google group, Yahoo and social networks such as WhatsApp, Telegram, etc. as well as sending the questionnaire via email to software engineering expert. Finally, the questionnaire was responded by 52 people in six different countries, in professional groups of software engineering. In analyzing data from the questionnaire the characteristics of the respondents to the questionnaire, such as work experience in traditional and Agile software development, role or job, methods used in the organization, country of activity, the scope of the project by the organization, the number of employees and scale of enterprise have been shown in Table I.

[1] Link to the questionnaire:
https://docs.google.com/forms/d/e/1FAIpQLScWC0RCuUUlAKVQZUnEkcNRf4oUdrgRqGJDyRqnoDayopRpCg/formResponse

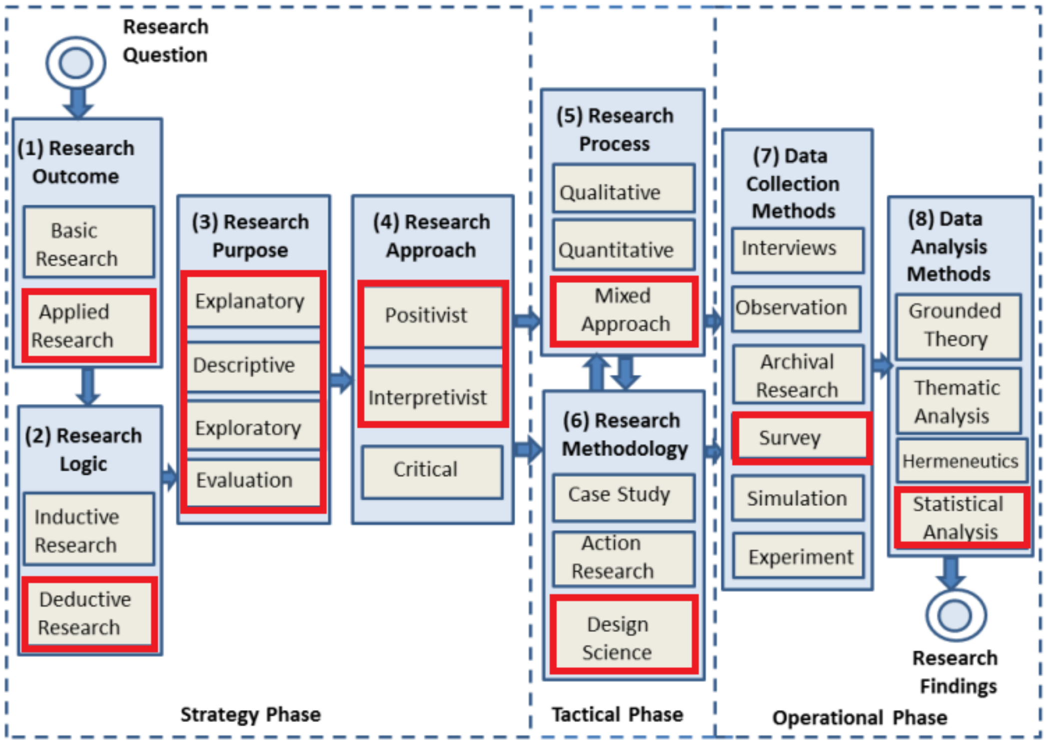


Fig. 1. Steps design in the current research, adopted from [29].

TABLE I. FREQUENCY DISTRIBUTION RELATED TO CHARACTERISTICS OF THE RESPONDENTS

Demography of the participants (Agile positions: Project Manager(PM), Scrum Master(SM), Product Owner(PO), Programmer(PRG), Developer (DEV), Agile Coach(AC), Customer(CUS), Advisor(ADV), Tester(TST))

Experts No.	Work experience (years)	Work experience in agile (years)	Role (job) in the project	Agile methods	Country of activity	Field project	Number of Employees
1	1-3	4-6	PM	SCRUM	USA	Accounting and Finance	55
2	7-10	6-10	SM , PO	SCRUM, DSDM	Iran	Office Automation	31
3	7-10	1-3	PM, DEV	SCRUM	Iran	Office Automation	30
4	4-6	1-3	PRG, DEV	SCRUM, XP	Iran	Applications	26
5	1-3	1-3	AC, DEV	SCRUM, XP , DSDM	Iran	All Areas	12
6	1-3	1-3	other	Others	Iran	Other	6
7	7-10	1-3	PRG, DEV	SCRUM	Iran	Applications Web Based	23
8	7-10	1-3	PRG	SCRUM	Iran	Applications Web Based	8
9	4-6	1-3	PRG	SCRUM, XP	Iran	Applications	21
10	1-3	1-3	CUS	Others	Iran	All Areas	28
11	4-6	4-6	SM	SCRUM, XP,DSDM	Malaysia	Mobile Application	25
12	4-6	4-6	SM , PRG, TST, DEV	SCRUM, XP,DSDM	Australia	All Areas	22
13	4-6	4-6	SM , PO, PM, PRG	SCRUM, XP	Australia	All Areas	20
14	4-6	1-3	PM, ADV, PRG	SCRUM, XP,DSDM	Iran	Applications Web Based	15
15	4-6	1-3	PRG	SCRUM, XP	Iran	All Areas	15
16	1-3	1-3	PRG	SCRUM	Iran	Applications	12
17	4-6	1-3	PRG, CUS	SCRUM, XP, Others	Turkey	Applications	22
18	4-6	4-6	SM , PO, PRG, TST	SCRUM, DSDM	Turkey	Accounting and Finance	40
19	4-6	4-6	SM , PO, PRG, TST	XP , SCRUM	Iran	Applications	22
20	4-6	4-6	SM , PO, AC	SCRUM, XP , DSDM	Turkey	Applications Web Based	20
21	4-6	1-3	PRGوSM, PM, ADV	SCRUM, XP , Others	Turkey	All Areas	12
22	1-3	1-3	SM , PO	SCRUM, DSDM	Turkey	Applications Web Based	13

23	4-6	4-6	PM	XP	Iran	Applications	22
24	1-3	1-3	PRG	SCRUM, XP	Iran	Accounting and Finance	8
25	1-3	1-3	SM	SCRUM	Iran	Applications	20
26	1-3	1-3	PRG	SCRUM	Turkey	Applications	14
27	7-10	6-10	SM	SCRUM	Malaysia	Applications	40
28	1-3	1-3	SM	SCRUM	Iran	All Areas	8
29	4-6	4-6	SM	SCRUM	Turkey	Office Automation	10
30	4-6	4-6	SM , PO, PRG	SCRUM, XP , TDD	Malaysia	Office Automation	20
31	7-10	6-10	SM , PRG, DEV	SCRUM, XP	Iran	Applications	80
32	1-3	1-3	SM	SCRUM	Iran	Office Automation	11
33	4-6	4-6	SM	SCRUM, XP	Iran	Applications	10
34	4-6	1-3	PM, PRG, TST	XP	Malaysia	Office Automation	22
35	4-6	4-6	SM , PRG, TST	SCRUM	Iran	Applications Web Based	9
36	1-3	1-3	SM , PRG, TST, CUS	SCRUM, XP	Iran	Office Automation	7
37	>10	>10	PM, PRG	SCRUM, Others	Australia	Applications Web Based	18
38	7-10	6-10	PRG, TST	SCRUM, XP	Iran	Applications	12
39	4-6	4-6	PRG	SCRUM	Iran	Applications Web Based	9
40	1-3	1-3	PRG	SCRUM	Iran	Applications Web Based	7
41	1-3	1-3	PRG	SCRUM	Iran	Applications	20
42	4-6	4-6	SM	SCRUM	Iran	Applications Web Based	8
43	4-6	4-6	SM , PRG	SCRUM	Iran	Applications	9
44	7-10	1-3	SM	SCRUM	Iran	Applications	18
45	1-3	1-3	SM	SCRUM	Iran	Applications	9
46	7-10	1-3	SM	SCRUM	Iran	Applications Web Based	8
47	>10	6-10	SM	SCRUM	Turkey	Applications Web Based	12
48	7-10	1-3	PRG	SCRUM, Others	Qatar	Applications Web Based	20
49	>10	>10	PRG	SCRUM, Others	Iran	Applications	12
50	7-10	1-3	SM	SCRUM	Qatar	Applications	15
51	4-6	1-3	Other	TDD	Iran	Office Automation	21
52	1-3	1-3	CUS	TDD	Iran	Mobile Application	6

V. Proposed Model

According to product quick release in an iterative and incremental mode in Scrum, the Scrum methodology can be used in product delivery [32]. In this idea, with using prince2 management features and scrum iterative delivery features can be introduced a combined model.

With this idea, processes are run in different layers. By applying the features of different parts of Prince2 and dimensions of the Scrum framework with a focus on risk management can be provided an integrated model to identify or improve risk management. The integrated model Prince2 project management processes and Scrum process with the risk management approach in agile projects can increase the success rate of software agile projects [33]. The proposed model has been shown in Fig. 2.

First, in the ***project charter,*** defined the obligations of individuals relative to the project can be defined. One weakness of Scrum is a severe attachment to team members leaving may result in the project stoppage and failure. These challenges of Scrum can be resolved by the project charter. The ***project manager*** begins the project with estimates of the needs and costs. ***Product Owner and Project Manager*** cooperate in the assessment and transparency requirements. In the process of directing the project, decisions are made to start the project and licenses are issued. Projects can be started when the project manager designs details and stages. After that, the project enters the initiation phase. One of the weaknesses of Scrum is a failure to identify area involved in the project. However, in the preparation phase (project initiation), the project scope can be partly identified. In determining the scope of the project, the risks can be better identified. The project manager must link process with Sprints. The product owner should also express their basic needs in the format of "Product Backlog". Scrum master and project manager design the sprints and facilitate the implementation of processes. Product owner and development team together, exchange ideas about "product Backlog" and the requirements that must be delivered in the next sprint.

In the ***boundary management process***, risks are recorded and reported and by the process of directing project, the next stage of the project is done step by step. Process and "product Backlog" are prioritized by the product owner and planned on Sprint by Scrum Master and processes in the format of "Product Backlog" delivered to the development team and the development team will release the product. After the release, review and retrospective meetings are held, and if recompletion is required, the sprints are planned and developed. After the sprint review meeting, the project manager should prepare a special report on the project progress. Product Owner and Project Manager by using the burndown charts and weekly meetings evaluate the sprints and identify the project risks. In the proposed model, risks are recorded and updated in the initiation step, process control, and boundary management steps. In the end stage, reports are presented and these processes are performed iteratively and incrementally.

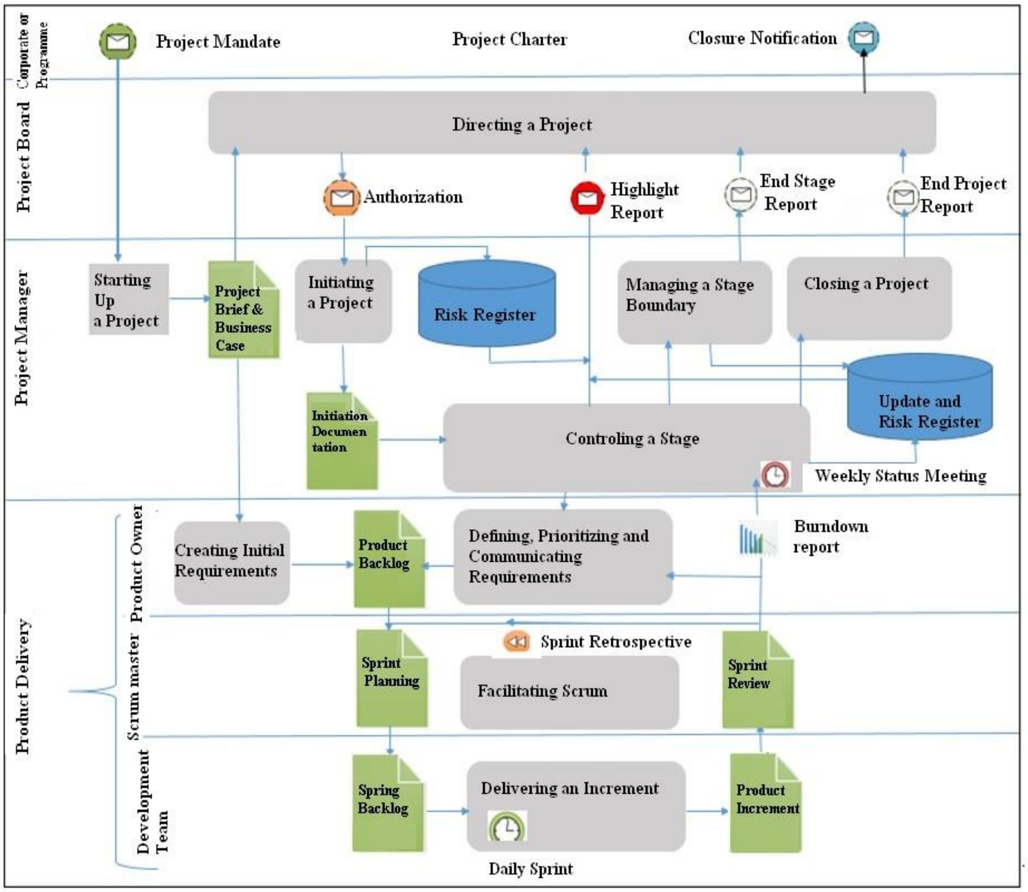


Fig. 2. The proposed model of risk management based on Scrum with using Prince2.

VI. Evaluation of the Proposed Model

A. Model Usability

Based on results in Table II, it is concluded that according to the Pearson correlation coefficient, usability between the model intelligibility and easy learning and the ability to implement the model gets the value of 0.698. Also, a decision criterion (Sig.) was close to 0.000, which is lower than 0.05. So there are significant correlations on the usability of the model.

TABLE II. Result of Pearson Correlation Coefficient between Model Intelligibility, Easy Learning of the Model, the Ability to Implement, and Execute the Model

	Ability to implement and execute the model	**Model intelligibility and easy learning model**
Model intelligibility and easy learning model Pearson correlation Sig. (2-tailed) N	0.698 0.000 52	1 52
Ability to implement and execute the model Pearson correlation Sig. (2-tailed) N	1 52	0.698 0.000 52

B. Flexibility of Model

The results in Table III show that the flexibility of the model, the ability of different process adoption, and interaction of the model gets the value of 0.864. Also, a decision criterion (Sig.) was close to 0.000, which is lower than 0.05. So there are significant correlations on the flexibility of the model.

TABLE III. Result of Different Processes, Adaption and Interaction of the Model

	Ability adapt and interact different processes	**The performance of the model**
Performance of the model Pearson correlation Sig. (2-tailed) N	0.864 0.000 52	1 52
Ability of adaption and interaction of different processes Pearson correlation Sig. (2-tailed) N	1 52	0.864 0.000 52

C. Model Performance

The results in Table IV show that according to Pearson's correlation coefficient between the ability of the product, quick delivery and timely delivery with the ability of optimal use of all the project resource gets the value of 0.434. Also, decision criterion (Sig.) is 0.001, which is lower than 0.05 and

Pearson correlation coefficient between the ability of the product, quick delivery and timely delivery with the ability to create a commercially valuable product gets the value of 0.309. Also, decision criterion (Sig.) is 0.026, which is lower than 0.05 and Pearson's correlation coefficient between the ability to optimal use all resources of the project with the ability to create a commercially valuable product gets the value of 0.596. Also, decision criterion (Sig.) was close to 0.000, which is lower than 0.05 Therefore, between these three factors, there are significant correlations effective on the model performance.

TABLE IV. RESULT OF PEARSON CORRELATION COEFFICIENT BETWEEN THE ABILITY OF THE PRODUCTS QUICK DELIVERY, AND TIMELY DELIVERY, ABILITY OF OPTIMAL USE OF ALL THE PROJECT RESOURCES AND ABILITY TO CREATE A COMMERCIALLY VALUABLE PRODUCT

	Ability to create a commercially valuable product	**Ability to optimal use all resources of the project**	**Ability to quick delivery and timely delivery of products**
Ability to quick delivery and timely delivery of products Pearson correlation Sig. (2-tailed) N	0.309 0.026 52	0.434 0.001 52	1 52
Ability to optimal use all resources of the project Pearson correlation Sig. (2-tailed) N	0.596 0.000 52	1 52	0.434 0.001 52
Ability to create a commercially valuable product Pearson correlation Sig. (2-tailed) N	1 52	0.596 0.000 52	0.309 0.026 52

D. *Reliability of the Model*

Based on the results in Table V, it is concluded that the reliability of the model according to the Pearson correlation coefficient between satisfaction with a product created in the software life cycle model with confidence in the risk identification, analysis, and control gets the value of 0.323. Also, a decision criterion (Sig.) was close to 0.000, which is lower than 0.05. So، there are significant correlations on the reliability of the model.

TABLE V. RESULT OF PEARSON CORRELATION COEFFICIENT BETWEEN SATISFACTION WITH PRODUCTS CREATED IN THE SOFTWARE LIFECYCLE WITH CONFIDENCE IN THE RISKS IDENTIFICATION, ANALYSIS, AND CONTROL

	Confidence in the identification, analysis, and control of the risks	Satisfaction with products created in the software lifecycle
Satisfaction with products created in the software lifecycle Pearson correlation Sig. (2-tailed) N	0.323 0.019 52	1 52
Confidence in the identification, analysis, and control Pearson correlation Sig. (2-tailed) N	1 52	0.323 0.019 52

VII. DISCUSSION

After reviewing and evaluating the model parameters some of the most important qualitative results are shown in Fig. 3.

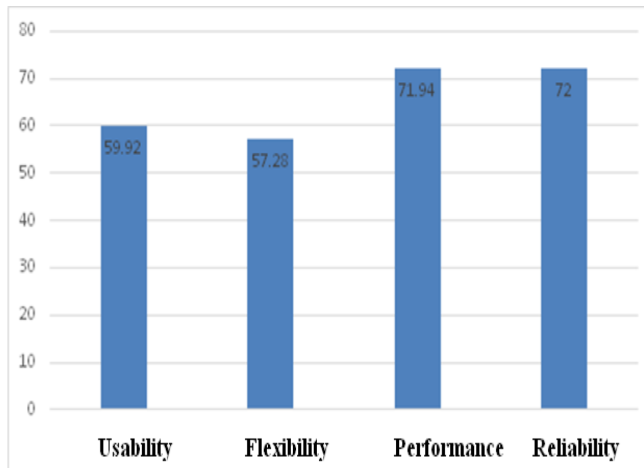


Fig. 3. The most important results of the qualitative parameters of the proposed model.

The results showed that the proposed model can have usability in the projects at 59.92 percent. If teams and companies have more information on the model to increase the model intelligibility and learnability and ability to implement and execute it, usability will also increase. Results showed that the proposed model can be flexible at 57.28 percent. By increasing, the correctness of performance and different process compatibility, the flexibility of the model also will be increased.

The results showed that the proposed model has a 71.94 percent performance. With increase the ability on the timely and quick delivery of product, and the most optimal use of all the project resources, and increase the ability to establish a more commercially valuable product, the rate of efficiency as well.

The results showed that the proposed model has 72 per cent reliability. With increase satisfaction of products created in the software life cycle and confidence for identification, analysis, and control of the risks by the proposed model, the reliability will be increased.

VIII. Research Limitations

One of the limitations of this study is the lack of similar work in this field. Also, for evaluation of the proposed model through questionnaire constantly had faced the lack of the cooperation and follow up. Many items of the questionnaire may prolong the research conduction time, some respondents may have provided false responses and responses precision has been influenced. Another limitation of the research is the lack of expertise and lack of their knowledge in the field of risk management in Prince2 and Scrum. However, this study does not claim that the results could the global model for all the teams because team's organizational cultures are different.

IX. Conclusion and Future Work

According to the studies in the field of risk management and Agile Scrum methodology in future research can be combined with project management standards. Such as PMBOK, P2M, OPM3, PRINCE2 and Agile methods such as Scrum, DSDM, XP, ASD that have more focus on project management processes to create hybrid models to identify and reduce the risks of software projects. Also, using project management standards and agile methods can be created the models and ways to improve project management and assurance quality in agile software development.

In this study, using risk management techniques with Scrum framework and project management standard Prince2, suggested a model to increase in the success rate projects. There are some of scrum weak points causing risk increase in software development. The proposed model covered the weaknesses and this covering reduces the risk of the project failures. Some of the important results include: covering of risk management at 67.4%, increasing project success about 75.4%, the ability to provide a quality product at 75%, and the reliability to identify, analyze, and control the risks at 85.71%.

References

[1] Project Management Institute (PMI) .Available from: http:// www.pmi.org [Accessed 6 February 2007].

[2] Santos, Flávio Roberto Souza dos, and Sandro Cabral. "FMEA and PMBOK applied to project risk management." *JISTEM-Journal of Information Systems and Technology Management* 5.2 (2008): 347-364.

[3] Chawan, P. M., Jijnasa Patil, and Radhika Naik. "Software risk management." *International Journal of Computer Science and Mobile Computing* 2.5 (2013): 60-66.

[4] Chaudhari, S., Chouksey, P., & Lonkar, P. traditional approach to agile approach in software development. 13th International Conference on Recent Innovations in Science, Engineering and Management, pp. 212–218, February 2018.

[5] Standish Group. 2016, Available from http://www. standishgroup. com /news/archive. [Accessed 26 May 2016].

[6] Cadle. J, and Yeates. D, Project Management for Information Systems (5th edition), England: Pearson Education Ltd,(2008).

[7] Kaur, N., & Singh, G. Critical Success Factors in Agile Software Development Projects: A Review. International Journal on Emerging Technologies, 7(1), 1.(2016)

[8] Hansenne, R., & Hibner, A.Overcoming Organisational Challenges related to Agile Project Management Adoption, Master's Thesis in Business Administration, MBA programme,page 73, (2011)

[9] Mahalakshmi, M., & Sundararajan, M. Traditional SDLC Vs Scrum Methodology–A Comparative Study. International Journal of Emerging Technology and Advanced Engineering, 3(6), 192-196, (2013).

[10] Cerpa, N., & Verner, J. M. Why did your project fail?. Communications of the ACM, 52(12), 130-134, (2009).

[11] Zheng, G. Implementing a business process management system applying Agile development methodology: A real-world case study (Doctoral dissertation, Erasmus Universiteit Rotterdam), (2012).

[12] Beck. K, Beedle. M, Van Bennekum. A, Cockburn. A, Cunningham. W, Fowler. M, ... & Thomas. D, Manifesto for agile software development, (2001).

[13] Augustine, S., Payne, B., Sencindiver, F., & Woodcock, S, Agile project management: steering from the edges. Communications of the ACM, 48(12), 85-89 (2005).

[14] Boehm, B., & Turner, R. Management challenges to implementing agile processes in traditional development organizations. IEEE software, 22(5), 30-39,(2005)

[15] Ceschi, M., Sillitti, A., Succi, G., & De Panfilis, S. Project management in plan-based and agile companies. IEEE software, 22(3), 21-27, (2005)

[16] Salameh, Hanadi. "What, When, Why, and How? A Comparison between Agile Project Management and Traditional Project Management Methods." *International Journal of Business and Management Review* (2014).

[17] Kaur, Attinder, and Kailash Bahl."Analysis of Agile Project Management practice and Comparison with Traditional Project Management.", International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 5, (2015)

[18] Chang, H. F., & Lu, S. C. Toward the Integration of Traditional and Agile Approaches. arXiv preprint arXiv:1303.1238, (2013).

[19] El-Licy, F. A. Paired Scrum for Large Projects. Egyptian Computer Science Journal (ISSN-1110-2586), 40(1), (2016).

[20] Uikey, Nitin, and Ugrasen Suman."Risk Based Scrum Method: A Conceptual Framework." Proceedings of the 9th INDIACom; INDIACom-2015; IEEE Conference ID: 35071, 2015 2nd International Conference on "Computing for Sustainable Global Development", (March, 2015).

[21] Al-Zoabi, Z. Introducing discipline to XP: Applying PRINCE2 on XP projects. In Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on (pp. 1-7). IEEE

[22] Saad, S., Ibrahim, A., Asma, O., Khan, M. S., & Abdul, A. PRINCE2 Methodology: An Innovative Technique of Project Management growing progressively across the globe, Proceedings of 3rd International Conference on Business Management (ISBN: 978-969-9368-07-3),(2012).

