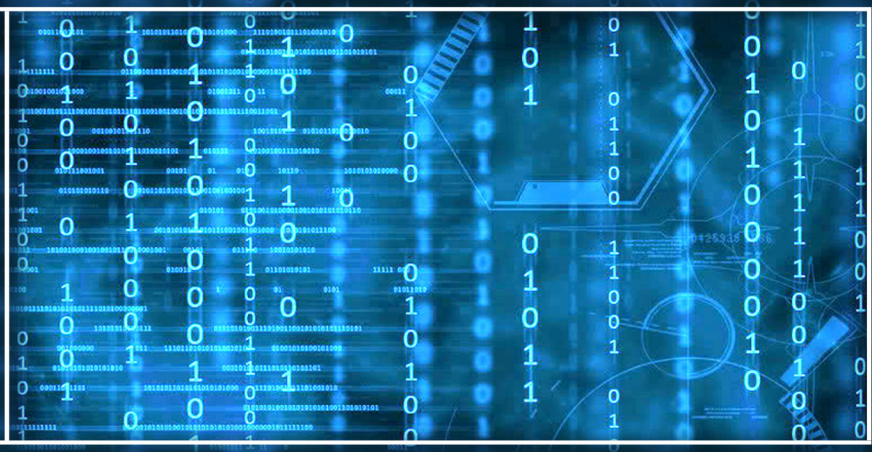
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 9 Issue 9
September 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 9 September 2018**
**ISSN 2156-5570 (Online)**
**ISSN 2158-107X (Print)**

# Editorial Board

# Reviewer Board Members

Aakash Ahmad Abbasi
Abbas Karimi
Abbas M. Al-Ghaili
Abdelghni Lakehal
Abdul Aziz Gill
Abdul Hamid Mohamed Ragab
Abdul Karim Assaf ABED
Abdul Razak
Abdul Wahid Ansari
Abdur Rashid Khan
Abeer Mohamed ELkorany
ABRAHAM VARGHESE
Adebayo Omotosho
ADEMOLA ADESINA
Aderemi A. Atayero
Adi A. Maaita
Adnan Ahmad
Adrian Nicolae Branga
Ahmad A. Al-Tit
Ahmad A. Saifan
Ahmad Hoirul Basori
Ahmad Mousa Altamimi
Ahmed Boutejdar
Ahmed Nabih Zaki Rashed
Ahmed S.A AL-Jumaily
Ahmed Z. Emam
Ajantha Herath
Akram Belghith
Alaa F. Sheta
Albert Alexander S
Alci-nia Zita Sampaio
Alexane Bouënard
ALI AMER ALWAN
Ali Asghar Pourhaji Kazem
Ali Hamzeh
Ali Ismail Awad
Ali Mustafa Qamar
Alicia Menchaca Valdez
Altaf Mukati
Aman Chadha
Amin Ahmad Shaqrah
amine baina
Amir HAJJAM EL HASSANI
Amirrudin Kamsin
Amitava Biswas
Amjad Gawanmeh
Anand Nayyar
Anandhi Mohanraj Anu
Andi Wahju Rahardjo Emanuel
Anews Samraj
Anirban Sarkar
Anita Sofia V S
Anju Bhandari Gandhi
Anouar ABTOY
Anshuman Sahu
Anthony Nosike Isizoh
Antonio Dourado
Antonio Formisano
ANUAR BIN MOHAMED KASSIM
Anuj Kumar Gupta
Anuranjan misra
Appasami Govindasamy
Arash Habibi Lashkari
Aree Ali Mohammed
Arfan Jaffar
ARINDAM SARKAR
Aris Skander Skander
Arun D Kulkarni
Arun Kumar Singh
Arvind K Sharma
Asadullah Shaikh
Asfa Praveen
Ashok Matani
Ashraf Hamdy Owis
ASIM TOKGOZ
Asma Cherif
Asoke Nath
Athanasios Koutras
Ayad Ghany Ismaeel
Ayman EL-SAYED
Ayman Shehata Shehata
Ayoub BAHNASSE
Ayush Singhal
Azam Moosavi
Babatunde Opeoluwa Akinkunmi
Bae Bossoufi
Balasubramanie Palanisamy
BASANT KUMAR VERMA
Basem M. ElHalawany
Basil Hamed
Basil M Hamed
Basim Almayahi
Bestoun S. Ahmed
Bhanu Kaushik
Bhanu Prasad Pinnamaneni
Bharti Waman Gawali
Bilian Song
Binod Kumar
Bogdan Belean

Marwan Alseid

Mazin S. Al-Hakeem

Md Ruhul Islam

Md. Al-Amin Bhuiyan

Mehdi Bahrami

Mehdi Neshat

Messaouda AZZOUZI

Milena Bogdanovic

Miriampally Venkata Raghavendra

Mirjana Popovic

Miroslav Baca

Moamin Mahmoud

Moeiz Miraoui

Mohamed AbdelNasser Mohamed Mahmoud

Mohamed Salah SALHI

Mohamed A. El-Sayed

Mohamed Abdel Fatah Ashabrawy

Mohamed Ali Mahjoub

Mohamed Eldosoky

Mohamed Hassan Saad Kaloup

Mohamed Najeh LAKHOUA

Mohamed SOLTANE Mohamed

Mohammad Abdul Qayum

Mohammad Ali Badamchizadeh

Mohammad Azzeh

Mohammad H. Alomari

Mohammad Haghighat

Mohammad Jannati

Mohammad Zarour

Mohammed Abdulhameed Al-shabi

Mohammed A. Akour

Mohammed Ali Hussain

Mohammed Sadgal

Mohammed Shamim Kaiser

Mohammed Tawfik Hussein

Mohd Ashraf Ahmad

Mohd Helmy Abd Wahab

Mokhtar Beldjehem

Mona Elshinawy

Monir Kaid

Mostafa Mostafa Ezziyyani

Mouhammd sharari sharari alkasassbeh

Mounir Hemam

Mourad Amad

Mudasir Manzoor Kirmani

Mueen Uddin

Muhammad Adnan Khan

Muhammad Abdul Rehman

Muhammad Asif Khan

Muhammad Hafidz Fazli Bin Md Fauadi

Muhammad Naeem

Muhammad Saeed

Muniba Memon

MUNTASIR AL-ASFOOR

Murphy Choy

Murthy Sree Rama Chandra Dasika

MUSLIHAH WOOK

Mustapha OUJAOURA

MUTHUKUMAR S SUBRAMANYAM

N.Ch. Sriman Narayana Iyengar

Nadeem Akhtar

nafiul alam siddique

Nagy Ramadan Darwish

Najeed Ahmed Khan

Najib A. Kofahi

Namrata Dhanda

Nan Wang

Naseer Ali Alquraishi

Nasrollah Pakniat

Natarajan Subramanyam

Natheer Gharaibeh

Nayden V. Nenkov

Nazeeh Ghatasheh

Nazeeruddin Mohammad

Neeraj Kumar Tiwari

NEERAJ SHUKLA

Nestor Velasco-Bermeo

Nguyen Thanh Binh

Nidhi Arora

NILAMADHAB MISHRA

Nilanjan Dey

Ning Cai

Niraj Singhal

Nithyanandam Subramanian

Nizamud Din

Noura Aknin

Obaida M. Al-Hazaimeh

Olawande Justine Daramola

Oliviu Matei

Om Prakash Sangwan

Omaima Nazar Al-Allaf

Omar A. Alzubi

Omar S. Gómez

Osama Ali Awad

Osama Omer

Ouchtati Salim

Ousmane THIARE

P.V. Praveen Sundar

Paresh V Virparia

Parminder Singh Kang

PAUL CELICOURT

Peng Xia

Ping Zhang

Piyush Kumar Pareek
Poonam Garg
Prabhat K Mahanti
PRASUN CHAKRABARTI
Praveen Kumar
PRISCILLA RAJADURAI
PROF DURGA PRASAD SHARMA ( PHD)
Purwanto Purwanto
Qaisar Abbas
Qifeng Qiao
Rachid Saadane
Radwan R. Tahboub
raed Kanaan
Raghuraj Singh
Rahul Malik
Raja Ramachandran
raja sarath kumar boddu
Rajesh Kumar
Rakesh Chandra Balabantaray
Rakesh Kumar Dr.
Ramadan Elaiess
Ramani Kannan
RAMESH MUTHUSAMY
RAMESH VAMANAN
Rana Khudhair Abbas Ahmed
Rashad Abdullah Al-Jawfi
Rashid Sheikh
Ratnesh Litoriya
Ravi Kiran Varma P
Ravi Prakash
RAVINA CHANGALA
Ravisankar Hari
Rawya Y. Rizk
Rayed AlGhamdi
Reshmy Krishnan
Reza Fazel-Rezai
Reza Ghasemy Yaghin Dr Reza Ghasemy Yaghin
Riaz Ul-Amin
Ricardo Ângelo Rosa Vardasca
Ritaban Dutta
Rodica Doina Zmaranda
Rohini Ravi
Rohit Raja
Roopali Garg
roslina ibrahim
Ruchika Malhotra
Rutvij H. Jhaveri
SAADI Slami
Sachin Kumar Agrawal
Sagarmay Deb
Sahar Abd El_RAhman Ismail
Said Ghoniemy
Said Jadid Abdulkadir
Sajal Bhatia
Saman Hina
SAMSON OLUWASEUN FADIYA
Sanam Shahla Rizvi
Sandeep R Reddivari
Sangeetha SKB
Sanskruti V Patel
Santosh Kumar
Sasan Adibi
Sattar Bader Sadkhan
Satyena Prasad Singh
Sebastian Marius Rosu
Secui Dinu Calin
Seema Shah
Seifedine Nimer Kadry
Selem Charfi
SENGOTTUVELAN P
Senol Piskin
SENTHIL P Prof
Sérgio André Ferreira
Seyed Hamidreza Mohades Kasaei
Shadi Mahmoud Atalla
Shafiqul Abidin
Shahab Shamshirband
Shahanawaj Ahamad
Shaidah Jusoh
Shaiful Bakri Ismail
Shailesh Kumar
Shakir Gayour Khan
Shashi Dahiya
Shawki A. Al-Dubaee
Sheeraz Ahmed Dr.
Sheikh Ziauddin
Sherif E. Hussein
Shishir Kumar
SHOBA MOHAN
Shriniwas Vasantrao Chavan
Shriram K Vasudevan
Siddeeq Ameen
Siddhartha Jonnalagadda
Sim-Hui Tee
Simon L. R. Vrhovec
Simon Uzezi Ewedafe
Siniša Opic
Sivakumar Poruran
sivaranjani reddi
Slim BEN SAOUD
Sobhan Roshani
Sofien Mhatli
sofyan Mohammad Hayajneh
Sohail Jabbar
Sri Devi Ravana

Sudarson Jena

Sudipta Roy

Suhail Sami Owais Sami Owais Owais

Suhas J Manangi

SUKUMAR SENTHILKUMAR

Süleyman Eken

Sumazly Sulaiman

Sumit Goyal

Sunil Phulre

Suparerk Janjarasjitt

Suresh Sankaranarayanan

Surya Narayan Panda

Susarla Venkata Ananta Rama Sastry

Suseendran G

Suxing Liu

Syed Asif Ali

T C.Manjunath

T V Narayana rao Rao

T. V. Prasad

Taghi Javdani Gandomani

Taiwo Ayodele

Talal Bonny

Tamara Zhukabayeva

Taner Tuncer

Tanvi Banerjee

Tanweer Alam

Tanzila Saba

TAOUFIK SALEM SAIDANI

Tarek Fouad Gharib

tarig ahmed

Taskeed Jabid

Tasneem Bano Rehman

thabet Mohamed slimani

Totok R. Biyanto

Touati Youcef

Tran Xuan Sang

TSUNG-CHUAN MA

Tsvetanka Georgieva-Trifonova

Uchechukwu Awada

Udai Pratap Rao

Urmila N Shrawankar

V Baby Deepa

Vaidas Giedrimas

Vaka MOHAN

Venkata Raghavendran Chaluvadi

VENKATESH JAGANATHAN

Vijay Bhaskar Semwal

Vijayarani Mohan S

Vijendra Singh

Vinayak K Bairagi

VINCE PAUL A

Visara Urovi

Vishnu Narayan Mishra

Vitus S.W. Lam

VNR SAIKRISHNA K

Voon Ching Khoo

VUDA SREENIVASARAO

Wali Khan Mashwani

Wei Wei

Wei Zhong

Wenbin Chen

Wenzhao Zhang

Wichian Sittiprapaporn

Xi Zhang

Xiao Zhang

Xiaojing Xiang

Xiaolong Wang

Xunchao Hu

Y Srinivas

Yanping Huang

Yao-Chin Wang

Yasser M. Alginahi

Yaxin Bi

Yi Fei Wang

YI GU

Yihong Yuan

Yilun Shang

Yu Qi

Zacchaeus Oni Omogbadegun

Zaffar Ahmed Shaikh

Zairi Ismael Rizman

Zarul Fitri Zaaba

Zeki Yetgin

Zenzo Polite Ncube

ZHENGYU YANG

Zhigang Yin

Zhihan Lv

Zhixin Chen

Zia Ur Rahman Zia

Ziyue Xu

Zlatko Stapic

Zne-Jung Lee

Zuraini Ismail

# CONTENTS

# A Proposal for A High Availability Architecture for VoIP Telephone Systems based on Open Source Software

Alejandro Martin[1]
School of Computing
Central University of Venezuela
Caracas, Venezuela

Eric Gamess[2]
MCIS Department
Jacksonville State University
Jacksonville, AL, USA

Dedaniel Urribarri[3], Jesús Gómez[4]
School of Computing
Central University of Venezuela
Caracas, Venezuela

***Abstract*—The inherent needs of organizations to improve and amplify their technological platform entail large expenses with the goal to enhance their performance. Hence, they have to contemplate mechanisms of optimization and the improvement of their operational infrastructure. In this direction arises the need to guarantee the correct operation and non-degradation of the services provided by the platform during the periods with a significant load of work. This type of scenario is perfectly applicable to the field of VoIP technologies, where users generate elevated loads of work on critical points of the infrastructure, during the process of interaction with their peers. In this research work, we propose a solution for high availability, with the goal of maintaining the continuity of the operation of communication environments based on the SIP protocol in high load. We validate our proposal through numerous experiments. Also, we compare our solution with other classical VoIP scenarios and show the advantages of a high availability and fault tolerance architecture for organizations.**

***Keywords*—Cluster; high availability; load balancer; VoIP; SIP; kamailio; corosync; asterisk; SIPp**

## I. Introduction

The rise of new technologies in the field of communications through the usage of computer networks has driven the growth of organizations. Motivated by their interest in an efficient application of resources, these organizations invest on innovative communication mechanisms, to establish connections between devices for the exchange of information, allowing the communication of well-identified entities. Usually, these communication technologies follow well-regulated operating schemes, which clearly define intermediate points, states, processes, and possible behaviors of communication, also known as communication protocols.

Extrapolating these concepts, organizations interested in optimally diversifying their communication capacities are inclined to use these state-of-the-art technologies and protocols of communication. However, there are cases in which the magnitude of some organizations requires network infrastructures that prioritize the performance and stability of the communication platform, and it is at this point that the following topics of interest arise:

*1) Optimization of the network infrastructure at the level of performance*

*2) Scalability and longevity of the network infrastructure*

*3) Adequate performance of the communication system in stressful conditions*

*4) Backup plan to tackle faults and malfunction events.*

In this work, we propose a Voice over IP (VoIP) solution based on the Session Initiation Protocol [1][2][3] (SIP), and the usage of tools that allow the establishment of a network infrastructure of high availability and good performance, such as Kamailio [4], Corosync [5], Asterisk [6][7], among others. To validate our work, we test our suggested architecture under high stress and report the obtained results.

The following document is organized as follows. Section II presents the communication problems faced by large organizations when it comes to scalability and fault tolerance. Section III introduces some general solutions for high availability. The related works are discussed in Section IV. In Section V, we propose our high availability architecture for VoIP based on open source software. The test scenarios and definitions are presented in Section VI and Section VII, respectively. In Section VIII, we analyze the results of our tests to validate the proposed architecture. Finally, Section IX concludes the paper and gives some directions for future work.

## II. Organizational Problems

From a technological perspective, there are some limitations in existing IP telephony architectures aimed at the provision of a high-quality telephony service, particularly in high-volume conditions where a significant workload must be supported, such as organizational environments, in which users rely on services in a sustained and repetitive manner. These limitations can lead to a point in which an organization, for the sake of the improvement of the quality of service and non-interruption of the business, opts for proprietary solutions that in many cases represent a significant monetary investment. The main problems faced by organizations that use open-source software solutions for telephony are those that prevent the continuity of the service in environments of high concurrency and high demand.

Many of these problems are due to the fact that some organizations ignore the expected requirement parameters in the environments in which the service is provided within their networks. Therefore, these organizations do not evaluate certain topological and architectural solutions with open-source software tools to face their limitations, which could be solved by offering proactive mechanisms of high availability and contingency, in case of malfunction of their systems. In addition, if the possible growth in the requirements of the services managed by the organization is anticipated, a good approach in the management and configuration of the services to be offered within the architecture, plus load balancing solutions, would allow greater scalability in the service.

The concept of high availability and continuity of the operation in computer networks is a solution for environments of high workload. However, the high cost associated with implementing proprietary solutions to improve the quality of the communication service makes them unfeasible for some organizations. Therefore, our interest in proposing an architecture of high availability through the usage of open-source software tools. Our solution considers that the particular limitations of each piece of software can be overcome by the coupling of multiple tools synchronized with each other to offer a standout service, that is, by using the famous saying "Unity is Strength."

## III. High Availability Solutions

In any type of communication solution, faults are highly probable. Therefore, contingency mechanisms are required to resolve these faults automatically, with the smallest possible response time, and when possible, in a transparent way for users. To achieve this goal, a high availability solution must have the following aspects:

- high availability mechanisms
- load balancing mechanisms
- division of services
- monitoring and management of services and load.

### A. High Availability Mechanisms

From the computational point of view, the term "availability" refers to the period of time a service is available. When seen as an expression, "availability" can be either the response time of the service or whether it is accessible to be consumed by the users [8]. Hence, "high availability" refers to "... a system that is continuously operational and available for the use of the services provided to end-users" [9]. Providing high availability to a system or service is not an easy task, since not every service manages mechanisms to support it. Therefore, offering high availability in a system requires a specific design and implementation, mitigating all the points of failure that a system may have.

Redundancy is important in a system that offers high availability. Redundancy generally consists of backup components that automatically "kick in" if one component fails. It can be achieved at the level of the nodes that provide the service, i.e., the servers that are managing the service. The two most commonly used high availability cluster configurations are active-passive and active-active.

*1) Active-Passive:* In an active-passive scheme, one of the nodes is processing the requests received by the critical service (thus considered as the active node), while the other node is monitoring the active node and ready to take over as soon as the active node gets disconnected or is unable to serve [10].

*2) Active-Active:* In an active-active scheme, both nodes are actively providing critical services simultaneously. One of the main purposes of an active-active cluster is to achieve load balancing, that is, distribute the workload. In case of a failure of either node, the other node will be responsible for the provision of all the critical services [10].

### B. Load Balancing Mechanisms

Let consider two servers or nodes with similar characteristics that are setup as telephone exchanges, one active and one passive. The basic idea is that the passive server will take over when the active server faces a failure. Now, if the active server were to fail due to very high service requirements (numerous controls for calls managed simultaneously: user registration, start of calls, end of calls, etc), it is very likely that the same will happen to the other server after the failover, since they are servers with similar characteristics. To solve this kind of problems, it is advisable to distribute the load among multiple servers (load balancing).

In the concept of load balancing, multiple servers are offering the same service and do the same tasks. Usually, an appliance (specialized physical device) or software solution is in charge of the load balancing management. It receives all the requests that are directed to the servers and is in charge of redirecting the traffic to one of the servers that it considers appropriate, through different techniques and mechanisms to balance the workload.

Load balancers can be divided into stateless and stateful load balancers:

- **Stateless Load Balancer:** Stateless load balancers typically use hashing algorithms to transform part of the data taken from the requests or packets into a low hash value, for the selection of a server from the farm [11]. This method can permit a certain level of persistence; For example, if the data to be hashed is the source IP address of the client, all requests made by a single client will be sent to a specific server. There are multiple hashing methods: hash based on IP source, hash buckets, etc.
- **Stateful Load Balancer:** Stateless load balancers have certain limitations in a fair distribution of workloads. Stateful load balancers maintain the state of the sessions. To accomplish this, the balancer must be able to determine when a session starts and when it ends. When a session is started, the target server is determined using load distribution mechanisms. Once determined, all subsequent packets that belong to the same session will be sent to the same destination server, until the end of the session [11].

There are many methods for load distribution, such as:

- **Round-Robin:** The Round-Robin method is one of the most popular and simple methods of load distribution.

Its distribution is based on granting the incoming requests to a different server, until all servers have received one. After reaching this point, the process is repeated with the subsequent requests. Some of the advantages of Round-Robin are its simplicity and the few resources needed by the balancer. Also, the method is very fast, even in large architectures with numerous servers in the farm. However, it has the disadvantage of providing a rather poor mechanism of load distribution, because it has no mechanisms to evaluate the load of each element of the farm.

- **Choice by Load:** In this load balancing technique, each server of the farm is monitored by a program known as a server-side agent. The server-side agent provides real-time information about the load of the server where it is running, such as: CPU load, memory usage, disk I/O, etc. Hence, the load balancer obtains information about the load conditions of each server in the farm with a very high level of detail [11], before it determines which server is the most suitable to take the next request.

Many parameters can be used to assess the load of the servers in the farm. The parameters to choose will depend on the application. Some of the most common load parameters are: CPU utilization, RAM utilization, quantity of connections, and quantity of transactions or active calls (very common in solutions based on SIP). In many cases, parameters are combined for a better load evaluation.

### C. Division of Services

When there are multiple tools with the ability to offer the same functionality, it is important to consider which one can provide it in the best way. Throughout this paper, the advantages of the division of functionalities of certain services will be described, with the corresponding evidence to support the decisions.

### D. Monitoring and Management of Services and Load

The establishment of monitoring and management service mechanisms within a high availability solution allows the identification of failure points, as well as the efficient determination and management of the components involved in its operation. The monitoring mechanisms permit to have an instant state of the situation, which facilitates the decision to be taken on the exposed services.

## IV. Related Works

Some studies and solutions related to providing mechanisms of contingency and scalability in IP telephony systems currently exist. Throughout this section, some related researches will be described and analyzed, highlighting their advantages and weaknesses.

### A. High Availability for SIP: Solutions and Real-Time Measurement Performance Evaluation

This research proposes the creation of a transparent and practical failover solution for proxy servers, for both SIP and RTP [12]. The objective is to increase the availability, stability, and scalability of multimedia systems based on SIP by using active-passive failover mechanisms with floating IP addresses, for both SIP and RTP.