[23] Kruger, W., & Rudman, R. Strategic alignment of application software packages and business processes using PRINCE2. The International Business & Economics Research Journal (Online), 12(10), 1239, (2013).

[24] Al Matari, A. PRINCE2 and PMBoK: Towards a Hybrid Methodology for Managing Virtual Projects Master–Thesis, (2014).

[25] Turley, Frank.2016. Available from: http:// http://prince2.wiki/PRINCE2 [Accessed 6 May 2016].

[26] Maarleveld, S., & Silvius, A. G. Adapting PRINCE2® to a Multi Vendor Outsourcing context. International Journal of Computer Science and Management Research, (2), 115-122, (2012)

[27] Hughes, R. T. Project management process ontologies: a proof of concept. In Proceedings of the 15th Annual Conference of UK Academy for Information Systems, Oxford (pp. 23-24), (2010)

[28] S. de Koning. Complementing agile software project management methods with plan-driven components. MSc Thesis Software Project Management, Tilburg University.53, (2011)

[29] Wohlin, C., & Aurum, A. Towards a decision-making structure for selecting a research design in empirical software engineering. Empirical Software Engineering, 20(6), 1427-1455, (2015)

[30] Likert, R. A technique for the measurement of attitudes. Archives of Psychology. New York: Columbia University Press, (1931).

[31] Cronbach, L. J. Coefficient alpha and the internal structure of tests. psychometrika, 16(3), 297-334, (1951)

[32] Pieroni, alessandra, noemi scarpato, and matteo scorza. "affective agile design a proposal for a new software development model." *Journal of Theoretical & Applied Information Technology* 96.1 (2018).

[33] Aquino Jr, Plinio Thomaz. "Small Brazilian Business and IT Governance: Viability and Case Study." Information Technology for Management. Ongoing Research and Development: 15th Conference, AITM 2017, and 12th Conference, ISM 2017, Held as Part of FedCSIS, Prague, Czech Republic, September 3-6, 2017, Extended Selected Papers. Vol. 311. Springer, 2018.

Load Balancing based on Bee Colony Algorithm with Partitioning of Public Clouds

Pouneh Ehsanimoghadam
Department of Computer,
Germi Branch, Islamic Azad University
Germi, Iran

Mehdi Effatparvar*
Department of Computer
Ardabil Branch, Islamic Azad University
Ardabil, Iran

***Abstract*—Cloud computing is an emerging trend in the IT industry that provides new opportunities to control costs associated with the creation and maintenance of applications. Of prevalent issues in cloud computing, load balancing is a primary one as it has a significant impact on efficiency and plays a leading role in improved management. In this paper, by using a heuristic search technique called the bee colony algorithm, tasks are balanced on a virtual machine such that their waiting time in the queue is minimized. In the proposed model, the cloud is partitioned into several sectors with many nodes as resources of distributed computing. Furthermore, the indices of speed and cost are considered to prioritize virtual machines. The results of a simulation show that the proposed model outperforms prevalent algorithms as it balances the prioritization of tasks on the virtual machine as well as the entire cloud system and minimizes the waiting times of tasks in the queue. It also reduces the completion time of tasks in comparison with the HBB-LB, WRR, and FCFS algorithms.**

***Keywords*—Cloud computing; load balancing; bee colony algorithm; public cloud; cloud partitioning**

I. Introduction

Cloud computing provides ways of presenting IT services in a similar manner to public utility companies. In simple terms, cloud computing is a new approach to use computing resources. The cloud is a group of distributed nodes that supply resources, hardware, and software over the network based on user demand. The increasingly strong presence of such companies as Microsoft, Google, and Amazon in the arena of cloud computing indicates its rapid development and influence on IT.

New changes or concepts in the technological world can lead to problems and complications, and cloud computing is no exception. It poses many challenges to experts in the field. One of these is load balancing in the cloud. Load balancing in is crucial in computer science, and has attracted a considerable amount of research. Many techniques have been employed for load balancing, such as the genetic algorithm, bees' algorithm, neural networks, and distributed research. Load balancing algorithms make decisions about allocating resources to tasks and coordinating among them. The aim of load balancing is to share resources among tasks within the system such that for every resource, there are an equal number of tasks to be completed, and this minimizes the total time needed [3]. Algorithms for load balancing and resource management can be categorized into three groups (Wu, Wang & Xie, 2013 [7]; Yan, Wang, Chang & Lin, 2007 [9]). The first consists of algorithms for static load balancing. In such algorithms, decisions concerning load balancing are made at compile time. The advantage of static load balancing is its simple implementation and low overhead, as there is no need to permanently monitor nodes to assess the efficiency of the system. These algorithms work well when there are small changes in loads in virtual machines. Therefore, they are not appropriate for cloud and grid computing environments because the load on the network is variable at every point of time in such environments. The second category of load balancing and resource management techniques consists of algorithms of dynamic load balancing. In these algorithms, the distribution of load among nodes changes, and they use the given information to make decisions about load distribution. The third category consists of hybrid algorithms, which involve the hybrid use of static and dynamic algorithms, and switch between them when necessary.

There are many papers have been proposed based on optimization in them [10]-[19] and [24]-[26]. These papers tried to optimize their problems by presented some formal methods and fitness functions in them. In [18] authors presented new distributed method to reducing the energy for the communication between nods and coordinator. Also in [24], [25] and [26] authors used the optimization methods in Meso-Scaled material. Saffar Ardabili and Aghayi [20] evaluated efficiency score of decision making units (DMUs) by the undesirable outputs. Aghayi et al. [22] measured efficiency measure using common set of weights in present of uncertainty based on robust optimization. Aghayi [23] proposed the approach to obtain cost efficiency of DMUs by fuzzy data. Aghayi and Maleki [21] measured the efficiency of bank branches of Ardabil, Iran using robust optimization theory and undesirable outputs. Rostamy-malkhalifeh, and Aghayi [27] suggested the method for calculating overall profit efficiency using uncertainty as fuzzy in data. Aghayi and Ghelejbeigi [28] presented the improvement of cost efficiency based on resource allocation. Aghayi [29] computed revenue efficiency of DMUs with undesirable and fuzzy data. Salehpour and Aghayi [30] calculated the most revenue efficiency with price uncertainty.

In this study, a method for cloud partitioning is proposed that is also used to investigate the load on systems in heterogeneous environments, with the aim of reducing the time needed for scheduling and other tasks. The proposed method is dynamic. In this method, the algorithms of load balancing can represent each resource according to its capabilities and accessibility to tasks services, which enhances the efficiency of the cloud system. To balance the load, load balancing as used in cloud computing environments is used, inspired by the HBB-LB algorithm (Babu & Krishna, 2013 [1]).

The rest of this paper is organized as follows: in Section 2, prevalent scheduling algorithms are introduced. Section 3 describes the proposed scheduling algorithm, Section 4 details the simulation to test it and the results and Section 5 contains the conclusions of this study and recommendations for future work in the area.

II. REVIEW OF LOAD BALANCING ALGORITHMS

One method for load balancing involves the efficient use of virtual machines. It was proposed by Domanal and Reddy (2013) [2], and allows for load distribution on accessible virtual machines to guarantee that stable use of resources or virtual machines in the cloud system. This is in contrast to active load balancing, which involves the proper distribution of load within the system to solve the problem of inefficient use of virtual machines in other algorithms. However, the service time following load balancing has not been yet studied. In 2013, Babu and Krishna [1] proposed a load balancing algorithm inspired by the food-finding behavior of honey bees, and is used on the Web. The aim is to reach a balanced load within virtual machines by maximizing capability. Moreover, it balances the prioritization of tasks in virtual machines such that the waiting times of tasks in the queue are minimized. However, this algorithm is impractical for dependent tasks. Ren, Lan, and Yin (2012) [5] proposed a dynamic load balancing algorithm according to the migration of virtual machines within the cloud computing environment. This algorithm contains a unit to monitor excessive loads, one for diagnosis, and a unit for load scheduling. The unit of load monitoring is used to collect the load information pertaining to a group of virtual machines and the resources' server (calculating the load and updating it). The database information for this algorithm is collected according to the trigger strategy based on fractal methods. It determines the time of migration from an overloaded virtual machine in the system. In this method, operating capability is maximized by using unemployed nodes in the system. Moreover, the overload in load balancing systems is minimum. However, in this method, only the load is studied. TeraScaler ELB was proposed by Wu et al. (2013) [7] based on the prediction of elastic load balancing for resource management in cloud computing. In this algorithm, virtual machines are added or removed according to the analysis and prediction of the given load and its history. The algorithm of ELB resource management is regularly implemented through two events:

- The load balancer regularly collects resource information from the back-end server.
- The load balancer determines whether there is a request to remove the back-end virtual machine according to the collected resource information from the back-end server.

In 2012, Nishant and et al. [4] proposed an algorithm for the load distribution of workloads among the nodes of a cloud using ant colony optimization. According to this study, ants can move in two directions: forward and backward. In the forward direction, if an ant faces overloaded nodes, they will move forward. In the backward direction, if an ant faces an overloaded node, which has faced an under-loaded node, it will move backward. The main duty of ants is to redistribute tasks. In this approach, the ants repeatedly update their pheromones during all their moves. They also identify the tasks of nodes and find their way among different types of nodes. In 2013, Xu, Pang, and Fu [8] proposed a load balancing model based on cloud partitioning for public clouds in different geographical locations. This method renders load balancing easier in extremely large and complicated environments. Clouds have a main controller to choose the proper sectors of tasks, and the balancer chooses the best strategy for load balancing per sector of the cloud. Sectors can be unemployed, normal, or overloaded. The sector of load balancing decides how to assign tasks to nodes of normal or unemployed sectors. In this algorithm, the features of virtual machines are not considered. Soni and Kalra (2014) [6] proposed a central load balancer to balance loads among virtual machines in a cloud data center. This algorithm distributes the load among heterogeneous virtual machines based on hardware configuration and their states in the cloud data center. This method can balance the load quickly and reliably in cloud computing environments by using all virtual machines based on their calculation capacities. The central load balancer communicates with all users and virtual machines, which are presented in cloud data center through a data center controller, which also analyzes the values' table containing the identifies, states, and priorities of virtual machines. It searches for the virtual machine with the highest priority to allocate user requests. The data center controller allocates the requests to the identifier of the virtual machine as presented by the central load balancer.

By studying research on the load balancing of tasks and resources, it can be concluded that more research has been conducted on load balancing in heterogeneous environments. Existing algorithms have some drawbacks in the cloud sample, and this study attempts to alleviate some of them with solutions for responding to requests quickly and managing virtual machines properly.

III. PROPOSED METHOD

In this method, load balancing in cloud computing environments, inspired by the algorithm that mirrors the food-finding behavior of honey bees (HBB-LB), is used. This algorithm not only balances load, but also considers the priority of tasks removed from virtual machines due to overload. The techniques of load balancing are effective for reducing the time needed to answer and service requests. The load balancing of non-exclusive independent tasks on virtual machines is an important aspect of task scheduling in cloud computing. The load on virtual machines must be distributed on balance, so that the machine is used efficiently.

In the ABC algorithm, several species of bees act in a research atmosphere. The bee that is randomly chosen to act to search is called the scout bee. It determines the location of sources of food and nectar.

A. Scheduling System Model

As in the cloud computing environment, we encounter a large space with several users and service providers, tasks are not predictable, and the capacity of each virtual machine is different. In the proposed method, the cloud is partitioned into several sectors. When the environment is very large, balancing the load within the entire system is difficult. In this method, to balance the load in smaller sectors, the load balancing algorithm inspired by the behavior of honey bees is used.

The aim of load balancing algorithms is to balance the load among virtual machines to maximize operating capability. The proposed algorithm balances tasks on virtual machines as well as the entire cloud system, so that the waiting time of tasks in the queue is minimized.

Fig. 1 depicts the proposed method. The user presents tasks to the cloud system, which contains a data center for independent tasks. The tasks are presented to the system and, prior to execution, the processing time of each task is calculated (or the processing times of tasks are estimated through mathematical models) and the characteristics of all tasks are identified. The service provider has a controller and a state table that records the features of all virtual machines. Moreover, in the controller sector of the load of each virtual machine, the system is divided into three general sectors (overloaded, under-loaded, and balanced sectors). If a virtual machine loses its load while performing tasks, it can be moved from one sector to another. Tasks are assigned based on the magnitude of loads of the virtual machines and processing times. Moreover, the indices of speed and cost are considered for the prioritization of the virtual machines. In the following, all stages of this process are discussed.

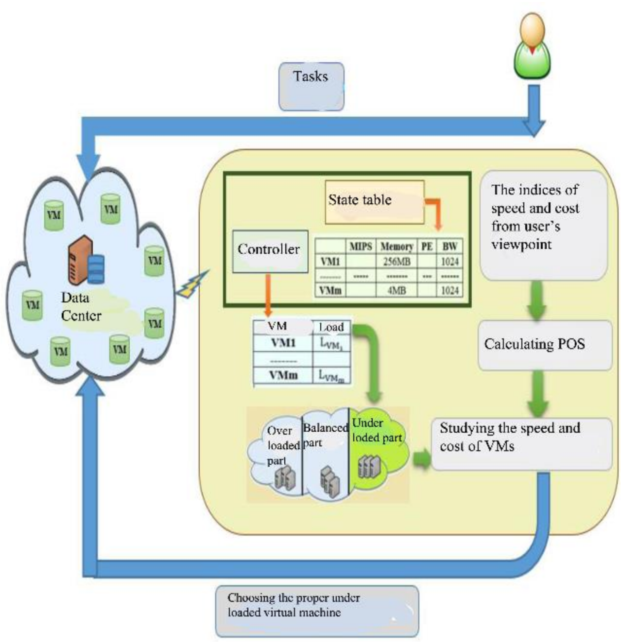


Fig. 1. The architecture of tasks.

B. Mechanism of Partitioning the Cloud System

In the cloud model, an infrastructure is considered an IaaS service that provides users with virtual resources. The cloud is partitioned into several sectors. A cloud may contain a large number of nodes in different geographical locations. Partitioning the cloud leads to better use. Fig. 2 lists details of a cloud system that has been partitioned into several separate sectors. It is worth mentioning that each area can be partitioned into several sectors. Heterogeneous virtual machines have been used in various areas, and are managed by a central controller. The methods of management may vary by cloud service provider.

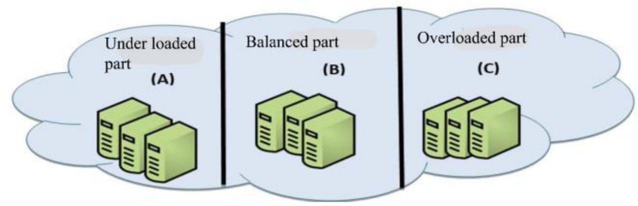


Fig. 2. Partitioning the cloud into three sectors.

When the environment is large, partitioning based on load balancing limits the search environment for assigning tasks. The cloud has a main controller that selects appropriate sectors for input tasks. The balancer per sector of the cloud selects the best strategy for load balancing (Xu et al., 2013) [8]. The load states of all virtual machines per data center are stored in the central controller, which controller deals with information for each sector, collects the information of each node, and selects the best strategy for partitioning virtual machines and assigning tasks. This information is updated repeatedly. When a task is entered into the cloud, the first stage involves selecting the proper sector. To partition the data center, it is necessary to calculate the load of each virtual machine. The controller searches sectors of the cloud and investigates the overloaded, underloaded and balanced states. Therefore, the cloud environment is partitioned into overloaded, underloaded, and balanced sectors. The method of calculating load per virtual machine is explained in Table I. Following partitioning, the tasks are presented to the underloaded sectors. To prioritize virtual machines, speed and cost are used. It is worth mentioning that after each stage of assigning tasks, the load on the system is updated. In this cloud system, for each sector, a processor is used to monitor and investigate the load on the virtual machines. If the magnitude of load on a virtual machine changes, it is moved to another sector. The values of $L_{VMk,t}$ and $L_{VMk,t+1}$ indicate the loads on virtual machines at times t and t + 1, respectively. Load differences in a time interval may show a change in the load on the virtual machine in overloaded or balanced sectors. When diffL is zero, there are no load changes in virtual machines and no changes in the sectors. However, if it is lower than zero, virtual machines are moved from one sector to another.

TABLE I. PSEUDO CODE USED TO DETERMINE CHANGES IN LOAD

The pseudo code used to determine changes in load
1) diffL = $L_{VMk,t+1} - L_{VMk,t}$
2) If (diffL = 0) no change in the magnitude of load
3) Else if (diffL < 0) the virtual machine can be moved from one sector to another.
4) End

In Table II, the parameters used in this study are introduced.

TABLE II. DEFINITIONS OF THE PARAMETERS IN THE EQUATIONS

Variables	Definitions of the Variables
CT_{max}	Maximum completion time of task
CT_{ij}	Completion time of task i on machine j
P_{ij}	Processing time of task T_i by virtual machine VM_j
Length T_i	Length of task i
Pe_{numj}	Number of processors in VM_j
Pe_{mipsj}	Million instructions per second of all processors in VM_j
P_j	Processing time of all tasks in VM_j
C_j , C	Processing capacity of VM_j and optimal processing capacity
VM_{bwj}	Communication bandwidth capability of VM_j
$L_{VMj,t}$	Load of VM_j at time t
N(T,t)	Number of requests per period
$S(VM_j,t)$	Speed of service
L , L_{VMj}	Load on all virtual machines in a data center, load on VM_j
PT_j , PT	Processing time of VM_j and processing times of all virtual machines
δ	Standard deviation
Pos_j	Priority per node
$Speed_j$	Speed of VM
$Cost_j$	Cost of VM
α	Importance of speed index
β	Importance of cost index

C. Calculating Processing Time

In HBB-LB, it is hypothesized that VM = $\{VM_1, VM_2, VM_3 \ldots VM_m\}$ is a collection of m virtual machines with no links and in parallel, where they must process n tasks. Tasks are shown as a collection T = $\{T_1, T_2, T_3, \ldots , T_n\}$. Independent tasks are non-exclusive, and are scheduled on virtual machines. A collection of virtual machines for processing tasks are an underloaded collection of virtual machines in the data center. Makespan is the time taken for task completion, and is shown in (1) (Babu & Krishna, 2013):

1. MakeSpan = max{CTij|i ϵ T , I = 1,2,…,n and j ϵ VM , j = 1,2,…,m }

The processing time of task i on virtual machine j is P_{ij}, and is calculated through (2) (Li, Xu, Zhao, Dong & Wang, 2011). The processing time of all tasks on virtual machine j is P_i, and is obtained through (3):

2. $P_{ij} = \frac{\text{length } T_i}{Pe_{numj} \times Pe_{mipsj}}$ I = 1, 2, … , n j = 1, 2, … , m
3. $P_j = \Sigma^n_{i=1} P_{ij}$ I = 1, 2, … , n j = 1, 2, … , m

The processing times of all tasks on a virtual machine must be smaller than or equal to their completion times. Therefore, by minimizing CT_{max}, (4) is obtained:

4. $\Sigma^n_{i=1} P_{ij} \le CT_{max}$ j = 1, 2, … , m

According to (3) and (4), (5) is obtained as

5. 3 and 4 $\Rightarrow P_j \le CT_{max}$ j = 1, 2, … , m
6. $CT_{max} = \{\max^n_{i=1} CT_i, \max^m_{j=1} \Sigma^n_{i=1} P_{ij}\}$

D. Calculating Capacity of Virtual Machines

The capacity of a given virtual machine as well as all virtual machines is calculated through (7) and (8) in conjunction with the HBB-LB algorithm. The total capacity of all virtual machines is equal to the capacity of the data center. C_j is the capacity of virtual machine j and C the capacity of all virtual machines. Moreover, Pe_{numj}, Pe_{mipsj}, and VM_{bwj} respectively indicate the number of processors in virtual machine, millions of instructions for all VM_j and the bandwidth of VM_j.

7. $C_j = Pe_{numj} \times Pe_{mipsj} \times VM_{bwj}$
8. $C = \Sigma^m_{j=1} C_j$

E. Calculating Load on Virtual Machines

Load includes all tasks assigned to a virtual machine. A problem in cloud systems that has negative effects on them is unbalanced loads. Heterogeneous and unequal distributions of loads among virtual machines of cloud systems can create this problem. As some processors may be overloaded and others unemployed, load balancing increases the efficiency of distributed systems. This happens when nodes with heavy loads are moved to other nodes for processing. The proposed algorithms for load balancing are inventive and varied.

The load on a virtual machine can be calculated as the number of tasks at time t in a queue in virtual machine j divided by the servicing speed of virtual machine j at time t. The magnitude of the load on all virtual machines in a data center is calculated through (10) (Babu & Krishna, 2013) [1]:

9. $L_{VMj,t} = \frac{N(T,t)}{S(VM_j,t)}$
10. $L = \Sigma^m_{j=1} L_{VMj}$

The processing time of a virtual machine, the processing times of all virtual machines, and the standard deviation of load are calculated through (11), (12), and (13), respectively:

11. $PT_j = \frac{L_{VMj}}{C_j}$
12. $PT = \frac{L}{C}$
13. $\delta = \sqrt{\frac{1}{m}\Sigma^m_{j=1}(PT_j - PT)^2}$

If δ for a virtual machine is equal or lower than [0–1] (δ ≤ Ts), the system is balanced. Otherwise, it is unbalanced, and may or not have extra load. If the magnitude of load in a virtual machine is greater than the permissible capacity, the virtual machine is overloaded and load balancing is impossible. When the load on a system is balanced, tasks are assigned to virtual machines. Table III shows the pseudo code of the load balancing algorithm.

TABLE III. PSEUDOCODE OF LOAD BALANCING ALGORITHM

Pseudocode of load balancing algorithm

```
Input: the set task; the set VM.
While there are tasks in the list do
    for all VMs of a host do
        The Ci and LVMj,t of every VMs are calculated.
        The PTi and PT of every VMs are calculated.
        The σ of every VMs is calculated.
            If σ ≤ Ts Then
                System is balanced, and Send Task to Partition;
                    The Posj of every VMs is calculated;
                    The Pj and CTi of every tasks are calculated.
                    If Posjmax and CTmax
                        Assign taski to VM
            Exit
        If L > maximum capacity
            Load balancing is not possible
        Else
            Trigger load balancing
    end for
end while
```

F. Calculating Priority of Virtual Machines

Following partitioning, the priority of under loaded virtual machines is calculated through (14). The priority per virtual machine is set according to speed and cost. POS, Speed, and Cost, respectively, indicate the priority per node in the virtual machine, the speed per node in the virtual machine, and the cost per node in it. The coefficients α and β indicate the importance of the speed and cost indices, respectively. As users assign varying priorities to indices of α and β, their values change in the interval [1, –1].

14. $Pos_j = \alpha * Speed_j + \beta * Cost_j$

IV. EVALUATION AND RESULTS OF SIMULATION

The accuracy of the proposed method was tested (according to a CloudSim simulation) and the performance of the HBB-LBP algorithm assessed in comparison with the HBB-LB, FCFS, and WRR algorithms. For the simulation, a data center and four groups of tasks of 10, 20, 30, and 40 were used. Table IV lists the values of the simulation.

TABLE IV. VALUES OF SIMULATION

Type	Parameters	Value
Data center	Number of data centers	1
	Number of hosts	3
Virtual machine	Number of virtual machines	25
	Value of misp in every processor	250–2000
	Bandwidth	10000–1000
Tasks	Number of tasks	10–40
	Length of tasks	Random
	Number of needed processors	1–4

Fig. 3 shows the average time for task completion before and after load balancing in the HBB-LBP algorithm (the proposed algorithm) with different numbers of tasks. The x-axis shows the number of tasks and the y-axis their completion times in seconds. The completion times of tasks continued to decrease. If the number of tasks increased, the algorithm exhibited better performance. Fig. 4 shows a comparison between the completion times of tasks for HBB-LBP, HBB-LB, FCFS, and WRR. The x-axis shows the number of tasks and the y-axis their completion times. According to the results, the HBB-LBP algorithm was the most efficient. Fig. 5 shows a comparison between the average times of response of HBB-LBP, HBB-LB, FCFS, and WRR.