*1) SIP redundancy architecture:* In order to have redundancy through SIP, it is essential to have two or more SIP servers. Also, each SIP server must be aware of all the SIP transaction made [12]. This is achieved by replicating all messages received by the active server to all the backup servers, even if these last ones are not available to the public. To accomplish this, the authors proposed the creation of a daemon called High Availability Daemon (HAD), which is run in each SIP server, including a SIP proxy.

*2) RTP redundancy architecture:* The approach of the RTP architecture is very similar to the one for SIP. However, the RTP proxy allocates two port numbers for each forwarding relation. Data which are received on one port are forwarded to the other peer through the other port number. This prior knowledge by the RTP proxy is done through a mapping list that is shared by HAD.

To begin the RTP message exchange of a SIP session, the RTP proxy opens the ports and virtually interconnects with the peer. At this point, it does not know what are the IP addresses and ports of the peer and learns them when receiving the first RTP message of the peer (the mapping list is updated) [12]. This mapping list must be exchanged with the other RTP proxy. HAD is in charge for this, thus storing the mapping list and replicating it with the other HAD of the other RTP server. HAD receives mapping requests from the SIP proxy and forwards them to the local RTP proxy. Additionally, HAD intercepts the local RTP proxy responses and forwards the chosen port to the backup system. Figure 1 and Figure 2 [12] refer to the operation mechanism of the HAD daemon, for SIP and RTP, respectively.

*3) Load Balancing scheme:* In this work, a DNS load balancing mechanism based on SRV registers was used. The load balancing module is responsible for requesting in static time interval the number of SIP servers available to the DNS server (DNS records) and resolves its Fully Qualified Domain Name (FQDN). This means that for each SIP domain, the DNS server has multiple DNS records for the DNS proxies added by it.

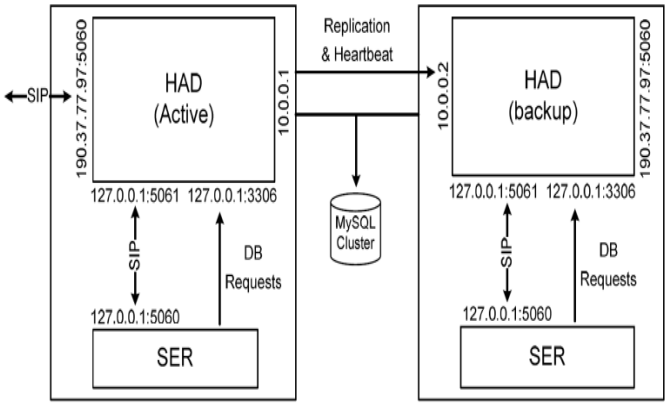



Fig. 1. SIP Replication Architecture.

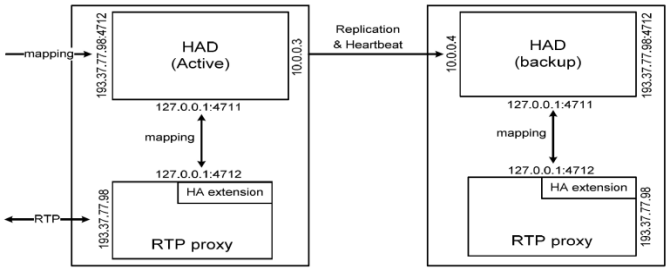



Fig. 2. RTP Relay Replication.

The balancer is responsible for querying the DNS and maintaining the DNS records. The SIP client, when making a SIP request, will first reach the balancer and, based on the information provided by the DNS and the weight of the different services, forward the request to the SIP server with more priority. This, however, is only required for certain SIP messages. Once a SIP client has established the connection (sending an INVITE and receiving an OK from the SIP server), it is not required to pass through the balancer again.

*4) Differences with our proposed architecture:* In contrast to our work, the authors wrote their own HAD daemon from scratch, and did not use a well-known and debugged software developed by the community. This solution is error-prone, since their software has only been exposed to a very small group of people. Finally, another clear difference is the use of a DNS-oriented balancer, which is stateless oriented, resulting in unfair balancing.

### B. On The Reliability of Voice over IP (VoIP) Telephony

The architecture proposed by Pal, Gadde, and Latchman [13] consists of:

- Two or more virtual servers with Kamailio.
- Two or more virtual servers with FreeSWITCH [14][15].

Calls are routed from the clients to the Kamailio servers using a distributed DNS-based ENUM (E.164 Number to URI Mapping) system with priority settings [13]. If at any moment a particular Kamailio virtual server is down or running at its full capacity, the calls will be rerouted to another Kamailio server. If none of these servers is active, the call will be rerouted directly to any of the FreeSWITCH that are responsible for the voice messaging, music on hold, and automated attendant (see Figure 3 [13]).

To provide hardware redundancy, the authors propose Ultra Monkey [16], that uses Linux Virtual Server (LVS), for the creation of high availability in network services. Ultra Monkey is a framework that uses the Heartbeat protocol to monitor if two servers are operating properly or not (presence of periodic messages). In this proposed architecture, there are two balancers configured in an active-passive clustering scheme with Ultra Monkey. At any moment, if the active balancer stops responding requests, the balancer that was in passive mode will pass to active mode and will process the requests. Heartbeat uses a plugin called IPFail that helps determine, at layer 3 (via ICMP messages), whether the balancers are working properly or not.

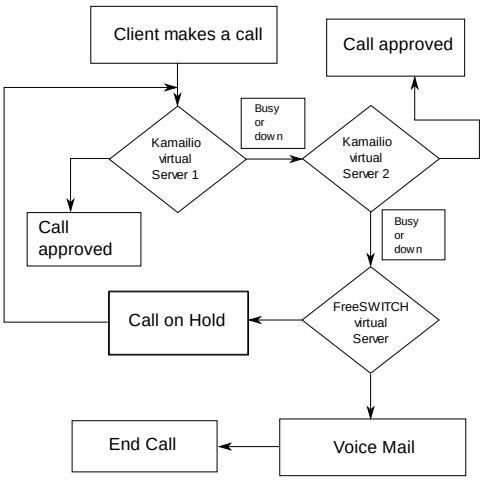



Fig. 3. Operating Scheme.

*1) Differences with our proposed architecture:* The use of Heartbeat as a communication and membership protocol differentiates this work from our proposed architecture. It is worth to remember that many developments in this field are migrating from Heartbeat to Pacemaker/Corosync, that is, Heartbeat can be considered as a project that is losing popularity. Also, the selected balancer, Ultra Monkey, does not maintain the state of SIP, resulting in a lower quality balancer than Kamailio.

### C. Design and Implementation of a System to Interconnect VoIP Services and CERN's Telephony Networks

The scope of this work [17] is broad and is intended to provide an interface to the telephone networks of the European Organization for Nuclear Research (CERN), specifically services based on SIP. The implemented system serves as an entry point for calls originated outside the CERN's telephony network, allowing users using this service to communicate with the CERN telephone network (landline or mobile).

*1) Logic topology and components:* It consists of multiple components based primarily on common existing SIP servers. Among them are Media Servers and Proxy Servers, interconnected as can be seen in Figure 4 [17]. The Media Servers provide functionalities to SIP users. They tend to use many hardware resources (since they process multimedia information by software). Proxy Servers are used to protect them from the signaling part [17]. For high availability requirements, the topology consists of two Proxy Servers and two Media Servers (active-passive and active-active, respectively), thus preventing single fault points in the topology.

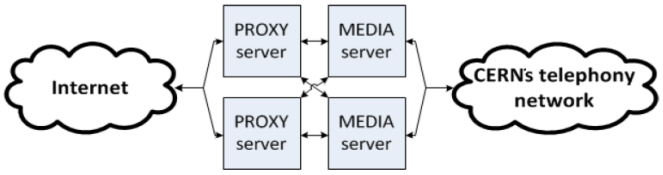



Fig. 4. Logical Topology.

The system has load balancing mechanisms served by Kamailio, including health checks or keepalive mechanisms to the different FreeSWITCH with failover capability, as well as active-active clustering techniques through Pacemaker and Corosync among Media Servers. Also, it makes use of active-passive clustering techniques between Proxy Servers.

Each Media Server is configured to manage a limited amount of channels while interrupting the others if the limit is already set. The load balancing mechanism allows the distribution of calls between the different Media Servers, thus increasing the scalability and the call limit [17].

*2) Functions of Immersed components:*

*a) Proxy Server:*

The Proxy Server must accept connections from both TCP and UDP. It must only accept calls destined for the CERN PBX; i.e., it should not act as Proxy Server among SIP users [17]. The call setup process keeps track of the progress of the call once answered. This is achieved by storing the dialog information using the Dialog module provided by Kamailio in a database [17].

*b) Media Server:*

The Media Server must reserve one channel for each incoming call and then proceed to authenticate the user of the incoming call. Users are not authenticated by using "Mod_directory" module, making it a Registrar Server, but by configuring an access list that only accepts incoming calls originated by the Proxy Server and, in addition, a custom mechanism that will be explained in the following paragraph. The SIP profiles were completely eliminated except for the internal profile in which incoming calls from Kamailio were assigned to this profile.

In addition, a method is used to perform an extra layer of users' authentication. This functionality varies depending on how the call flow is performed. This mechanism is created by scripts written in Lua where a PIN is offered to the user who is calling through an IVR. The user must set the PIN correctly, using DTMF tones, for the authentication to be complete.

*3) Differences with our Proposed Architecture:* This architecture was designed with the purpose of allowing external entities to communicate through VoIP (using the SIP protocol for session establishment) with the CERN private segments, which are legacy telephone networks. In our proposed solution, the architecture is totally based on the SIP protocol.

## V. Proposed Architecture and Solution

Figure 5 depicts the conceptual architecture of the proposed solution. The subsystem that supports high availability telephony is in network A, on the left side (inside the green box), which the following elements:

- **Proxy/Registrar:** This is the main component in charge of the services provided by the SIP signaling protocol. It acts as a stateful SIP proxy to allow load balancing mechanisms towards the PBX entities. In addition, it functions as an RTP proxy to solve problems related to NAT Traversal.
- **PBX:** These components act as multimedia servers. They are used as secondary SIP traffic receivers (in case of call establishment requirements) and these servers are being balanced by the Proxy/Registrar servers.
- **DB/SAN:** This component acts as a shared storage for the two components defined above and maintains the information centralized.

Each instance of the different components, in Figure 5, indicates whether it is active (working perfectly and providing services), passive (working perfectly but not providing services), or failed (out of order). This information refers to the implementation of the clustering schemes that are used as contingency mechanisms within the architecture, consisting of multiple instances of the same component.

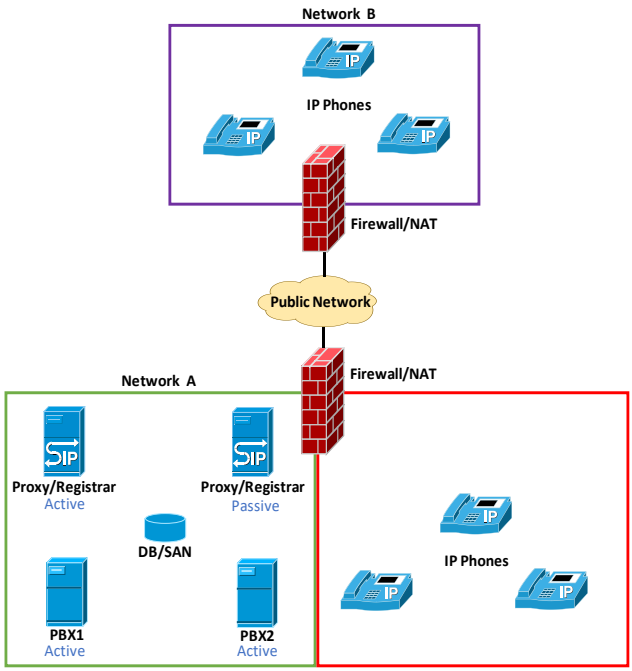



Fig. 5. General Architecture of the Proposed Solution.

The rest of the components that are part of the solution are the following:

- **Firewall/NAT:** These components act as delimiters between the internal networks and the public networks. Their main purpose is to achieve the simulation of the NAT Traversal phenomenon by having SIP users outside of network A, where the telephony architecture is implemented.
- **IP Phones:** These components act as the endpoints that require communication services using the implemented telephony architecture.

### A. Open Source Components Used in our Architecture

As mentioned previously, our solution is based on open source software such as:

*1) Kamailio:* Component previously mentioned as Proxy/ Registrar provides:

- **Kamailio v4.3:** Tool that turns this component into a stateful SIP proxy, offering SIP transaction failover mechanisms and load balancing mechanisms. It also acts as a Registrar Server and a Location Server.
- **RTPProxy v2.0.0:** Tool used to convert the component into an RTP proxy if necessary (against NAT Traversal phenomenon).

*2) Asterisk:* The component mentioned as PBX is running Asterisk v13.1. It acts as a multimedia server with the ability to offer multiple functionalities such as: transcoding, voicemail service, IVR, call transfer, etc.

*3) MySQL/SAN:* Component mentioned as DB/SAN, that acts as a shared storage for the Proxy/Registrar and PBX servers. Since some information is stored in the database and other at the file system level, there are two tools to support these services:

- **MySQL Community Edition v5.7:** MySQL is used as a database management system to maintain shared information between the different instances of the different components that are part of the architecture.
- **Linux-IO Target (targetcli v2.1):** This tool is used to convert the SAN server into a server from which an iSCSI Target is configured to allow to different instances of the Asterisk servers to create iSCSI sessions against that Target and be able to manage this storage in a shared way. This is necessary because certain Asterisk information (some configuration files, multimedia files such as voicemail, music on hold, etc.) must be shared by both instances for the proper operation of the architecture.

## B. Design of the Clustering Schemes

In this architecture, there are two clustering schemes of two nodes each, one for the different instances of the components that run Kamailio and RTPProxy, and another for the different instances of the components that execute Asterisk. As explained earlier, the reason behind these schemes is to provide contingency mechanisms in case of errors, manage shared storage, and monitor the status of the various relevant services that are part of the different schemes. These clustering schemes are formed by Pacemaker v1.1.13 [18] for the resources management and Corosync v2.3.4 [5] for the communication layer and the creation of the cluster infrastructure.

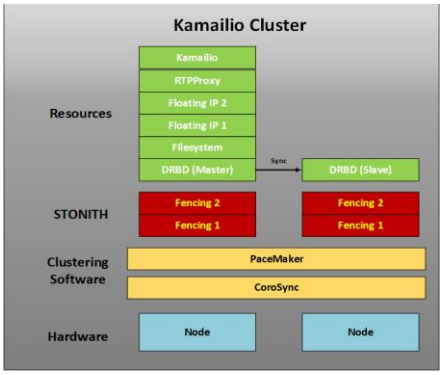



Fig. 6. Logical Organization for the Proxy/Registrar Clustering Scheme.

*1) Clustering Scheme for the Proxy/Registrar Components:* This is an active-passive clustering scheme, which means that one of the instances is the one that will be in charge of providing all the services, while the other instance will be monitoring and taking over when some error or problem occurs in the active instance. Figure 6 shows the logical organization of the components within the Proxy/Registrar cluster.

*2) Clustering Scheme for the PBX Components:* This is an active-active clustering scheme, which means that both instances are responsible for having all services active and for providing them. This scheme has the primary function of preventing data corruption of shared storage between both instances of the PBX. That is, the monitoring system is made such as that when a PBX fails while providing a service, the other PBX must terminate this service, to avoid inconsistencies in the data and a malfunction of the architecture. Figure 7 depicts the logical organization of the components within the PBX cluster.

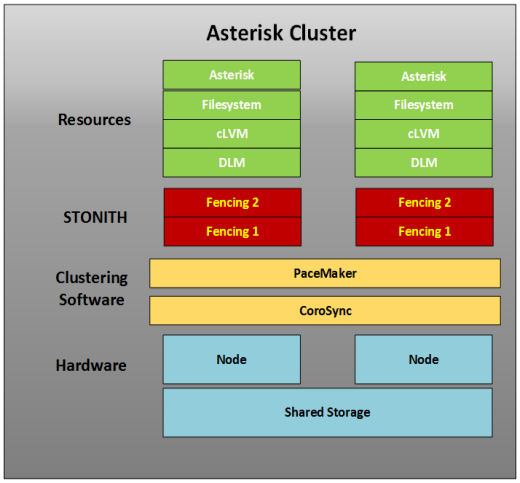



Fig. 7. Logical Organization for the PBX Clustering Scheme.

# VI. Test Scenarios

In order to validate our architecture, we developed two kinds of test scenarios: (1) scenarios to assess the contingency mechanisms and (2) scenarios to study the behavior of the proposed architecture under strong stress. It is worth mentioning that a PBX server has two possible states: (1) **active** when it is working properly and providing services and (2) **failed/passive** when it is out of order. A Proxy/Registrar component has three possible states: (1) **active** when it is working properly and providing services, (2) **passive** when it is working properly and not providing services, and (3) **failed** when it is out of order.

## A. Scenarios to Assess the Contingency Mechanisms

*1) Scenario 1 - Recovery from the Failure of a PBX Component:*

Figure 8 depicts the situation when one of the Asterisk servers fails. In this case, all the load of the services must pass to the other Asterisk server.

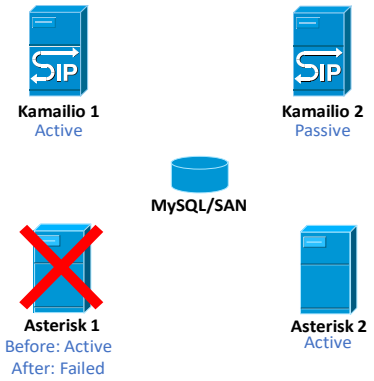



Fig. 8. Recovery from the Failure of a PBX Component.

*2) Scenario 2 - Recovery from the Failure of a Proxy/ Registrar Component:*

Figure 9 illustrates this scenario. Before the failure, Kamailio 1 was active and Kamailio 2 was passive. As soon as Kamailio 2 notices the failure of the other server, it switches from passive to active mode.

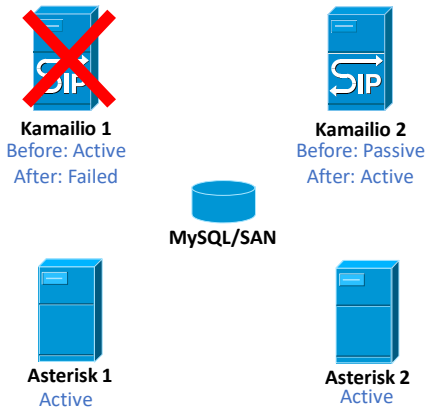



Fig. 9. Recovery from the Failure of a Proxy/Registrar Component.

*3) Scenario 3 - Recovery from the Failure of a Proxy/Registrar Component and a PBX Component:*

Figure 10 depicts this test scenario, where there are two failures: (1) failure of an instance that is providing the service of Kamailio and RTPProxy, and (2) failure of an instance that is offering the service of Asterisk.

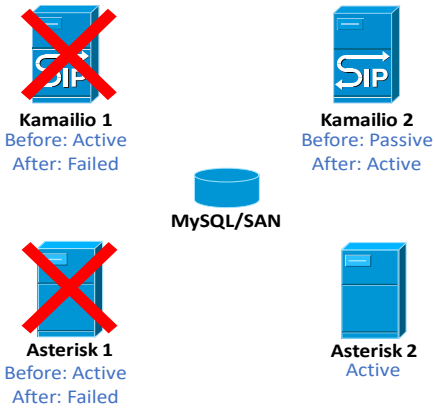



Fig. 10. Recovery from the Failure of a Proxy/Registrar Component and a PBX Component.

## B. Scenarios to Assess the Architecture under Stress

To assess the behavior of our proposed architecture in scenarios of high concurrency of SIP services (INVITE and REGISTER transactions) and RTP traffic flows, the following scenarios were proposed:

*1) Scenario 1 - System under Stress in a Conventional VoIP Implementation:*

In this scenario, there is a single PBX component, as shown in Figure 11. This is a classical VoIP system, without redundancy. The idea is to study the scalability of a conventional VoIP implementation, by exposing it to a high volume of requests.


Fig. 11. Conventional VoIP Implementation.

*2) Scenario 2 - System under Stress with the Proposed Architecture:*

Figure 12 depicts our proposed architecture with redundancy at the level of the Proxy/Registrar servers, and load balancing at the level of the PBXs. The idea is to submit the proposed architecture under a high volume of requests, and compare its behavior with the one shown in Figure 11 (conventional VoIP implementation).

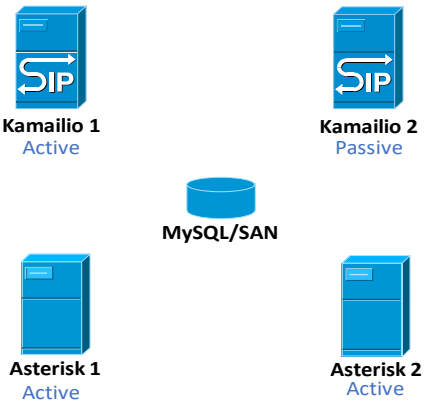



Fig. 12. Proposed Architecture when all the Components are Operating Properly.

# VII. Definitions of Tests and Particular Situations

In this section, we define the performed assessments for the proposed scenarios.

## A. Assessment of the Contingency Mechanisms

To assess the contingency mechanism, we defined three tests, that can be applied to any of the three scenarios, as specified below:

- **Test 1:** Send an INVITE transaction just after a failure and before the application of contingency mechanisms.
- **Test 2:** Send an INVITE transaction before a failure and the application of contingency mechanisms,

without the completion of the INVITE session (i.e., without multimedia flows).

- **Test 3:** Send an INVITE transaction before a failure and the application of contingency mechanisms, with the completion of the INVITE session (i.e., with multimedia flows).

Due to the ability of Kamailio to maintain the state of SIP transactions and make decisions based on the lack of response or error messages by the Asterisk nodes, we also defined the following two situations:

- **Situation 1:** Generation of an INVITE request that attempts to be transmitted to an asterisk node that has just failed, as explained in Test 1. If there is no response within 4 seconds, Kamailio will retransmit the INVITE request to the other active Asterisk node.
- **Situation 2:** It occurs after receiving the message "100 Trying". If Kamailio does not receive any further message ("180 Ringing") after 60 seconds (usually due to a crash of the Asterisk server), Kamailio will retransmit the INVITE transaction to the other active Asterisk node.

Tables I, II, and III represent the expected results (behavior and response time) for the different failover mechanisms.

TABLE I. EXPECTED RESULTS OF TRANSACTIONAL FAILOVER MECHANISMS

| Scenario | Behavior | Response Time |
|---|---|---|
| **1** | **User Interaction:** The IP phone does not participate in the contingency<br>**Contingency**: The active Kamailio retransmit the INVITE message after a timeout of the response | **Situation 1**: 4 seconds<br>**Situation 2**: 60 seconds |

TABLE II. EXPECTED RESULTS OF THE PROXY/REGISTRAR CONTINGENCY MECHANISMS

| Scenario | Behavior | Response Time |
|---|---|---|
| **2** | **User Interaction:** The IP phone replicates the initial request in case of being in the establishment of a session<br>**Contingency:** Service Reestablishment | **Fencing Application:** A few seconds<br>**Services or Nodes Errors:** A few seconds |
| **3** | **User Interaction:** The IP phone replicates the initial request in case of being in the establishment of a session<br>**Contingency:** Service Reestablishment | **Fencing Application:** A few seconds<br>**Services or Nodes Errors:** A few seconds |

TABLE III. EXPECTED RESULTS OF THE PBX CONTINGENCY MECHANISMS

| Scenario | Behavior | Response Time |
|---|---|---|
| **1** | **User Interaction:** The IP phone replicates the initial request in case of being in the establishment of a session<br>**Contingency 1:** In case of a first failure when monitoring the resources of an Asterisk node, Pacemaker restarts the service.<br>**Contingency 2:** In any other cases, Pacemaker stops the service on the node where it fails. | **Fencing Application:** A few seconds<br>**Service or Nodes Errors:** A few seconds |
| **3** | **User Interaction:** The IP phone replicates the initial request in case of being in the establishment of a session.<br>**Contingency 1:** In case of a first failure when monitoring the resources of an Asterisk node, Pacemaker restarts the service.<br>**Contingency 2:** In any other cases, Pacemaker stops the service on the node where it fails. | **Fencing Application:** A few seconds<br>**Services or Nodes Errors:** A few seconds |

### *B. Assessment of the Architecture under Stress*

To assess the behavior of the proposed architecture when suffering stress, we defined the tests shown in Table IV and Table V, which were performed with SIPp [19], a free open source traffic generator for the SIP protocol. The idea was to flood the proposed architecture with SIP requests (INVITE and REGISTER) and study how well it can manage the stress and the load balancing among the Asterisk servers. Table IV is focused on INVITE messages, while the emphasis of Table V is on REGISTER messages. From the first to the last column, Table IV contains (1) the test that we are defining, (2) the scenario used for this test, (3) the desired calling rates (number of calls made by SIPp per second), (4) the total limit of active calls at any time, (5) the duration of each call, and (6) the duration of the test, respectively. The last four parameters are set through SIPp. It is worth to clarify that the total limit of active calls is the maximum number of calls that can be active at any time during the experiment. If SIPp reaches this limit, then it will not initiate another new call until one of the active calls is completed.

From the first to the last column, Table V contains (1) the test that we are defining, (2) the scenario used for this test, (3) the desired transaction rates (number of REGISTER messages sent by SIPp per second), (4) the total number of SIP users in the system, and (5) the duration of the test, respectively. The last three parameters are set through SIPp. Note that we did not define tests for Scenario 1, since the registration of users is done by the Proxy/Registrar servers (Kamailio), and just having one of them instead of an active-passive cluster will not change the observed behavior.

TABLE IV. PARAMETERS USED IN SIPP FOR INVITE MESSAGES

| Test | Scenario | Calling Rate | Limit of Active Calls | Duration of Calls | Duration of Test |
|---|---|---|---|---|---|
| **Test 1** | **1** | 175 cps | 1000 calls | 6 seconds | 1 hour |
| **Test 2** | | 200 cps | 1000 calls | 5 seconds | 1 hour |
| **Test 3** | | 200 cps | 1200 calls | 5 seconds | 1 hour |
| **Test 4** | | 200 cps | 1200 calls | 6 seconds | 1 hour |
| **Test 1** | **2** | 300 cps | 2000 calls | 6 seconds | 1 hour |
| **Test 2** | | 300 cps | 2000 calls | 7 seconds | 1 hour |
| **Test 3** | | 350 cps | 2000 calls | 5 seconds | 1 hour |
| **Test 4** | | 350 cps | 2000 calls | 6 seconds | 1 hour |

TABLE V. PARAMETERS USED IN SIPP FOR REGISTER MESSAGES

| Test | Scenario | Transaction Rate | Number of SIP Users | Duration of Test |
|---|---|---|---|---|
| **Test 1** | **2** | 500 tps | 10000 users | 1 hour |
| **Test 2** | | 750 tps | 10000 users | 1 hour |
| **Test 3** | | 1000 tps | 10000 users | 1 hour |

## VIII. Results And Analysis

To do the validation experiments, we used a server with Xen Project [20][21] for the management of virtualized environments. All elements that are part of the architecture, as well as the network topology of the architecture, were virtualized in this server.

Since the elements that are running in the proposed architecture have different requirements at the level of CPU resources, we did not use the default configuration provided by the Xen hypervisor with a single vCPU per device. We did multiple attempts of different hardware resource distributions, and selected the one shown in Table VI.

TABLE VI. Resource Assigned to Each Device of the Architecture Through the Xen Hypervisor

| Machines | vCPUs Management | Weight Management (Schedule-Credit) |
|---|---|---|
| Domain-0 | 4 vCPUs, No affinity | 256 |
| DB/SAN | 1 vCPU, No affinity | 512 |
| Asterisk 1 | 2 vCPUs, No affinity | 1000 |
| Asterisk 2 | 2 vCPUs, No affinity | 1000 |
| Kamailio 1 | 2 vCPUs, No affinity | 1000 |
| Kamailio 2 | 2 vCPUs, No affinity | 1000 |

### A. Tests for Contingency Mechanisms

For this case, the tests are divided according to the types of failover mechanisms, which are:

*1) Tests for the Transactional Failover Mechanisms:*

Table VII represents the results obtained by our tests in Situation 1 and Situation 2, of Scenario 1. We repeated the experiments 200 times and the response time presented is an average. Our experiments showed an adequate response of the transactional failover mechanisms, with an average response time of 4.174 seconds in Situation 1 (as shown in Table I, it should be greater than 4 seconds) and 63.049 seconds in Situation 2 (as shown in Table I, it should be greater than 60 seconds).

TABLE VII. Results of the Transactional Failover Mechanisms Tests

| Scenario | Test | Situation | User Interaction | Contingency | Response Time |
|---|---|---|---|---|---|
| 1 | 1 | 1 | None | Retransmission of the INVITE message after a timeout of the response | 4.17 sec |
| | 2 | 2 | None | INVITE session retransmission after meeting response times | 63.04 sec |

*2) Tests for the Failover Mechanisms at Services and Nodes Levels:*

Table VIII shows the results obtained by our tests in the case of failover mechanisms at the levels of services and nodes. It shows the average response time at the level of users and services (we repeated the experiments 200 times and the response time presented is an average). Our experiments showed that our architecture is working well with an adequate reaction of the failover mechanisms at the level of services and nodes, confirmed the expected results of Table II and Table III.

TABLE VIII. Results of the Failover Mechanism Tests At Services and Nodes Levels

| Scenario | Test | User Interaction | Contingency | Fencing | Response Time |
|---|---|---|---|---|---|
| 2 | 1 | None | Services are migrated from one node that is part of the clustering scheme to the other. | No | User - 4.50 sec<br>Service – 3.49 sec |
| | | | | Yes | User - 7.51 sec<br>Service - 7.38 sec |
| | 2 | None | Services are migrated from one node that is part of the clustering scheme to the other. | No | User - 4.21 sec<br>Service - 3.44 sec |
| | | | | Yes | User - 7.51 sec<br>Service - 7.41 sec |
| | 3 | Replicate call establishment | Services are migrated from one node that is part of the clustering scheme to the other. | No | User – Interaction Required<br>Service - 3.48 sec |
| | | | | Yes | User - It may or may not affect the service<br>Service - 7.39 sec |
| 3 | 1 | None | Services are migrated from one node that is part of the clustering scheme to the other. | No | User – 7.47 sec<br>Service - 3.43 sec |
| | | | | Yes | User - 11.49 sec<br>Service - 7.34 sec |
| | 2 | Replicate call establishment | Services are migrated from one node that is part of the clustering scheme to the other. | No | User - Interaction Required<br>Service - 3.38 sec |
| | | | | Yes | User - Interaction Required<br>Service - 7.45 sec |
| | 3 | Replicate call establishment | Services are migrated from one node that is part of the clustering scheme to the other. | No | User - Interaction Required<br>Service - 3.44 sec |
| | | | | Yes | User - Interaction Required<br>Service - 7.39 sec |

### B. Tests of Stress

Unlike the previous tests, the results of the tests of this section are highly dependent on the hardware resources and the topology. It is worth to mention that the SIPp tool has tests for both the SIP and RTP protocols. In the case of the RTP tests, SIPp sends media (RTP) traffic through RTP echo and RTP/pcap replay. We faced some difficulties with these RTP tests. The problem is based on the fact that SIPp has a single threaded architecture, where events are repeated through a loop with the same thread, which causes the .pcap files to be managed by a single core on the SIPp machine. Thus, in our case, when reaching a maximum of 800 active calls approximately, the core reached 100% of its capacity, resulting in a malfunction of SIPp for the RTP tests. Hence, we did not use the RTP tests to assess our architecture.

The objective of these tests is to study the ability of the proposed architecture to manage a big number of calls or registrations through multiple Asterisk servers, by doing load balancing.

*1) Tests of Stress Performed with INVITE Messages:* The results are shown in Table IX (conventional VoIP implementation) and Table X (our proposed architecture). For each test, we run the experiment twice. From the first to the last column of Table IX and Table X, we have (1) the test

performed, (2) the actual duration of the test, (3) the average call rate performed by SIPp in one second, (4) the total number of calls during the experiment, (5) the average number of actives calls in the system at any moment, (6) the total number of successful calls during the experiment, (7) the total number of failed calls during the experiment, and (8) the total number of retransmissions made by SIP during the experiment, respectively.

TABLE IX. RESULTS OF STRESS TESTS PERFORMED BY INVITE MESSAGES IN SCENARIO 1 (CONVENTIONAL VOIP IMPLEMENTATION)

| Test | Duration | Avg Call Rate | Generated Calls | Avg Active Calls | Successful Calls | Failed Calls | Retransmissions |
|---|---|---|---|---|---|---|---|
| 1 | 1 hour | 142.42 | 512733 | 998.33 | 512730 | 3 | 0 |
| | | 142.42 | 512731 | 998.37 | 512726 | 5 | 0 |
| 2 | 1 hour | 166.00 | 597613 | 997.79 | 512730 | 3 | 0 |
| | | 165.97 | 597530 | 997.79 | 597522 | 8 | 0 |
| 3 | 1 hour | 197.70 | 307064 | 509.27 | 98073 | 207796 | 23530 |
| | | 199.10 | 717763 | 1197.62 | 717658 | 105 | 104 |
| 4 | 1 hour | 170.86 | 615960 | 1199.04 | 615859 | 101 | 0 |
| | | 170.85 | 615953 | 1198.83 | 615830 | 123 | 2 |

As we can see from Table IX and Table X, the proposed architecture significantly improves the VoIP system. The number of successful calls processed by our solution is almost the double of the one handled by a conventional VoIP architecture.

TABLE X. RESULTS OF STRESS TESTS PERFORMED BY INVITE MESSAGES IN SCENARIO 2 (OUR PROPOSED ARCHITECTURE)

| Test | Duration | Avg Call Rate | Generated Calls | Avg Active Calls | Successful Calls | Failed Calls | Retransmissions |
|---|---|---|---|---|---|---|---|
| 1 | 1 hour | 284.75 | 1025107 | 1997.58 | 1025091 | 16 | 0 |
| | | 284.80 | 1025314 | 1997.66 | 1024246 | 1068 | 34 |
| 2 | 1 hour | 249.34 | 897640 | 1997.66 | 897626 | 14 | 13 |
| | | 248.77 | 897573 | 1997.75 | 897562 | 11 | 58 |
| 3 | 20 min | 345.76 | 418292 | 344.20 | 53614 | 364574 | 29308 |
| | | 346.87 | 420175 | 268.00 | 59533 | 360638 | 33832 |
| 4 | 1 hour | 284.87 | 1025566 | 1998.33 | 1025001 | 565 | 0 |
| | | 284.76 | 1025162 | 1998.79 | 1025143 | 19 | 4 |

In Figure 13, we have more details on the evolution of the experiment that is highlighted in Table X. The x-axis shows the time during the realization of the experiment (0, 600, 1200, 1800, 2400, 3000, 3600, and 3607 seconds). For each sample, there are four bars representing (1) the number of active calls in PBX1, (2) the number of active calls in PBX2, (3) the total number of processed calls by PBX1, and (4) the total number of processed calls by PBX2. This experiment confirmed the correct operation done by Kamailio 1 (the active Proxy/ Registrar server), when doing load balancing between the two PBXs.

*2) Tests of Stress Performed with REGISTER Messages:* The results are shown in Table XI. For each test, we run the experiment twice. From the first to the last column of Table XI, we have (1) the test performed, (2) the actual duration of the test, (3) the average number of transactions/registrations performed by SIPp in one second, (4) the total number of transactions during the experiment, (5) the total number of successful transactions during the experiment, (6) the total number of failed transactions during the experiment, and (7) the total number of retransmissions made by SIP during the experiment, respectively.

TABLE XI. RESULTS OF STRESS TESTS PERFORMED BY REGISTER MESSAGES IN SCENARIO 2

| Test | Duration | Avg Transaction Rate | Generated Transactions | Successful Transactions | Failed Transactions | Retransmissions |
|---|---|---|---|---|---|---|
| 1 | 1 hour | 498.81 | 1795733 | 1795733 | 0 | 72855 |
| | | 498.86 | 1795932 | 1795932 | 0 | 71952 |
| 2 | 1 hour | 748.15 | 2693403 | 2693385 | 18 | 232129 |
| | | 748.29 | 2693882 | 2693882 | 0 | 174833 |
| 3 | 1 hour | 996.05 | 3585832 | 3585831 | 1 | 301450 |
| | | 995.23 | 3582895 | 3582887 | 8 | 275637 |

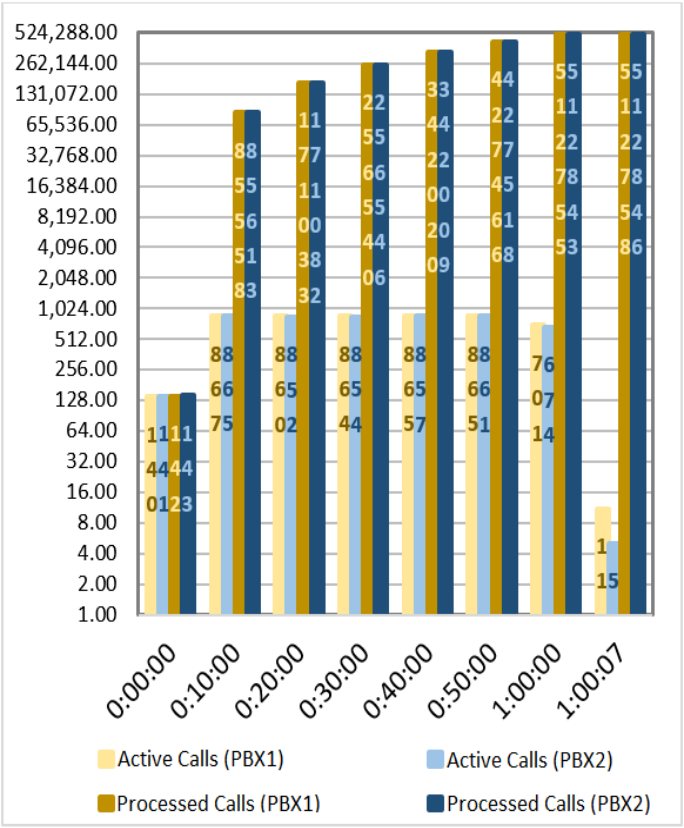



Fig. 13. Active and Processed Calls as the Time Passes - Confirmation of Load Balancing.

## IX. CONCLUSIONS AND FUTURE WORK

VoIP is a technology that has achieved a great impact on the communication of digital networks. It takes advantage of existing networks and different standard protocols such as SIP, Inter-Asterisk eXchange [22] (IAX), RTP, and RTCP. On the one hand, many institutions have implemented VoIP solutions based on proprietary software, which tends to increase the implementation cost. On the other hand, other organizations

have decided to go for open source solutions like Asterisk. Asterisk is focused on the inclusion and development of multiple services and the support of multiple signaling protocols. Hence, per see, it provides a high-quality service for telephony, but has weak aspects such as scalability and contingency mechanisms in case of failures.

In this work, we proposed a telephony architecture based on Asterisk and the SIP signaling protocol that covers those aspects not well supported by Asterisk per see, such as scalability and fault tolerance. Our solution is totally focused on open source software. In our architecture, we inserted different tools that allow high availability, and clustering concepts to offer contingency mechanisms, as well as load balancing technics to add greater scalability to the provided telephony service.

As future work, we are interested in developing other high availability VoIP telephony architectures based on the SIP protocol, using SIP servers such as OpenSIPS [23], in conjunction with clustering technology tools such as Pacemaker, Corosync, CMAN, etc. Another direction of research that we are planning to investigate is to construct cluster solutions based on different software, that is, instead of having several copies of the same server for redundancy, we are interested in using different implementations of software servers so that the system can be tolerant to bugs.

## REFERENCES

[1] A. Johnston, SIP - Understanding the Session Initiation Protocol, Fourth Edition, Artech House, October 2015.

[2] R. Ranjan Roy, Handbook on Session Initiation Protocol: Networked Multimedia Communications for IP Telephony, First Edition, CRC Press, March 2016.

[3] N. Leonard, Understood SIP and IP Telephony in 3 Days: For the Beginner, Independently published, August 2017.

[4] Kamailio, https://www.kamailio.org.

[5] S. Dake, The Corosync High Performance Shared Memory IPC Reusable C Library, in Proceedings of the Linux Symposium, Montreal, Canada, July 2009.

[6] V. Stanislovaitis, How to Start a VoIP Business: A Six-Stage Guide to Becoming a VoIP Service Provider, February 2016.

[7] R. Bryant, Asterisk: The Definitive Guide: The Future of Telephony Is Now, O'Reilly, Fourth Edition, June 2013.

[8] P. Weygant, Cluster for High Availability: A Primer HP Solution, Second Edition, Prentice Hall, 2001.

[9] M. Resman, CentOS High Availability, Packt Publishing, April 2015.

[10] E. Marcus, Blueprints for High Availability, Second Edition, Wiley Publishing, September 2003.

[11] C. Kopparapu, Load Balancing, Servers, Firewalls and Caches, Wiley Computer Publishing, January 2002.

[12] G. Kambourakis, D. Geneiatakis, S. Gritzalis, C. Lambrinoudakis, T. Dagiuklas, S. Ehlert, and J. Fiedler, "High Availability for SIP: Solutions and Real-Time Measurement Performance Evaluation," International Journal of Disaster Recovery and Business Continuity, vol. 1, no. 1, pp. 11-29, February 2010.

[13] S. Pal, R. Gadde, and H. Latchman, On the Reliability of Voice Over IP (VoIP) Telephony, University of Florida, 2011.

[14] A. Minessale and G. Maruzzelli, FreeSWITCH 1.8 - VoIP and WebRTC with FreeSWITCH: The Definitive Source, Packt Publishing, July 2017.

[15] A. Minessale and G. Maruzzelli, Mastering FreeSWITCH, Packt Publishing, July 2016.

[16] Ultra Monkey, Load Balancing and Highly Available Solutions, http://www.ultramonkey.org.

[17] M. Pohančeník, "Design and Implementation of a System to Interconnect VoIP Services and CERN's Telephony Network," University of Žilina, Ginebra, 2013.

[18] Pacemaker, http://clusterlabs.org/pacemaker.

[19] R. Gayraud, O. Jacques, R. Day, C. P. Wright, et al., SIPp Reference Documentation, http://sipp.sourceforge.net/doc/reference.pdf.

[20] T. Mackey and J. K. Benedict, XenServer Administration Handbook: Practical Recipes for Successful Deployments, First Edition, O'Reilly Media, April 2016.

[21] M. Reed, Mastering Citrix XenServer, Packt Publishing, December 2014.

[22] M. Spencer, B. Capouch, E. Guy, F. Miller, and K. Shumard, IAX: Inter-Asterisk eXchange Version 2, RFC 5456, February 2010.

[23] F. Goncalves and B.-A. Iancu, Building Telephony Systems with OpenSIPS - Second Edition, Packt Publishing, January 2016.

# Aesthetics Versus Readability of Source Code

Ron Coleman
Computer Science Department
Marist College
Poughkeepsie, NY, 12601, United States

***Abstract*—The relationship between programming style and program readability has never been examined empirically, although the association has substantial importance for both pedagogical and industry best practices. This paper studies a fractal, relativistic measure of programming style called the beauty factor or "beauty" and puts forward two new hypotheses of beauty. First, code with increasing beauty tends to be more readable. Second, beauty measures a unique property in code called aesthetic value distinct from readability. These hypotheses are tested on a corpus of 53,000 lines of open source system codes written by experienced Linux programmers. Statistical correlation analysis is used on 11 different beauty factors versus eight different readability models (i.e., 88 experiments total). As the primary finding, the data show the maximum absolute statistically significant correlation is $|\rho|=0.59$ whereas the absolute median correlation is $|\rho|=0.33$. In other words, at least 65% of statistically significant variations in beauty cannot be explained by variations in readability; approximately 90% of statistically significant variations in beauty cannot be explained typically by variations in readability. These results lend support to both hypotheses. The data further shows indentation is more reliably correlated with readability than mnemonics or comments and GNU style is more correlated with readability than K&R, BSD, or Linux styles.**

***Keywords—Programming style; fractal geometry; readability***

"We have seen that computer programming is an art, because it applies accumulated knowledge to the world, because it requires skill and ingenuity, and especially because it produces objects of beauty." ~Donald E. Knuth, 1974 ACM Turing Award Lecture [1]

## I. Introduction

In 1979 AT&T Bell Labs released Unix version 7 which included *cb*, the C beautifier [2]. A unique tool and the first of its kind, *cb* "beautified" programs, that is, reformatted them, according to rules prescribed by the K&R style [3]. In creating *cb*, software engineers did not address the ontology of *what is* beauty in code. They instead focused on the epistemology of *what is knowable* about such beauty which could be automated, demonstrated, practiced, and taught. They had hoped paying attention to sensori-emotional or aesthetic values in code might also promote readability of code. However, they were not explicit about this and left the precise nature of the relationship between aesthetics and readability assumed and open. The literature frequently confounds aesthetics and readability as if they are interchangeable (see for instance [4-6]). Indeed, the long-held, widely taught, and often repeated justification for good style is to make programs more readable and presumably, more maintainable [7].

One problem is that this view has never been tested empirically. Another problem is ontological. Readability is about understanding code and aesthetics is about appreciating it, *l'art pour l'art*. Thus, in principle, conflating aesthetic and readability is a category mistake. Finally, how different styles affect cogntive ease or difficulty of grasping code remains poorly understood. In other words, programmers may assume on the basis of preference, experience, etc., that one style is more readable than another but is personal taste as such supported by the data? The question has never been investigated empirically.