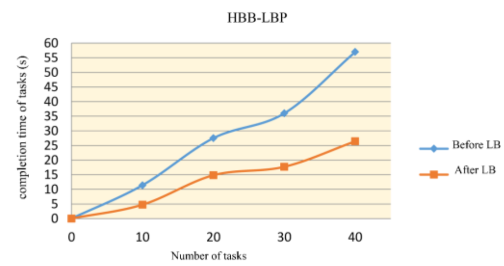


Fig. 3. Completion times of tasks before and after load balancing in the proposed algorithm.

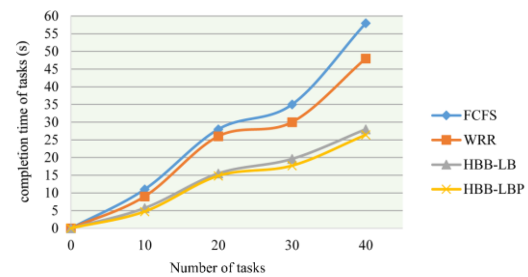


Fig. 4. Comparison among the completion times of tasks in the algorithms.

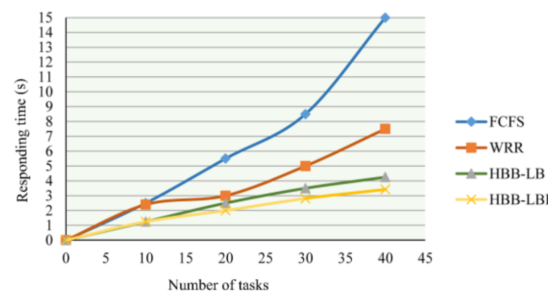


Fig. 5. Comparison among average times of response by the algorithms.

In general, the HBB-LBP method reduced the completion times of tasks and their response times in comparison with the three other algorithms. According to Table V, it reduces completion times by 9.02%, 44.08%, and 51.97% compared with HBB-LB, WRR, and FCFS, respectively.

Moreover, HBB-LBP reduced response time by 13.80%, 44.34%, and 63.94% in comparison with HBB-LB, WRR, and FCFS, respectively.

TABLE V. INDICES USED TO ASSESS THE PROPOSED ALGORITHM (HBB-LB) AND THE PERCENTAGE OF IMPROVEMENT

Algorithms / Evaluations	FCFS	WRR	HBB-LB
Completion times of tasks	51.97	44.08	9.02
Response times	63.94	44.34	13.80

V. Conclusion and Future Research

In this study, a method to partition a cloud system and investigate system load in heterogeneous environments was proposed. In the proposed model, the cloud is partitioned into several sectors and a load balancing method is used in the smaller sectors. This was inspired by the food-finding behavior of honey bees. This load balancing algorithm can consider each resource based on its capabilities and accessibility to tasks service providers, which increases the efficiency of the cloud system. Moreover, the indices of speed and cost are considered for the prioritization of virtual machines. Therefore, the proposed algorithm selects efficient resources for performing tasks based on the indices of speed and cost for the prioritization of virtual machines and the amount of load on the resources, which minimizes the time needed to service all tasks and balances system load. For future resources, the effect of fixed cost as well as the time can be evaluated, and a pricing model for user payments to cloud service providers can be organized. Moreover, cloud system modeling can be accomplished through reliable models and a hierarchical scheduler that considers the reliability of an application. This kind of load balancing can be expanded for independent tasks and the algorithm can be improved to include other factors pertaining to service quality. Moreover, the magnitude of load within the entire cloud system can be investigated, and migration can be used to distribute load between under loaded and overloaded resources.

References

[1] D. Babu, P.V. Krishna, Honey bee behavior inspired load balancing of tasks in cloud computing environments. *Applied Soft Computing, 13*(5), 2292–2303, 2013.

[2] S. Domanal, G. R. M. Reddy, Load balancing in cloud computing using modified throttled algorithm. *Cloud Computing in Emerging Markets.* Doi:10.1109/CCEM.2013.6684434, 2013.

[3] K. Li, G. Xu, G. Zhao, Y. Dong, D. Wang, Cloud task scheduling based on load balancing ant colony optimization. *Sixth Annual Chinagrid Conference* (pp. 3–9). Liaoning, 2011.

[4] K. Nishant, P. Sharma, V. Krishna, C. Gupta, Load balancing of nodes in cloud using ant colony optimization. *14th International Conference on Computer Modeling and Simulation* (pp. 3–8). Cambridge, 2012.

[5] H. Ren, Y. Lan, C. Yin, The load balancing algorithm in cloud computing environment. *2nd International Conference on Computer Science and Network Technology* (pp. 925–928). Changchun, 2012.

[6] G. Soni, M. Kalra, A novel approach for load balancing in cloud data center. *IEEE International Advance Computing Conference* (pp. 807–812). Gurgaon, 2014.

[7] H-S. Wu, C-J. Wang, J-Y. Xie, Terascaler ELB-an algorithm of prediction-based elastic load balancing resource management in cloud computing. *27th International Conference on Advanced Information Networking and Applications Workshops* (pp. 649–654). Barcelona, 2013.

[8] G. Xu, J. Pang, X. Fu, A load balancing model based on cloud partitioning for the public cloud. *Tsinghua Science & Technology, 18*(1), 34–39, 2013.

[9] K. Q. Yan, S. C. Wang, C. P. Chang, J. S. Lin, A hybrid load balancing policy underlying grid computing environment. *Computer Standards & Interfaces, 29*(2), 161–173, 2007.

[10] M. Effatparvar, M. Dehghan, A. M. Rahmani, A comprehensive survey of energy-aware routing protocols in wireless body area sensor networks. *Journal of medical systems*, *40*(9), 201, 2016.

[11] M. Effatparvar, M. S. Garshasbi, A genetic algorithm for static load balancing in parallel heterogeneous systems. *Procedia-Social and Behavioral Sciences*, *129,* 358-364, 2014.

[12] S. Molaiy, M. Effatparvar, Scheduling in Grid Systems using Ant Colony Algorithm. *International Journal of Computer Network and Information Security*, *6*(2), 19, 2014.

[13] M. Effatparvar, M. Dehghan, A. M. Rahmani, Lifetime maximization in wireless body area sensor networks. *Biomedical Research*, *28*(22), 2017.

[14] M. Effatparvar, F. Rezanezhad, S. Aghayi, Load balancing and resource allocation management with Data Envelopment Analysis method for cloud computing environment. *Recent Applications of Data Envelopment Analysis*, *978*(1), 92, 2016.

[15] M. Effatparvar, S. S. Madani, Evaluation of Fault Tolerance in Cloud Computing using Colored Petri Nets. *Evaluation*, *7*(7), 2016.

[16] A. Jalalat, B. Mahdavi, M. Salemi, M. Effatparvar, Comparative Study of Heuristic Algorithms and Nature-Inspired Algorithms for Scheduling in Grid Computing

[17] B. Azizpour, M. Effatparvar, M. S. Garshasbi, A New Fuzzy-based Job Scheduling Algorithm for Cluster Computing. *International Journal of Computer Applications*, *77*(1), 2013.

[18] M. EffatParvar, A. Bemana, M. Dehghan, Determining a central controlling processor with fault tolerant method in distributed system. In *Information Technology, 2007. ITNG'07. Fourth International Conference on* (pp. 658-663). IEEE. 2007.

[19] M. S. Garshasbi, M. Effatparvar, High performance scheduling in parallel heterogeneous multiprocessor systems using evolutionary algorithms. *International Journal of Intelligent Systems and Applications*, *5*(11), 89, 2013.

[20] J. S. Ardabili, N. Aghayi, A. Monzali, New efficiency using undesirable factors of data envelopment analysis. *Adv. Modeling & Optimization*, *9*(2), 249-255, 2007.

[21] N. Aghayi, B. Maleki, Efficiency measurement of DMUs with undesirable outputs under uncertainty based on the directional distance function: Application on bank industry. *Energy*, *112*, 376-387, 2016.

[22] N. Aghayi, M. Tavana, M. A. Raayatpanah, Robust efficiency measurement with common set of weights under varying degrees of conservatism and data uncertainty. *European Journal of Industrial Engineering*, *10*(3), 385-405, 2016.

[23] N. Aghayi, Cost efficiency measurement with fuzzy data in DEA. *Journal of Intelligent & Fuzzy Systems*, *32*(1), 409-420, 2017.

[24] P. Shankar, M. Fazelpour, J. D. Summers, Comparative study of optimization techniques in sizing mesostructures for use in NonPneumatic tires. *Journal of Computing and Information Science in Engineering*, *15*(4), 041009, 2015.

[25] M. Fazelpour, P. Shankar, J. D. Summers, Developing design guidelines for meso-scaled periodic cellular material structures under shear loading. In *ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (pp. V02BT03A002-V02BT03A002). American Society of Mechanical Engineers. 2016.

[26] M. Yoder, Z. Satterfield, M. Fazelpour, J. D. Summers, G. Fadel, Numerical Methods for the design of meso-structures: a comparative review. In *ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (pp. V02BT03A003-V02BT03A003). American Society of Mechanical Engineers, 2015.

[27] Rostamy-malkhalifeh, M., & Aghayi, N. (2011). Measuring Overall Profit Efficiency with Fuzzy Data. Journal of Mathematical Extension.

[28] Aghayi, N., & Ghelejbeigi, Z. (2016). Improving Cost Efficiency by Resource Allocation. Far East Journal of Mathematical Sciences, 99(11), 1633.

[29] Aghayi, N. (2016, January). Revenue Efficiency Measurement with Undesirable Data in Fuzzy DEA. In Intelligent Systems, Modelling and Simulation (ISMS), 2016 7th International Conference on (pp. 109-113).

[30] Salehpour, S., & Aghayi, N. (2015). The most revenue efficiency with price uncertainty. International Journal of Data Envelopment Analysis, 3(1), 575-592.

Combating the Looping Behavior: A Result of Routing Layer Attack

David Samuel Bhatti[1], Kinza Sardar[2], Meh Jabeen[3], Umair B. Chaudhry[4]
[1, 2, 3, 4]Department of Computer Science and Information Technology,
University of Lahore,
Lahore, Pakistan
[1]School of Electrical Engineering and Computer Science,
National University of Science and Technology,
Islamabad, Pakistan

***Abstract*—Routing layer is one of the most important layers of the network stack. In wireless ad hoc networks, it becomes more significant because nodes act as relay nodes or routers in the network. This characteristic puts them at risk of routing attacks. A wormhole is the most treacherous attack on a routing layer of wireless ad hoc networks. The present proposed techniques require extra hardware, clock synchronization; or they make restrict assumption to deal with this attack. We have proposed a simple behavior-based approach which uses a small amount of memory for recording a few packets received and sent by the neighboring nodes. From this information, a behavior of these nodes is detected, that is, whether the behavior is benign or malicious. Nodes exhibiting malicious behavior are placed in the blocked node list. Malicious nodes are broadcasted in the network. None of the legal nodes in the network entertains any packet from these nodes. This approach has been simulated and verified in ns2.30 which detects and isolates wormhole nodes successfully. The current study focuses on the looping behavior of this attack.**

***Keywords*—Wormholes attack; wireless routing layer attack; detection time; throughput**

I. Introduction

Wireless ad hoc networks are very economical, simple and flexible. Easily, they can be deployed even in the hostile environments. Due to these features, they are being used extensively in civilian as well as in the defense domains. Their importance can be envisioned from the range of deployments; as they are deployed in large enterprise organizations, corporate sector, offices, and homes. Even, in the bodies of different living and non-living creatures like guided flying objects, these networks are operating now. Major categories of ad hoc networks are PANs, MANETs, VANETs and WSNs [1], [2].

But, now the wireless ad hoc networks are taking new turns and transforming themselves to Internet of Things [3], Internet of Vehicles [4] and mobile cloud computing [5]. With all these advantages, they are inherently broadcast networks and attract different types of attacks at different layers of the underlying network stack. The major attacks these networks encountering are the sinkhole, rushing, Byzantine, black hole, wormhole, and Sybil. Attacks like spoofing, dropping of routing traffic, selective forwarding, resource consumption are also important to be mentioned here [6], [7], [8], [9], [10], [11]. Besides this, the safe use of wireless devices like mobiles and smart-phones has been well discussed in [13]. All the attacks mentioned above are equally significant, but, the wormhole is one of the trickiest attacks observed in the literature [12]. The aim of this study is to devise a flexible and extendable mechanism to detect attacker nodes involved in the creation of wormhole link involved in looping of data packets.

In rest of the article, we have discussed wireless networks in Section II; routing protocols, wormhole attack and looping in Sections III, IV and V, respectively. Sections VI, VII and VIII shed light on wormhole classification, related work, and proposed approach. Under Sections IX, X, XI, results, future directions, and the conclusion is discussed.

II. Wireless Networks

This section has been added in this study to introduce the readers to different types of network and their applications; helping them focus on the particular type being discussed in this study.

A. Wireless Personal Area Network (PAN)

These are the networks which are within a reach of a person. These are the replacement of peripheral devices. NFC, Bluetooth, and Zigbee are the examples of such networks [46].

B. Wireless Local Area Network (LAN)

The range of this network can be campus, building, home or office. It is the extension of the wired network. WiFi (IEEE-802.11) types of networks fall into this category [46].

C. Wireless Metropolitan Area Network (MAN)

It's the type of network which covers an area of a city or a town. WiMax (IEEE 802.15) network are basically metropolitan area networks. They are used to connect wireless network with one another [46]. In other words, these provide inter-network connectivity.

D. Wireless Wide Area Network (WAN)

These are the big networks which provide wireless access beyond the range limits of LAN and MAN. LTE UMTS, GSM, and satellites networks all are the types of wide area networks [46]. The type of network discussed in this study is basically an 802.11 WiFi ad hoc mode where no central control (Access Point) is applied rather the wireless nodes are providing relaying function themselves.

III. Routing Protocols

It has been observed that due to the ready availability of wireless technology, wireless networks are becoming very popular in all areas of life. Mobile ad hoc networks are among popular types of wireless networks have specific features like dynamic topology, open network boundaries and hop by hop communication. These networks are facing a lot of routing challenges too; some of these are the limited wireless range, constraints in battery power, heterogeneity in devices hardware and software, hidden terminals, etc.

Actually, it is the routing which finds and maintains routes between nodes, so that data can be transferred from the source to the destination. This process itself is considered to be the most challenging because nodes do not have familiarities with the underlying topology which is very dynamic in nature due to mobility. Messages are sent and received according to the predefined set of rules. These set of rules are referred to as protocols. Routing protocols are categorized into flat routing and geographical position assisted routing. In flat routing, each network identity is represented individually. Flat routing is divided into proactive, reactive and hierarchical/hybrid routing. Proactive routing protocols are also known as table driven routing protocols. DSDV, WAR, CGSR, WRP, QDRP, TBRPF, and OLSR are the well-known examples of proactive routing protocols [47], [48]. Whereas, AODV, DSR, LMR, TORA, and LQSR are the famous examples of reactive routing protocols [47], [48]. In proactive routing, the entire network routing information is maintained continuously by using the routing tables.

On the contrary, the reactive routing nodes only maintain information of the active routes to the destination nodes. In reactive routing searching for the path starts only when a node is required to send a data to another node in the network. In flat routing, wireless ad hoc networks start facing additional overhead when the network size increases. In these scenarios, hybrid/hierarchical routing protocols like ZRP, BGP, EIGRP, CGSR routing protocols are chosen [47], [48].

In the geographical position assisted routing, performance of the routing algorithm increases by using the moving node. Global positioning systems help in determining the location information of nodes in the network. LAR, DREAM, GPSR, and EGR are the protocols of geographical position assisted routing. Fig. 1 shows the taxonomy of these routing protocols [47], [48], [49].

In the current study, AODV has been used as a routing protocol in ns2.30.

IV. Wormhole Attack

Wormhole is one of the most problematic attacks in routing layer of the network stack [8]. It is a very intelligent attack that creates a tunnel whose one end opens at the sender side and other end opens at the receiver side.Whereas, these senders and receivers are very distant nodes. All traffic travels through this tunnel is controlled by the wormhole attackers. These attacker

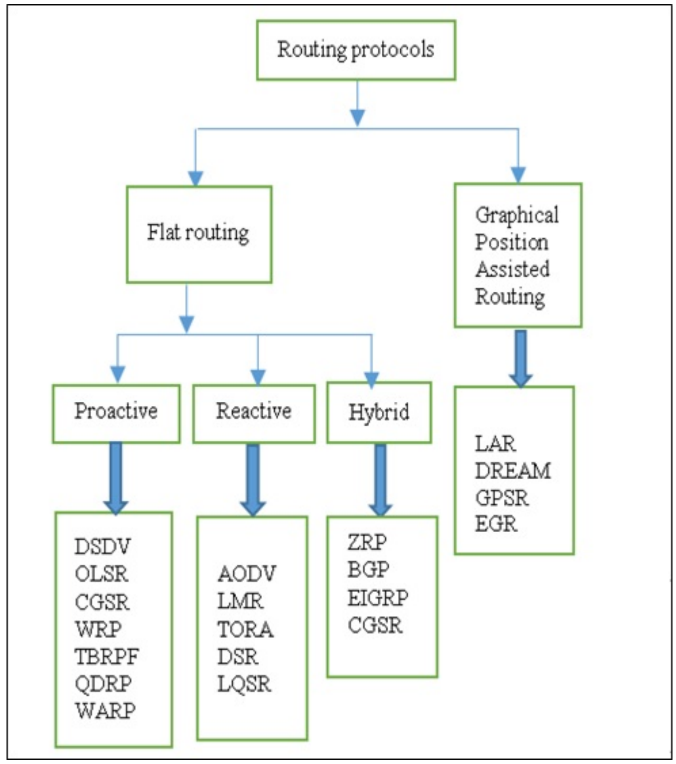


Fig. 1. Routing protocols in MANET.

nodes are sitting in the vicinity of sender and receiver which are many hops away from each other [10].

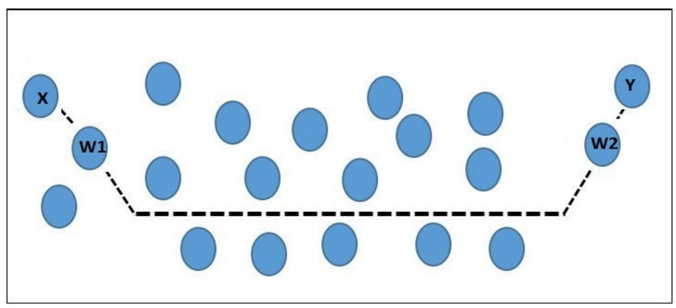


Fig. 2. Wormhole attack.

Usually, these attacker nodes have low latency wired or high radio range wireless link. This low latency link attracts the maximum traffic of the region covered by wormhole nodes. This is a 2 hops illusion created by wormhole which seduces the traffic to allure toward this malicious link in shortest path routing protocols. Wormhole captures traffic from one part of the network that is from sender side and replays it in another part of the network, that is, a receiver side. As shown in Fig.2, once the wormhole link has been established, wormhole node W1 captures the packets from the source X, relays them to the attacker node W2 which is resided in the vicinity of destination Y. Attackers W1 and W2 will take over the route $X \leftrightarrow W1 \leftrightarrow W2 \leftrightarrow Y$.

When source will send data, this packet will be entrapped into a wormhole link $W1 \leftrightarrow W2$. Due to, different malicious behaviors like replay and drop or loop-back of this link, a data packet will not be able to escape from this trap. The scenario becomes worst when intermediate nodes are compromised they also start taking part in relaying traffic from one wormhole node to another. Wormhole can disrupt 30% to 90% of network [14]. Wormhole can also create sinkhole which result in the drawn of all traffic from the surrounding provided alternative routes are less attractive than the wormhole link [10]. State of the systems may become poor, when wormholes combine with Sybil attack, which becomes hard to detect then [11].

V. Looping Behavior of Wormhole

For the onward discussion we follow the conventions as given in Table I, as well as assuming, a wormhole link has already been established as shown in Fig. 2

TABLE I. Conventions to be Used

Node	IP-Address	MAC-Address
Source node ”S”	IP-S	MAC-S
Destination node ”D”	IP-D	MAC-D
Wormhole Attacker ”W1”	IP-W1	MAC-W1
Wormhole Attacker ”W2”	IP-W2	MAC-W2

In distance vector routing protocols, when the node “S” (Source Node) sends a data packet to “D” (destination node), “S” will insert its Source-IP “IP-S”, Destination IP “IP-D” in place of Original-Src-IP and Original-Dest-IP, respectively. In the place of Src-MAC, it will insert its MAC “MAC-S” and in place of “Dst-MAC”, it will place next node MAC “MAC-W1”. “S” will transmit this packet. When this packet reaches the wormhole node “W1”, then in the place of “Src-MAC”, “W1” will insert its MAC “MAC-W1” and in place of “Dst-MAC”, it will place next node MAC “MAC-W2”. “W1” will transmit this packet through high radio range link established at route discovery time.

When this packet reaches the wormhole node “W2”, then in the place of “Src-MAC”, “W2” will insert its MAC “MAC-W2” and in place of “Dst-MAC”, it will place “W1” MAC “MAC-W1”. “W2” will change the direction of this packet while transmitting. This packet will come back to “W1”. In this way, this packet not will reach the destination at all as shown in the Fig. 3. For the simplicity packets with fewer fields have been shown moving between two wormholes nodes W1 and “W2” in this figure.

VI. Classification of Wormholes Attacks

More robust solutions can be devised if the wormhole attacks are made classified according to their mode of behavior, that is, whether the attacker nodes are visible or not in the route. These are classified as open wormhole attack, closed wormhole attack, half open wormhole attack [50].

A. Open Wormhole Attack

In the open wormhole, source S, attacker X and Y and destination D are visible in the network. So, the path formed would be S↔X↔Y↔D. Attackers make themselves a part of the header following route discovery algorithm. A packet will be tunneled from one end to the other where it will be broadcasted. In this case, attackers do not let intermediate nodes A, B and C to make themselves visible in this network.

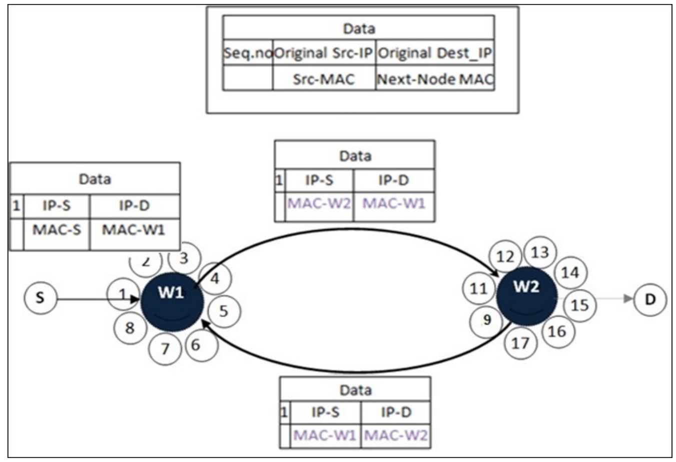


Fig. 3. Looping behavior of wormhole.

B. Closed Wormhole Attack

While in half open wormhole one of the attackers keeps itself hidden while other visible. Wormhole attacker close to the source will receive the packet from the source and will broadcast it in the other end. So the path formed would be S↔X↔D.

C. Half Open Wormhole Attack

Whereas, both of the malicious nodes X and Y nodes along with intermediate A, B and C are kept hidden in the closed wormhole. Source and destination think that they are one hop away from each other. Thus fake neighbors are created in this case. These three modes can be visualized from Fig. **??** Wormhole attacks can project themselves using packet encapsulation, high-quality/out-of-band link, high-power transmission capability, using packet relay and protocol distortion [50].

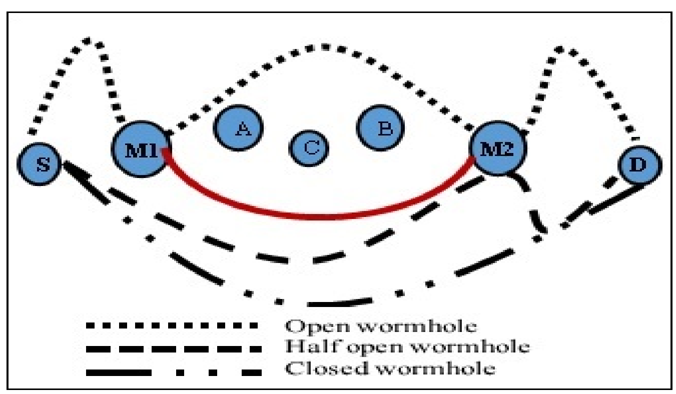


Fig. 4. Wormhole modes.

VII. Related Work

There are different types of mitigation approaches which have been discussed under different sections for the better understanding of the reader. They have been categorized like hardware-based, cryptographic-based, guard-node-based, packet-leashes-based, etc.

A. Reply Count-based Approaches

WARP [15] is a very promising approach for the mitigation of wormhole attacks. It is based on different abnormalities observed in the system. It modifies the RREQ message of AODV [17], [18], [19] protocol by introducing an additional field 'first-hop'. It adds three fields in AODV routing table named first-hop, RREP-count, and RREP-DEC-count. It uses an additional RREP-DEC message with the same fields as that of the RREP of AODV. 'Type' field of the RREP has been used to distinguish them. The addition of such fields in the packet header results in a considerable overhead of bandwidth. Upon receiving the RREP, originator always sends the RREP-DEC message. The routing table is also overwhelmed by storing the information of the first-hop, RREP-Count, RREP-DEC-count. Periodic broadcasts also result in the increase of bandwidth required. Nodes continuously observe the behavior of other nodes. If they notice some irregularities beyond a certain threshold, then these nodes will be cut off from the network.

DAWSEN [16] is a wormhole mitigation scheme which makes use of reply counts. DAWSEN is simple, but, requires a high power base station, so that, a request can be received in one hop.

B. Guard Nodes-based Approaches

Honey-pots are the computer nodes which are deployed in the network to attract the attackers, so that; they may be detected and kept isolated from the network. They are also used to study the unauthorized attempts to get access to different business information systems. Honey-pots have been proposed in [20] for catching wormhole nodes.

Authors in [20] suggest placing some honey-pot nodes with some vulnerability in the network. These honey-pot nodes will allure the attackers which become exposed ultimately and can be removed from the network subsequently. LITEWORP [21] is another approach which places guard nodes in the network to identify the wormhole nodes. These are simple approaches, but, in bigger networks, deployment of additional nodes may result in some scalability issues [22].

C. Cryptographic Approaches

Cryptographic approaches are husbanded with complex computational operations, but, still, some authors suggest them for the detection and isolation of wormhole links in wireless ad hoc networks.

Some of them are TESLA [27] "Timed Efficient Stream Loss-tolerant Authentication" and TIK [28] "TESLA with instant key", MOBIWORP [23], SOADV [24], SPINS [25], TrueLink [26], Ariadne [29], etc. These all approaches are computation and bandwidth hungry and some even require precise clock synchronization. For instance, TESLA, which cannot be achieved without additional dedicated hardware. TESLA [27] is good for laptop class, but, it does not suit network scenarios with low resources [25]. Since wormhole attack projects itself at the time of route discovery; so, it does not require cryptographic information to relay or forward routing packets. That is why; these approaches are still vulnerable to this attack [30].

D. Additional Hardware based Approaches

There are certain approaches which believe in, that there should not be any objection if some extra hardware is brought into use for the handling of this treacherous attack. Among these approaches [11], [31], [32] are the very famous one. They make use of the directional antenna. In these scenarios, receiver detects the direction of the signal and it can bypass the wormhole. But, all these scenarios put forward an assumption that sender and receiver must be carefully aligned with one another.

Moreover, [31] and [11] assume that secret keys are pre-shared and secure discovery of neighbor nodes is already running. Approach [31] can only partially mitigate the wormhole attack and scalability factor also rules out the use of such approaches according to other researchers [21], [26]. SECTOR [33], MOBIWORP [23], DAWSEN [16], Leashes [28] are some of the other approaches which fall in this category.

E. RF Based Techniques

Radio Frequency (RF) based approach has been proposed at the physical layer in [34]. A waveform with a special pattern is projected and if any of the wormhole nodes cause a change in this special pattern that can be detected easily, and, the route is discarded. This is fine where malicious nodes cause a change in the pattern, but, if they exactly replicate the wave form they can be bypassed [21]. It is also not feasible to provide every node with RF capability.

F. Hop-Count Based

The approach proposed in [36] suggests the pre-distribution of secret keys pairwise. According to them, secret keys can be generated by using the one-way hash functions. They also make use of a hop-count parameter to differentiate between normal and wormhole nodes. The approach proposed in [35] is fairly simple and based upon the number of hops traversed by the routing packet. In this approach, routes with a higher number of hops are considered to be fair, whereas, routes with relatively fewer hops are treated as the malicious or corrupted ones. This idea works well for high transmission-power based wormhole attackers, but, not equally suitable for low range attackers. Authors of [37] suggest detecting distant nodes whose messages arrive quite earlier than the messages sent by other nodes. This behavior predicts the suspiciousness of such nodes. These techniques are simple and free of extra hardware cost and complex cryptographic operations.

G. Graph Theory-based Approaches

Graph theory has been leveraged in this regard and approach [40], [44], [45] make use of this field for the detection and isolation of multiple occurrences of the wormholes. These are the more advanced versions of this attack which work together for projecting the collaborative attacks. These are new classes of attack; Evil Twin is one of the famous attacks of this category in wireless ad hoc networks.

H. Other Secure Routing Protocol Approaches

The authors of [38] make use of changes occurring in the network along with routing information to detect the wormhole attacker nodes. It is a simple approach which does not require additional hardware and even does not impose any strict assumption. Secure routing protocols used in [8], [39]project a quite reasonable defense layer against the nodes which collude with one another for the launch of wormhole link. They achieve this by using end-to-end authentication with the help of hashed message authentication codes.

SEAD [41] is the secure routing protocol which provides authentication for routing processes. SEAD is basically an improvement of DSDV [42]. It is a good guard against attack which result in the modification of the packet, and hence unable to catch the wormhole nodes. In other words, it can only resist against the illegal increase in the sequence-no or illegal decrease in the hop-count. An approach proposed in [43], suggests to make use of Timed Colored Petri Net for the formal verification of the proposed approach, whether, the proposed approach works or not under different network conditions and scenarios. The technique in [43] is based upon the round trip time (RTT). The value of RTT is always very short over the routes with wormhole link and very high over the route which are free from this malicious link.

All these approaches are promising; each has its own pros and cons. The aim of this study is not to reject other proposed approaches but to bring attention of the research community toward simpler solutions, because they are equally as effective and efficient as compared to the complex solutions.

VIII. Proposed Approach

Referring to Section-V (Looping Behavior of Wormhole), data packets will not be able to escape from this wormhole. Neighbor nodes of both the attackers maintain suspicious-node-list by storing packet sequence-no, previous-hop and next-hop. If for the same packet, its previous hop is becoming its next hop and next hop is becoming its previous hop, then node will be suspected as wormhole attacker node. Neighbor nodes will send alert message for this node as suspicious node and a node on getting alerts beyond the certain thresh-hold, against this node, will broadcast the ID of this node as malicious. RERR message will be generated. On receiving RERR message, source sends new RREQ message, where the malicious nodes will be blocked to take part in new route discovery process. We have discussed the scenario where the tunnel is created with the help of compromised nodes in Fig. 5. In this case, W1 and W2 are real attackers whereas node-5 and node-11 are compromised nodes.

As an example, list of packets to be cached at node-1 has been shown in this figure. Nodes involved in looping can be detected from the cache list. If we look at the detection mechanism node-5 and node-11 will also be detected as malicious nodes and will be isolated.

Carefully, we have observed this event and decided to remove this limitation of isolation of compromised nodes in future. This whole mechanism has been formulated in the form of a flow chart given in 6, for the better understanding of the reader. Algorithm (Pseudo Code) which is close to implementation has been elaborated in Algorithm 1. For successful implementation of this algorithm, two cache lists were created one is the suspicious node list whose fields are packet-ID, previous hop, next hop and the other is blocked node list whose fields are malicious node-ID and witness-count. The purpose of the first list is to store the packets along with node-IDs and the purpose of the second list is to hold the malicious nodes Ids respectively. These lists are given in from Fig. 5 and 6.

Node Ids given in flowchart are ones used in the simulation, where 'from' and 'to' have been used for source-MAC and Destination-MAC, respectively. CBR is constant bit rate data traffic. Entries in the suspicious node list help to detect the looping behavior of malicious and compromised nodes. From Fig. 5 it can be observed that W1 is giving data packet to node-5, node-5 to node-11 and node node-11 to W2. W2 instead of delivering the packet to destination node "D" sends the packet back to node-11 and node-11 to node-5 and node-5 to W1. In this way, packets are kept in a loop until they drop down.

Detection algorithm processes the suspicious list and detects this behavior. Nodes showing this behavior are captured and are broadcasted in the whole network. Every node in the network on receiving broad alert against newly detected malicious node, adds that node in its list of blocked nodes. None of the nodes in the network accepts any route request packets from nodes placed in the blocked node list. This results in secure route discover.

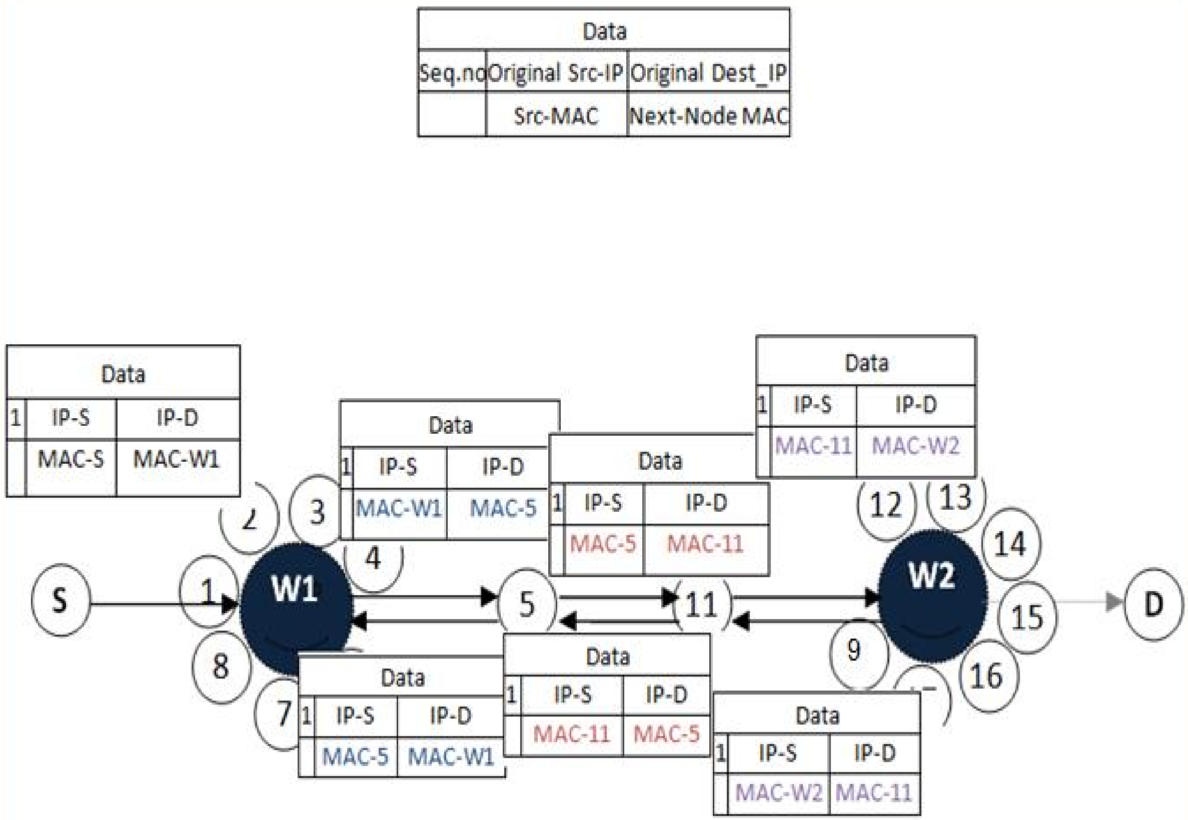
IX. Results and Discussion

Simulations have been carried out in ns2.30 with 802.11 MAC, at 64 Kbps data rate, with an omnidirectional antenna, in 1000 m x 1000 m topology with 200 wireless legal and 2 wormhole nodes. In this simulation attacker nodes radio range is 400 m and legitimate nodes 80 m. Static as well as mobile scenarios with 1000 J as the initial energy of the nodes have been simulated for 5-100 seconds using AODV protocol. The packet size has been kept as 512 Bytes in all cases. The malicious behavior of the wormhole node-200 and node-201 referred in Fig. 7. In the absence of new scheme, these wormhole nodes capture data packets going from source node-34 to destination node-100. Wormhole nodes keep on relaying these data packets. Ultimately, their time to live (TTL)value reaches zero and they dropped down. Wormhole nodes can disrupt only selective packets just to make their detection more problematic.

When we apply the proposed approach, wormhole node-200 and node-201 were successfully detected and isolated. The new route discovered was free from wormhole attacker nodes. And a successful communication was made possible between the source node 34 and destination node-100 as shown in Fig. 7. Whereas, prior applying this new algorithm, the wormhole nodes did not allow the data packets to move from node-34 to node-100.

A. Throughput & Loss Ratio

The proposed approach has a quite satisfactory throughput, which has been observed ranging from 78% to 96%. In static scenarios, the average throughput was about 96%, whereas, in mobile scenarios, it was about 78%. These observations were made with two wormhole nodes. For brevity, results of one of the experiment have been shown in Fig. 8.



Suspicious Node List at Node 1		
Packet-Id	Previous-Hop	Next-Hop
1	MAC-S	MAC-W1
1	MAC-W1	MAC-5
1	MAC-5	MAC-11
1	MAC-11	MAC-W2
1	MAC-W2	MAC-11
1	MAC-11	MAC-5
1	MAC-5	MAC-W1

Fig. 5. Detection mechanism with intermediate node.

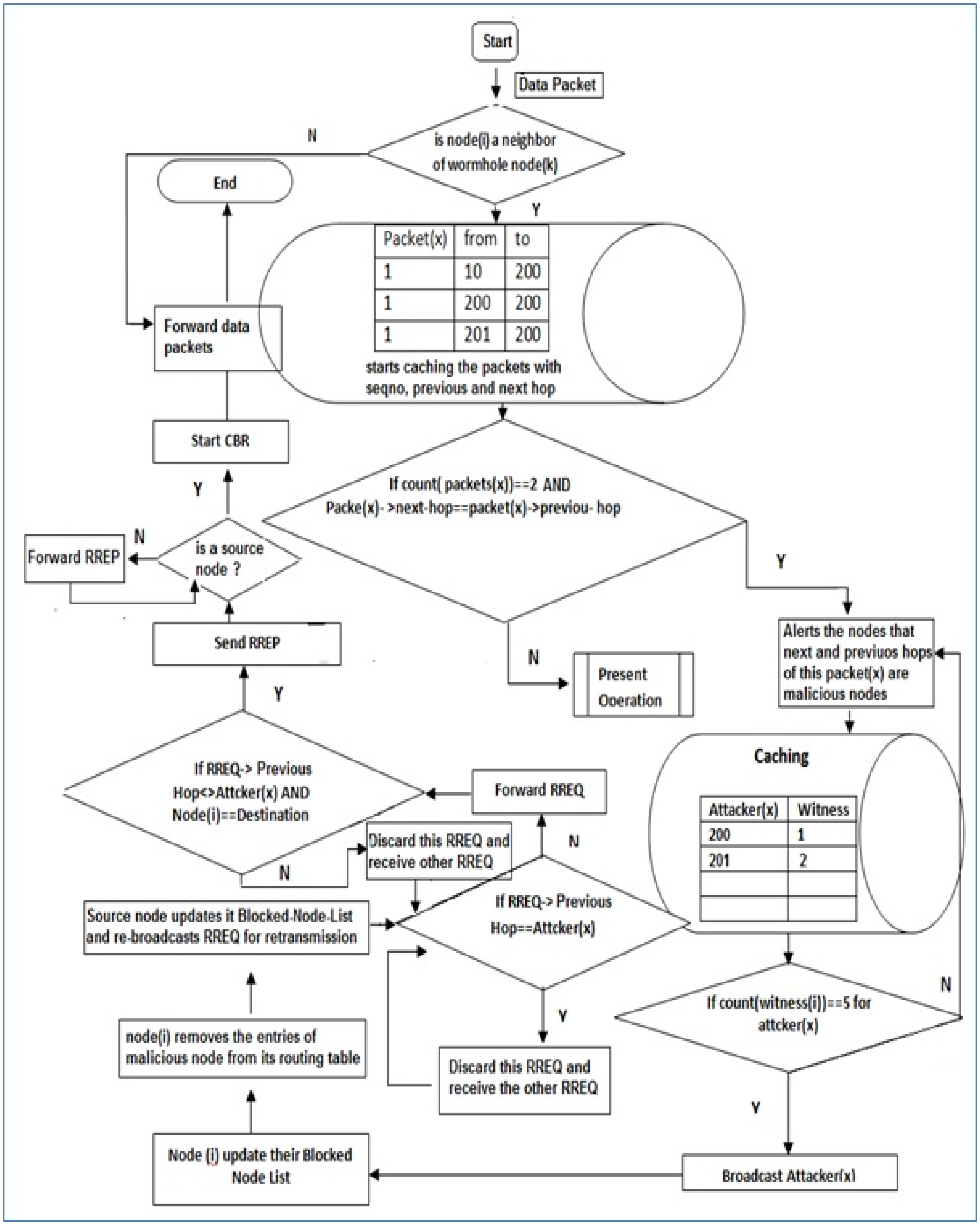


Fig. 6. Detection mechanism with intermediate node.

Algorithm 1: Wormhole Detection

begin
- **if** $Packet\ Type = Data\ Packet$ **then**
  - $Store\ packet - id,\ previous - hop,\ next - hop\ in\ suspected\ neighbor\ list$;
- **else**
  - $Forward\ Packet$;
- **end**
- **while** $NOT\ END\ OF$ "$SuspectedNeighborList$" **do**
  - **if** $same\ packet\ occurs\ twice\ whose\ next\ and\ previous\ hops\ are\ same$ **then**
    - $Generate\ 1 - hopalert\ against\ this\ suspected\ node$;
- **end**
- **Processing at Wormhole Neighbors**
  The following operating will be performed at neighboring nodes wormhole
  1.Receive suspicious alert message
  2.Insert the suspected-nod in the blocked-node-list
  3.Increments witness-count
  Threshold of the witness-count is decided wisely
  if $witness - count > Threshold$ **then**
  - $Broadcast\ this\ malicious\ node - idusingmaliciousalertmessage$;
- $Generate\ route\ error\ message$;
- **Processing at All Nodes**
  All nodes will remove the entries of malicious nodes from their routing tables when they will receive the malicious node-id from malicious alert message. Moreover, nodes will update their blocked-node-list by adding this malicious node on receiving the malicious node-id
  Processing at Source Node
  Source node will broadcast RREQ again
  Processing at Intermediate Node
  Node Receive RREQ
  if $RREQ\ previous - hop \in BlockedNodeList$ **then**
  - $DiscardtheRREQ$;
- **else if** $RREQ - destination - id = -node - id$ **then**
  - $Send\ RREP$;
- **else**
  - $Forward\ the\ RREQ$;
- **Processing at Destination Node**
  Node Receive RREQ **if** $RREQprevious - hop \in BlockedNodeList$ **then**
  - $Discard\ the\ RREQ$;
- **else if** $RREQ - destination = node - id$ **then**
  - $send\ RREP$;
- **else**
  - $Forward\ the\ RREQ$;
- **Processing at Source Node**
  In this way, wormhole nodes will be filtered out because none of the nodes from the network will accept RREQ from the nodes listed in Blocked-Node-List. In this, there is very high probability the route established between source and destination will be free from the wormhole. Thus, the Source Node receives RREP from the path which has dispelled out wormhole nodes.It starts the retransmission of data packets. Data packet reaches a destination through newly established path successfully.

end

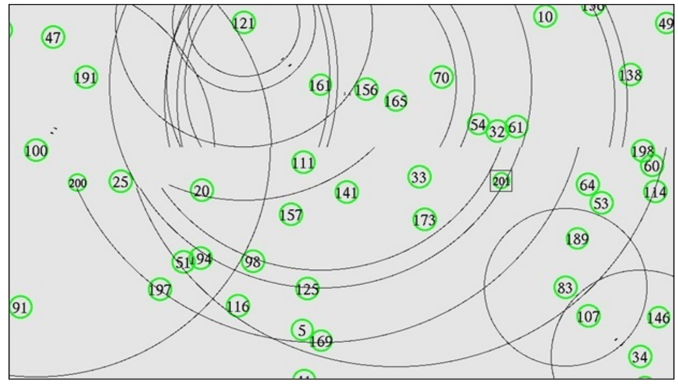


Fig. 7. Wormhole isolation in simulated environment.

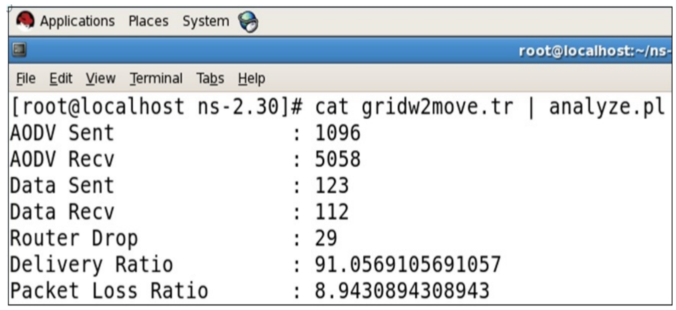


Fig. 8. Throughput in mobile scenario with 2 wormhole nodes.

B. Detection Time

Satisfactory detection and recovery time has been observed during the simulation. Simulation started at 1.0 second; wormhole nodes were detected at 1.149 seconds of the 10-100 second. It also depends upon the witness threshold. We suggest witness-count threshold to be kept low as one or two. However, if a witness-count threshold is kept low, detection latency will be decreased respectively. If it is kept high, detection latency would be increased, respectively. Therefore, it is important to wisely keep both the thresholds.

It was observed during simulation, the destination sends RREP at 1.248 second and data packets then successfully received at the destination at time 1.3150 second. Wormhole detection time, as well as system recovery time, is quite satisfactory. We calculated that system recovered within 0.15 second after the wormhole detection, as shown in Fig. 9

X. Future Work

We have planned to extend the existing approach for detecting multiple wormhole links equipped with other malicious behaviors like packet drop, packet selective drop, manipulation of TTL, replaying in wireless ad hoc networks. We aim that the extended strategy should be efficient with respect to memory, computation, and bandwidth.

XI. Conclusion

The proposed technique is simple, scalable and does not require any additional hardware. It does not impose any strict assumption of loose or tight clock synchronization for the proposed approach to work. The solution is not computation,

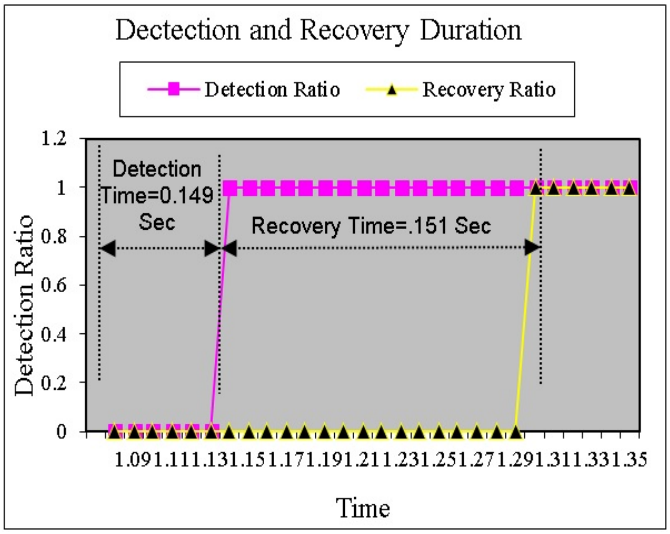


Fig. 9. Detection & recovery time.

memory, and bandwidth hungry. Size of cache lists is small enough to be easily processed by the wireless nodes including wireless sensor nodes. Thus, the solution is equally deliverable in small as well as in large wireless ad hoc networks in static as well as in mobile scenarios.