This paper studies a relativistic, fractal measure of programming style called the beauty factor (or "beauty" for short) and puts forward two new hypotheses. First, code with increasing beauty tends to be more readable. Secondly, beauty measures a unique property of code, call it aesthetic appeal, which is distinct from readability. To test these hypotheses, the paper assesses the beauty and readability scores of a statistically meaningful number of files and analyzes the correlations. No correlation denies the first hypothesis. Strong correlation denies the second hypothesis. Thus, both hypotheses can be logically true simultaenously only with weak or moderate correlation. In other words, the hypothesized aesthetic value of code is not necessarily completely orthogonal to nor a proxy for readability but a spectrum of potential.

## II. Related Work

Kokol, et al, [8-10] showed that programs contain long-range correlations in characters and tokens. These researchers were searching for a fractal metric of software complexity using lexical analysis of a small sample of randomly generated Pascal programs. This paper uses a larger corpus of production C codes and image analysis. Coleman and Gandhi [11] hypothesized programming style might be related to fractals since fractals are often associated with beauty [12]. This effort showed that changes in style were systematically correlated with changes in fractal dimension. Coleman and Gandhi [13] proposed a fractal, relativistic model and showed beauty was weakly to moderately correlated with software complexity in directions that comported with style recommendations. In other words, there is virtue in brevity and simplicity in code. Coleman and Boldt [14] investigated disorder that is often introduced in code through maintenance and showed beauty was weakly to moderately anti-correlated with entropy. Coleman and Rahtelli [15] showed the beauty model predicts aesthetic value in scientific libraries; in this paper we study system codes. These investigations of beauty resemble efforts by researchers who used fractal geometry to assess aesthetic

values in paintings, including Pollock's "action paintings" [16-20]; the main technical differences are that this paper studies code as opposed to fine art and programming style as opposed artistic style. *Beautiful Code* [21] deals with conceptual beauty in the design and analysis of algorithms, testing, and debugging, topics which are outside the scope of this paper. Also outside the scope of this paper is the more general topic of program comprehension, which focuses on cognitive models to explain how programmers understand code and the development of tools to aid them [22, 23].

## III. THEORY

### A. Aesthetic Theory

#### 1) Working Definitions

The subject of aesthetics has a long history of thoughtful consideration of matters of style and taste dating to ancient. According to the Stanford Encyclopedia of Philosophy the term "aesthetic" refers to, among other things, a kind of heuristic judgement or value called beauty, appeal, appreciation, virtue, goodness, etc. [24] We use the term "aesthetic appeal" when referring sensori-emotional judgements as opposed to other forms of appeal. Thus, the working definition of "beauty," namely, the measure of programming style, is a special case of the aesthetic definition. We say "beautiful" code has measurably more style than indecorous code. To "beautify" code is to measurably improve its style according to what we know about well-written code in terms of layout and structure. To "debeautify" code is to do the opposite; in fact, there are tools, known as "obfuscators" that use anti-style techniques usually for information security purposes [25, 26].

#### 2) Immediacy and Disinterest theses Applied to Code

The immediacy thesis maintains that judgements of appeal are immediate or straightforward through sensory discernment. [24] For code, this suggests the possibility of appreciating its form as a work of art, that is, without attempting to "read" or "understand" its function. This paper argues (see below) that the rules of judging good programming style in this way are widely known and firmly established. So much that these rules can be automated by programs like *cb* which reformat target code without regard for how the target works or even what it does. What matters foremost, aside from preserving the semantics, is how the target looks in the end.

The disinterest thesis claims that judgements of appeal are not self-interested. This is the sense of "art for its own sake." For code, a programmer could appreciate the rules of a style like K&R for "intellectual and emotional satisfaction," as Knuth described it. [1] Whether K&R or one of the other styles we study makes the code more readable and how much is the subject of this paper.

#### 3) Basic Tenets

The beauty model presupposes we know or can know epistemologically what programmers think—or better, how they *feel*—about style. This knowledgebase already exists in a mature and rich form that can be observed directly, repeatedly, and systematically in style guides, coding standards, organizational coding policies, textbooks, research reports, example codes, blogs, etc. It can also be observed indirectly through use and side-effects of tools like *cb*, functions embedded in modern IDEs for reformatting code code, and online sites that have codified this knowledge for a variety of styles and languages. In an observational study, Coleman and Gandhi [13] surveyed this knowledgebase and identified three general principles they called "basic tenets" of good programming style: namely, 1) use white space judiciously; 2) choose mnemonic names; and 3) include documentation.

#### 4) Beauty Factors

Let S be some source called the control or "baseline." Then, we have S' such that

$$S' = T(S) \tag{1}$$

where $T$ is a semantic-preserving transformation or treatment. That is, S and S' differ only in style. There are two modalities of $T$ with respect to the basic tenets: beautification and debeautification. We encode $S$ and $S'$ separately as an in-memory bitmap called an *artefact*. (The fact that it is in-memory only serves to say it is independent of file format althought it may for some reasons reside in secondard storage.) Finally, we measure the fractal dimension of these artefacts using reticular cell counting (a.k.a., the box counting dimension), $D$ [27]:

$$D(S) = \lim_{r \to 0} \frac{\log N_r(S)}{\log 1/r} \tag{2}$$

where $N_r(S)$ is the number of components (i.e., cells) covered by the ruler of size $r$. (Note that $S$ in the above equation refers to the artefact of the baseline and not the baseline itself; we use this form only to simplify the notation.) Thus, $D(S)$ is the slope of the regression line over different ruler sizes, $r$. We similarly measure $D'=D(S')$. The beauty factor model, $B$, is given by the following equation:

$$B(S \mid T) = k \log (D / D') \tag{3}$$

where $k$ is a constant. When $k$=10 and the logarithm is base ten, the units are decibels. $B$ is indicated as follows:

1. If B < 0, the style of $S$ might be improved by $T$.
2. If B ≥ 0, the style of $S$ probably won't be improved by $T$.

A *contrary indication* implies $T$ is categorized in a modality mathematically with $-B$.

#### 5) Semantic-Preserving Transformations

The tables 1 and 2 give the semantic-preserving treatments we use in this paper.

TABLE I. BEAUTIFYING TREATMENTS

| T | Tenet | Regime |
|---|---|---|
| GNU | 1 | Apply GNU style [28]. |
| K&R | 1 | Apply K&R style [3]. |
| BSD | 1 | Apply BSD style [29]. |
| LIN | 1 | Apply Linux style [30]. |
| MNE | 2 | Refactor names to be more mnemonic |
| REC | 3 | Add one or more comments. |

TABLE II. DE-BEAUTIFYING TREATMENTS

| T | Tenet | Regime |
|---|---|---|
| NOI | 1 | Remove indents. |
| R2 | 1 | Randomize indent with 1-2 spaces. |
| R5 | 1 | Randomize indent with 1-5 spaces. |
| NON | 2 | Refactor names to be less mnemonic |
| DEC | 3 | Remove all comments. |

These transformations are not, nor intended to be, exhaustive. Rather, they are indicative of transformations in general and sufficient to test the main hypotheses.

*6) Block Artefact Method*

The literal artefact method (or LAM) encodes the bitmap using a fixed width font as S (or S') literally looks like itself. The figure 1 shows the file, hello.c, with its LAM encoding. We do not use LAM in this paper.

```
#include <stdio.h>
int main(int argc, char** argv)
  printf("Hello, world!");
  return 0;
}
```

Fig. 1. LAM encoding of hello.c

The block artefact method (or BAM) which we use in this paper encodes the bitmap with block characters (e.g., ■ ) in place of the regular, textual characters and tabs. It leaves spaces as blanks. The figure 2 shows hello.c with its BAM encoding

Fig. 2. BAM encoding of hello.c

From a fractal point of view, LAM and BAM are strongly correlated with $r$=0.95 and thus, only one of them is needed for the purposes of this paper [10]. The primary advantage of BAM over LAM is the former is more robust against language dependencies, both programming and the cultural languages in the case of names and comments. Furthermore, BAM effectively destroys the source readability in favor of the text's spatial-visual pattern in direct support of the immediacy thesis. We use only BAM in this paper.

## B. Readability Theory

The literature on program understanding presupposes the existence of a posited cognitive load associated with comprehending code through an understanding hierarchy. The "bottom-up" theory of understandability maintains that the first step in grasping the code is to read it. Higher mental models of program flows, organization themes, abstraction, design patterns, etc. develop from this first and arguably, necessary step. Readability models hypothesize that the cognitive load is measurable and furthermore, correlated with density of operators and operands, logic complexity, lines of code (LOC), statement length, number of statements, etc. Advocates apprehend these quantities as constituting metrics they define as "readability". While it is clear that readability models reside at the lowest level in the understand hiearchy, they do not, nor are they designed to, capture all aspects of program understanding. According to Borstler, et al [31], readability models have value precisely because they "catch" some aspects of program understanding rather than attempting to measure it entirely.

Readability research is perhaps represented variously in the literature by three distinct generations: Halstead, machine-learned, and prose-inspired. They are similar in that they each use syntactic featuires (e.g., line length, number of identifiers, length of identifiers, etc.) to assess the readability of source.

*1) Halstead Statistics*

Halstead [32] was interested in predicting programming effort to which he related physical quantities like volume and gas pressure. He defined *program length*, *N*, to be the sum of operators, $N_1$, and operands, $N_2$,

$$N = N_1 + N_2 \tag{4}$$

*Program vocabulary* is the sum of unique is the sum of unique operators, $n_1$, and unique operands, $n_2$,

$$n = n_1 + n_2 \tag{5}$$

*Program volume* is a measure similar to Shannon entropy, namely,

$$V = N \log n \tag{6}$$

*Difficulty* combines information about unique operators and operands and total operations,

$$Df = N_1 \, n_1 / \; 2 \, n_2 \tag{7}$$

*Effort* is difficulty as a multiplier of volume:

$$E = Df \times V \tag{8}$$

A higher value in any of these statistics predicts the code is more difficult to read.

*2) Machine-Learned Metrics*

Buse and Weimer (or BW) [33] used supervised learning to train a Bayesian classifer to associate Halstead-type measurements with human judgements of the same code. After training is complete, the classifier operation is $X \rightarrow Y$ where $X$ is a code snippet and $Y$ is its predicted readability score.

Posnet, Hindle, and Devanbu (or PHD) [34] endeavored to simplify and improve the BW model using classical software engineering and information theory defined as

$$PHD = 1 / (1 + e^{-z}) \tag{9}$$

where

$$z = 8.87 - 0.033 \, V + 0.40 \, LOC - 1.5 \, H \tag{10}$$

where $V$ is the Halstead volume, LOC is lines of code, and H is the Shannon information of tokens, namely,

$$H(S) = -\sum_{i=1}^{n} p_i \log p_i \tag{11}$$

and $p_i$ is the fraction of tokens, $s_i$, in a source, $S$. In this case, the learning was derived through regression analysis.

A higher value of BW or PHD predicts the source code is more difficult to read.

*3) Prose-Inspired Metrics*

These metrics are inspired by Flesch-Kincade [35] readability score for prose. Abbas [36] with Borestler, Cspersend, and Nordstrin [31] developed the "average sentence length" (ASL) metric which is the average number of tokens per statement. They also formulated the "average word length" (AWL) metric, which is the average length of tokens adusted for special operators like ".". The authors integrated these measures into a model they called the "software readability ease score or SRES,

$$SRES = ASL - 0.1\ AWL \quad (12)$$

A higher SRES value predicts the source code is more difficult to read.

*4) Readability Indications*

The table 3 summarizes the readability indications each model.

TABLE III. READABILITY MODEL INDICATIONS

| Model | Predicts |
|---|---|
| *N* | Difficulty |
| *n* | Difficulty |
| *V* | Difficulty |
| *Df* | Difficulty |
| *E* | Difficulty |
| BW | Ease |
| PHD | Ease |
| SRES | Difficulty |

## IV. METHODS

The experimental design is conceptually very simple: generate the beauty and readability scores for each file in the corpus and using robust methods, analyze correlations.

### A. C language

The choice of C was motivated by a few factors, one being that C is one of the most widely used languages, frequently at or near the top of popularity and for a number of years [37, 38]. Another is many modern language designs have been significantly influenced by and borrow from C and consequently, what may be valid for C could have implications broadly for other languages. Finally, C has a production-quality beautifier tool, *indent* [39], which succeeds *cb*. It generates GNU, K&R, BSD, and Linux styles in Table 1.

### B. Corpus

The GNU Core Utilities [40] is the testbed. It consists of 114 open source programs that comprise of the standard shell commands like *cat*, *ls*, *sort*, etc. of the Linux operating system. In total, there are nearly 70,000 LOC. However, we don't use the C programs directly. We instead follow Coleman and Gandhi [14] and strip compiler and preprocessor directives, prototypes, typedefs, copyright notices, and the like, then decompose the remains into 1,043 single-function C files or approximately 53,000 LOC. Essentially, anything outside a function definition gets filtered except for comments that immediately precede the function definition; we keep those. We believe single-function C files as such reduces the number of confounding variables and simplifies the study. Furthermore, single-function C files lend the experimental results greater generalization for languages like Java, Python and others that are similar to C except in those aspects we remove.

The figure 3 (Chart) below gives summary statistics of the corpus of 1,043 baselines. The minimum length is 2 LOC; the median is 32 LOC; and the maximum is 1,034 LOC.

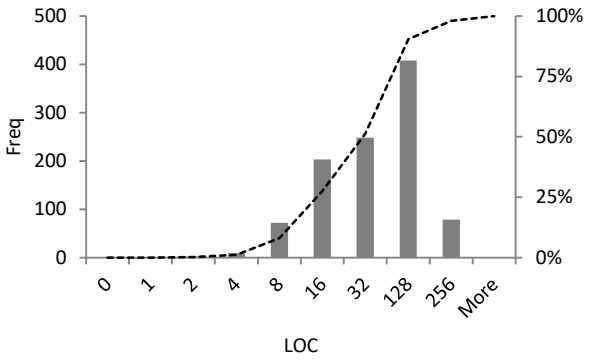



Fig. 3. Distribution of baseline sizes of the test bed corpus.

### C. Fractal Dimension

To estimate the fractal dimension we repurpose a library, Fractop [41], which had been originally designed to estimate the fractal characteristics of nerve tissue images. For ruler, *r*, in Equation 2, we use 2, 3, 4, 6, 8, 12, 16, 32, 62, and 128 pixels which is the default grid in Fractop.

### D. Result Matrices

There are 11 beauty factors and eight readadbility scores, respectively. For each file in the corpus, calculate these 19 scores. Since we are studying only aesthetics versus readability, not aesthetics by itself nor readability by itself, we only compute the correlation coefficients to generate the 11×8 matrix. However, for simplicity, we present the results in three separate, grayscale-encoded matrices: 11×5 for the Halstead models, 11×2 for the machine-learning models, and 11×1 for prose-inspired model. We intend the use of grayscale only to explicate and explain general patterns in the data.

### E. Statistical Methods

Preliminary analysis of results using the Kolmogorov-Smirnov test of normality suggest the distributions of beauty factors and readability scores are not Gaussian. Thus, we use instead Spearman's rho (ρ), the rank-based correlation coefficient. Since the degree of correlation is important for our study, we use widely accepted definitions of "weak", "moderate" and "strong" correlation per the table 4.

TABLE IV. CORRELATION DEFINITIONS WITH RANGES, UPPER-BOUND P-VALUES, AND GRAYSCALE INDICATOR SHADES

| Correlation | $\lvert\rho\rvert$ | $\lceil$ *P*-value $\rceil$ | Indicator |
|---|---|---|---|
| Strong | [0.70, 1.00] | $< 10^{-153}$ | |
| Moderate | [0.30, 0.70] | $< 10^{-22}$ | |
| Weak | [τ, 0.30] | <0.05 | (clear) |
| None | [0, τ] | ≥0.05 | |

where $\tau$=0.051 for the one-tailed test (i.e., where direction matters) versus $\tau$=0.061 for the two-tailed test (i.e., where direction doesn't matter).

According to our hypotheses, we expect readability ease (see Table 3) to be positively correlated with beautification and negatively correlated with debeautification. We expect readability difficulty to be negatively correlated with beautification and positively correlated with debeautification. If beauty is contraindicated, then these correlations are reversed.

### F. Code Analysis Code

The codes that analyze the corpus for beauty and readability are freely available on GitHub.com [42]. The remainder of statistical methods are implemented in Microsoft Excel.

## V. Results

This section gives the results of experiments.

### A. Beauty versus Halstead metrics

The table 5 gives rank correlations of beauty with Halstead statistics.

TABLE V. Spearman's Rank Correlation Coefficients of Beauty Vs. Halstead Statistics

| B | | Halstead statistics | | | | |
|---|---|---|---|---|---|---|
| Modality | T | N | n | V | Df | E |
| Debeautify | NOI | -0.55 | -0.55 | -0.56 | -0.53 | -0.56 |
| | R2 | 0.15 | 0.15 | 0.15 | 0.14 | 0.15 |
| | R5 | 0.10 | 0.10 | 0.10 | 0.09 | 0.10 |
| | NON | 0.48 | 0.47 | 0.48 | 0.49 | 0.50 |
| | DEC | 0.18 | 0.18 | 0.18 | 0.24 | 0.21 |
| Beautify | GNU | -0.58 | -0.57 | -0.59 | -0.56 | -0.59 |
| | K&R | -0.46 | -0.45 | -0.46 | -0.46 | -0.47 |
| | BSD | -0.46 | -0.45 | -0.46 | -0.37 | -0.43 |
| | LIN | -0.51 | -0.50 | -0.51 | -0.53 | -0.53 |
| | MNE | -0.31 | -0.28 | -0.30 | -0.31 | -0.31 |
| | REC | 0.09 | 0.07 | 0.09 | 0.10 | 0.09 |

### B. Beauty Versus Machine-Learned Metrics

The table 6 gives rank correlations of beauty with machine-learned statistics.

TABLE VI. Spearman's rank correlation coefficients of beauty vs. machine-learned metrics

| B | | Machine-learned metrics | |
|---|---|---|---|
| Modality | T | BW | PHD |
| Debeautify | NOI | 0.36 | 0.33 |
| | R2 | -0.12 | -0.12 |
| | R5 | -0.05 | 0.01 |
| | NON | -0.37 | -0.36 |
| | DEC | -0.18 | -0.07 |
| Beautify | GNU | 0.44 | 0.31 |
| | K&R | 0.31 | 0.31 |
| | BSD | 0.38 | 0.29 |
| | LIN | 0.29 | 0.38 |
| | MNE | 0.17 | 0.18 |
| | REC | -0.06 | -0.00 |

### C. Beauty Versus Prose-Inspired Metric

The table 7 gives rank correlations of beauty with prose-inspired statistics.

TABLE VII. Spearman's rank correlation coefficients of beauty vs. prose-inspired metric

| B | | Prose-inspired metric |
|---|---|---|
| Modality | T | SRES |
| Debeautify | NOI | 0.38 |
| | R2 | 0.04 |
| | R5 | 0.02 |
| | NON | 0.29 |
| | DEC | 0.05 |
| Beautify | GNU | -0.44 |
| | K&R | -0.35 |
| | BSD | -0.41 |
| | LIN | -0.29 |
| | MNE | -0.16 |
| | REC | 0.04 |

## VI. Discussion

We summarize patterns in the data in a series of points. Note that in every case except where explicitly noted, the results are statistically significant with the ceiling of corresponding *P*-values in Table 4.

### A. Prevalence of Weak-to-Moderate Correlations

The data in Tables 5-7 clearly beauty and readability are related in statistically significant ways. The correlations are all weak or moderate (i.e., the cells are clear or shaded as ■; there are no cells shaded as ■.) There is one There are a few cases of no correlation but these are not statistically significant. These data support our first hypothesis.

Furtheremore, beauty and readability are correlated in the directions we would expect with the exception of NOI and REC (see below). For instance, in Table 5 the de-beautifying treatments are positively correlated with Halstead statistics while all the beautifying treatments are anti-correlated with Halstead statistics. A similar pattern exists in Table 6—and appropriately since SRES, like the Halstead statistics, indicate difficulty of readability. In Table 7, we note the opposite pattern: BW and PHD indicate ease of readability and they are thus, positively correlated with beauty factors.

### B. Absence of Strong Correlations

Tables 5-7 show no evidence of strong correlations. If we ignore direction, the range of statistically significant correlations is $|\rho|=[0.05,0.59]$. Thus, according to $R^2$ analysis, at least 65% of variations in beauty cannot be explained by variations in readability. However, the median correlation which is more representative, $R^2=0.0961$. In other words, less than 10% of variations in beauty can be explained by variations in readability. These data support our second hypothesis: beauty is a unique property in so far as readability is concerned. In other words, beauty and readability are not proxies.

### C. Infrequency of Zero Correlation

There is scant evidence of zero correlations (i.e., cells boxed as □) in Tables 5-7. There are some correlations that are near zero but they are not statistically significant.

### D. NOI as a Contrary Indicator

Although categorically NOI is a debeautifying regime, statistically it behaves like a beautifying regime. This offers some insight into how beauty factors work. When a decrease in readability is accompanied by a increase in surface texture (i.e., B>0), there is negative correlation between beauty and readability. However, NOI decreases the average surface texture since more text is collected in the left edge of the artefact. This results in a positive correlation between de-beauty and readability. Generally, the debeautifying regimes in this paper tend to decrease surface texture while the beautifying regimes tend to increase surfacr texture.

### E. REC as a Possible Contrary Indicator

In the cases, of Halstead statistics, REC appears to be a reasonable candidate as a contrary indicator. The correlation coefficient has the opposite direction implied by its modality for all readability scores. However, for PHD and SRES, the correlations are not statistically significant.

### F. R2 and R5

Unlike NOI and to a less extent REC, the correlations for R2 and R5 are in the appropriate directions for their implied modality. R2 is not statistically significant for just one of the readability models (i.e., SRES) while R5 is not statistically significant for three of them (i.e., BW, PHD and SRES).

### G. A case for "self-documenting" Code

Mnemonics are a form of documentation. The data for MNE suggests perhaps a more reliable approach to comment (i.e., as opposed to REC) to improve style is through mnemonics. That is make the code more "self-documenting" by choosing symbol names that reflect their use and aids in memory. While these data are not in any way intended to settle controversies concerning comments in code [43-45], it lends support to the "no comments" school, at least as far as aesthetics are concerned.

### H. Improving Readability Through White Space

Tables 5-7 shows that all beauty treatments that affect basic tenet #1 (i.e., GNU, K&R, BSD, AND LIN) correspond consistently to improvements in readability. These treatments are more strongly correlated with readability than MNE or REC. In other words, it appears the most efficient means to improve readability through beautification is through basic tenet #1. NOI further supports this conclusion which is stronger than both NON, though the difference is not statistically significant, and DEC, where the difference is statistically significant.