REFERENCES

[1] Y. Xiao and Y. Pan, ”Emerging Wireless LANs, Wireless PANs, and Wireless MANs: IEEE 802.11, IEEE 802.15, 802.16 Wireless Standard Family,” Wiley Publishing, 1st ed., 2009.

[2] I. Al Shourbaji, ”An overview of wireless local area network (WLAN),” CoRR, vol. abs/1303.1882, 2013.

[3] M. A. Razzaque, M. Milojevic-Jevric, Palade, and S. Clarke, ”Middleware for the internet of things: A survey,” IEEE Internet of Things Journal, vol. 3, pp. 70-95, Feb 2016.

[4] J. Contreras-Castillo, S. Zeadally, and J. Guerrero-Ibanez, ”Internet of vehicles: Architecture, protocols, and security”, IEEE Internet of Things Journal, vol. pp. 1-1, Apr 2017.

[5] L. Lei, Z. Zhong, K. Zheng, J. Chen, and H. Meng, ”Challenges on wireless heterogeneous networks for mobile cloud computing,” IEEE Wireless Communications, vol. 20, pp. 34-44, Jul 2013.

[6] S. R. Surya and G. A. Magrica, ”A survey on wireless networks attacks,” in Computing and Communications Technologies (ICCCT), 2017 2nd International Conference on, (Chennai India), IEEE, 23-24 Feb. 2017.

[7] S. Mavoungou, G. Kaddoum, M. Taha, and Matar, ”Survey on threats and attacks on mobile networks,” Included in Special Section in IEEE Access: Security in Wireless Communications and Networking, vol. 4, pp. 4543 - 4572, August 2016.

[8] P. Papadimitratos and Z. Haas, ”Secure routing for mobile ad hoc networks,” in Proceedings of the SCS Communications Networks and Distributed Systems, Modeling and Simulation Conference (CNDS), (San Antonio, TX, USA), pp. 193-204, January 2002.

[9] B. Zhu, Z. Wan, M. S. Kankanhalli, F. Bao, and R. H. Deng, ”Anonymous secure routing in mobile ad-hoc networks,” in Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks, LCN 04, pp. 102- 108, 2004.

[10] C. Karlof and D. Wagner, ”Secure routing in sensor networks: Attacks and countermeasures,” in First IEEE International Workshop on Sensor Networks Protocols and Applications, May 2003.

[11] L. Hu and D. Evan, ”Using directional antennas to prevent wormhole attacks,” in Proceedings of the IEEE Symposium on Network and Distributed System (NDSS), pp. 131-141, 2004.

[12] V. Kumar and R. Kumar, ”Mitigation of wormhole attack using SOA in MANET,” Global Journal of Pure and Applied Mathematics, vol. 13, no. 2, pp. 431-452, 2017

[13] D. S. Bhatti , N. A. Saqib, Z. Anwar,” SCEAMS: Secure corporate environment adhered to mobile & smartphones”, in 2016 Sixth International Conference on Innovative Computing Technology (INTECH), IEEE, Dublin Ireland, Aug.2016

[14] M. Khabbazian, H. Mercier, and V. K. Bhargava, ”Severity analysis and countermeasure for the wormhole attack in wireless ad hoc networks,” IEEE Transactions On Wireless Communications, vol. 8, no. 2, 2009.

[15] M.-Y. Su, ”WARP: A wormhole avoidance routing protocol by anomaly detection in mobile ad hoc networks,” Computer & Security, ELSEVIER, pp. 208-224, 2010.

[16] A. C. Z. D. Rouba El Kaissi, Ayman Kayssi, ”DAWSEN: a defense mechanism against wormhole attacks in wireless sensor networks,” in Second International Conference on Information Technology, (Dubai, UAE), 2005.

[17] C. E. Perkins and E. M. Royer, ”Ad hoc on demand distance vector routing,” in Proceedings of the Second IEEE Workshop on Mobile Computer Systems and Applications, WMCSA 99, 1999.

[18] C. Perkins, E. Belding-Royer, and S. Das, ”Ad hoc on-demand distance vector (AODV) routing,” 2003.

[19] C. E. Perkins and Belding-Royer, ”Ad hoc on demand distance vector (AODV) routing,” In IETF RFC 3561, (Mountain View, CA, USA), July 2003.

[20] P. V. T. Divya Sai Keerthi, ”Locating the attacker of wormhole attack by using the honey pot,” Liverpool, United Kingdom United Kingdom, pp. 1175-1180, 2012.

[21] I. Khalil, S. Bagchi, and N. B. Shroff, ”Liteworp: Detection and isolation of the wormhole attack in static multi-hop wireless networks,” Comput. Netw., vol. 51, pp. 3750-3772, Sep. 2007.

[22] Y.-C. Hu, A. Perrig, and D. B. Johnson, ”Rushing attacks and defense in wireless ad hoc network routing protocols,” in Proceedings of the 2nd ACM Workshop on Wireless Security, WiSe 2003, (New York, NY, USA), pp. 30-40, ACM, 2003.

[23] I. Khalil, S. Bagchi, and N. B. Shroff, ”MOBIWORP: Mitigation of the wormhole attack in mobile multi-hop wireless networks,” Ad Hoc Networks, vol. 6, no. 3, pp. 344-362, 2008.

[24] M. G. Zapata, ”Secure ad hoc on- demand distance vector (SAODV) routing,” in internet draft, Aug. 2001.

[25] A. Perrig, R. Szewwczyk, V. Wen, D. Culler, and D. Tygar, ”SPINS: Security protocol for sensor networks,” in Mobile Computing and Networking, (Rome, Italy), 2001.

[26] J. Eriksson, S. V. Krishnamurthy, and Faloutsos, ”Truelink: A practical countermeasure to the wormhole attack in wireless networks,” in Proceedings of the 14th IEEE International Conference on Network Protocols, ICNP 2006, November 12-15, 2006, Santa Barbara, California, USA, pp. 75-84, 2006.

[27] A. Perrig, R. Canetti, J. D. Tygar, and D. Song, ”Tesla: Timed efficient stream loss tolerant authentication, broadcast authentication protocol,” in CryptoBytes, pp. 2-13, 2002.

[28] Y.-C. Hu, A. Perrig, and D. B. Johnson, ”Wormhole attacks in wireless networks,” vol. 24, (Piscataway, NJ, USA), pp. 370-380, IEEE Press, Sep. 2006.

[29] A. P. Y. Hu and D. Johnson, ”Ariadne: A secure on-demand routing protocol for ad hoc networks,” in MOBICOM, (Atlanta), pp. 12-23, ACM, September 2002.

[30] S. Choic, D. Kim, D. Lee, and J. Jung, ”Wap: wormhole attack algorithm in MANETs,” in 2008 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, Taichung (Taiwan), 2008.

[31] Romit Roy Choudhury, Xue Yang, Ram Ramanathan, Nitin H. Vaidya, ”Using directional antennas for medium access control in ad hoc networks”, MobiCom ’02, (New York, NY, USA), ACM, 2002.

[32] Y.-B. Ko, V. Shankarkumar, and N. H. Vaidya, ”Medium access control protocols using directional antennas in ad hoc networks,” in (INFOCOM) Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE, vol. 1, pp. 13-21, IEEE, 2000.

[33] S. Capkun, L. Buttyan, and J.-P. Hubaux, and E. M. Belding-Royer, ”A secure routing Sector: Secure tracking of node encounters in multi-hop wireless networks,” in Proceedings of the 1st ACM Workshop on Security of Ad Hoc and Sensor Networks, SASN ’03, New York, pp. 21-32, ACM, 2003.

[34] R. C. Merkle, ”Protocols for public key cryptosystems,” in Proc. of the IEEE Symposium on Security and Privacy, 1980.

[35] S.-M. Jen, C.-S. Laih, and W.-C. Kuo, ”A hop count analysis scheme for avoiding wormhole attacks in MANETs,” Sensors, vol. 9, no. 6, pp. 5022- 5039, 2009.

[36] A. S. M. E. A. G. M. K. K. X. L. Mehdi Sookhak, Adnan Akhundzada and X. Wang, ”Geographic wormhole detection in wireless sensor networks,” PLoS ONE, vol. 10, Jan 2015.

[37] H. Chen, W. Chen, Z. Wang, Z. Wang, and Y. Li, ”Mobile beacon based wormhole attackers detection and positioning in wireless sensor networks,” International Journal of Distributed Sensor Networks, vol. 10, no. 3, 2014.

[38] L. Lu, M. J. Hussain, G. Luo, and Z. Han, ”PWORM: Passive and real-time wormhole detection scheme for WSNS,” International Journal of Distributed Sensor Networks, vol. 11, no. 11, p. 356382, 2015.

[39] K. Sanzgiri, B. Dahill, B. N. Levine, C. Shields,” protocol for ad hoc networks,” in Proceedings of the 10th IEEE International Conference on Network Protocols, ICNP 02, pp. 78-89, 2002.

[40] R. Maheshwari, J. Gao, and S. R. Das, ”Detecting wormhole attacks in wireless networks using connectivity information,” in INFOCOM 2007, 26th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, (Anchorage, Alaska, USA), pp. 107-115, May 2007.

[41] Y.-C. Hu, D. B. Johnson, and A. Perrig, ”SEAD: Secure efficient distance vector routing for mobile wireless ad hoc networks,” in Proceedings of the Fourth IEEE Workshop on Mobile Computing Systems and Applications, WMCSA 02, 2002.

[42] C. Perkins and P. Bhagwat, ”Highly dynamic destination-sequenced distance vector routing (DSDV) for mobile computers,” in SIGCOMM Conference on Communication Architecture, Protocols, and Applications, ACM, 1994.

[43] L. Chen, C. Liu, and H. Huang, ”Secure routing against wormhole attack and its formal verification based on timed colored petri net,” in Proceedings of the 11th ACM Symposium on QoS and Security for Wireless and Mobile Networks, Q2SWinet ’15, (New York, NY, USA), pp. 157-164, ACM, 2015.

[44] T. R. R. Revathi Venkataraman, M. Pushpalatha1 and R. Khemka, ”A graph-theoretic algorithm for detection of multiple wormhole attacks in mobile ad hoc networks,” International Journal of Recent Trends in Engineering, May 2009.

[45] L. A. Radha Poovendran, ”A graph-theoretic framework for preventing the wormhole attack in wireless ad hoc networks,” Wireless Networks (Springer), pp. 27-59, 2006

[46] Ilya Grigorik, ”Introduction to Wireless Networks: Performance of Wireless Networks, Chapter 5”, O’Reilly Media, Inc , 2013

[47] Anuj K. Gupta, Harsh Sadawarti, and Anil K. Verma, ”Review of Various Routing Protocols for MANETs”, International Journal of Information and Electronics Engineering, Vol. 1, No. 3, November 2011

[48] A. Chauhan and V. Sharma, ”Review of performance analysis of different routing protocols in MANETs,” 2016 International Conference on Computing, Communication and Automation (ICCCA), Noida, pp. 541-545, 2016

[49] Harminder Kaur1, Harsukhpreet Singh, Anurag Sharma, ”Geographic Routing Protocol: A Review”, International Journal of Grid and Distributed Computing Vol. 9, No. 2, pp.245-254, 2016

[50] R. Mudgal and R. Gupta, ”Study of various wormhole attack detection techniques in mobile ad hoc network,” 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, pp. 3748-3754, 2016

Automated Segmentation of Whole Cardiac CT Images based on Deep Learning

Rajpar Suhail Ahmed, Dr. Jie Liu*
School of Computer & Information Technology
Beijing Jiaotong University,
Beijing, China 100044

Muhammad Zahid Tunio
Department of Computer Systems Engineering
Dawood University of Engineering and Technology
Karachi, Pakistan

Abstract*—Segmentation of the whole-cardiac CT image sequence is the key to computer-aided diagnosis and study of lesions in the heart. Due to the dilation, contraction and the flow of the blood, the cardiac CT images are prone to weak boundaries and artifacts. Traditional manual segmentation methods are time-consuming and labor-intensive to produce over-segmentation. Therefore, an automatic cardiac CT image sequence segmentation technique is proposed. This technique was employed using deep learning algorithm to understand the segmentation function from the ground truth data. Using the convolution neural network (CNN) on the central location of the heart, filtering ribs, muscles and other contrasting contrast are not an obvious part of the removal of the heart area. Staked denoising auto-encoders are used to automatically deduce the contours of the heart. Therefore, nine cardiac CT image sequence datasets are used to validate the method. The results showed that the algorithm proposed in this paper has best segmentation impact to such cardiac CT images which have a complex background, the distinctness between the background and the target area which is not obvious; and the internal structure diversification. It can filter out most of the non-heart tissue part, which is more conducive to the doctor observing patient's heart health.

Keywords*—Cardiac CT; segmentation; deep learning; automatic location; contour inference

I. Introduction

Cardiac disease is the leading cause of death in Human. The heart as a substantial organ, having only Imaging data to understand the location of its internal lesions, is an effective means of noninvasive diagnosis of cardiac disease [1]. Cardiac CT image sequence segmentation is the key to the diagnosis of cardiac diseases using CT images. Therefore, improvement of the segmentation accuracy of whole-heart CT image sequences has got a major concentration of cardiac disease research [2].

Heart and other soft tissue images are hard to get segmented because of its target and background contrast are small as compared to a large scale of noise. The traditional approaches mainly obtain the edge points of the target manually by the anatomical knowledge and experience by the clinician [3]. These methods have a high accuracy but time-consuming, labor-intensive, and the doctor's experience surely affect the accuracy of the segmentation strongly. With the development of computer and image processing technology, the researchers have developed a series of methods which semi-automatically extract the target through input of specified calculation parameters and human-computer interaction. These methods can combine a prior knowledge such as anatomical knowledge, which let it be more popular. In recent years, some full automatic segmentation methods were developed, which do not need to enter any parameters and interaction, make the segmentation more efficient. However, the accuracy of the segmentation cannot be ensured and the calculation is extremely expensive [4].

In order to increase the precision and segmentation efficiency regarding the automatic segmentation algorithm, a novel method based on deep learning used for whole cardiac CT image sequences segmentation is proposed here. The procedure is done in three stages: First, the convolution neural network technique used to locate the center of the heart. Then the stacked denoising auto-encoder was used to automatically deduce the contours of the heart. Finally, the Gaussian smoothing and cardiac segmentation were employed to further improve the efficiency of the segmentation. Experiments display that the preparation method based on deep learning has high segmentation efficiency and accuracy.

II. Methods

The Architecture of the prepared segmentation method is shown in Fig. 1. The segmentation method mainly includes the following three steps: 1) Region of interest consisting the heart is located using a convolutional neural network [5]-[9]. 2) The cardiac contour is inferred using staked denoising auto-encoders, which trained to delineate the heart [10]-[16]. 3) Smoothing cardiac contour using Gaussian Smoothing Filter and using smoothed cardiac contour image to segment cardiac CT image [17]-[19]. Each step is individually trained during an offline training process to obtain its optimum values of parameter. After training, the automatic segmentation to learn deep learning network employee is used. This way, three-steps are described below.

A. Automatic Location

CT cardiac tissues of the heart and chest raw image sequence usually contain tissues. In order to reduce the impact of the surrounding tissue on the segmentation while improve the accuracy of segmentation, the first step of the algorithm is to locate the region of interest (ROI) containing the heart and filter out ribs, vertebrae and other interference. Fig. 2 shows the diagram of the convolution neural network to locate the specific location of the heart of the process.

**National Natural Science Foundation of China under Grant Nos. 81571836.*

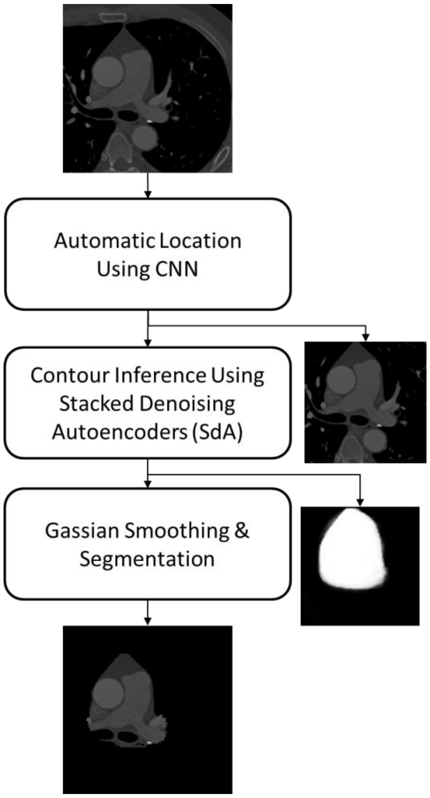


Fig. 1. The architecture of the developed algorithm.

In order to reduce the complexity of the algorithm and computational complexity, the original image size of 512×512 is down-sampled to 64×64, which is used as the input data of the neural network.

Then, the filter $F_l \in \Re^{11\times11}$, $b_0 \in \Re^{100}$ is employed to convolution operation with the input image to obtain a convolution characteristic map. I signify an input grayscale image having a size of 64×64, $I[i,j]$ represents the gray value of the point with the coordinate $[i,j]$ in the image. The convolution characteristic is calculated by this formula $C_l[i,j] = f\left(Z_l[i,j]\right)$, where

$$Z_l[i,j] = \sum_{k_1=1}^{11}\sum_{k_2=1}^{11} F_l[k_1,k_2]\, I[i+k_1-1, j+k_2-1] + b_0[l] \quad (1)$$

In the formula $1 \le i, j \le 54$, $l = 1,\ldots,100$. This results in 100 convolved features $Z_l \in \Re^{54\times54}$. Here, $x[i]$ signify the ith number of the vector x and $X[i,j]$ signify the number of ith row and jth column of the matrix X.

Next, the convolution feature map is sub-sampled by maximum pooling, and each pooled feature value calculates the maximum of the adjacent overlapping.

7×7 size convolution map area is calculated through (2).

$$P_l[i_1, j_1] = \max\left(\sum_{i=7i_1-6}^{7i_1}\sum_{j=7j_1-6}^{7j_1} C_l[i,j]\right) \quad (2)$$

For $1 \le i_1, j_1 \le 8$. The result in 100 reduced-resolution features, where $P_l \in \Re^{8\times8}$, $l = 1,\ldots,100$.

Finally, the pooled feature expands to the vector $p \in \Re^{6400}$, a logical layer that associates with cell output 1024. 32 x 32 size picture of the specific region of interest, logical regression layer will create a mask.

The output of the formula is $y_c = f\left(w_1 p + b_1\right)$, where $w_1 \in \Re^{1024\times8100}$ and $b_1 \in \Re^{1024}$ are trainable matrices.

Because the size of the original input CT image is 512×512, so the output mask image is up-sampled to the size of the original image, and finally, according to the mask image. To produce the ROI image for the next step, and the size is 400×400. It is necessary to train the network parameters to confirm the accuracy of the location.

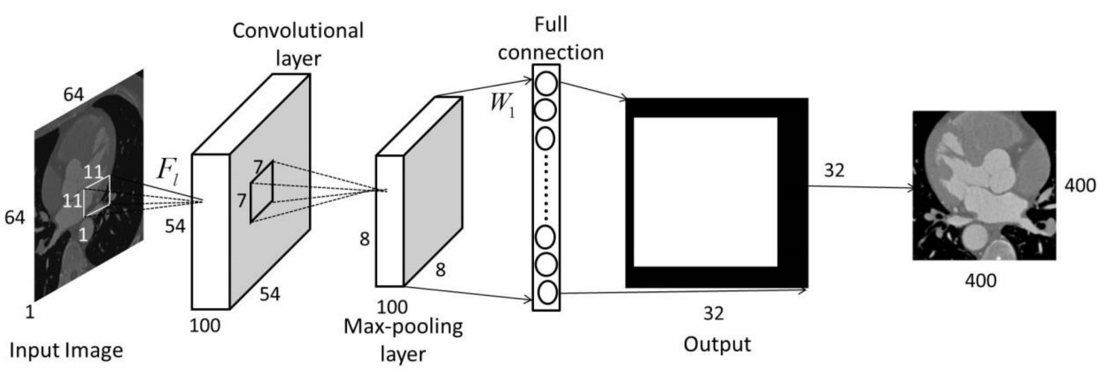


Fig. 2. The architecture of the automatic location for the heart in CT dataset.

The training and optimization of the parameters in location operation will introduce as follows:

1) Training Convolutional Network

The CNN training mainly focuses on obtaining the parameters $F_l, l = 1,...,100$ b_0 of the filter, and the optimal values of the other parameters w_1, b_1. In the case of large enough data sets, these parameters can obtain through the training of the ideal parameter values; therefore, in the normal case, the parameters are randomly initialized and then trained to obtain the desired parameter values. When using the unsupervised method of hierarchical pre-training depth network weights, random noise is introduced into the visual layer (e.g. the input layer of data) of the network, which is called denoising auto-encoder (dA). As shown in Fig. 3, the random noise is added to the sample x from the QD distribution instead of adding the Gaussian noise, the input value (as many as half of them) is cleared to zero with a certain probability，this is known as stochastic corrosion process. This successfully solves the problem of few data sets and avoids the resulting overfitting. At here the stochastic corrosion process ratio is 0.5. As shown in Fig. 4, the number of input and output units of dA is 121 and the number of hidden layer units is 100. Randomly select $N_1 = 10^4$ blocks of size 11×11 from the input image sequence, each block is expanded into a vector $x^{(i)} \in \Re^{121}$, $i = 1,..., N_1$ and is used as input to the dA.

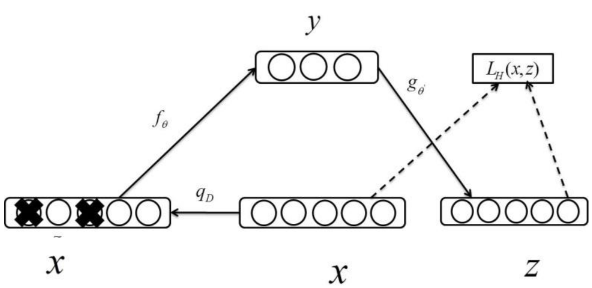


Fig. 3. The architecture of denoising auto-encoder.

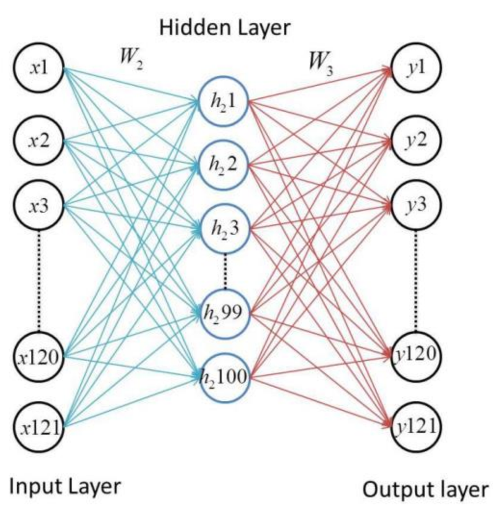


Fig. 4. Train denoising auto-encoder is to initialize filters.

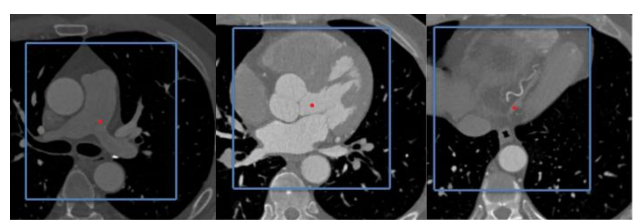
The weight parameter matrix between input layer and hidden layer is $W_2 \in \Re^{100\times121}$, and the weight parameter within the implicit layer and the output layer is $W_3 \in \Re^{121\times100}$. Hidden layer and output layer are calculated through the formula $a_2^{(i)} = f\left(w_2 x^{(i)} + b_2\right)$ and $y^{(i)} = f\left(w_3 a_2^{(i)} + b_3\right)$, here the data $x^{(i)\prime}$ and $a_2^{(i)\prime}$ are the stochastic corruption process of $x^{(i)}$ and $a_2^{(i)}$. The sigmoid activation function is $f(x) = 1/\left(1+e^{-x}\right)$, $b_2 \in \Re^{100}$, $b_3 \in \Re^{121}$ is the bias vector.

The purpose of the dA is to recompose $x^{(i)}$ at the output layer over the output of the hidden layer. Therefore, the input data is used as the mark data without requiring the actual mark data. The dA parameter optimization is achieved through the low-cost function.

$$J\left(w_2, b_2\right) = \frac{1}{2N_1}\sum_{i=1}^{N_1}\left|y^{(i)} - x^{(i)}\right|^2 + \frac{\lambda}{2}\left(\left\|w_2\right\|^2 + \left\|w_3\right\|^2\right) + \beta\sum_{j=1}^{k} KL\left(\rho \cdot \hat{\rho}_j\right) \quad (3)$$

The first term is the direct mean square error within the output $y^{(i)}$ and the ideal output $x^{(i)}$. Moreover, in order to prevent overfitting, the second terminology is added; a weight attenuation term whose purpose is to reduce the magnitude of the weights. At the same time in order to be able to learn more features from the input data, the hidden unit added sparse constraints. In this way, a dA network is built. Here, KL divergence limits the activation mean $\hat{\rho}_j = \left(1/N_1\right)\sum_{i=1}^{N_1} a_2^{(i)}[j]$, $j = 1,...,100$ of the hidden layer, and ρ_j is usually small in order to maintain equilibrium with the discrete parameter ρ .

The weight attenuation coefficient λ and the discrete coefficient controls the relative of the three conditions is the importance of the cost function. The optimal values are $\lambda = 10^{-4}$, $\rho = 0.1$, and $\beta = 3$. After the dA is trained, W_2 is the initial value of $F_l \in \Re^{11\times11}$, $l = 1,...,100$ and $b_0 = b_2$ for the next step.



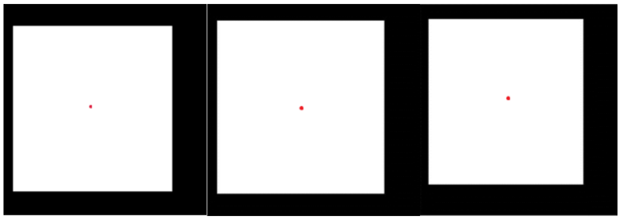


Fig. 5. Input images (top) and the use of automatic location for training related binary mask image (bottom).

The feed-forward operation, then (1) and (2) through to the output layer is performed, and the output layer is already trained through the minimal cost function.

$$J\left(w_1, b_1\right) = \frac{1}{2N_2}\sum_{i=1}^{N_2}\left|y_c^{(i)} - I_{roi}^{(i)}\right|^2 + \frac{\lambda}{2}\left(\left\|w_1\right\|^2\right) \qquad (4)$$

Where I_{roi} is described data matching to the ith input image and N_2 is the number of training images? The described data of the output layer is a binary mask image which generates through the described manual. The binary mask is a black background and the white foreground image matching to the ROI.

As shown in Fig. 5, the midpoint of the foreground corresponds to the midpoint of the heart contour, the binary mask is down-sampled to 32×32 and expanded into vectors I_{roi} for the network training.

Finally, the fine tuning of the entire network achieved is by minimizing the cost function represented by (5).

$$\begin{aligned} J\left(F_l, b_0, w_1, b_1\right) = {} & \frac{1}{2N_2}\sum_{i=1}^{N_2}\left|y_c^{(i)} - I_{roi}^{(i)}\right|^2 \\ & + \frac{\lambda}{2}\left(\left\|w_1\right\|^2 + \sum_{l=1}^{100}\left\|F_l\right\|^2\right) \end{aligned} \qquad (5)$$

The cost function can be minimized by back propagation aggregate, where $\lambda = 10^{-4}$, the training method only runs once.

B. Contour Inference

Unlike the literature [6], to train and use the stacked dA to infer the contour of the heart is shown in Fig. 6. The Stacked dA has an input layer, two hidden layers, and an output layer. The sub-image obtained by the convolution network is down-sampled and expanded into a vector $x_s \in \Re^{6400}$ and feedback to the input layer. The output of the two hidden layers is calculated by $a_1 = f\left(w_4 x_s{}' + b_4\right)$ and $a_2 = f\left(w_5 a_1{}' + b_5\right)$. Finally, the output layer computes $y_s = f\left(w_6 a_2{}' + b_6\right)$ to produce a binary mask image with a pixel value of 0 or 1, where the heart segment corresponds to 1. Here, the parameters $w_4 \in \Re^{100\times6400}$, $b_4 \in \Re^{100}$, $w_5 \in \Re^{100\times100}$, $b_5 \in \Re^{100}$, $w_6 \in \Re^{6400\times100}$, $b_6 \in \Re^{6400}$ are training matrices and vectors, respectively, obtained by the training program, the data $x_s{}'$, $a_1{}'$ and $a_2{}'$ are the result of stochastic corruption process of data x_s, a_1 and a_2.

1) Training stacked-dA

Training divides into two steps, similar to previous convolution neural network training, include pre-training and fine-tuning. Because of the limited data available in the experiment, a layer-to-layer approach is used for pre-training parameters. This method effectively reduces the overfitting. In the layer-to-layer pre-training process, parameters w_4, w_5 of the staked-dA are achieved layer-to-layer with no described data. The parameter is achieved applying described data. This Valuation is as follows.

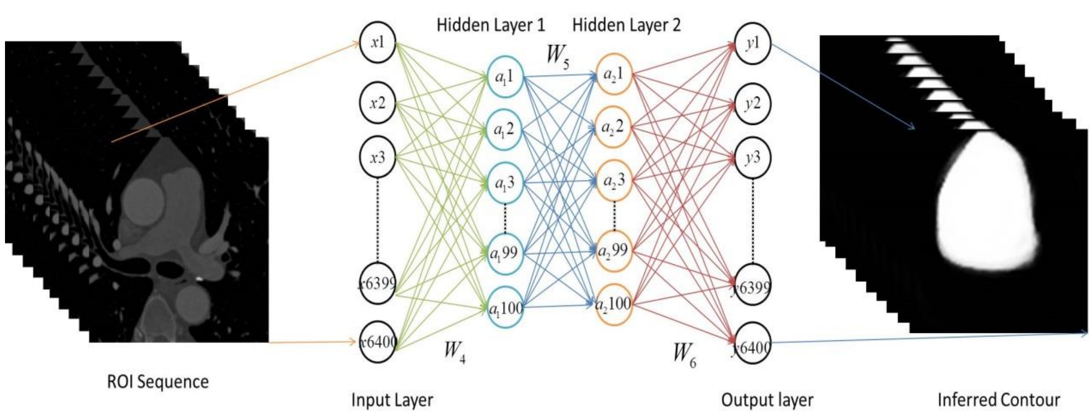


Fig. 6. Stacked dA for inferring the cardiac contour. An input is sub-image and the binary mask is an output.

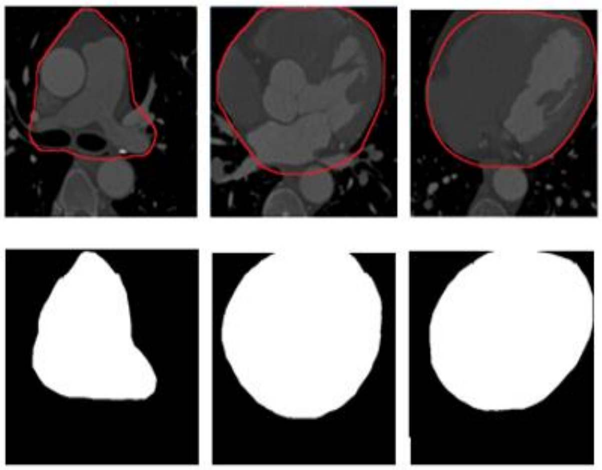


Fig. 7. Typical sub-image with manual heart segmentation and the matching binary mask applied training the stacked denoising auto-encoder.

Initially, an input layer and from the H_1 hidden layer is distinct from the stacked-dA by calculating the output layer of the similar size as an input layer to form a sparse dA (similar to Fig. 4). An unsupervised way to get the w_4 parameter is trained to sparse dA, as described in Section 2.1.1. The optimal parameter is set as $\lambda = 3\times10^{-3}$, $\rho = 0.1$, and $\beta = 3\times10^{-1}$.

The training input and output data of reserve dA are sub-images of the size 400×400 drawn out from the full-size training data, image midpoint matching to the center of the heart. The input image is down-sampled to the size of 80×80 to be compatible with an input size 6400 of the stacked-dA. Once training of reserve dA is complete, it's an output layer that is deserted. An output of the dA hidden layer unit is used as input to the next hidden layer H_2.

Next, separate the H_1 and H_2 from the stacked dA and add a layer of the same size as the H_2 output to compose a sparse dA. In addition, the next sparse dA is used to get w_5, with no described data. The input data is same as the input data of the previous dA hidden layer.

An output of the last hidden layer is the input of the last layer, it is trained in the unsupervised mode to obtain w_6. The cost function used to train the last layer is calculated as follows:

$$J(w_6, b_6) = \frac{1}{2N_2}\sum_{i=1}^{N_2}\left|y_s^i - I_{lv}^i\right|^2 + \frac{\lambda}{2}\|w_6\|^2 \tag{6}$$

Where, $I_{lv}^i \in \Re^{6400}$ is the described dataset matching to the *i*th image. The described dataset is a binary mask image generated by hand segmentation. Fig. 7 depicts three input images and a matching binary mask image for the training of the stacked-dA. It should note that the binary mask image is expanded into the vector I_{lv} during optimization.

The purpose of layer-by-layer pre-training is to obtain the appropriate initial values for parameters w_4, w_5, w_6. Finally, the complete architecture is fine-tuned through reducing the cost function as shown in (2).

$$J(w_4, w_5, w_6, b_4, b_5, b_6) = \frac{1}{2N_2}\sum_{i=1}^{N_2}\left|y_s^i - I_{lv}^i\right|^2 + \frac{\lambda}{2}\left(\|w_4\|^2 + \|w_5\|^2 + \|w_6\|^2\right) \tag{7}$$

The minimization process of the cost function uses a supervised back-propagation algorithm. The automatic detection, as well as the training process, perform only once. The optimal parameter is $\lambda = 10^{-4}$.

C. Smoothing and Segmentation

The final step of the algorithm is smoothing and segmentation, because the contour inference network output image size is 80×80. It has a clear jagged edge after the image is up-sampled to the size of 400×400. To resolve this issue, a two-dimensional Gaussian filtering algorithm is followed to smooth the contour inference image which is up-sampled to the size of 400×400, the algorithm formula is:

$$G(x, y) = \frac{1}{2\pi\sigma^2}e^{-\frac{(x^2+y^2)}{2\sigma^2}} \tag{8}$$

Finally, a binary mask image with contour inference is obtained after high-pass filtering and binarization. The size of the smoothing filter is 10, after the experiment, when σ is 2.5, the smoothing effect is the best. Finally, the heart contours are registered by a high pass filter, and the segmentation result will be got.

III. EXPERIMENT

In the proposed method the training dataset from TaiZhou City People Hospital was used. The training dataset consisted of 9 groups, each group of dataset matching to a patient's cardiac CT images, the number of each group ranging from 241 to 344. The size of cardiac CT image is 512×512. The number of available experimental image of data is 2500.

In order to enhance the efficiency of the segmentation algorithm, the training dataset are linearly interpolated, and the image data is expanded to 5000 pieces. The described dataset consists of two parts; the described image is binary mask image. The first part is for convolution neural network training, the center of the region of interest corresponds to the cardiac center of the cardiac CT image. The second part is for stacked denoising auto-encoder training, the foreground white outline of the described image corresponds to the contour of the heart. The training is performed only once.

Considerable overfitting may happen in deep learning network due to the large number of parameters to be learned. Mainly focus a lot on the network to intercept the problem overfitting, to find the solution to this problem, we adopted layer-wise pre-training, l_2 regularization and denoising constraints as detailed in Section 2.1 and 2.3. The challenge was the deficit of data and the use of layer-wise pre-training and denoising constraints is very useful. Hence, also kept the number of the layer in the location networks do not go before three and the number of the layer in the contours inferring do not go before four to make sure that the number of parameters is manageable. While training of the proposed method, cross-validation was accomplished through dividing the training dataset into 8 subjects for training and 2 subjects for validation and also early stopping to supervise and intercept overfitting. Furthermore, the training dataset is increased as described in the beginning. The hyperparameters of the networks, the number of layers and units, filter and pooling size, are persistent empirically throughout the cross-validation method.

The performance of the segmentation algorithm is evaluated by its accuracy, regional statistical characteristics, and reliability [20]. The manual segmentation result is taken as the reference standard, and the performance of the segmentation algorithm can realize by calculating the segmentation algorithm accuracy. Accuracy is the degree of consistency between the segmentation results and the manual segmentation results [21], where expressing the accuracy of the segmentation by calculating the percentage of correct classified pixels as a percentage of the number of reference pixels [22]. The four measures of accuracy are True Positives (TP), False Positive (FP), True Negative (TN) and False Negative (FN), the matching percentages are TPR, FPR, FNR, and TNR. The correct rate (TTR) is calculated by (9).

$$\begin{aligned} TPR &= \frac{TP}{T}\times\frac{100}{100} \\ FNR &= \frac{FN}{T}\times\frac{100}{100} \\ FPR &= \frac{FP}{I-T}\times\frac{100}{100} \\ TNR &= \frac{TN}{I-T}\times\frac{100}{100} \\ TTR &= (1-\frac{(FP+FN)}{\mathrm{T}})\times\frac{100}{100} \end{aligned} \tag{9}$$

In the recent study, this program designed in MATLAB 2015 a, accomplished on a Dell OptiPlex 7020 machine, with Intel(R) Core(TM) i5-4150 CPU 3.5GHz, 16GB RAM, on a 64-bit Windows 10 platform. The process was training working on the training dataset and tested on the testing dataset, the testing dataset includes three sequences of cardiac CT images.

IV. Results

A. Illustrative Results

In order to determine the aspect of every step, the effect of CNN (heart location, Step 1) for three image segments close the base/middle and the point of the heart as shown in Fig. 8, illustrates the location results of the different cardiac segment. Fig. 9 illustrates the location center of a whole sequence of cardiac CT images.

The deep learning method (contours inference, Step 2) outcome for three image segments close the base/middle and the point of the heart is shown in Fig. 10.

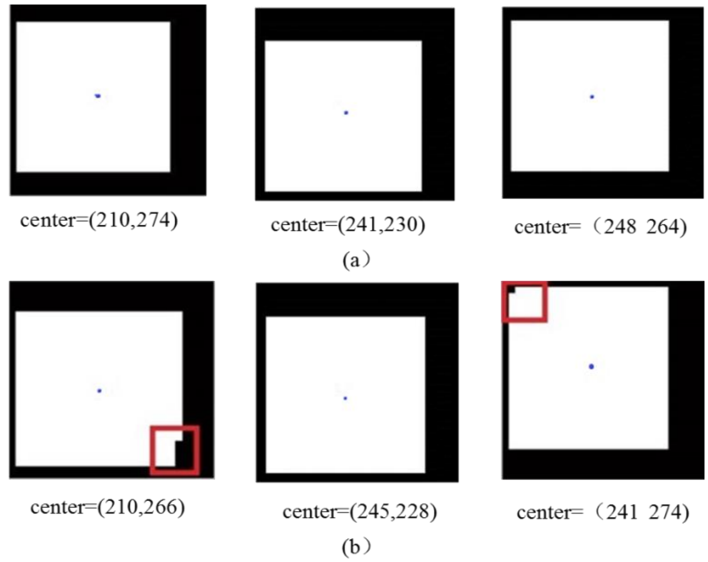


Fig. 8. CNN and manual location result from base to point of the heart.

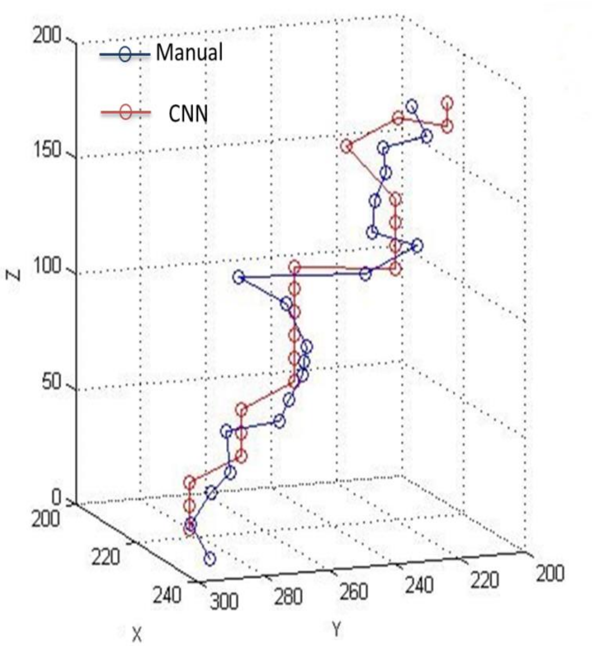


Fig. 9. Location centers from base to the apex of a whole cardiac CT image sequences.

The region growing (RG) segmentation algorithm is used as the contrast algorithm. Fig. 11 illustrates the original, region growing automated, and manual whole heart segmentation results for a typical cardiac CT dataset.

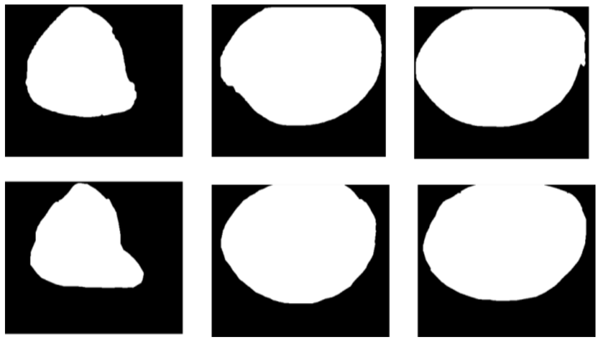


Fig. 10. Automatic (top) and manual (bottom) contours inference from base to the point of the hear.

From the base to Apex, in Fig. 11, original images (a), (b), (c) are shown in the first row, region growing segmentation results (d), (e), (f) are shown in the second row, automatic segmentation results (h), (i), (j) are shown in the third row, and the ground truth manual segmentations (k), (l), (m) are shown in the last row.

B. *Quantitative Results*

Table I shows the evaluation of method prepared and region growing segmentation results, from which we can understand that the method prepared in this paper has a higher TTR (average 93.77%), TNR (average 98.15%) and lower FPR (average 1.84%); it is almost completely close to the manual segmentation.

Table II compare the running time of method proposed in this paper and RG spent in the segmentation. The average time of proposed method is 0.1244s, it is far less than the average time of RG (4.2385 s).

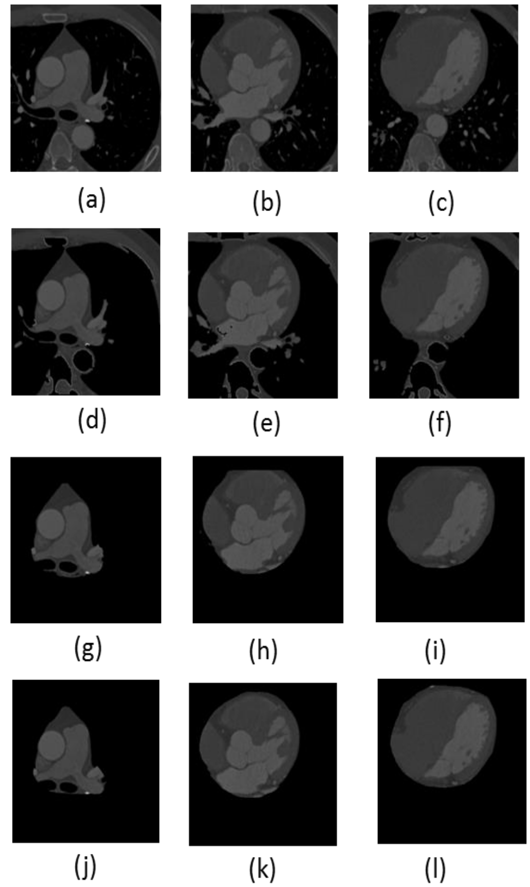


Fig. 11. Segmentation results of heart.

TABLE I. OBJECTIVE EVALUATION OF DEEP LEARNING AND REGION GROWING SEGMENTATION METHODS

Segment	Method	TPR (%)	FPR (%)	TNR (%)	FNR (%)	TTR (%)
Base	DL	98.67	1.30	98.69	1.32	92.23
	RG	99.43	25.03	74.97	0.56	-23.86
Mid	DL	97.78	1.53	98.46	2.22	94.8
	RG	99.56	23.29	76.71	0.43	54.26
Apex	DL	99.25	2.69	97.31	0.74	94.30
	RG	99.95	19.65	80.35	0.04	63.80
Average	DL	98.56	1.84	98.15	1.42	93.77
	RG	99.65	22.65	77.34	0.34	31.4

TABLE II. COMPARISON OF RUNNING TIMES

Segment	Base	Mid	Apex	Average
DL(s)	0.1227	0.1229	0.1276	0.1244
RG(s)	1.8297	4.6046	6.2820	4.2385

Approximated lapsed times of the training method were as follows: Training denoising auto-encoder to get filters: 123.4s, training convolutional neural network: 4.1 h, training stacked denoising auto-encoder: 3.2h. Once trained, the elapsed times of the heart segmentation in a typical CT image were as follows: ROI location (convolution, pooling, and the logistic regression):0.23s, contour inferring (stacked -dA):0.12s, and segmentation (smoothing and segmentation): 0.15s.

V. DISCUSSION AND CONCLUSION

In this research, we suggested and validated an automated heart segmentation process placed on deep learning algorithms. In this issue, cover the localization, contour inferring and segmentation tasks. Convolutional neural networks were preferred for the localization and extracting an ROI through which they are matched to special translation and variation in scale and pixels' density [23]. Also, cover a stacked dA for contour inferring as its simplicity in training and implementation still shown to become effective in different vision tasks [24]. Meaningfully, a perfect deep learning network is desired. However, this was impossible due to the limited number of training dataset. Hence, we kept the number of the layer in the location networks not to go beyond three and the number of the layer in the contours inferring not to go beyond four to make sure that the number of parameters are manageable. Also, employ the linear interpolation to enlarge the training dataset.

As seen in Fig. 3 and 4, the denoising auto-encoder and layer-wise pre-training were adopted in CNN training to ensure the accuracy of localization. And the outcome of CNN in Fig. 8 and 9 also verify the effectiveness of the measures. The CNN and manual location of the heart are pretty close and the trend of location center is consistent, the average distance error is below 10 pixels. Although, there is still some error in Fig. 9, but the error is within the allowable range.

From Fig. 10, this could be recognized that the heart was accurately segmented from the base to point. In Fig. 11, the first row is the original cardiac CT image, the slices near the ground truth from a small base cane showed leakage. This position is one of the most difficult cases that due to the fuzziness of the heart border, contour tends to leak to surrounding areas in normal segmentation methods. In this study, the used staked-dA is perfect to solve this problem.

The quantitative results in Table I show that the method prepared in this study has a pretty high average accuracy (93.77%) of segmentation. Table II compare the average running time of method to the region growing algorithm, the result showed that the average running time of methods proposed is much faster, and can complete fully automatic segmentation tasks of whole heart CT image sequence in a short time.

In summary, a novel segmentation process for a fully automatic cardiac CT is presented. The process places on deep learning algorithm broke down into automatic localization and inferring the heart contour. And the contour inferred was smoothing by the Gaussian filter. It could be recognized that the segmentation process prepared in this paper has better segmentation efficiency than the regional growth segmentation method and successfully filters out most of the interfering tissue; the result is the closest to the manual segmentation and can replace the manual segmentation and realize the Automatic Segmentation of cardiac CT image sequences.

In the future, other depth learning algorithms may be introduced to realize heart image segmentation. In the medical image field, the application of 4D images is the mainstream trend. The basic data of the 4D image is a large volume of information contained in the three-dimensional volume data. The existing segmentation algorithm can be extended to the three-dimensional space to directly segment the volume data.