### I. GNU as a More Readable Style

Note further in Tables 5-7 that $|\rho|$ generally tends to be greater for GNU and the difference is statistically significant ($P$ <0.05). The one exception is SRES versus GNU ($\rho$=-0.44) and BSD ($\rho$=-0.41): the pattern persists but it is not statistically significant ($P$=0.40). In other words, GNU tends to be more readable. Although the corpus is a GNU project, presumably written to the GNU standard, that fact should in theory should reduce the correlation because more files will have $B$=0. However, this is not the case which argues GNU, at least from a readability perspective, is a better style if we just go by the data.

## VII. Conclusions

The data suggests that while some variations in beauty can be explained by variations in readability, most cannot be explained as such, at least not on the corpus in this study. In other words, beauty and readability are related as we hypothesized and beauty appears to measure a unique property in code called aesthetic appeal. The data further suggests that indentation is reliably correlated with readable code, more than mnemonics or comments and of the four styles, GNU style is the most correlated with readability. Future research needs to confirm these findings for different repositories, different languages and different styles. We believe this is a worthwhile endeavor with potential to inform certain deeply felt and passionately argued beliefs about style.

## Acknowledgment

The author thanks Brendon Boldt for assistance with the ANTLR specification and Maria Luisa Coleman for reading drafts.

## References

[1] D. E. Knuth, "Computer Programming as An Art," *CACM*, 17 (12), 1974

[2] Bell Telephone Laboratories, *Unix Seventh Edition Manual*, volume 1, 1979, http://plan9.bell-labs.com/7thEdMan/, last accessed: 10 Jul 2018

[3] B. Kernighan, D. Ritchie, *The C Programming Language*, Prentice Hall, 1978.

[4] R. Green and H. Ledgard, "Coding Guidelines: Finding the Art in the Science," *Communications of the ACM*, vol 52, issue 2, December 2011, p57-63, doi :10.1145/2043174.2043191

[5] M. J. Black, "The Art of Code," Ph.D. thesis, University of Pennsylvania, 1 Jan 2002

[6] D. Boswell and T. Foucher, *The Art of Readable Code*, O'Reilly, 2011

[7] H. Abelson and G. Sussman, *The Structure and Interpretation of Computer Programs*, 2nd ed., MIT Press, 1996

[8] P. Kokol, J. Brest, and V. Zumer, "Long-range correlations in computer programs," *Cybernetics and systems*, 28 (1), 43-57, 1997

[9] P. Kokol, and J. Brest, "Fractal structure of random programs," *SIGPLAN notices*, 33 (6), 1998

[10] P. Kokol, V. Podgorelec, and J. Brest, "A wishful complexity metric," in H. Combes (ed.), FESMA, p. 235-246, 1998

[11] R. Coleman, and P. Gandhi, "Fractal Analysis of Good Programming Style", *Proc. Second International Conference on Computer Science & Engineering*, Dubai, UAE, 28-29 Aug 2015

[12] H-O.Peltgen and P.H. Richter, P.H., *The Beauty of Fractals*, Springer, 1986

[13] R. Coleman and P. Gandhi, "Fractal Beauty of Programming Style," *Proc. of the 14th International Conference on Software Engineering Research and Practice*, SERP '16, CSREA Press, 2016

[14] R. Coleman and B. Boldt, "Aesthetics Versus Entropy in Source Code," *Proc. of the 15th International Conference on Software Engineering and Practice*, SERP '17, CSREA Press, 2017

[15] R. Coleman and M. Rahtelli, "A Fractal Geometry Approach to Quantifying Aesthetic Values in Scientific Codes", *Proceedings of the 2017 International Conference on Computational Science and Computational Intelligence*, IEEE Computer Society, Editors: Hamid R. Arabnia, Leonidas Deligiannidis, Fernando G. Tinetti, Q-N. Tran, M. Qu Yang, ISBN-13: 978-1-5386-2652-8; BMS Part # CFP1771X-USB; DOI 10.1109/CSCI.2017.313

[16] R.P. Taylor, A.P. Micolich, and D. Jonas, "Fractal analysis of Pollock's drip paintings," *Nature* 399, 422 (3 June 1999), doi:10.1038/20833

[17] P. Gerl, Schönlieb, C., and K.C. Chieh Wang, "The Use of Fractal Dimension in Arts Analysis," *Harmonic and Fractal Image Analysis*, 2004, p70-73

[18] R.P. Taylor, R. Guzman, T.P. Martin, G.R.D. Hall, A.P. Micolich, D. Jonas, D., Scannell, M.S. Fairbanks, and C.A. Marlow, "Authenticating Pollock paintings using fractal geometry," *Pattern Recognition Letters*, Volume 28, Issue 6, 2007, p695-702

[19] J. Coddington, J. Elton, D. Rockmore, and Y. Wang, "Multifractal analysis and authentication of Jackson Pollock paintings," *Proc. Computer Image Analysis in the Study of Art* (SPIE 6810), 2008, doi: 10.1117/12.765015

[20] M. Irfan and D. Stork, "Multiple visual features for the computer authentication of Jackson Pollock's drip paintings: Beyond box counting and fractals," *Proc. SPIE 7251*, Image Processing: Machine Vision Applications II, 72510Q, 2009

[21] A. Oram and G. Wilson (eds.), *Beautiful Code: Leading Programmers Explain How They Think*, O'Reilly, 2007

[22] F. Brooks, "Towards a theory of the comprehension of computer programs," *International Journal of Man-Machine Studies*, 18, 543-554, 1984

[23] M.A., Storey, "Theories, tools, and research methods in program comprehension: past, present, and future," Software Quality J., 14:187-208, doi 10.1007/s11219-006-9216-4, 2006

[24] Stanford Encyclopedia of Philosophy, "The Concept of the Aesthetic," https://plato.stanford.edu/entries/aesthetic-concept/, last accessed 10 Jul 2018

[25] E. Avidan and D.G. Feitelson, "From obfuscation to comprehension," ACM SIGSOFT Software Engineering Notes, 23(3):75-77, 1998, Proceeding ICPC '15 Proceedings of the 2015, IEEE 23rd International Conference on Program Comprehension, pp. 178-181

[26] J.. Elshoff and M. Marcotty, "Improving Computer Program Readability to Aid Modification," *Communications of the ACM*, 25(8):512-521, 1982

[27] B. Mandelbrot, "How Long is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension," *Science*, 156 (3775)

[28] R. Stallman, *GNU Coding Standards*, Samurai Media Limited, 2015

[29] FreeBSD, "FreeBSD Kernel Developer's Manual," 2015 https://www.freebsd.org/cgi/man.cgi?query=style&sektion=9, last accessed: 10 Jul 2018

[30] L. Torvalds, "Linux Kernel Coding Style," http://slurm.schedmd.com/coding_style.pdf, last accessed: 10 Jul 2018

[31] J. Borstler, M.E. Caspersen, and M. Nordstrom, "Beauty and the Beast: Toward a Measurement Framework for Example Program Quality," *Software Quality Journal*, June 2006, Vol 24, Issue 2, pp. 231-246

[32] M. Halstead, *Elements of software science*, Elsevier, New York, 1977

[33] R.P.L. Buse and W. Weimer, "Learning a Metric for Code Readability," *IEEE Transactions on Software Engineering*, vol 4, issue 4, July 2010, p 546-558

[34] D. Posnett, A. Hindle, and P. Devanbu, "A Simpler Model of Software Readability," *Proceeding MSR '11 Proceedings of the 8th Working Conference on Mining Software Repositories*, Pages 73-82

[35] R. Flesch, "A New Readability Yardstick," *The Journal of Applied Psychology*, 32(3):221, 194

[36] A. Abbas, "Properties of Good Java Programs," Master's Thesis, Umea University, Sweden, 2009

[37] Cass, S. "The 2017 Top Popular Languages," *IEEE Spectrum*, 17 Jul 2017

[38] TIOBE: The Software Quality Company, "TIOBE Index for 2018," https://www.tiobe.com/tiobe-index/, last accessed: 11 Jul 2018

[39] C. Wood, J. Arceneaux, J. Kingdon, and D. Ingamells, "Indent," edition 2.2.10, for Indent Version 2.2.10, 23 July 2008, https://www.gnu.org/software/indent/manual/indent.pdf, retrieved 11 Jul 2018

[40] Free Software Foundation, Coreutils – GNU core utilities, http://www.gnu.org/software/coreutils/coreutils.html, last accessed: 11 Jul 2018

[41] D. Cornforth, H. Jelinek, and L. Peichl, "Fractop: A Tool for Automated Biological Image Classification," Proc. Sixth Australia-Japan Joint Workshop on Intelligent and Evolutionary Systems, 2002, p1-8

[42] R. Coleman, https://github.com/roncoleman125/Pretty, last accessed: 11 Jul 2018

[43] J. Kunk, "To Comment or Not Comment," *Visual Studio Magazine*, 1 June 2011, https://visualstudiomagazine.com/articles/2011/01/06/to-comment-or-not-to-comment.aspx, last accessed: 13 June 2017

[44] J. Raskin, "Comments are more important than code," *ACM Queue*, vol 3, issue 2, 18 Mar 2005

[45] P. Vogel, "Why You Shouldn't Comment (or Document) Code," *Visual Studio Magazine*, 27 June 2013, https://visualstudiomagazine.com/articles/2013/06/01/roc-rocks.aspx, last accessed: 13 June 2017

# User-Defined Financial Functions for MS SQL Server

Jolana Gubalova, Petra Medvedova
Department of Quantitative Methods and Information systems
Faculty of Economics, Matej Bel University
Banska Bystrica, Slovakia

***Abstract*—The paper deals with mathematical preparation and subsequent programming of various types of financial functions with using of Transact-SQL in Database Management System MS SQL Server. Financial functions are used to automate calculations in the area of Financial Economics. In MS SQL Server, any financial functions are not offered for financial data processing, how such as in program MS Excel. We emphasize that we have used a different calculation methods to create financial formulas, not those used in Excel. If users want to work with some special functions, there is a possibility to prepare User-Defined Functions (UDFs). The use of UDFs will make it easier to work on financial calculations in large databases.**

***Keywords*—*Financial economics; user-defined functions; financial functions; database management system; structure query language; transact-SQL***

## I. Introduction

Aggregate queries over big economic relational databases, prepared with using of Structure Query Language (SQL) and special programs belong to the most used tools in the area of Business Intelligence (BI) and Data Analytics (DA). One of the most important Relational Database Management Systems (RDBMS) for data saving, processing and analyzing, in the area of huge corporate or financial databases, is MS SQL Server. SQL Server runs on Transact – SQL (T-SQL), a set of programming extensions, that add several features to standard SQL, including transaction control, error handling, row processing and declared variables. SQL Server also allows stored procedures to be defined. Functions are a special type of stored procedures. They accept parameters, perform some sort of action and return a result. Functions do all of this with no side effects [14]. As Simhadri, V., at all said [12], queries containing user-defined functions (UDFs) are widely used, since they allow queries to be written using a mix of imperative language constructs and SQL, thereby increasing the expressive power of SQL; further, they encourage modularity, and make queries easier to understand. Writing user-defined functions or stored procedures presents common way in application development using a relational database management system. It allows to embed application code inside of RDBMS [15]. SQL Server provides numerous types of built-in scalar functions, for example, there are many built-in mathematical functions, date functions, string functions or aggregate functions. The types of user-defined functions (UDFs), that SQL Server supports, are scalar (return a single value) and table-valued (return a table). Most commercial SQL database systems support user-defined functions that can be used in WHERE clause filters, SELECT list items, or in sorting/grouping clauses. Often, user-defined functions are used as inexact search filters and then the filtered rows are sorted by a relevance measure [8]. Running analytics computation inside a database engine through the use of UDFs (User Defined Functions) has been investigated, but not yet become a scalable approach due to several technical limitations. One limitation lies in the lack of generality for UDFs to express complex applications and to compose them with relational operators in SQL queries. Another limitation lies in the lack of systematic support for a UDF to cache relations initially for efficient computation in multi-calls. Further, having UDF execution interacted efficiently with query processing requires detailed system programming, which is often beyond the expertise of most application developers [4]. Ordonez, C., at all described in [10], [11] vector and matrix operations programmed with UDFs in a relational DBMS and a data mining system based on SQL queries and UDFs for relational databases. Sousa, M., at all [13] dealt with consolidation of queries with UDFs.

UDFs can also be used in Excel. Lester in [6] recommended them as alternative methods to perform duct calculations.

## II. Objective and Methods

In SQL Server, any financial functions are not offered for financial data processing, how such as in program MS Excel. This fact we felt like a big shortage in processing of financial data. Because of this problem, we decided to prepare main financial functions, available in program MS Excel, also in MS SQL Server with using of program extension T-SQL. There were particularly financial functions: for calculation of the future value of an investment based on a constant interest rate, for returning the number of periods for an investment based on periodic, constant payments and a constant interest rate, for calculation of the payment for a loan based on constant payments and a constant interest rate, or for calculation of the present value of a loan or an investment, based on a constant interest rate. The concepts of financial mathematics are described in sources [1] – [3], [9].

Finally we compared the speed and efficiency of work with classical formulas and UDFs in SQL Server 2012 with using of special tools - Execution plan and Client Statistics. Execution plans display how the database engine navigates tables and uses indexes to access or process the data for a query or other DML (Data Manipulation Language) statement, such as an update [7]. This graphical approach is very useful for understanding the performance characteristics of a query.

Client Statistics is SQL Server data tool which is very helpful in determining the statistics that how much data received from server to the client side. It means, client statistics helps in analyzing the traffics load like packets/bytes sent and received at client – server side. When we run a script or query in T-SQL editor, we can enable Client statistics to collect statistics like application profile, time statistics and network statistics which help in checking the efficiency of the script.

## III. Results and Discussion

In the following, we approached the basic knowledge and relationships of interest and rent numbers that we later used. We only dealt with compound interest when the interest is added to the original capital and the sum is further capitalized. In all considerations, we considered overdue (decursive) capitalization, in other words - interest paid at the end of the interest period. Throughout the text, we used the following symbols in Table I.

TABLE I. List of arguments

| **Financial mathematics** | **T-SQL** | **Meaning** |
|---|---|---|
| *FV* | @FV | Future Value of a capital |
| *PV* | @PV | Present Value of a capital |
| *PMT* | @PMT | Payment based on regular constant payments |
| *i* | @RATE | Interest Rate per year |
| *m* | @NPER | Number of conversions per year |
| *n* | @YEARS | Number of years |
| *p* | @NPAY | Number of payments per year |
| *T* | @TYPE | Value representing the timing of payment<br>T=1 payment at the beginning of the period<br>T=0 payment at the end of the period |

We calculated the future value of the initial capital at compound interest over $n$ years based on the following formula by [5].

$$FV = PV \cdot (1+i)^n. \tag{1}$$

In practice, it is common that the interest rates are considered more often than once a year, and then we talk about compound interest capitalization with conversions. We call the period between the two following interest charges conversion. Interest is generally charged $m$ -times annually. The future value of capital in compound interest with conversions in $n$ years was determined based on the following formula by [5]

$$FV = PV \cdot \left(1+\frac{i}{m}\right)^{m \cdot n}. \tag{2}$$

In real-life economics, we often encounter a system of regularly repeated payments. This sequence of regularly repeated payments is called a rent or cash flow. In this analysis, we dealt only with constant, unconditional, temporary, immediate-term rents (cash flows). With constant rent, the amount of the individual payments does not change (remains unchanged). Unconditional or sure rent is a rent, where individual rent payments are not subject to any conditions. Temporary or terminal rent has the finite number of payment. We are talking about immediate rent, if the first payment is made at the beginning or end of the first rent period. When considering a $p$ -term rent, $p$ determines the number of payments per year. If the payments are always made at the end of the time period, we are talking about a strenuous (overdue, post-term) rent, if the payments are always made at the beginning of the period, we are talking about the pre-term rent. The future value of an annuity (rent payments, cash flow) is the sum of the future values of all annuity payments calculated at the end of the $n$ -th year.

For the future value of a $p$ -term strenuous (overdue, post-term) rent after years, the following formula applies by [5]

$$FV = PMT \cdot \frac{\left(1+\frac{i}{m}\right)^{m \cdot n} - 1}{\left(1+\frac{i}{m}\right)^{\frac{m}{p}} - 1}. \tag{3}$$

The future value of a $p$ -term rent, provided a pre-term (pre-paid) payment after $n$ years was determined based on the following formula by [5]

$$FV = PMT \cdot \left(1+\frac{i}{m}\right)^{\frac{m}{p}} \cdot \frac{\left(1+\frac{i}{m}\right)^{m \cdot n} - 1}{\left(1+\frac{i}{m}\right)^{\frac{m}{p}} - 1}. \tag{4}$$

We were interested in the future value of the initial capital provided, that we regularly deposited payments to the initial capital $p$ -times per year for $n$ years with annual interest rate $i$ and $m$ conversions per year.

### A. Future Value of a Series of Payments

At first we considered the overdue (post-term) rent, so we paid the payments at the end of $p$ -th of the year each year. Next we considered the pre-term rent, so we paid the instalments at the beginning of $p$ -th of the year each year. The future value in our case we determined as the sum of the future value of compound interest with conversions and the future value of the $p$ -term rent by using formulas (2) and (3) for post-term rent and by using formulas (2) and (4) for pre-term rent, which resulted in the following formulas in Table II.

TABLE II. FUTURE VALUE – FORMULAS AND PROGRAM CODE

T=0

$$FV = PV \cdot \left(1+\frac{i}{m}\right)^{m \cdot n} + PMT \cdot \frac{\left(1+\frac{i}{m}\right)^{m \cdot n} - 1}{\left(1+\frac{i}{m}\right)^{\frac{m}{p}} - 1}$$

T=1

$$FV = PV \cdot \left(1+\frac{i}{m}\right)^{m \cdot n} + PMT \cdot \left(1+\frac{i}{m}\right)^{\frac{m}{p}} \cdot \frac{\left(1+\frac{i}{m}\right)^{m \cdot n} - 1}{\left(1+\frac{i}{m}\right)^{\frac{m}{p}} - 1}$$

```
--FV
CREATE FUNCTION  FV (@RATE FLOAT, @YEARS FLOAT, @PMT FLOAT,@PV FLOAT, @NPER FLOAT, @NPAY FLOAT, @TYPE INT)
RETURNS MONEY
AS
BEGIN
DECLARE @FV MONEY
IF @TYPE = 0
SET @FV = @PMT*((POWER(1+@RATE/@NPER,@YEARS*@NPER)-1)/(POWER(1+@RATE/@NPER,@NPER/@NPAY)-1))+
@PV*POWER(1+@RATE/@NPER,@NPER*@YEARS)
IF @TYPE = 1
SET @FV = @PMT*((POWER(1+@RATE/@NPER,@YEARS*@NPER)-1)/(POWER(1+@RATE/@NPER,@NPER/@NPAY)-
1))*POWER(1+@RATE/@NPER,@NPER/@NPAY)+ @PV*POWER(1+@RATE/@NPER,@NPER*@YEARS)
RETURN @FV
END
```

### B. Present Value of a Capital

From the formulas for the future value $FV$ (2), (3), (4), we expressed the initial deposit which is needed in the periodic payments to accumulate the future value using the following expressions in Table III.

TABLE III. PRESENT VALUE OF A CAPITAL – FORMULAS AND PROGRAM CODE

T=0

$$PV = FV \cdot \left(1+\frac{i}{m}\right)^{-m \cdot n} - PMT \cdot \frac{1-\left(1+\frac{i}{m}\right)^{-m \cdot n}}{\left(1+\frac{i}{m}\right)^{\frac{m}{p}} - 1}$$

T=1

$$PV = FV \cdot \left(1+\frac{i}{m}\right)^{-m \cdot n} - PMT \cdot \left(1+\frac{i}{m}\right)^{\frac{m}{p}} \cdot \frac{1-\left(1+\frac{i}{m}\right)^{-m \cdot n}}{\left(1+\frac{i}{m}\right)^{\frac{m}{p}} - 1}$$

```
CREATE FUNCTION PV (@RATE FLOAT, @YEARS FLOAT, @PMT FLOAT, @FV FLOAT, @NPER FLOAT, @NPAY FLOAT, @TYPE INT)
RETURNS MONEY
AS
BEGIN
DECLARE @PV MONEY
IF @TYPE = 0
SET @PV =(@FV*POWER(1+@RATE/@NPER,-@NPER*@YEARS) - @PMT*(1-POWER(1+@RATE/@NPER,-
@NPER*@YEARS))/(POWER(1+@RATE/@NPER,@NPER/@NPAY)-1))
IF  @TYPE = 1
SET @PV = @FV*POWER(1+@RATE/@NPER,-@NPER*@YEARS) - @PMT*POWER(1+@RATE/@NPER,@NPER/@NPAY)*(1-
POWER(1+@RATE/@NPER,-@NPER*@YEARS))/(POWER(1+@RATE/@NPER,@NPER/@NPAY)-1)
RETURN @PV
END
```

### C. Payment based on Regular Constant Payments

We determined the amount of a regular instalment $PMT$ which would give us with the initial deposit $PV$ after $n$ years a future value $FV$. It was enough to express from formula for future value $FV$ (2), (3), (4), $p$-term payment after $n$ years to express the unknown $PMT$ and we have got the following formulas in Table IV.