ACKNOWLEDGMENTS

This work is partially supported by National Natural Science Foundation of China under Grant Nos. 81571836. Authors would also like to thank the Taizhou City People Hospital and Beijing Jiaotong University for giving us the opportunity to do some furthermore research on CT & MRI image feature evaluation.

REFERENCES

[1] Organization. W H. Global status report on non-communicable diseases 2014.[J]. Women, 2011, 47(26):2562-2563.

[2] S.Mendis, P.Puska, B.Norrving. Global Atlas on cardiovascular disease prevention and control.World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization.2011, pp.3-18.

[3] J.Rogowska.”Overview and fundamentals of medical image segmentation”. Handbook of Medical Imaging Processing and Analysis. Pp. 69-85, 2000.

[4] N.R Pal,S.K.Pal,”A Review on Image Segmentation Techniques”. Pattern Recognition,vol. 26, pp:1277-1294, 1993.

[5] Lecun Y, Kavukcuoglu K, Farabet C. Convolutional Networks and Applications in Vision[J]. 2010, 14(5):253-256.

[6] Szegedy C, Toshev A, Erhan D. Deep Neural Networks for object detection[J]. Advances in Neural Information Processing Systems, 2013, 26:2553-2561.

[7] Sermanet P, Eigen D, Zhang X, et al. Over Feat: Integrated Recognition, Localization and Detection using Convolutional Networks[J]. EprintArxiv, 2014.

[8] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in Neural Information Processing Systems, 2012, 25(2):2012.

[9] Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 79(10):1337-1342.

[10] Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(8):1798-828.

[11] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising auto-encoders[C]// International Conference. 2008:1096-1103.

[12] Bengio Y. Learning Deep Architectures for AI[J]. Foundations & Trends® in Machine Learning, 2009, 2(1):1-55.

[13] Baldi P, Guyon G, Dror V, et al. Auto-encoders, Unsupervised Learning, and Deep Architectures Editor: I[J]. Journal of Machine Learning Research, 2012.

[14] Deng L, Yu D. Deep Learning: Methods and Applications [J]. Foundations & Trends® in Signal Processing, 2013, 7(3):197-387.

[15] Vincent P, Larochelle H, Lajoie I, et al. Stacked Denoising Auto-encoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion.[J]. Journal of Machine Learning Research, 2010, 11(12):3371-3408.

[16] Avendi M R, Kheradvar A, Jafarkhani H. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI.[J]. Computer Science, 2016, 30:108-119.

[17] Xiao K, Ho S H, Salih Q. A Study: Segmentation of Lateral Ventricles in Brain MRI Using Fuzzy C-Means Clustering with Gaussian Smoothing[C]// Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, International Conference, Rsfdgrc 2007, Toronto, Canada, May 14-16, 2007, Proceedings. DBLP, 2007:161-170.

[18] Xiao K, Ho S H, Salih Q. A Study: Segmentation of Lateral Ventricles in Brain MRI Using Fuzzy C-Means Clustering with Gaussian Smoothing [M]// Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Springer Berlin Heidelberg, 2007:161-170.

[19] Izquierdo M. E, Ghanbari M. Texture smoothing and object segmentation using feature-adaptive weighted Gaussian filtering[C]// Telecommunications Symposium, 1998. ITS '98 Proceedings. SBT/IEEE International. 1998:650-655 vol.2.

[20] Chalana V, Kim Y. A methodology for evaluation of boundary detection algorithms on medical images.[J]. IEEE Transactions on Medical Imaging, 1997, 16(5):642-52.

[21] Zhang Y J. A survey on evaluation methods for image segmentation ☆ [J]. Pattern Recognition, 1996, 29(8):1335-1346.

[22] Zhang Y. A Classification and Comparison of Evaluation Techniques for Image Segmentation[J]. Journal of Image & Graphics, 1996.

[23]]Bruna J, Zaremba W, Szlam A, et al. Spectral Networks and Locally Connected Networks on Graphs[J]. Computer Science, 2014.

[24] Vincent P, Larochelle H, Lajoie I, et al. Stacked Denoising Auto-encoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion[J]. Journal of Machine Learning Research, 2010, 11(12):3371-3408.

Energy Efficient Algorithm for Wireless Sensor Network using Fuzzy C-Means Clustering

Abhilasha Jain
CSE Department
GZS Campus College of Engineering and Technology
Maharaja Ranjeet Singh, PTU, Punjab, India

Ashok Kumar Goel
ECE Department
GZS Campus College of Engineering and Technology
Maharaja Ranjeet Singh, PTU, Punjab, India

***Abstract*—Energy efficiency is a vital issue in wireless sensor networks. In this paper, an energy efficient routing algorithm has been proposed with an aim to enhance lifetime of network. In this paper, Fuzzy C-Means clustering has been used to form optimum number of static clusters. A concept of coherence is used to eliminate redundant data generation and transmission which avoids undue loss of energy. Intra-cluster and inter-cluster gateways are used to avoid nodes from transmitting data through long distances. A new strategy has been proposed to select robust nodes near sink for direct data transmissions. The proposed algorithm is compared with LEACH, MR-LEACH, MH-LEACH and OCM-FCM based upon lifetime, average energy consumption and throughput. From the results, it is confirmed that the performance of the proposed algorithm is much better than other algorithms and is more suitable for implementation in wireless sensor networks.**

Keywords*—*WSN; clustering; sleep-awake; virtual grids; multi-hop; routing

I. Introduction

Wireless Sensor Networks (WSNs) consist of thousands of micro-sized low power Sensor Nodes (SNs) randomly deployed in the Sensor Field (SF). These SNs sense local environmental statistics, aggregate and communicate sensed information to sink. For each operation SN consumes its battery power. As SFs are hostile in most of the applications of WSN, it is not possible to replace batteries of SNs [1]. In order to enhance lifetime of network, the routing algorithms in WSN mainly focus on energy efficiency.

Clustering is one of the most effective techniques in routing to preserve the energy of the network. The whole network is organized into small groups of SNs called clusters. In each cluster, one node is elected as Cluster Head (CH) which performs the task of collecting data from all Cluster Member (CMs) nodes, aggregation of data and forwarding it to the sink directly or in multi-hop manner [2], [3]. The aggregation may or may not be perfect depending upon the relation between the sensed data. Perfect aggregation means many k-bit data packets are compressed to a single k-bit packet. Perfect-fusion, a simple technique has attracted many researchers' interest [2], [4]. LEACH (Low Energy Adaptive Clustering Hierarchy) [2], EEUC (Energy Efficient Unequal Clustering) [3], IB-LEACH (Intra-Balanced LEACH) [5], MR-LEACH (Multi-hop Routing LEACH) [6], MH-LEACH (Multi-Hop LEACH) [7], EMRP (Energy efficient multi-hop routing protocol) [8], ACEEC [9], OCM-FCM (Optimal Clustering Mechanism Fuzzy-C Means) [10], SEECP (Stable Energy Efficient Clustering Protocol) [11] and MLRC (Multi-Level Route-aware Clustering) [12] are some of the popular clustering protocols. In these protocols, all nodes in the network actively sense from the environment and continuously generate data. The nodes which are placed close to each other tend to generate redundant data due to phenomenon of coherence. A major amount of network energy is wasted in transmitting redundant data to sink which directly affects the network lifetime. Protocols like Span [13] and LEACH-SM (LEACH Spare Management) [14] keep a small subset of SNs active in such a way that these nodes cover the whole network while other nodes in the network are kept in sleep mode. In [15] an energy efficient sleep scheduling has been proposed in order to maintain network coverage and connectivity. Consequently, the energy consumed per round is reduced, thus, increasing the WSN lifetime.

Major part of energy is consumed by SNs in transmitting data which is proportional to the distance between sensor nodes raised to power $\beta(\beta \geq 2)$. Therefore, long-distance transmissions should be minimized for optimizing the usage of network energy [16]. Multi-hop data transmission between Gate Way nodes (GWs), CHs and sink can be used to reduce energy consumption, but SNs (may be CHs or GWs) near sink are more probable to be selected for transmitting the data to sink, resulting in their early death and thus, effecting the overall performance of the network.

In many clustering algorithms, number of clusters in each round is not fixed and employs poor CH election and cluster formation techniques. As a result, total inter and intra cluster distance becomes large, resulting in high energy consumption of the network. Therefore, soft computing techniques can be beneficially employed in cluster based routing protocols [16], [17]. Fuzzy-C means clustering groups the SNs based upon their degree of membership in each cluster. The aim is to minimize the sum of distances between the SNs and the centroid of their cluster. OCM-FCM uses Fuzzy-C means clustering to form more uniform and correlated clusters in order to improve the lifetime of WSN [10].

In this paper, a multi-hop efficient routing algorithm for WSN has been proposed**.** The algorithm emphasizes to optimize the energy usage in the network in order to enhance the network lifetime. It eliminates redundant data generation in the network by keeping only one node active from the set of nodes that are sending redundant data and also renders full network coverage. Initially, the whole network is divided into optimal number of static clusters using FCM. CH in each cluster changes deterministically during the network operation. In order to reduce long distance transmissions and balance load, multi-hop routes are established using gate way nodes. In multi-hop routing, the nodes which lie near the sink are more probable to be selected for establishing direct link to the sink despite of having lower residual energy. Therefore, SNs near the sink will die soon and create energy hole in the network. To avoid it, a strategy is designed which selects the nearest eligible SN for transmitting data to sink, therefore consuming least amount of energy. It also ensures that this SN has sufficient residual energy so that it may not become dead.

The remainder of the paper is organized as follows: Section 2 gives brief review of the related work. Network model and assumption made for the proposed work has been presented in Section 3. Section 4 describes the proposed algorithm for an energy efficient multi-hop adaptive routing using FCM clustering. The results and their comparison with other algorithms have been discussed in Section 5. Finally concluding remarks are given in Section 6.

II. Related Research

Many clustering protocols have been explored in literature for WSN in the last few years to increase network lifetime. LEACH is forerunner decentralized clustering protocol uses random rotation of CH in each round to evenly distribute the load in the network. CHs are responsible to transmit cluster information to the sink [2]. CHs far away from the sink consume lot of the energy to transmit data to sink. Randomized selection of CHs in each round adds extra overhead to the network. The algorithm does not assure even distribution of CHs in the network which may result into uneven energy consumption in different parts of the network. Moreover, the clusters formed in each round may or may not be correlated which increases the intra cluster distances. Therefore, LEACH is not suitable for large scale networks.

IB-LEACH is an improvement over LEACH [5]. CH selection of method is complex and involves lot of computations. Some nodes in the network are overburdened.

MH-LEACH and MR-LEACH are the variants of LEACH, which follow the same CH selection methods as proposed in LEACH [6], [7]. These algorithms establish the multi-hop routes to transmit data to the sink and reduce number of long distance transmissions in the network. CH nodes which are near the sink (inner CH) are responsible for transmitting data to the sink and other CHs (outer CH) will transmit their data to the nearby CH. When sink lies outside the sensor field, inner CH will consume higher energy than outer CHs and may die soon.

EEUC and EMRP divide the network into unequal sized clusters [3], [8]. The clusters that are formed near the sink are smaller in size as compared to clusters far away from the sink. There is an additional overhead involved for CH selection in each round. Nodes near sink are responsible to transmit data directly to sink even if they have low energy.

ACEEC and THCEEC [9] are the centralized routing algorithms which results in better network lifetime as compared to the conventional distributed routing protocols like LEACH. The algorithm divides the whole network in regions which act as static clusters for WSN. The sink finds CH in each region and efficient path to route region data to sink. This centralized scheme is helpful in planning power-aware, well-organized routes in the network. Network status needs to transmit to the sink in each round which increases energy consumption in all SNs and overall performance of network degrades.

Soft computing methods like Fuzzy C-Means clustering (FCM), K-means clustering, Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) are recently used to overcome problems of non-correlated clusters [10], [17]. It helps to prolong the life of WSN. However, all these methods run at centralized node i.e. at sink and need physical location of SNs. In OCM-FCM reduces the cluster formation overhead on SNs as sink divides WSN in K optimal number of static clusters. After first round, the current CH selects a highest energy node in the cluster and elects it as CH for the next round. This algorithm uses multi-hop technique by finding Secondary Cluster Heads (SCHs) from already chosen CHs if sink is far away from the sensor field [18], [19]. It will result in reduction of long distance transmissions. The drawbacks of the algorithm are: 1) Poor CH selection; 2) CHs are overburdened as compared to other SNs; 3) SCHs consume more energy as compared to primary CHs. In all the algorithms discussed, all nodes are regularly sensing and transmitting data to sink. Thus, large amount of network energy is wasted in redundant data transmissions which need to be reduced.

III. Network Model and Assumptions

Total N sensor nodes in WSN are randomly deployed in of X m by Y m rectangular sensor field. The assumptions made for the implementation of the proposed work are as given under:

1) All nodes are stationary and have unique ID.

2) It is assumed that all SNs find their locations either through GPS system or using some localization algorithm.

3) All nodes are static and have same initial energy E_{init}

4) All the active sensor nodes regularly generate information from the environment.

5) All SNs are capable to perform data aggregations in perfect fusion mode.

6) In the cluster formation process, each SN can be a member of only one cluster and can act as CH for the same cluster.

7) The sink has unlimited energy and computational resources.

8) First order radio energy model described in [2] is utilized for sensor nodes during their communications procedure.

IV. Proposed Energy Efficient Algorithm

The paper proposes an energy efficient algorithm for WSN using multi-hop routing. Initially it runs a network dimensioning phase where sink divides the whole network into desired number of static clusters using FCM technique and turns off all SNs producing redundant data. Thereafter, it organizes the network operation into a series of rounds where each round is divided into network setup phase and network communication phase. In network set up phase, new CHs are elected in static clusters if required and energy efficient multi-hop paths are established from each SN to the sink. The network communication phase deals with intra-cluster and inter-cluster multi-hop routing and data transmissions. In network communication phase every CH allocates TDMA slots to CMs in each cluster and for inter-cluster communication is achieved through CDMA codes.

A. Network Dimensioning Phase

a) Division of SF in granules: WSN is initially divided uniformly into granules. The size of granule may vary according to degree of coherence in sensed data. In a SF, if information to be sensed is more coherent, then size of granule may be taken large and vice-versa. The entire SF is divided into n_g granules where n_x granules are along x-axis and n_y granules along y-axis. The size of granule is x_w by y_w subject to constraint that x_w and y_w are integer multiple of width and height of SF respectively. Any i^{th} SN with co-ordinates (x_i, y_i) finds its associated granule ID using formula given as under:

$$GranID = round\left(\frac{x_i}{x_w}+0.5\right)+\left(round\left(\frac{y_i}{y_w}+0.5\right)-1\right)\times y_n \quad \ldots\ldots(1)$$

All SNs save this information for further use.

b) Initial Active-Sleep Methodology: Once SNs find their associated granule ID, each SN will broadcast a message $S(NodeID, GranID)$ to all SNs in its radius R_{com}. Where R_{com} is equal to the diagonal of a granule and is given by following equation:

$$R_{com} = \sqrt{x_w^2 + y_w^2} \quad \ldots\ldots\ldots (2)$$

Initially, SNs do not need to broadcast their residual energy as all SNs have same E_{init}. Each SN saves $NodeID$ of the received message if it belongs to its granule node and discards otherwise. In this way, all SNs maintains $NodeID$ and E_{resi} of its granule SNs. Once, all SNs save information of their associated $GranID$, granule SNs. SN with highest node ID in a granule will activate itself and broadcast its location and ID information to the sink. Remaining nodes in the granule shall be in sleep mode.

c) Clusters Formation: At the end of active-sleep process, the sink will receive location information of all n active nodes. The sink first calculates their associated *GranIDs* using (1), then divides whole WSN into *K* optimal number of clusters using FCM clustering algorithm. The SN is associated with the cluster in which it has the highest membership value based upon its distance from the centroids of the clusters. The sink broadcasts messages $S(ClusterID, <List\ of\ associated\ GranIDs>)$ for each cluster to all SNs. SNs belonging to $GranID$ specified in the message will save this message in its cluster table.

d) Initial Selection of Cluster Heads: The sink divides the whole network into *K* clusters. Thereafter, the sink evaluates the fitness of all SNs in each cluster on the basis of an objective function given by (3). The SN with highest fitness value will be elected as CH in each cluster. The sink broadcasts this information to each cluster.

$$f(i) = \sum_j \frac{1}{dis(i,j)^{\beta}}, \forall j \in c^{th}\ cluster, j \neq i, 1 \le c \le K \qquad (3)$$

where $f(i)$ is the fitness of i^{th} member of c^{th} cluster. $dis(i, j)$ is the Euclidean distance between i^{th} and j^{th} member of c^{th} cluster. The value of β is taken as 2 for $dis(i, j)$ less than threshold distance, otherwise 4.

e) Logical Region Division: In many multi-hop routing algorithms proposed by various researchers, SNs nearer to sink carry heavier traffic loads and are responsible for routing data to the sink. Therefore, SNs around sink deplete their energy at faster rate and *may* create energy hole near the sink. To avoid the problem, the SF under consideration is divided into three logical regions as Low Energy Area (LEA), Medium Energy Area (MEA) and High Energy Area (HEA) depending upon the energy required by SNs in these regions to transmit data directly to the sink. The sink assigns region IDs to all SNs. The logical region division has been shown in Fig. 1. The SNs from the region specified by the sink will only establish direct communication to it. Initially, LEA becomes region of interest.

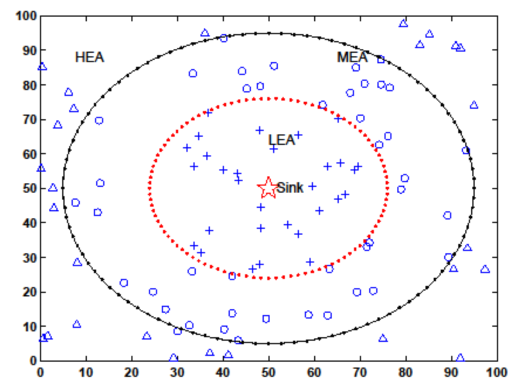


Fig. 1. Logical Region Division of SF in LEA, MEA and HEA (this figure is created for this research work).

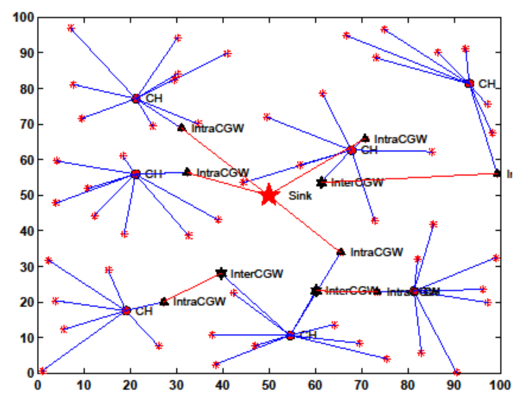


Fig. 2. Multi-hop topology for proposed system (this figure is created for this research work).

B. Network Setup Phase

After initial set up of WSN by sink, multi-hop energy efficient routes are established for each SN in distributive manner. Re-election of CHs may start from next round at each cluster level.

a) CH Election: A CH of a cluster continues to be CH for the next rounds provided its residual energy E_{resi} is greater than threshold energy of its cluster, E_{Th}^{C} . The technique eliminates the overhead of forming CHs in each round and saves energy. For its implementation, threshold energy of each cluster E_{Th}^{C} begins with 60% of E_{init} and decreased by 20% when energy of all nodes in the cluster falls below the threshold level. When E_{resi} of CH goes below E_{Th}^{C}, it prompts new CH election process and activates the active-sleep process in the cluster. Each active SN in the cluster scans its granule to finds a SN (in sleep mode) with highest E_{resi} . Make specified SN active and goes itself in sleep mode. Before going to sleep mode it transmits all its routing information to newly selected active node. Then, present CH finds candidate CH nodes, having residual energy above E_{Th}^{C} . Candidate CH with maximum fitness calculated according to (3) is selected as new CH. If there are two or more nodes with same fitness value, then, a node with higher ID number is chosen.

b) Multi-hop route establishment: CH in a cluster receives information from its CMs, aggregates the received data with its own sensed data and forwards it to sink either in single hop or in multi-hop. CMs are accountable to sense environmental data and transmit sensed information to CH. Therefore, CH in a cluster does more work than other nodes and consumes much higher energy. To reduce its work, the task of inter cluster communication is shifted to another node known as Intra Cluster GW (*IntraCGW)*. The cluster CH will choose *IntraCGW* in the direction of logical region announced by sink. If the selected *IntraCGW* is in the announced logical region, link is established directly to sink depending upon a probability function described in (4), otherwise inter cluster path is devised. Then, *IntraCGW* finds a nearest inter cluster SN in a cluster, currently not connected in order to avoid loops in communication path. The selected inter cluster node called "Inter Cluster GW (*InterGW*)" forms the link between the two clusters. The multi-hop topology formed by the proposed algorithm is shown in Fig. 2.

$$T(c)=\begin{cases}\dfrac{P}{1-P\times(r \bmod \dfrac{1}{P})}\times\dfrac{E_{avg}^{R}}{E_{avg}}, if\ c\in \mathrm{R} \\ 0, otherwise\end{cases} \quad \ldots\ldots\ldots (4)$$

where P is the desired percentage of direct links to the sink, r is the current round, E_{avg}^{R} is the average energy of CMs of the cluster in announced region *R*, E_{avg} is the average energy of the active nodes in the network. Each CH in c^{th} cluster in announced region generates a random number between 0 and 1. The CH will establish a link to the sink through selected *IntraGW*, if generated random number is less than threshold value of the cluster as calculated in (4). The proposed algorithm to establish multi-hop routes is given in Fig. 3.

c) Region Announcement by Sink: In order to attain nearly balanced energy depletion and to increase the reliability of WSN, it is desirable that all nodes remain alive for larger time period. SNs in LEA region consume lowest energy in direct data transmissions than MEA and HEA region nodes. Therefore, nodes from LEA region are desirable for establishing direct link to sink, but if these nodes continue to transmit information directly to sink, they will consume energy continuously and ultimately become dead. To avoid such situation, a shifting strategy between three logical regions has been proposed in this paper. A threshold parameter E_{th}^{R} is taken to decide this shifting strategy. Initially, E_{th}^{R} is set to 30% of E_{init} energy. Depending upon its value, the sink announces the region of interest in each round. If residual energy of all SNs in the SF goes below E_{th}^{R} then E_{th}^{R} is decreased to 10% of E_{init} and announce LEA as region of interest.

Algorithm: Network Setup Phase

1. For each CH check $\left(E_{resi}^{CH} < E_{Th}^{C}\right)$
2. Broadcast a query "What is the highest residual energy of the granule?" to all CMs and activate active sleep process.
3. Each active SN in the cluster finds other SN in its granule with highest E_{resi} . Activate newly selected SN. Transfer its routing information and goes in sleep mode.
4. All the activated nodes send reply message to $S\left(NodeID, GranID, E_{resi}\right)$ current CH.
5. Current CH finds a set CCH of candidate CHs such that their $\left(E_{resi} \geq E_{Th}^{C}\right)$
6. Check if no CCH is selected

   E_{resi}^{CH} is decreased by 20% and current CH remains CH for next the rounds.

   Otherwise

   Select a candidate CH with maximum fitness value and broadcast it to all SNs in the cluster.
7. Multi-Hop Route Establishment to Sink by Selected CH
   a. CH chooses GW within its cluster in the direction of announced region of interest with highest E_{resi} .
   b. Check selected GW is in Region of Interest and $\left(E_{resi} \geq E_{th}^{R}\right)$

      CH generates a random number <1; if it is less than threshold of its cluster then establish connection to the sink.

   Otherwise

      Check if all other clusters in the network are connected directly or indirectly to it.

      Establish connection to the sink

      Otherwise GW selects another closest SN from other clusters not connected to it and

   Establish connection to it.