TABLE IV. PAYMENT BASED ON REGULAR CONSTANT PAYMENTS – FORMULAS AND PROGRAM CODE

T=0

$$PMT = \left[ FV - PV \cdot \left(1+\frac{i}{m}\right)^{m \cdot n} \right] \cdot \frac{\left(1+\frac{i}{m}\right)^{\frac{m}{p}} - 1}{\left(1+\frac{i}{m}\right)^{m \cdot n} - 1}$$

T=1

$$PMT = \left[ FV - PV \cdot \left(1+\frac{i}{m}\right)^{m \cdot n} \right] \cdot \frac{\left(1+\frac{i}{m}\right)^{\frac{m}{p}} - 1}{\left(1+\frac{i}{m}\right)^{m \cdot n} - 1} \cdot \left(1+\frac{i}{m}\right)^{-\frac{m}{p}}$$

```
--PMT
CREATE FUNCTION PMT (@RATE FLOAT,@YEARS FLOAT, @PV FLOAT,@FV FLOAT, @NPER FLOAT, @NPAY FLOAT, @TYPE BINARY)
RETURNS MONEY
AS
BEGIN
DECLARE @PMT MONEY
IF @TYPE = 0
SET @PMT = (@FV-@PV*POWER(1+@RATE/@NPER,@NPER*@YEARS))*(POWER(1+@RATE/@NPER,@NPER/@NPAY)-
1)/(POWER(1+@RATE/@NPER,@NPER*@YEARS)-1)
IF @TYPE = 1
SET @PMT = (@FV-@PV*POWER(1+@RATE/@NPER,@NPER*@YEARS))*POWER(1+@RATE/@NPER,-
@NPER/@NPAY)*(POWER(1+@RATE/@NPER,@NPER/@NPAY)-1)/(POWER(1+@RATE/@NPER,@NPER*@YEARS)-1)
RETURN @PMT
END
```

TABLE V. NUMBER OF YEARS – FORMULAS AND PROGRAM CODE

T=0

$$n = \frac{1}{\ln\left(1+\frac{i}{m}\right)^{m}} \cdot \ln\left\{ \frac{FV \cdot \left[ \left(1+\frac{i}{m}\right)^{\frac{m}{p}} - 1 \right] + PMT}{PV \cdot \left[ \left(1+\frac{i}{m}\right)^{\frac{m}{p}} - 1 \right] + PMT} \right\}$$

T=1

$$n = \frac{1}{\ln\left(1+\frac{i}{m}\right)^{m}} \cdot \ln\left\{ \frac{FV \cdot \left[ \left(1+\frac{i}{m}\right)^{\frac{m}{p}} - 1 \right] + PMT \cdot \left(1+\frac{i}{m}\right)^{\frac{m}{p}}}{PV \cdot \left[ \left(1+\frac{i}{m}\right)^{\frac{m}{p}} - 1 \right] + PMT \cdot \left(1+\frac{i}{m}\right)^{\frac{m}{p}}} \right\}$$

```
--YEARS
CREATE FUNCTION YEARS (@RATE FLOAT,@NPay FLOAT,@PV FLOAT,@FV FLOAT, @PMT FLOAT, @NPER FLOAT, @TYPE INT)
RETURNS FLOAT
AS
BEGIN
DECLARE @YEARS FLOAT
IF @TYPE = 0
SET @YEARS = LOG((((@FV*(POWER(1+@RATE/@NPER,@NPER/@NPay)-1)+@PMT))/(((@PV*(POWER(1+@RATE/@NPER,@NPER/@NPay)-
1)+@PMT)))
/LOG(POWER(1+@RATE/@NPER,@NPER))
IF @TYPE = 1
SET @YEARS = LOG((((@FV*(POWER(1+@RATE/@NPER,@NPER/@NPay)-
1)+@PMT*POWER(1+@RATE/@NPER,@NPER/@NPAY)))/(((@PV*(POWER(1+@RATE/@NPER,@NPER/@NPay)-
1)+@PMT*POWER(1+@RATE/@NPER,@NPER/@NPAY))))
/LOG(POWER(1+@RATE/@NPER,@NPER))
RETURN @YEARS
END
```

### D. Number of Years

The same, on the basis of previous formulas (2), (3), (4), we have expressed a number of years $n$ which are needed on obtaining of the future value $FV$ . The corresponding formulas are as follows in Table V.

### E. Number of Payments per Year

By expressing the number of payments per year $p$ from future value formulas (2), (3), (4), we obtained the following formulas in Table VI.

### F. The Comparison of Classical Calculations and Calculations with using UDFs

For comparison the speed and efficiency of work with classical formulas and UDFs we decided to use a training financial database with 1 048 575 records in DBMS MS SQL Server 2012. Firstly we run query with using of classical formula for calculation of Future Value and next with UDF Future Value. We also controlled Execution plans and Clients Statistics for this queries.

The Execution Plan consists of different operations and each operation has one output which is called the result set. The operations can have one or more inputs. There are many potential ways to execute a query thus SQL Server has to choose the most beneficial one. Client statistics helps in analyzing the traffics load like packets/bytes sent and received at client – server side. When we run a script or query in T-SQL editor, we can enable Client statistics to collect statistics like application profile, time statistics and network statistics which help in checking the efficiency of the query.

TABLE VI. NUMBER OF PAYMENTS PER YEAR– FORMULAS AND PROGRAM CODE

| T=0 | T=1 |
|---|---|
| $p = \frac{\ln\left(1+\frac{i}{m}\right)^{m}}{\ln\left\{\frac{PMT\cdot\left[\left(1+\frac{i}{m}\right)^{m\cdot n}-1\right]}{FV-PV\cdot\left(1+\frac{i}{m}\right)^{m\cdot n}}+1\right\}}$ | $p = \frac{\ln\left(1+\frac{i}{m}\right)^{m}}{-\ln\left\{1-\frac{PMT\cdot\left[\left(1+\frac{i}{m}\right)^{m\cdot n}-1\right]}{FV-PV\cdot\left(1+\frac{i}{m}\right)^{m\cdot n}}\right\}}$ |

```
-- Number of payments per year
CREATE FUNCTION NPay (@RATE FLOAT,@YEARS FLOAT,@PV FLOAT,@FV FLOAT, @PMT FLOAT, @NPER FLOAT, @TYPE INT)
RETURNS FLOAT
AS
BEGIN
DECLARE @NPay FLOAT
IF @TYPE = 0
SET @NPay = LOG(POWER(1+@RATE/@NPER,@NPER))/LOG(((@PMT*(POWER(1+@RATE/@NPER,@YEARS*@NPER)-1))/(@FV-
@PV*POWER(1+@RATE/@NPER,@YEARS*@NPER))+1)
IF @TYPE = 1
SET  @NPay = LOG(POWER(1+@RATE/@NPER,@NPER))/-LOG(1-(@PMT*(POWER(1+@RATE/@NPER,@YEARS*@NPER)-1))/(@FV-
@PV*POWER(1+@RATE/@NPER,@YEARS*@NPER)))
RETURN @NPay
END
```

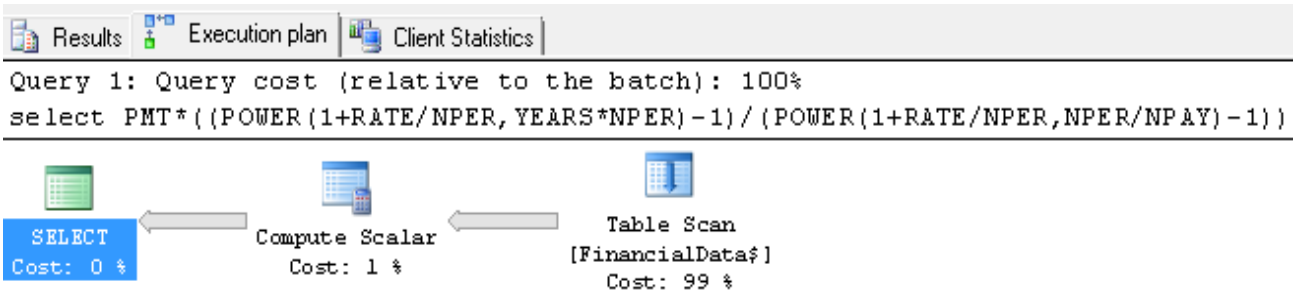



Fig. 1. Execution Plan of Classical Formula Future Value.

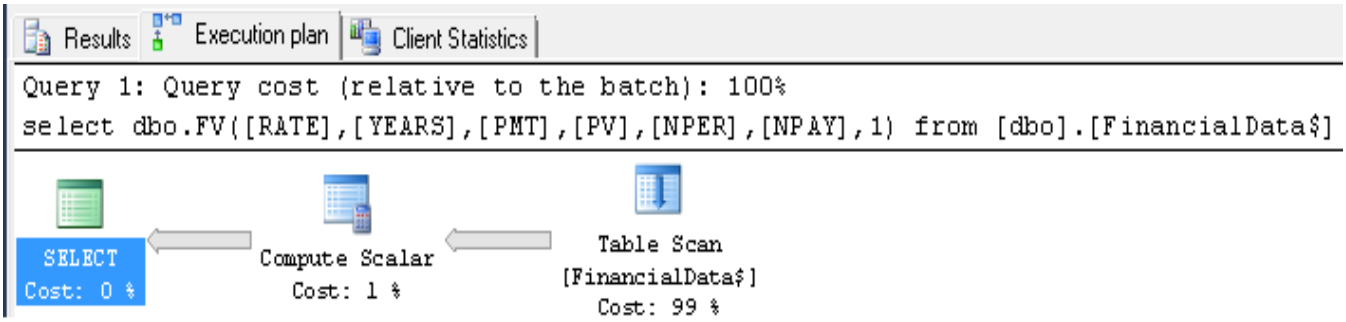



Fig. 2. Execution Plan of UDF Future Value.

As we can see at Figure 1, in case of classical formula Future Value, totally amount Query cost was spent by Table Scan of FinacialData (99%) and the rest (1%) was belong to Compute Scalar. The same situation was in case of UDF Future Value (Figure 2).

TABLE VII. COMPARISON OF CLIENT STATISTIC FOR CLASSICAL FORMULA AND UDF

| | **Classical Formula** | **UDF** |
|---|---|---|
| Client Execution Time | 16:43:38 | 16:52:49 |
| ***Query Profile Statistics*** | | |
| Number of SELECT statements | 1 | 1 |
| Rows returned by SELECT statements | 1048575 | 1048575 |
| ***Network Statistics*** | | |
| Number of server roundtrips | 3 | 1 |
| TDS packets sent from client | 3 | 1 |
| TDS packets received from server | 2572 | 2569 |
| Bytes sent from client | 504 | 210 |
| Bytes received from server | 1,052478E+07 | 1,050631E+07 |
| ***Time Statistics [ms]*** | | |
| Client processing time | 1178,6 | 2797,8 |
| Total execution time | 1185,4 | 2806,8 |
| Wait time on server replies | 6,8 | 9,0 |

If we compared Client Statistics of classical formula Future Value and Client Statistics of UDF Future Value (Table VII.), we can stated, that *Query* Profile *Statistics* is the same in both cases.

But the difference is in Network Statistics – item *Bytes sent from client* is bigger at Classical Formula Future Value, because in case of UDF, client sent only data and formula were prepared on server side.

But there is significant difference in Client processing time. This attribute is almost 2,37 times higher for the UDF than for the Classical formula.

## IV. CONCLUSIONS

The comparison classical calculations and calculations with using UDFs showed, that UDFs don't bring acceleration of the computation process opposite to the classical formulas. On the contrary, with using of the UDF the calculation process takes longer. However, working with them is simpler and more comfortable than with classical formulas, because they encourage modularity and make queries easier to understand. This is main reason they are popular among users in financial area. Stored procedures and UDFs can be prepared also in other extended DBMS, such as Oracle (Pragma UDFs), mySQL, PostgreSQL, DB2, Informix, etc.

## REFERENCES

[1] Baz, J., Chacko, G., (2004). Financial derivatives: Pricing, applications, and mathematics. Publisher: Cambridge University Press, ISBN: 978-051180664-3;978-052181510-9.

[2] Dworsky, L.N., (2009). Understanding the Mathematics of Personal Finance: An Introduction to Financial Literacy. Publisher: John Wiley and Sons, ISBN: 978-047049780-7.

[3] Hirsa, A., Neftci, S.N., (2013). An Introduction to the Mathematics of Financial Derivatives. Publisher: Elsevier Inc. ISBN: 978-012384682-2.

[4] Hsu, M., at all, (2010). Generalized UDF for analytics inside database engine. 11th International Conference on Web-Age Information Management, WAIM 2010; Jiuzhaigou; China, ISBN: 3642142451;978-364214245-1.

[5] Huťka, V., & Peller, F. (1999). Finančná matematika v Exceli, Elita, Bratislava, 1999, ISBN 80-8044-064-6

[6] Lester, T.G., (2009). Using excel for duct calculations user defined functions. ASHRAE Journal, Volume 51, Issue 8, August 2009, Pages 42-46. ISSN:0001-2491E-ISSN:0364-9962.

[7] Microsoft: Analyze Script Performance https://docs.microsoft.com/en-us/sql/ssdt/analyze-script-performance?view=sql-server-2017

[8] Murthy, R., at all, (2003). Supporting ancillary values from user defined functions in Oracle. Nineteenth International Conference on Data Ingineering; Bangalore; India; Conference Proceeding, Pages 151-162.

[9] Mishura, Y., (2016). Financial Mathematics. Publisher: Elsevier Inc. ISBN: 978-178548046-1.

[10] Ordonez, C., at all, (2006). Vector and matrix operations programmed with UDFs in a relational DBMS. 15th ACM Conference on Information and Knowledge Management, CIKM 2006; Arlington, VA; United States. Proceedings, Pages 503-512.

[11] Ordonez, C., Garcia-Alvaro, C., (2011). A data mining system based on SQL queries and UDFs for relational databases. 20th ACM Conference on Information and Knowledge Management, CIKM'11; Glasgow; United Kingdom, ISBN: 978-145030717-8.

[12] Simhadri, V., at all (2014). Decorrelation of user defined function invocations in queries. 30th IEEE International Conference on Data Engineering, ICDE 2014; Chicago, IL; United States, ISBN: 978-147992554-4.

[13] Sousa, M., at all, (2014). Consolidation of Queries with User-Defined Functions. 35th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI), Edinburgh, Scotland. Volume: 49, Issue: 6, Pages: 554-564. DOI: 10.1145/2594291.2594305.

[14] User defined function. Available at: https://docs.microsoft.com/en-us/sql/relational-databases/user-defined-functions/user-defined-functions

[15] Vagač, M., Melicherčík, M., (2015). Improving image processing performance using database user-defined functions. 14th International Conference on Artificial Intelligence and Soft Computing, ICAISC 2015; Zakopane; Poland. Conference Proceeding, ISBN: 978-331919323-6.

# An Amplitude Modulation of Cerebral Rhythms based Method in a Motor Task BCI Paradigm

Oana Diana Eva, Anca Mihaela Lazar
Medical Bioengineering Faculty
"Grigore T. Popa" University of Medicine and Pharmacy
Iasi, Romania

***Abstract*—Quantitative evaluation based on amplitude modulation analysis of electroencephalographic signals is proposed for a brain computer interface paradigm. The method allows characterization of the interaction effects of different frequency bands in the electroencephalographic rhythms during motor tasks. A new index was proposed and computed to be a measure of the amplitude modulation. Built on this index, features vector are established for training different classification algorithms. Signals recorded from 50 subjects revealed important differences in amplitude modulations between motor tasks. Most notably, Theta modulation of the Theta and Alpha rhythms proved to be reliable discriminant features between different mental tasks.**

***Keywords*—Brain computer interface; motor tasks; electroencephalographic signal; amplitude modulation analysis; classifiers**

## I. Introduction

Brain computer interface (BCI) provides a new communication channel for people with motor disabilities using electroencephalographic signals (EEG) or other brain signals.

EEG is often used in BCI systems because it is not an invasive experiment, has no risk and has good time resolution.

The oscillatory activity in the EEG is classified according to rhythms: Delta (0.1-4 Hz), Theta (4-8 Hz), Alpha (8-12 Hz), Beta (12-30 Hz), and Gamma (30-60 Hz).

Motor imagery (MI) is a dynamic state during which a given action is mentally simulated by a subject [1]. Sensory stimulation, motor behavior, movement or only imagination of movement can modify the functional links in the cortex and cause an amplitude decrease in the EEG, called Event Related Desynchronization (ERD), or an increase in the amplitude of the signal, named Event Related Synchronization (ERS) of the Mu or Beta rhythm. The Mu rhythm is a normal central rhythm in Alpha frequency band. Mu waves appear in the sensorimotor (SM) cortex.

Different feature extraction methods are reported for discriminating the motor tasks in a BCI paradigm: spectral analysis [2], autoregressive methods [3], independent component analysis [4], [5], Itakura distance [4] and phase synchronization [4], [6].

The main objectives of the proposed method are to discriminate, extract and classify representative EEG features which can be used in brain computer interface motor imagery paradigm by means of an amplitude modulation analysis.

Studies conducted so far using amplitude modulation analysis have shown that there are differences between the EEG from healthy subjects and EEG from subjects diagnosed with Alzheimer's disease [7], [8]. In our case, the differences that appear during motor imagery tasks (left hand imagination, right hand imagination and relaxation) are studied.

It is researched the possibility of using a quantitative method of analyzing the EEG signal. Amplitude modulation analysis is computed for EEG signal rhythms and then some classification methods are applied.

Section II describes the dataset used, the methodology that includes the proposed method and the classification methods. The results are presented in section III, the discussions in section IV and the conclusions in section V.

## II. Materials and Method

### A. Dataset and Subjects

The database is formed by 50 EEG signals collected from 50 healthy volunteers, men and women, age range 19-59 years, untrained. The EEG signals were recorded with g.MOBIlab+ portable acquisition system provided by g.tec Guger Technologies together with the BCI2000 platform. Eight electrodes (CP3, CP4, P3, P4, C3, C4, PZ and CZ) were placed on the volunteers scalp according to International 10-20 System. The reference electrode was linked on the right earlobe. Experimental protocol consisted of randomly right or left motor tasks and relaxation periods. The volunteers were comfortably seated on a seat in front of a PC monitor that displays left or right arrows. They must look carefully at the arrows and try to imagine the hand movement indicated by the arrow. When the screen is white, the subject needs to relax. Each arrow appears for 30 times. The time interval between visual stimuli was 2 seconds and the sampling frequency used was 256 Hz. Before the experiment, the subject is instructed not to speak, to move or to blink during the trials. The recordings took place on different days, under brightness conditions chosen by the volunteers. All volunteers provided written informed consent prior to the experiment.

### B. Data Processing

Three datasets were created associated to the motor imagery tasks: one representing the EEG during the left hand

imagination, one representing EEG during the right hand imagination and one representing EEG during the relaxation.

The EEG signal $x(n)$ is decomposed into three frequency bands (cerebral rhythms): 4 – 8 Hz (Theta rhythm), 8 - 12 Hz (Alpha rhythm), 12 -30 Hz (Beta rhythm).

$$x_i(n) = x(n) * \mathrm{h}_i(n), \quad (1)$$

Where $\mathrm{h}_i(n)$ , i=1,2,3 represents the impulse response of the applied bandpass filter corresponding to each frequency band. We justify later why there are only three rhythms used.

The Hilbert transform $H\{\cdot\}$ of $x_i(n)$signal is:

$$\mathcal{H}\{x_i(n)\} = \frac{1}{\pi} PV \int_{-\infty}^{+\infty} \frac{x_i(\tau)}{t-\tau} dt, \quad (2)$$

where $PV$ is the Cauchy principal value.

The analytical signal is denoted by:

$$\overline{x_\iota(n)} = x_i(n) + j\mathcal{H}\{x_i(n)\}. \quad (3)$$

The amplitude modulation $e_i(n)$ of each cerebral rhythm is the modulus of $x_i(n)$ [7], [8]:

$$e_i(n) = \sqrt{x_i(n)^2 + \mathcal{H}\{x_i(n)\}^2}. \quad (4)$$

The $e_i(n)$ is multiplied by a 5 s Hamming window which is shifted across the data set with a step of 0.5 s. The amplitude modulation for frame $m$ is represented as $e_i(m,n)$, where $n$ is the time variable.

For each rhythm $i$ the absolute value of the Fourier transform is computed for each sub band $j$.

$$E_i(m;f) = |\mathcal{F}\{e_i(m,n)\}|, \quad (5)$$

where $\mathcal{F}\{e_i(m,n)\}$ is the discrete Fourier transform of the temporal envelope for each frame $e_i(m,n)$ and $f$ denotes modulation frequency.

In order to quantify the rate of change of the temporal envelope and the potential frequency interactions, the modulation frequencies are subsequently arranged to coincide with the first four conventional frequency ranges from Delta to Beta.

Gamma band modulation frequencies (30-60 Hz) can only be present for gamma rhythm. Therefore, for this reason, but also to reduce the amount of data, the modulation band corresponding to gamma rhythm is not taken into account in this type of analysis. The Delta rhythm is not taken into consideration because there is only one situation (the modulation band corresponding to the Delta rhythm consisting of the lowest frequencies 0.1-4 Hz). Therefore, we work with three rhythms and four modulation bands, which lead to nine options. Assigning the *band modulation_rhythm* notation (e. g. *Delta_Theta* represents modulation in Delta band of the Theta rhythm), all the possibilities are depicted in Table 1.

With the aim of quantifying the weight of the different modulation frequencies in the cerebral rhythms, a new parameter called $Modulation\_index$ is proposed:

$$(Modulation\_index_m)_{i,j} = \frac{\overline{E_{\iota,j}(m)}}{\sum_{i=1}^{3}\sum_{j=1}^{4}\overline{E_{\iota,j}(m)}} \quad (6)$$

TABLE I. MODULATION BANDS CORRESPONDING TO THE CEREBRAL RHYTHMS

| **Rhythm** | **Modulation Band** | | | |
|---|---|---|---|---|
| | *Delta* | *Theta* | *Alpha* | *Beta* |
| Beta | Delta _Beta | Theta_Beta | Alpha_Beta | Beta_Beta |
| Theta | Delta_Theta | Theta_Theta | Alpha_Theta | ---------- |
| Alpha | Delta_Alpha | Theta_Alpha | ---------- | ---------- |

where $\overline{E_{\iota,j}(m)}$ is the average on all frames (resulting from the segmentation of the EEG signal with the Hamming sliding-window) of the Fourier transform moduli corresponding to the modulation frequency band $j$ of the rhythm $i$ , and $\sum_{i=1}^{3}\sum_{j=1}^{4}\overline{E_{\iota,j}(m)}$ is the sum of all Fourier transform module averages for all rhythms and for all possible modulation bands.

### C. Features Extraction, Selection and Data Classification

Discriminations between the left and the right motor activity (left-right), the left motor activity and relaxation (Left-Relaxation), the right motor activity and relaxation (Right-Relaxation) are evaluated with five classifiers: linear discriminant analysis (LDA) [9], quadratic discriminant analysis (QDA) [10], Mahalanobis distance (MD) [11], k nearest neighbor (kNN) [12] and support vector machine (SVM) [13]. A 10x10 fold cross validation method estimate the classification rates.

The aims of the research concern in finding the modulation band/s related to a specific cerebral rhythm and the classifier/s which best discriminate between the classes (left - right, left-relaxation or right - relaxation).

## III. Results

The implementation of the proposed method is in MATLAB.

There are already mentioned the three cases under study, namely left-right, left-relaxation and right-relaxation.

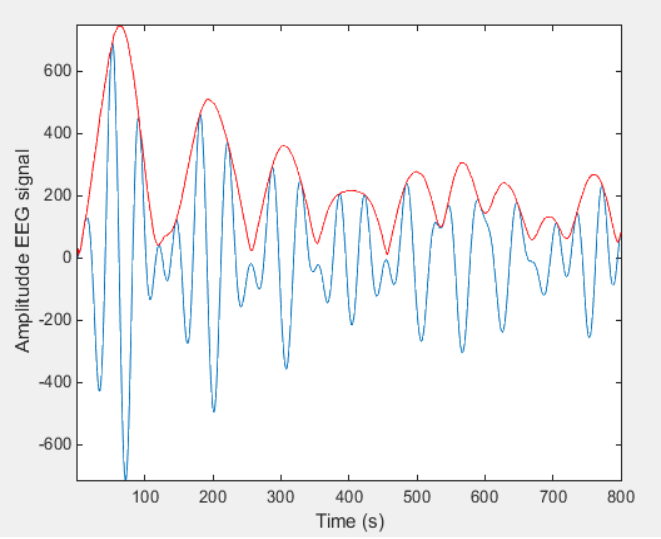



Fig. 1. EEG Alpha Rhythm from Channel C3 (blue) and Theta Amplitude Modulation (red) for Subject SL27i when he Imagines the left Hand Movement.

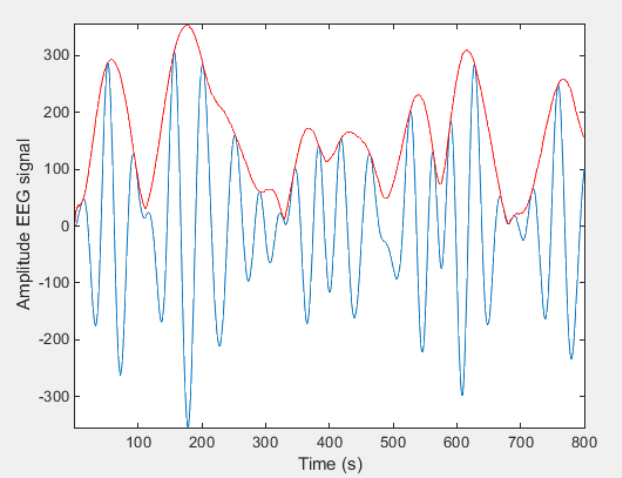



Fig. 2. EEG Theta Rhythm from Channel C4 (blue) and Delta Amplitude Modulation (Red) for Subject SL27i when he Imagines the left Hand Movement.

For each case, the recorded EEG signals are IIR bandpass filtered, on the corresponding frequency bands of the three rhythms under study (Theta, Alpha and Beta).

Then using (2), (3) and (4) the amplitude modulations are computed. The plots of Fig. 1 and Fig. 2 exemplify an EEG rhythm in blue and one of its envelopes in red.

For each case, the proposed new indexes, computed by means of (5) and (6) for the corresponding two classes, generate the consistent feature vector. For each subject from of all the 50 subjects the maximum of the classification rates estimated by all the mentioned classifiers is further extracted (for each rhythm and each corresponding modulation bands). So we get 50 maxima for each situation. Hereafter, there are taken into account only those values of maxima higher or equal to a high threshold considered to be 80%. The results are presented in Table II.

A result from the first cell from Table II is interpreted in the following manner: in the left-right case, there are 19 subjects from the all of 50 subjects whose classification rates are higher or equal to 80%.

TABLE II. THE NUMBER OF THE MAXIMA OF THE CLASSIFICATION RATES HIGHER OR EQUAL TO 80%

| Modulation_Rhythm | left-right | left-relaxation | right-relaxation |
|---|---|---|---|
| Delta_Theta | 19 | 19 | 21 |
| Theta_Theta | 20 | 19 | 19 |
| Delta_Alpha | 8 | 20 | 14 |
| Theta_Alpha | 8 | 20 | 23 |
| Alpha_Alpha | 9 | 14 | 19 |
| Delta_Beta | 8 | 12 | 7 |
| Theta_Beta | 14 | 13 | 16 |
| Alpha_Beta | 9 | 6 | 16 |
| Beta_Beta | 3 | 8 | 10 |

The green cells from Table II draw attention to the situations when the number of classification rates higher or equal to 80% exceeds a proper (high enough) value, considered by us being equal to 19. It is obvious that many cases are related to Theta rhythm, both for Delta and Theta modulation. There are no noteworthy differences for the three cases reported (left-right, left-relaxation and right-relaxation). Also, in the Alpha rhythm and modulation in Theta and Alpha bands, in the right-relaxation case, there are a significant number of classification rates greater than 80%. So, for Theta_Alpha, the classification rates are above 80% for almost a half of the subjects (23 from 50 subjects). In the left-relaxation case, high values are obtained for Theta and Delta modulation bands. It is worth noticing that low values are achieved for left-right (8 or 9 from 50 subjects). So we conclude that in left-right paradigm the Alpha rhythm modulations are presented in few persons, so we ignore it.

In order to evaluate the results attained for each classifier, there are considered only those green marked situations in Table II.

The results for left-right case are illustrated in Fig. 3 where only the Theta rhythm is under study.

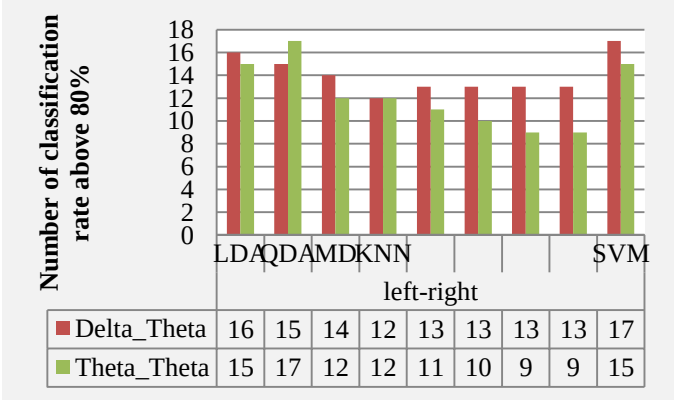



Fig. 3. Number of Classification Rates Higher than 80%, for all the Classifiers in the Left-Right Case

It is easy to observe that the best results are achieved by means of LDA, QDA and SVM classifiers.

The results for left-relaxation are gathered in Fig. 4. The modulations of Theta and Alpha rhythms are considered.

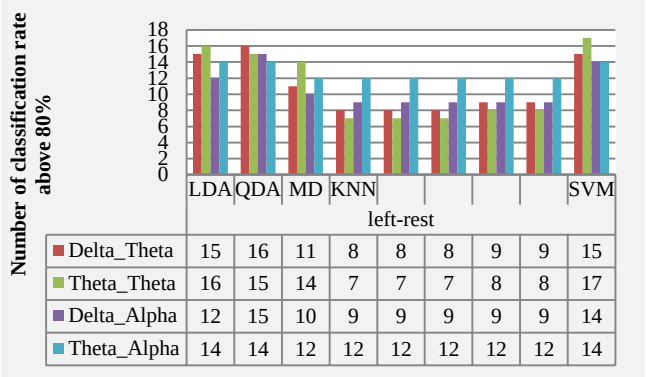



Fig. 4. Number of Classification Rates Higher than 80%, for all the Classifiers in the Left-Relaxation Case.

LDA, QDA and SVM are the best classifiers also in the case presented in Fig. 4.

The results for right-relaxation are drawn together in Fig. 5. As in the previous case, the modulations of Theta and Alpha rhythms are taken into consideration.

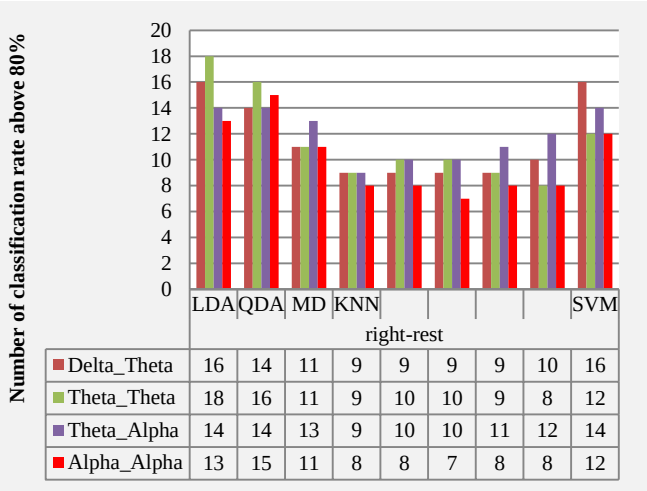
As in the two cases, the best results are found for LDA, QDA and SVM classifiers.

For further considerations, we have to mention that for Theta_Theta, in left-right case, there are 7 classification rates upper 90%, and one with 100%.

As for all the cases, LDA, QDA and SVM classifiers lead to high classification rates, in order to look deeper into and to study the distributional characteristics, the box plots are used.

Fig. 6 illustrates the box plots for LDA classifier for the left-right case, Left-relaxation and Right-relaxation cases (Delta_Theta and Theta_Theta).

In Fig. 7 the box plots for LDA classifier for the left-right case, left-relaxation and right-relaxation cases (Delta_Alpha Theta_Alpha and Alpha_Alpha) are presented.




| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Delta_Theta | 16 | 14 | 11 | 9 | 9 | 9 | 9 | 10 | 16 |
| Theta_Theta | 18 | 16 | 11 | 9 | 10 | 10 | 9 | 8 | 12 |
| Theta_Alpha | 14 | 14 | 13 | 9 | 10 | 10 | 11 | 12 | 14 |
| Alpha_Alpha | 13 | 15 | 11 | 8 | 8 | 7 | 8 | 8 | 12 |

Fig. 5. Number of Classification Rates Higher than 80%, for all the Classifiers in the Right-Relaxation Case.

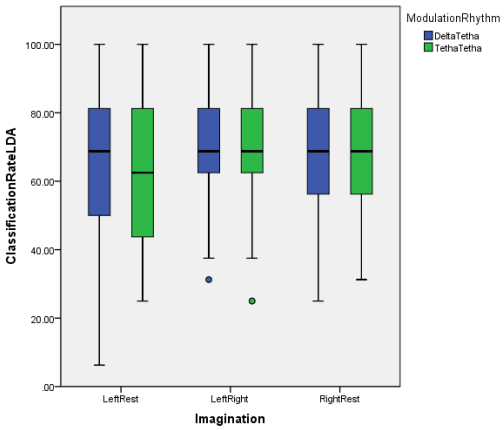



Fig. 6. The Box Plots for LDA Classifier for the Left-Right Case, Left-Relaxation and Right-Relaxation Cases (Delta_Theta and Theta_Theta).

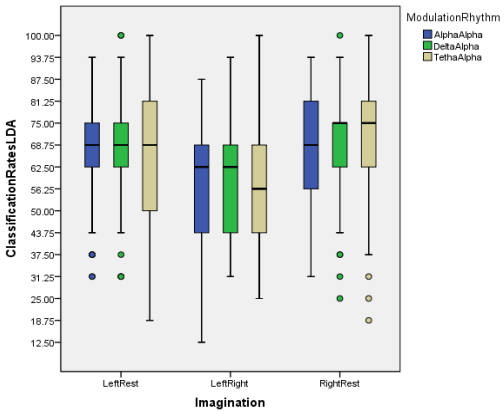



Fig. 7. The Box Plots for LDA Classifier for the Left-Right Case, Left-Relaxation and Right-Relaxation Cases (Delta_Alpha Theta_Alpha and Alpha_Alpha).

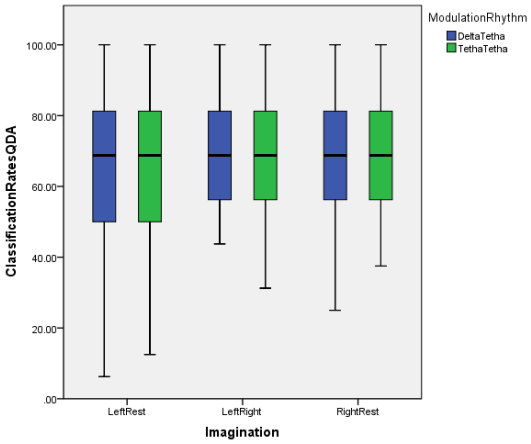



Fig. 8. The Box Plots for QDA Classifier for the Left-Right Case, Left-Relaxation and Right-Relaxation Cases (Delta_Theta and Theta_Theta).

Fig. 8 shows the box plots for QDA classifier for the Left-right case, left-relaxation and right-relaxation cases (Delta_Theta and Theta_Theta).

In Fig. 9 the box plots for QDA classifier for the left-right case, left-relaxation and right-relaxation cases (Delta_Alpha Theta_Alpha and Alpha_Alpha) are displayed.

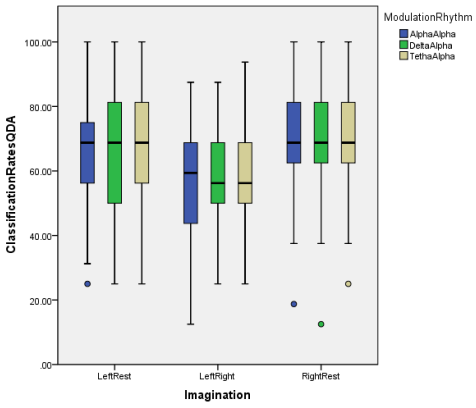



Fig. 9. The Box Plots for QDA Classifier for the Left-Right Case, Left-Relaxation and Right-Relaxation Cases (Delta_Alpha Theta_Alpha and Alpha_Alpha).

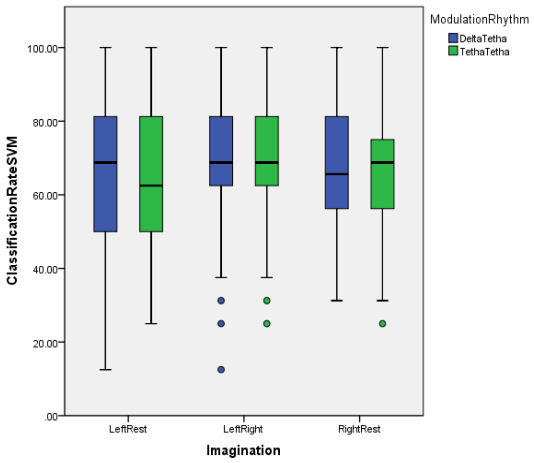



Fig. 10. The Box Plots for SVM Classifier for the Left-Right Case, Left-Relaxation and Right-Relaxation Cases (Delta_Theta and Theta_Theta).

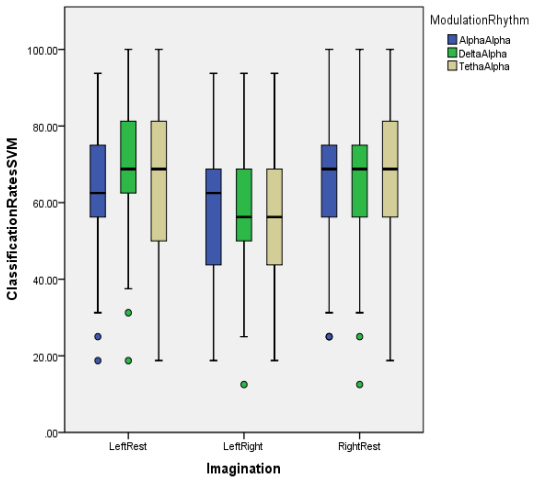



Fig. 11. The Box Plots for SVM Classifier for the Left-Right Case, Left-Relaxation and Right-Relaxation Cases (Delta_Alpha Theta_Alpha and Alpha_Alpha).

Fig. 10 and Fig. 11 indicate the box plots for SVM classifier for the left-right case, left-relaxation and right-relaxation cases (Delta_Theta and Theta_Theta) and (Delta_Alpha Theta_Alpha and Alpha_Alpha) respectively.

When the midpoint and spread of the classification rates are examined, we can observe that the lowest medians (lower than 60%) are for all the classifiers, for the left-right case in Alpha rhythm, for all amplitude modulation bands (Delta, Theta and Alpha). The medians are about 70% in almost all the other situations. The spread is evaluated by the interquartile range box (which represents the middle 50% of the data) and by the whiskers. Good results from this point of view are attained both for LDA and QDA classifiers, for right-relaxation situation, in Theta rhythm modulated by Theta band. Here the interquartile range is between 58 % and 80 % classification rate and there are no outliers. For LDA, even better results are attained for left-right case for Theta_Theta or Delta_Theta when the interquartile range is between 63 % and 80%, but there are two outliers. For all the situations where good results are reported, the whiskers extend to 100 %.

## IV. Discussions

We have to mention that it is impossible to make direct comparisons to other results reported on public EEG databases, because herein our own database has been handled. In a future work, we intend to validate the amplitude modulation analysis on the available public databases, especially as the EEG signals are recorded on more than 8 channels and better outcomes might be attained.

But comparisons must be done with related works. So, to assess the benefits of the proposed method by means of the amplitude modulation metric (Modulation_index) we do compare our findings with the results from [4] and [12], validated on the same database.

In [4], where Itakura distance based method, phase synchronization methods and independent component analysis are used, the classification rates are approximately the same with the actual results and most of classification rates greater than 80% are obtained with QDA, kNN and SVM. Phase synchronization metric conducted to the lowest classification rates (about 60%) and symmetric Itakura distance to the highest ones (about 90%).

By means of wavelet (Coiflet 4 and Daubenchies 2) coefficients and kNN classifiers, in [12], the results were between 68% and 91% (only one subject).

So, the herein results (7 classification rates better than 90% and one value of 100%) outperform those from [4], [12].

In order to improve the rate of classification for all the subjects, we will try to propose a further index/parameter which would better characterize the neuromodulatory activity of the brain, maybe to combine with other features and to employ also other classifiers such as Deep Neural Network (DNN) [14].

It is worth to mention that it is important to have trained subjects. So, we will develop a new database with many sessions of recordings.

## V. Conclusions

The method highlights the frequency bands the cerebral rhythms in a motor task based paradigm are modulated with and provides additional information over conventional methods.

EEG amplitude modulation analysis has revealed the best result concerning classification rate (discriminative patterns) between imagination of right hand movement and relaxation period when the Theta rhythm is modulated with Theta band using LDA and QDA classifiers and between imagination of right and left hand movement when the Theta rhythm is modulated with Theta or Delta bands, working with LDA classifier.

The results showed that EEG activity both for Theta and Alpha rhythms are modulated by the Theta band in a different manner when a person imagine the right hand movement, or left hand movement or he/she relax.

We can conclude that the proposed method is helpful in separating motor tasks for a BCI paradigm.

## REFERENCES

[1] R. Carabalona, P. Castiglioni P, F. Gramatica, "Brain-computer interfaces and neurorehabilitation", Stud Health Technol Inform.,145:160-76, 2009.

[2] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, T.M. Vaughan, "Brain computer interfaces for communication and control", Clinical Neurophysiology, 113(6), pp. 767–791, 2002.

[3] D.J. McFarland, J.R. Wolpaw, "Sensorimotor rhythm-based brain–computer interface (BCI): model order selection for autoregressive spectral analysis", Journal of neural engineering, 5(2), 155, 2008.

[4] O. D. Eva, A. M. Lazar, Feature Extraction and Classification Methods for a Motor Task Brain Computer Interface: A Comparative Evaluation for Two Databases, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No.8, 2017.

[5] A. Hyvärinen, "Independent component analysis: recent advances", Phil. Trans. R. Soc. A, 371(1984), p.20110534, 2013.

[6] V. Gonuguntla, Y. Wang, KC Veluvolu, "Phase synchrony in subject-specific reactive band of EEG for classification of motor imagery tasks", InEngineering in Medicine and Biology Society (EMBC), pp. 2784-2787, 2013.

[7] TH Falk, FJ Fraga, L. Trambaiolli, R. Anghinah, "EEG amplitude modulation analysis for semi-automated diagnosis of Alzheimer's disease", EURASIP Journal on Advances in Signal Processing, 1(192), 2012.

[8] FJ Fraga, TH Falk, PA Kanda, R. Anghinah, "Characterizing Alzheimer's disease severity via resting-awake EEG amplitude modulation analysis", PLoS One, 8(8):e72240, 2013.

[9] B. Blankertz, S. Lemm, M. Treder, S. Haufe, K.R. Müller, "Single-trial analysis and classification of ERP components—a tutorial", NeuroImage, 56(2), 814-825, 2011.

[10] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces" J. Neural Eng., 4, pp. R1-R13, 2007.

[11] A. Bashashati, M. Fatourechi, R.K. Ward, G.E. Birch, "A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals", Journal of Neural engineering, 4(2), R32, 2007.

[12] R. Aldea, M. Fira, A. Lazăr, "Classifications of motor imagery tasks using k-nearest neighbors", InNeural Network Applications in Electrical Engineering (NEUREL), IEEE, pp. 115-120, 2014.

[13] SM Zhou, JQ Gan, F. Sepulveda, "Classifying mental tasks based on features of higher-order statistics from EEG signals in brain–computer interface", Information Sciences, 178(6):1629-40, 2008.

[14] Abeer Al-Nafjan, Manar Hosny, Areej Al-Wabil, Yousef Al–Ohali, "Classification of Human Emotions from Electroencephalogram (EEG) Signal using Deep Neural Network", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 9, 2017.

# Financial Literacy of SME Managers' on Access to Finance and Performance: The Mediating Role of Financial Service Utilization

Juma Buhimila Mabula
Department of Business Administration
Harbin Institute of Technology
Harbin, China

Han Dong Ping
Department of Accounting
Harbin Institute of technology
Weihai, China

***Abstract*—Considering financial literacy as a central factor for consumer demand for financial services, we analyze its impact on access and actual use of financial services and its ultimate consequential reflections on SMEs performance in developing economies. By recognizing the important distinction between access and actual use of financial services this study uses the partial least square-structural equation modelling (PLS-SEM) to estimate the conceptual model. The study reveal significant positive impact of financial literacy to financial access and performance of the firm. It was also discovered that there is significant positive direct impact of access to financial services into actual use of financial services and positive significant effect of the use of financial service on firm performance. The firm use of financial services has a significant mediating role on firm access to financial services-firm performance relationship. The implications of these findings offers foretastes on the need to deepen and widen the scope of SMEs managers' financial literacy for effective financial management and financial financing decisions. We argue distinct contributions of access and actual use of financial services construct on firm performance has to be given attention in attempt to avoid generalizing the phenomenon.**

***Keywords*—Financial literacy; use of financial services; access to financial services; firm performance**

## I. INTRODUCTION

The significance of SMEs contribution to developed and developing economies is notably remarkable [1]-[4]. In developing economies having a small business has been a survival strategy for many individuals, due to thin base and inadequate public and private formal employment [5]. The development of SMEs is hampered by number of factors which may include access to finance, lack of managerial skills, poor equipment and technology, regulation system and absence of access to international markets [6]. The lack of managerial know how places a significant threat to SMEs development [7]. Despite the numerous initiatives to provides training to SMEs owners/managers' the skill gap is still large, this is because most of SMEs cannot afford the cost of deliberate skills upgrade through training and advisory services [8]. Moreover SMEs are considered to be informationally impervious due to their less external monitoring and thin reporting requirements as it can be distinguished from large firms [9].

Becoming an entrepreneurial manager of a venture can be exceedingly risky if there is little or no confidence in financial management [10]. The cost of information and the probability that people utilize a financial resource once they are aware of it depends on the incentives attached to it [11]. The emerging trends posit the low level of consumption of formal financial services and low level of participation into financial markets in developing economies as a function of financial literacy [12]. The firm's use of financial products may not just be an indicator for having access to such products but also having skills and knowledge of managing such products.

Situating access to financial services in developing economies, the global statistics for financial inclusion for 2014 data from Global Findex database stress that 62% of the adult own an account at a formal financial institution, this includes banks and mobile money services. With the innovation of the mobile banks the account ownership has substantially increased in developing economies reaching to 54% of the population in 2014 [13]. Beside the increase the share of excluded individual is still high in developing economies as compared to high income countries. Also [14] claim that despite the efforts that has been under move by the developed economies to enhance the provision of quality and sustainable formal financial services like credit, savings, payment systems, insurance and pensions to mention a few, most of the developing economies still have a large number of population who lack access to basic financial services.