C. Network Communication Phase

In communication phase, there are two types of communication takes place in the network: intra-cluster communication and inter cluster communication. For intra cluster communication CH node assigns TDMA schedule to its entire CMs. Each CM in a cluster transmits its sensed data to CH in its allotted slot. With TDMA scheme lot of energy of CMs is saved as CM goes in sleep mode when it is not transmitting any data. CH of each cluster receives data packets from CMs and aggregated the received information with its own sensed data and forwards it to *IntraGW*. It will then aggregate the received data packet with its own and forward it to *InterGW* using CDMA codes. In each round, active SN those go below E_{th}^{R} transmit its energy status to the sink. Depending upon information received, it announces the region of interest.

V. Results and Discussions

The proposed methodology has been implemented with MATLAB and executed on 100 randomly deployed wireless sensor networks in order to evaluate the performance parameters: network lifetime, energy consumption and throughput. Network lifetime is considered as the number of rounds until the complete network is dead. Energy consumption in each SN of the network is measured as the sum of energy consumed in receiving, aggregating and forwarding the data to other intermediate nodes or to the sink. Throughput is defined as sum of unique data packets generated in each round. For example, four data packets are generated from an active granule by four SNs in that granule. We consider it as one because other three carries same data as that of first packet. For simplicity, it is assumed that there is no data collision and packet loss in the wireless channel. Likewise, control packets have also been ignored for the evaluation of performance parameters.

In this simulation setup, one hundred static sensor nodes are randomly distributed in a SF of area 200m X 200m and considered as one WSN deployment. The sink is placed at location (100,100) and is immobile. Initial battery power for all SN is taken as 0.5J. Further, to ensure the generalisation of results, the simulation experiments were conducted for one hundred WSN deployments. Each simulation is run until all sensor nodes in the network become dead. The energy consumed in the network is calculated for transmitting, receiving and aggregating data packets. The energy consumed in aggregating and transmitting 1 bit data over distance d is calculated using (5).

$$e_t = \begin{cases} E_{elec} + E_{DA} + E_{fs}d^2, \forall d \le d_o \\ E_{elec} + E_{DA} + E_{mp}d^2, \forall d > d_o \end{cases} \quad \ldots\ldots\ldots\ldots.(5)$$

Energy consumed in receiving 1 bit data is given as under:

$$e_r = E_{elec} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(6)$$

where E_{elec} is energy dissipated per bit to run transmitter or receiver circuit, taken as 50nJ/bit. E_{DA} is energy consumed in aggregating one bit data and set to 5nJ/bit/signal. E_{fs} and E_{mp} are energy consumed in free space model and energy used in multipath fading model taken as 10 pJ/bit/m^2 and 0.0013 pJ/bit/m^4 respectively. The threshold distance d_o is equal to square root of ratio $\frac{E_{fs}}{E_{mp}}$. Total energy consumed in communication from SN_i to SN_j over distance d is given by (7).

$$E_{ij}(l,d) = l*(e_t + e_r) \quad (7)$$

Where l is the length of data packet taken as 4000 bits. To decide optimal number of clusters K_{opt} for the proposed algorithm, the algorithm is executed on all 100 simulations for varying K, i.e. number of clusters, from 1 to 20. The average energy consumption per round for each value of K is calculated. The minimum average energy consumption per round is found at K =10. Hence, the value of K_{opt} is taken as10.

In order to generalize the results, all the five algorithms (LEACH, MR-LEACH, MH-LEACH, OCM-FCM and EEA-FCM) were implemented on the same hundred deployments and averaged results obtained from them have been used to present the comparative account of the algorithms. The results are shown in Fig. 4, 5, 6, 7 and 8.

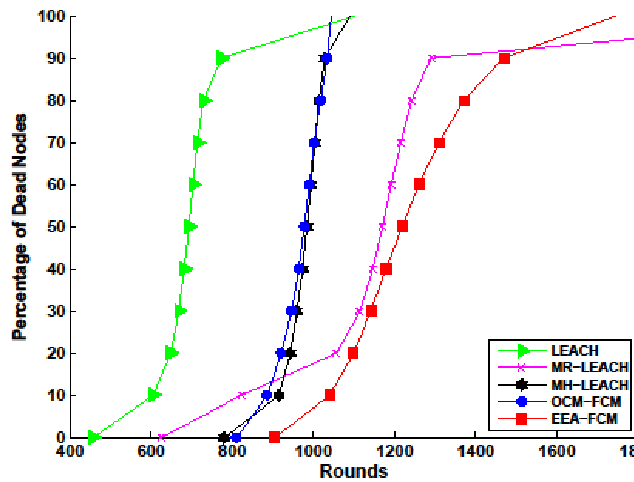


Fig. 3. Graph of number of operational nodes verses rounds in WSN for N=100.

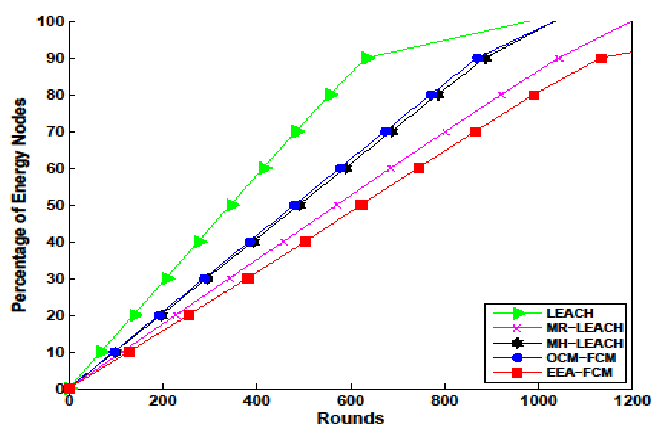


Fig. 4. Energy consumption in WSN for N=100.

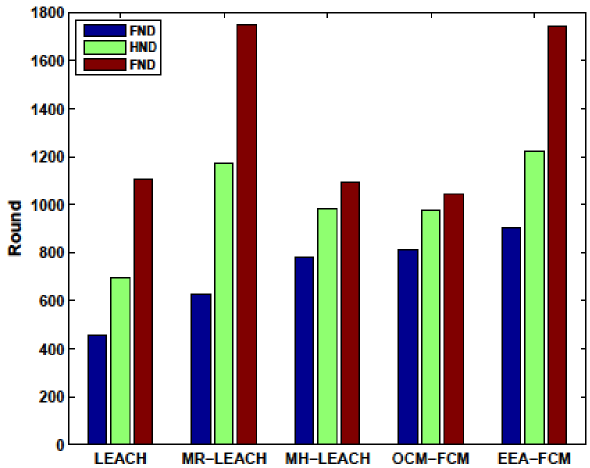


Fig. 5. FND, HND and LND for WSN for N=100.

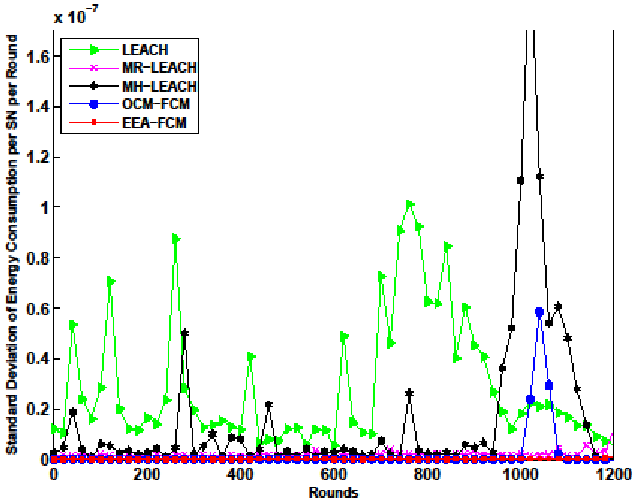


Fig. 6. Standard deviation of average energy consumption per SN per Round in WSN for N=100.

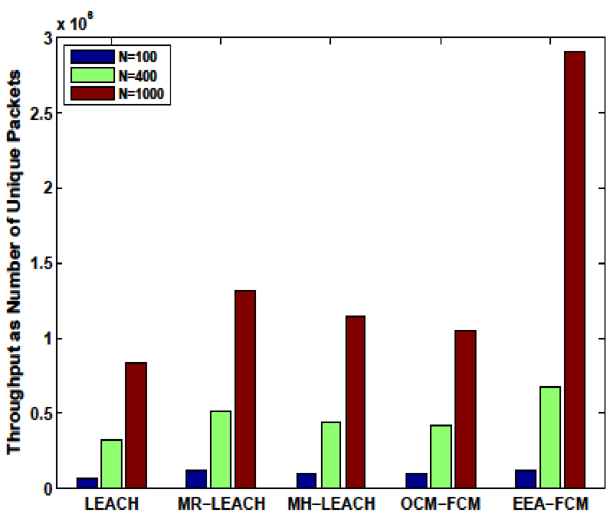


Fig. 7. Comparative plot of throughput for N=100, 400 and 1000.

The graph of number of operational nodes verses the number of rounds is shown in Fig. 4. It is clearly observed from the figure that EEA-FCM remains operational for longer period of time as compared to LEACH, MR-LEACH, MH-LEACH and OCM-FCM algorithms. At 900th round, the dead node percentage increases to 7, 10 and 14 for MH-LEACH, MR-LEACH and OCM-FCM algorithms respectively while it is 1% for EEA-FCM. Similarly, when EEA-FCM lost its 10% network operational ability (i.e. 10% nodes are dead), network in OCM-FCM and MH-LEACH almost ceases to exist. First 10% nodes in MR-LEACH died more quickly as compared to MH-LEACH, OCM-FCM and EEA-FCM. Network in MR-LEACH stabilizes to some extent after losing 20% of its operational ability. This shows EEA-FCM algorithm performs far better than the other algorithms under test at each point of network operation.

Percentage of total energy consumed in all the five protocols has been presented in Fig. 5. It is observed from the figure that the rate of energy consumption in LEACH is much higher as compared to EEA-FCM, OCM-FCM, MR-LEACH and MH-LEACH. Energy consumption in OCM-FCM and MH-LEACH is almost same. It is worth noting that OCM-FCM is transmitting cluster data to the sink in single hop whereas MH-LEACH transmits in multiple hops. Therefore, OCM-FCM clustering technique is much better as compared to MH-LEACH. At the point of network operation where EEA-FCM consumed fifty percent of its network energy, OCM-FCM and MR-LEACH consumed 65% and 60% respectively. Hence, energy usage of the network is more economical in EEA-FCM as compared to OCM-FCM and MR-LEACH.

From Fig. 4 and 5, it is clearly observable that distribution of network load amongst all the SNs is more balanced which will enhance the network lifetime. Round value for FND (First Node Dead), HND (Half Node Dead) and LND (Last Node Dead) has great significance to evaluate the performance of network. The round value of FND indicates the reliability and stability of the network operation. The round value of HND specifies rate of decay of performance of network after first node become dead. The round value for LND indicates the duration for which the network remains operational. Fig. 6 shows that FND, HND and LND for EEA-FCM is much higher as compared to compared to OCM-FCM, MH-LEACH, MR-LEACH and LEACH algorithms. In the proposed EEA-FCM algorithm, data collection, aggregation and data forwarding is divided amongst the CH and GW nodes of the cluster which results in healthier CH node for long period of time. Due to these factors, energy usage in the network and decay in reliability of WSN operation is reduced.

The standard deviation of node's average consumption per SN per round is shown in Fig. 7. Standard deviation value of average energy consumption in a SN at any round indicates load distribution amongst all SNs in the network. Large value of standard deviation means that there is large fluctuation from mean of average energy consumption in a SN. It indicates that some nodes in the network consume large amount of energy as compared to other nodes. OCM-FCM shows higher value of standard deviation, when less number of nodes in the network are alive. It is worth noting the standard deviation of energy consumption per SN in EEA is very small as compared to LEACH, MR-LEACH, MH-LEACH and OCM-FCM.

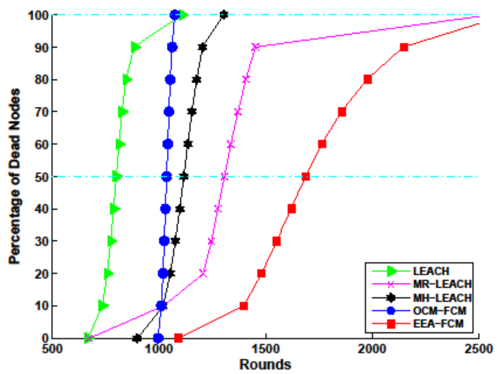


Fig. 8. Percentage of dead nodes verses rounds at number of nodes is 400.

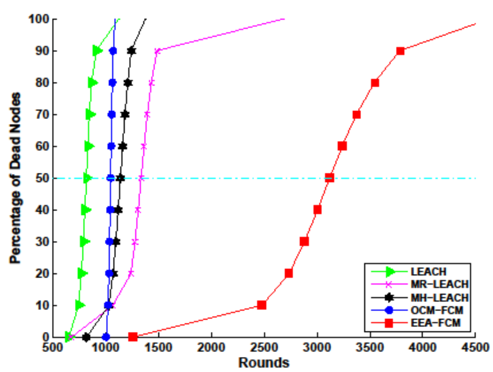


Fig. 9. Percentage of dead nodes verses rounds at number of nodes is 1000.

Therefore, network load is uniformly distributed amongst all SNs in EEA-FCM and consume almost equal amount of energy which leads to more stable and reliable network operation.

A comparative graph of throughput for all five algorithms is presented in Fig. 8. OCM-FCM and MR-LEACH acquire less amount data from the SF as their nodes have lesser lifetime as compared to EEA-FCM. In EEA-FCM, SNs are not generating redundant data from SF and remain alive for longer period. Therefore, throughput of the network is improved. Total throughput in case of EAA-FCM is 7.6% higher than MR-LEACH and 28% more than OCM-FCM. Thence, the proposed algorithm shows a significant improvement in throughput.

Node density of WSN may greatly affect SF coverage, network connectivity and network life time. Therefore, it is an important factor that needs to be considered while evaluating the performance of WSN. To verify the effect of node density on the performance of the network, additional simulation experiments are conducted by varying the number of nodes, N, in the WSN deployments as N=100, 400 and 1000. Please note that N=400 in SF of 200m X 200m is equivalent 100 nodes in an area of 100m X100m.

The percentage of dead nodes for varying number of rounds has been shown in Fig. 4 (N=100) and Fig. 10 (N=400) and Fig. 11 (N=1000). Comparison for all three algorithm for N=100 has already been discussed earlier. Taking the discussion further, it is seen that when network node density increased to 400, within first 1100 rounds, all nodes for LEACH, MH-LEACH and OCM-FCM become dead, but all the four hundred nodes for EEA-FCM are still alive. When 15% nodes of EEA-FCM are dead, 90% nodes in MR-LEACH has already become dead. Further, the slops of five curves in Fig. 6 have been observed. It is clearly seen that the slop of curve for EEA-FCM is much less than other two curves. This indicates that the rate of nodes becoming dead is much slower in EEA-FCM. Lifetime of WSN for the proposed algorithm is almost double than LEACH, MH-LEACH and OCM-FCM algorithms. Further, when number of nodes in WSN increased from 400 to 1000, similar behaviour of algorithms has been observed. There is no improvement in network lifetime of LEACH, MH-LEACH, MR-LEACH and OCM-FCM. Network lifetime of EEA-FCM is far improved over other two algorithms. It is approximately 4.5 times that of LEACH, MH-LEACH and OCM-FCM.

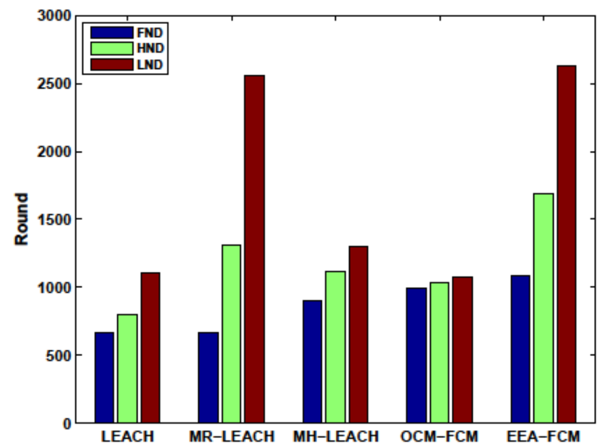


Fig. 10. FND, HND and LND for WSN for Varying Number of Nodes is 400.

The network throughput obtained in LEACH, MR-LEACH, MH-LEACH, OCM-FCM and EEA-FCM with varying N is shown in Fig. 9 and Table I. It is observed, for N=100, OCM-FCM, MR-LEACH and EEA-FCM generate 96641, 114973 and 123577 data packets in its lifetime, respectively. There is a 63% improvement in throughput of EEA-FCM over OCM-FCM for N=400. By varying nodes from 400 to 1000, throughput obtained in EEA-FCM is almost three times as compared to OCM-FCM. The throughput of EEA-FCM is improved in many folds compare to LEACH and OCM-FCM with the increase in node density.

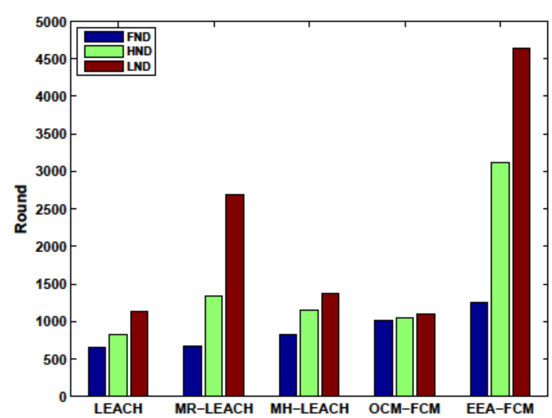


Fig. 11. FND, HND and LND for WSN for Varying Number of Nodes is 1000.

The value of rounds for First Node Dead (FND), Half Node Dead (HND) and Last Node Dead (LND) for all algorithms under examination at N=100, 400 and 1000 are shown in Fig. 10 and 11. For N=100, OCM-FCM improves the round number for FND by 4% over MH-LEACH. EEA-FCM improves FND of MH-LEACH further by another 11%. The round number for HND is 694th, 1171th, 984th, 977th and 1220th round in LEACH, MR-LEACH, MH-LEACH, OCM-FCM and EEA-FCM respectively.MR-LEACH improves the round number for HND for OCM-FCM by 19%. EEA-FCM improves it further by 5%. The LND round number is almost same for LEACH and OCM-FCM. The network lifetime for EEA-FCM is improved 58% over LEACH. It is also observed that by changing N from 100 to 400, FND values of all three algorithms are improved. EEA-FCM improves HND value of OCM-FCM by 63%. Network lifetime for EEA-FCM is almost double as compared to the other two algorithms. For N=1000, there is not much improvement in LEACH and OCM-FCM. The round number for FND, HND and LND in EEA-FCM is delayed by 182, 1396 and 2063 rounds respectively as compared to their values at N=400. Therefore, network operation got improved in EEA-FCM on increasing the node density.

TABLE I. LIFETIME AND THROUGHPUT FOR VARYING NODE DENSITY N=100, 400 AND 1000

Algorithm	**Lifetime**	**Throughput**	**Lifetime**	**Throughput**	**Lifetime**	**Throughput**
	Node Density N=100		Node Density N=400		Node Density N=1000	
LEACH	1102	70093	1109	323570	1129	829786
MR-LEACH	2375	114973	2559	512385	2684	1311447
MH-LEACH	1092	97638	1305	445357	1381	1139135
OCM-FCM	1044	96641	1075	414114	1093	1048670
EAA-FCM	1746	123577	2634	677378	4635	2906828

VI. CONCLUSION

In this paper, an energy efficient algorithm for WSN is presented. The results have been compared with LEACH, MH-LEACH, MR-LEACH and OCM-FCM algorithms. From the outcome, it is concluded that lifetime and throughput of proposed algorithm is 58% and 28% more as compared to the other algorithms under examination. Network load amongst all SNs is optimally distributed in the proposed algorithm. The rate of decay i.e. the nodes becoming dead is also minimum for the proposed algorithm as compared to other algorithms. Therefore, it is concluded that EEA-FCM achieves better energy efficiency and enhanced lifetime. Moreover, the performance of network has also been analyzed for varying N=100, 400 and 1000. The performance of EEA-FCM is far better for higher node density of WSN. From the results, it is concluded that the proposed EEA-FCM algorithm out performs other algorithms discussed in this paper.

REFERENCES

[1] Akyildiz, F., Su, W., Subramaniam, Y. S., & Cayirci, E. (2002).Wireless sensor networks: a survey. *Computer Networks, 38(4),* 393-422.

[2] Heinzelman, W. R., Chandrakasan, A., & Balakrishnan, H. (2000). Energy efficient communication protocol for wireless microsensor networks. In proceedings of 33rd annual Hawaii international conference on System Sciences, IEEE, http://doi.org/10.1109/HICSS.2000.926982.

[3] Chen, G., Li, C., Ye, M., & Wu, J. (2009). An unequal cluster-based routing protocol in wireless sensor networks. *Wireless Networks, 15(2),* 193–207.

[4] Shin, J., & Suh, C. (2011). CREEC: Chain routing with even energy consumption. *Journal of Communications and Networks, 13(1),* 17–25.

[5] Salim, A., Osamy, W., & Khedr, A. M. (2014). IBLEACH: Intra-balanced LEACH protocol for wireless sensor networks. *Wireless Networks, 20(6),* 1515–1525.

[6] Farooq, M. O., Dogar, A.B., & Shah, G.A. (2010). MRLEACH: Multi-hop Routing with Low Energy Adaptive Clustering Hierarchy. *In proceedings of fourth international conference on Sensor Technologies and Applications (SENSORCOMM),* (pp. 262 – 268).

[7] Neto, J. H. B., Rego, A. S., Cardoso, A. R., & Celestino Jr., J. (2014). MH-LEACH: A Distributed Algorithm for Multi-Hop Communication in Wireless Sensor Networks. *In proceedings of The Thirteenth International Conference on Networks,* (pp. 55-61).

[8] Huang, J., Hong, Y., Zhao, Z., & Yuan, Y. (2017). An energy-efficient multi-hop routing protocol based on grid clustering for wireless sensor networks. *Cluster Computing, 20(4)*, 3071–3083.

[9] Aslam, M., Munir, E. U., Rafique, M. M., & Hu, X. (2016). Adaptive energy-efficient clustering path planning routing protocols for heterogeneous wireless sensor networks. *Sustainable Computing Informatics & Systems, 12,* 57–71, http://dx.doi.org/10.1016/j.suscom.2017.08.001

[10] Su, S., & Zhao, S. (2017). An optimal clustering mechanism based on Fuzzy-C means for wireless sensor networks. *Sustainable Computing Informatics & Systems*, http://dx.doi.org/10.1016/j.suscom.2017.08.001

[11] Mittal, N., Singh, U., & Sohi, B. S. (2016) "A stable energy efficient clustering protocol for wireless sensor networks*," Wireless Networks*, pp. 1–13, 2016.

[12] Sabet, M. & Naji, H. (2016). An energy efficient multi-level route-aware clustering algorithm for wireless sensor networks: A self-organized approach. *Computers & Electrical Engineering, 56*, 399–417.

[13] Chen, B., Jamieson, K., Balakrishnan, H., & Morris, R. (2002). Span: An energy-efficient coordination algorithm for topology maintenance in ad hoc wireless networks. *Wireless Networks, 8*, 481–494.

[14] Bakr, B. A., & Lilien, L. T. (2014). Extending lifetime of wireless sensor networks by management of spare nodes. *Procedia Computer Science*, 34, 493–498.

[15] Bulut, E., & Korpeoglu, I.(2011). Sleep scheduling with expected common coverage in wireless sensor networks. *Wireless Networks*, 17(1), 19–40.

[16] Cheng, C.T., Tse, C. K., & Lau, F. C. M. (2011). A Clustering Algorithm for Wireless Sensor Networks Based on Social Insect Colonies. *IEEE Sensor Journal*, 11(3), 711–721.

[17] Rao, P. C. S., Jana, P. K., & Banka, H. (2017). A particle swarm optimization based energy efficient cluster head selection algorithm for wireless sensor networks. *Wireless Networks*, 23(7), 2005–2020.

[18] Smaragdakis, G., Matta, I., & Bestavros, A., (2004). Sep: A stable election protocol for clustered heterogeneous wireless sensor networks. Boston University Computer Science Department.

[19] Qing L., Zhu Q., & Wang M. (2006). Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks. *Computer communications*, 29(12), 2230–2237.