While the problem of SMEs access to finance has been overly studied and many studies has embarked on analyzing it by employing constructs on the supply side like number of accounts providers has, and number of ATM present in the area [15]. Such approach has a detriment of not harnessing the actual facts from individual/firms consumption or non-consumption of such services. The approach by surveying the demand side, where by consumers offers their ground views all else unchanged seems to be compelling. A plausible functioning financial market not only need good infrastructure but also informed consumers [16]. Few studies in developing economies has attempted to analyze the access to financial services based on consumers' financial literacy. [17] Analyzed the impact of individual financial literacy on access to financial services in Kenya using national financial access data. [18] Approached their cross countries analysis on the role of financial literacy education on financial inclusion. Even then these studies had has little to do with the distinction of access

and actual use of financial service, though connected these constructs would have different inferences.

Could the low level of participation in the formal financial market in developing economies be attributed to lack of financial literacy by SMEs? Do firm managers' utilizes their financial literacy to take advantage of available financial products? In the presence of actual use of financial products how does financial literacy and access to financial services impact firm performance? If the financial services available would financial literacy explain the use or the non-use of such services by consumers? What are the mediation roles of financial utilization on financial literacy and or access to financial services-firm performance relationship?

Therefore this study focuses on answering these questions, realizing the dearth of rigorous studies on the concept financial literacy, use and access to finance in developing economies, this scrutiny therefore analyze the impact of financial literacy on SMEs financial access and performance in developing economies. Further we attempt to analyze the mediation role of financial utilization on financial literacy-firm performance and financial access-firm performance relationship. Moreover this study contribute by examining access and use of financial services determinants separately contrary to most studies which studied financial access determinant and employ the use of as a proxy.

## II. Conceptual Framework and Hypothesis Development

This study is motivated by theory of bounded rationality which stress that individuals make decisions, whereby their rationality is limited by their ability to manage the decision problem, intellectual limitations and time available. Consumers in this view act as if seeking for satisfaction rather than focusing on an optimal solution. Bounded rationality impose the idea that individuals take shortcuts that may lead to suboptimal decision-making [19]. This study conceptually analyze the artifactual decisions taken by SME managers at firm level, based on their level of financial literacy which entails financial knowledge, attitudes and behavior will be a base for the decision to use or not to use a certain financial product/service even though is available. This phenomenon implicate supplier of financial services to not only invest of their services making accessible but also strategizing on initiatives that will inspire responsible consumers.

Financially literate consumer is able to identify financial products or services that meet his/her needs, can independently obtain and assess any financial advice and be more likely curb the impact of abusive exploitative practices and scams [20]. Financial literacy enhance consumer effective participation in financial market [21].

The general conceptual framework for this study is depicted in fig.1

Where by FL=Financial Literacy, UFS= Use of Financial Services, AC=Access to finance, PF=Firm performance, UFS'=Mediating variables on AW-FP relationship, UFS''=Mediating variable on AC-FP Relationship

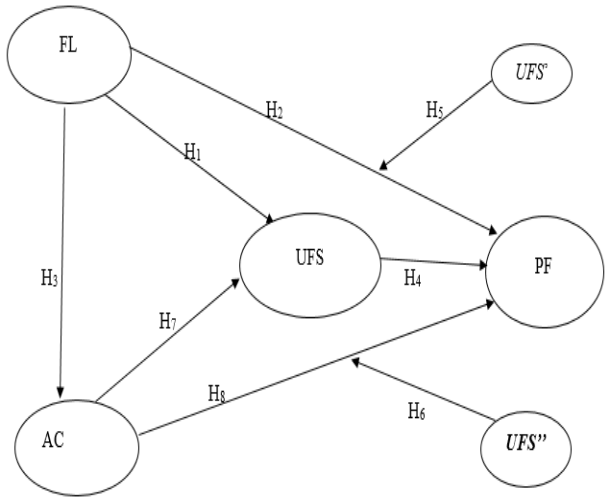



Fig. 1. The General study Model

Financial literacy has been theorized to promote active participation of individual and firms in the formal financial markets [11], [21]. Managers in firms with considerable financial literacy will be more active in making optimal choice of the funding sources by analyzing cost, benefits and risk associated with the potential funds. The British Business Bank (BBB) reveal that the firms' awareness to alternative funding sources is positively related to the participation of firms in the finance market [22]. Also literature evidence that there is a positive correlation between financial literacy to firm performance. [23] Stated that a major hindrance to performance growth of sustainable small and SMEs managers' across developing world is the lack of knowledge, skills, attitude and awareness to cope and manage finances. The dearth of business information and the incompetence of the SMEs management is singled out as the as one of the reason for firm not to access available finances from commercial banks [24]. Therefore we argued that:

Hypothesis $H_1$ Financial literacy has a positive impact on SMEs use of financial products in developing economies

Hypothesis $H_2$ Financial literacy has a positive impact on SMEs performance in developing economies

Hypothesis $H_3$ Financial literacy has a positive impact on firm financial access in developing economies

As it has been exemplified access to financial services doesn't necessarily amount to using it. Availability is inevitable but not a sufficient premise for use [25]. There is a clear difference between a mere access to and actual use of financial services [26]. Access to financial services is not to be equaled to the actual use of financial services. Individuals and firm might have access to financial services due to social or economic reasons decide not to use them. The outreach indicators employed to determine the actual uses of financial services and all these indicators showed a positive correlation on the firm performance indicators. Firms has claimed their non-use of certain financial services is due to cost and affordability. The high minimum deposits, high administrative fees, high fixed cost for loan application and high rejection rate

all these contribute problem. Financial institutions collateral demand to the poor borrowers and lack of financial literacy adds to the challenge. Therefore it is hypothesized that:

Hypothesis $H_4$: The use of financial services has positive impact on firm performance in developing economies.

The availability of financial services is compulsory condition for access from supply side and in addition to this convenience imply efficient access and use of such a service [27]. Use of finance refers to the actual consumption of the financial service. The usage can be referred at the point of intersection of the supply and demand curve of the financial services [25]. The use of financial services entails quality, regularity and the use duration which is subject to voluntary and involuntary exclusion probably more influenced by factors from the supply side connected to information asymmetry [28]. The mediating role of actual use financial services emanate from the fact that, the use of services is an action which is preceded by knowledge of such services and access to it. Without consumption of a particular financial services having knowledge and access to it won't make a resultant impact on SME. Therefore theoretically the use of financial services is supposedly to mediate the relationship of financial literacy/ access to financial services and firm performance. We therefore propose that:

Hypothesis $H_5$ Use of financial services mediate the financial literacy-firm performance relationship in developing economies

Hypothesis $H_6$ Use of financial services mediate the access to finance-firm performance relationship in developing economies

Access to financial services hypothetically has a positive correlation is into the actual use of such services. The more the financial products available at low cost the more the supposedly use by the economic agents. A consented theory propose that a deeper and wider financial sector can contribute to enhanced performance of SMEs. Firms which have access to finance has an opportunity to manage their risks, widen the menu of choices and the consumption patterns. For instance credit as an important instrument deemed to improve performance of firms by enhancing the productive capacity derived from the investment financed [29].

Hypothesis $H_7$ Access to finance has a positive impact on utilization of financial services in developing economies

Hypothesis $H_8$ Access to finance has a positive impact on firm performance in developing economies

## III. LITERATURE

First of all it is important to grasp the meaning of the key terms and their essence niceties in the context of this study.

Financial literacy refers to the possession of the right financial knowledge, skills, attitudes and behavior [30]. At OECD/INFE defines financial literacy as a combination of awareness, knowledge, skill, attitude and behavior necessary to make sound financial decisions and ultimately able to achieve financial wellbeing [30]. At another level financial literacy has been synonymously considered to be the same as financial literacy [31], [32].

Financial access to financial services refers to the availability of the supply of reasonable quality financial services. [33] Extend the definition of access to finance to mean the ability of individuals or enterprises to attain financial services, which may include credit, deposit, payments, insurance, and other risk management services. Financial utilizations means the actual use/consumption of financial services [12].

The terms utilization and access to financial services has been used conjointly and interchangeably by many studies even though they overlap they distinctly mean different things. To understand well the conceptualization of the differences between use and access to financial services [34] gives an illustration of worthy individuals in advanced financial markets, still they might not use some of the financial services even though they are readily available and they can afford to pay. Nevertheless every individual and firms at some points in time need a certain kind of financial product. The difference between access and actual utilization of financial services could be analyzed by a demand-supply structure. As access refers to supply while the use refers to the intersection between the supply and demand schedules [25]. Voluntary exclusion doesn't necessarily mean the service is not available or is rationed, the decision consumers' to utilize financial services is based on their financial attitude and behavior.

Financial inclusion is the term that exclusively embrace financial access and use together. [35] Define financial inclusion as "the process of insuring access to appropriate financial products and services needed by all sections of society including vulnerable groups such as weaker sections and low income groups at an affordable cost in a fair and transparent manner by main stream institutional players" .Financial inclusion focuses on making easy access to financial services to large disadvantaged people. The lack of financial access limit the array of services and credits to individual, household and enterprises. [36] Mention of factors that are limiting access to finance to include place of living, absence of legal identity and gender biasness, limited knowledge of financial services, level income and bank charges, rigid terms and conditions and the type of business. [37] Offers glimpses of the principle of how to measure countries progress towards enabling access to and driving usage of the different financial products. in attempt to define access and usage in accurate and comprehensive way, they identified levels of availability of financial product, actual ownership of the product, the use of the product over a specified time and how much the product is used to address the consumer's needs.

### *A. Financial Literacy and Financial Inclusion*

Financial literacy stem as an important constituent of consumer effective participation in the formal financial market. The benefits derived from enhancing it has dual benefit to consumers and suppliers of financial services [17]. While financial literacy provides the apt information for the consumer in the financial market. Provided need and access to certain financial product, the attitude and behavior will enable the consumer to be active partaker of the financial product. [16]

Argues that access to financial services and particularly having a bank account, the marginal benefits of financial literacy decreases with higher financial depth, indicating that financial access and financial literacy are supposedly substitutes. However if inclusion is all about access and use of financial services, financial literacy has a complementary effect on financial depth at higher level of comprehension and therefore they reinforce each other.

### B. Role of Financial Institutions on Financial Literacy and Financial Access

Banks has played a big role in providing awareness and appropriate use of financial services they offer to their clients. The role of banks in provision of financial education shore up their unique position in the financial system, as they act as a bridge in theoretical economic concepts such as scarcity and opportunity costs, with real money in the pocket services and supplement them with the necessary financial products. The access to low cost financial products is mainly valuable for unbanked and under-banked clients. Therefore connecting financial education and financial products allows all individuals to become full incorporated in the traditional financial system [38].

[14] Examined the impact of presence of mobile financial services in Sub Saharan Africa and found that the presence of mobile banking has positively enhanced the saving behavior, and it stands as an avenue for promoting financial inclusion especially to the poor and low income groups constrained to financial access. By reforming the banking sector and strengthening the financial sector legal and regulatory framework and modernizing the financial sector infrastructure and extending the range of basic financial products and services to underserved zones access to finance for households and firms will be enhanced.

### C. Financial Literacy for SMEs

The Bank Association of South Africa as cited by [39] defines financial literacy for SMEs as one which has the following qualities: first has an adequate level of personal entrepreneurial competencies, personal finance skills, and business management skills; has an appropriate level of understanding of functional financial management systems. Second, has appropriate level of understanding of SME life-cycle funding and other financial service requirements. third, understands legal, regulatory and tax issues as they relate to financial matters and lastly understands the range of legal recourses it can resort to when necessary, and namely, in case of bankruptcy or other situations of financial distress [40].

A financially literate SME owner/manager is considered to know the most suitable financing and financial management options for his/her business at various growth stages, knows where to attain the most appropriate financial products and services; and interacts confidently with the suppliers of these products and services. Is familiar with the legal and regulatory framework and his/her rights and recourses [41].

The ability of a small business owner/managers to gather information from financial statement has a big role to play in business success, because it supports them to detect signs of failure, and overall health status of the venture [42]. The general stance of firm financial literacy stems on manager's ability to translate their personal finance into a business venture context. Financial literacy in the new business reality is defined as the competence to adequately manage financial resources over the lifecycle and effectively link with financial products and services [43].

### D. Access to Finance and Firm Performance

[44] Using data from enterprise survey examined access to finance and performance of the firm in African context. Using the subjective measure they found that lack of access to finance exert a significant negative effect on firm growth. And the objective measures proved that firms that are not financially constrained has high rate of growth than the credit constrained ones. Also [45] similarly studying the impact of financial inclusion and bank concentration on firm performance, they found that the financial inclusion on distribution of financial service across firm has a positive impact onto firm growth.

### E. SME Definition in the Context

Since there is no universally accepted SME definition, the nomenclature in this study context is used to mean Micro, Small and Medium enterprises. It is sometimes referred to as micro, small and medium enterprises (MSMEs). The definition is derived from Tanzania mainland where the survey was relevantly conducted. The SMEs cover economic activities mainly manufacturing, mining, commerce and services. The commonly used benchmarks are total number of employees, total investment and sales turnover. In this standpoint, micro enterprises are those engaging up to 4 people, in most cases family members or employing capital amounting up to Tanzania Shilings (Tshs.) 5.0 million. The majority of microenterprises fall under the informal sector. Small enterprises are mostly formalized undertakings engaging between 5 and 49 employees or with capital investment from Tshs 5million to Tshs. 200million. Medium enterprises employ between 50 and 99 people or use capital investment from Tshs. 200 million to Tshs. 800 million [46].

## IV. METHODOLOGY

The sample for our study was composed of SMEs managers from region of Morogoro and Dar es Salaam from Tanzania. From previous research firm managers form an interesting group to analyze the level of financial literacy and its consequences to their firm operations in developing economies, considering the fact that they are involved in their daily financial management of the firm [12], [48].

We used a convenient sampling because of the locational and financial limitations. We used a self-administered questionnaire of which we had to physically distribute it to the premise of the firm and collet it thereon. The questionnaire comprehensively was designed to capture the SMEs managers' financial literacy at personal and firm level. And this study is part of such a comprehensive questionnaire. All questions used in this study has was 5 point likert scale questions. To encourage participation and minimize non responses the questionnaire was preceded by the introduction letter and personal expression for respondents to know that the study was solely academic based. The survey was piloted in the field and there after the instruments was refined by experts from the

field. The questionnaire completely filled was 311 even though we managed to reach 520 firms, a response rate of about 60%. Cronbach's alpha test was used to verify the quality and consistency of our instrument.

### A. Variable Definition and Sources

Table I presents definitions of the variables indicators and their respective sources.

TABLE I. VARIABLE DEFINITION AND SOURCES

| Variables | Indicators | Source |
|---|---|---|
| FL | General financial literacy (GFL), Financial knowledge (FK), Financial attitude (FA), and Financial Behavior (FB) | Atkinson, Adele<br>Messy, Flore-Anne 2013 |
| UFS | Easiness of having personal bank account (UFSi), using the firm bank account, banking of daily undertakings (UFSii), Reinvesting the cash balances (UFSiii), having credit from financial institution (UFSiv). | Olawale Fatoki, 2014 |
| AC | Easiness of getting funds from financial institutions (ACi), limited availability of funds for firm (ACii), and the cost of acquiring funds is too high (ACiii). | World Bank's enterprise survey |
| PF | Steady firm growth (PFi), steady increase number of employees (PFii), increase in productivity (PFiii) and ability to buy fixed assets (PFiv) | World Bank's Enterprise survey |

### B. Measurement of Variables

The constructs in this study were measured using a multi-item scale, all measures were adapted from the previous studies. Each measure was anchored on five-point scale. The survey instruments form the measures was pre-examined by three different business academics to make solid wording and clarity of the questions. To establish the consistent with the spirit of the original instrument we performed instrument factor analysis before including them in the PLS model.

#### 1) Financial Literacy

Financial literacy was adopted from [18] study, where by there is a question on self-rating financial literacy, financial knowledge, financial attitude and financial behavior. Then we adopted the questions and stylized them into a 5-point likert scale except for financial knowledge. Financial knowledge construct has a multiple choice questions to test the apt knowledge of consumers of financial literacy, correct answer and categorized them into low, medium and high scores.

#### 2) Use of Financial Services

We also adapted part of [47] questions on understanding financing sources because they were fit into the use of financial services. From thence we made a five point likert scale to actually draw a behavioral conclusions from the managers. The principal factor analysis yielded a single factor solution with eigenvalues exceeding 1 and factor loading ranging from 0.874 to 0.964. This factor explained 69.66 % of the total variance.

#### 3) Access to Finance

Coupled with the details from the World Bank enterprise survey, we created questions for access to finance. We created five-point scale subjective measures on easy access, availability and cost of funds to the firm. The principal factor analysis results were single factor solution factor loading ranging from 0.962 to 0.989. This factor explained 34.954% of the total variance.

#### 4) Firm Performance

We adapt construct from the Word Bank enterprise survey for performance of the firm. Then we created a five point likert scale measures to capture our respondents responses on the firm performance. The construct focused on sales growth, number of employees' growth, productivity and ability to purchase fixed assets. After performing the principal factor analysis our results gave a single factor solution, factor loading ranging from 0.698 to 0.826. The factor explained 53.139 of the total variance.

### C. Descriptive Statistics and Control Variables

We also collected information about the individual and firm industry for our respondents to add to the foundational constructs of our study interest. These variables were included in our statistical model to partial out their effects on our dependent variables. Table II shows demographic information at personal level includes gender and level of education. The reminder of the variables are at firm level with includes type of business, business location, age, and number of employees. None of our control variables were significantly related to our dependent variables.

TABLE II. SUMMARY OF RESPONDENTS' DEMOGRAPHIC CATEGORIES

| | | Number | Percentage |
|---|---|---|---|
| Gender | Male | 164 | 52.9% |
| | Female | 146 | 47.1% |
| Education Level | Primary | 11 | 3.5% |
| | Secondary | 87 | 28.1% |
| | College | 134 | 43.2% |
| | University | 78 | 25.2% |
| Business type | Wholesale and Retail | 101 | 32.2% |
| | Agricultural | 56 | 18% |
| | Construction | 31 | 10.0% |
| | Food and Accommodation | 31 | 10.0% |
| | Manufacturing | 30 | 9.7% |
| | Others | 61 | 19.1% |
| Location | Rural | 127 | 41% |
| | Urban | 183 | 59% |
| Firm Age | Below 1 Year | 39 | 12.6% |
| | 1-5 Years | 132 | 42.6% |
| | 6-10 Years | 100 | 32.3% |
| | More than 10 Years | 39 | 12.6% |
| Number of employees | 1-5 Micro | 144 | 46.5 |
| | 6-49 Small | 147 | 47.4 |
| | 50-99 Medium | 19 | 6.1 |

### D. Results

We use SPSS version 23 to obtain the descriptive statistics and then SmartPLS for estimation of Partial Least Square-Structural Equation Modelling (PLS-SEM).We use PLS-SEM approach because of its exceptional ability and fewer

assumption than covariance-based structural modelling [49]. PLS recognized to be a component based approach that attempt to maximize variance explained and minimize error. With PLS we are able to examine theory and measures simultaneously. PLS-SEM has an advantage of handling small sample sizes and no assumptions for particular scale and normality of the data distribution [50]. [51]. Basing on the rule of sampling that the minimum sample size for PLS modeling have to be ten times of the portion of the model that require multiple regression. In our model we had 3 independent variables, allowing us to use a sample of size of 30. Therefore our sample of 310 respondents was more than adequate for PLS modelling.

### *E. The Measurement Model*

The model estimate the relationships between manifest variables and latent variables hence we have the inner and outer model. In the model we estimate and evaluate the reliability and validity of the model. Table III the indicator's reliability, the composite reliability, convergent validity, and the descriptive statistics of our study model. For internal consistency of our model. [52] Suggests using factor loading instead of Cronbach alpha. the factor loading of 0.7 is always preferred but 0.4 or higher loading is acceptable for exploratory research [53], [54] in our model the minimum factor loading is 0.438 and the maximum is 0.911 all being in the acceptable range. Complementing with the factor loading our result shows that all factor were loaded significantly at $p< 0.001$ and $p< 0.05$ one-tailed significant levels. With these results we confirm that all our constructs has satisfactory reliability.

#### *1) Model Internal Consistency*

The model internal consistency or reliability is calculated by the formula in equation 1.

$$(\sum \boldsymbol{\lambda_i})^2 / [(\sum \boldsymbol{\lambda_i})^2 + \sum(\mathbf{1} - \boldsymbol{\lambda_i}^2)] \quad (1)$$

Whereby: $\boldsymbol{\lambda_i}$ represent the ith component loading and $\mathbf{1} - \boldsymbol{\lambda_i}2$ refer to the ith error variance. In this stance composite reliability measure is used as an alternative to Cronbach's alpha. This is undertaken on the basis that composite reliability weighs each factor loading. The threshold of ≥ 0.7 composite reliability is regarded to be suitable [55]. In this study model result shown in table III provides all composite reliability to be above 0.7 confirming that the model possesses enough internal consistency.

#### *2) Convergent and Discriminant Validity*

To assess the convergent and discriminant validity we consider the Average Variance Extracted (AVE) calculated for each component. The AVE measures the degree to which scale items for each items are correlated. 0.5 or above AVE is considered to be adequate [54], [56].

AVE can be calculated by the equation 2:

$$(\sum \boldsymbol{\lambda_i})^2 / [\sum \boldsymbol{\lambda_i}^2 + \sum(\mathbf{1} - \boldsymbol{\lambda_i}^2)] \quad (2)$$

Whereby: $\boldsymbol{\lambda_i}$ represent the ith component loading and $\mathbf{1} - \boldsymbol{\lambda_i}2$ refer to the ith error variance

In our model the minimum AVE is 0.503 and the maximum AVE is 0.603 as reported in table III. Therefore our results demonstrates adequate convergent validity.

TABLE III. RELIABILITY AND CONVERGENT VALIDITY (AVE)

| Latent Variable | Mean | SD | Loading | t-statistics |
|---|---|---|---|---|
| ***Financial Literacy (composite reliability = 0.726; AVE = 0.503)*** | | | | |
| GFL | 0.417 | 0.154 | 0.438 | 2.845** |
| FK | 0.593 | 0.140 | 0.607 | 4.341*** |
| FA | 0.718 | 0.099 | 0.731 | 7.409*** |
| FB | 0.716 | 0.100 | 0.730 | 7.267*** |
| ***Use of financial services (composite reliability = 0.773; AVE = 0.511 )*** | | | | |
| UFSi | 0.853 | 0.137 | 0.885 | 6.449*** |
| UFSii | 0.702 | 0.169 | 0.720 | 4.297*** |
| UFSiii | 0.513 | 0.194 | 0.551 | 2.845** |
| UFSiv | 0.496 | 0.180 | 0.530 | 2.947** |
| ***Access to financial services (composite reliability = 0.761; AVE = 0.522)*** | | | | |
| ACi | 0.718 | 0.174 | 0.753 | 4.318*** |
| ACii | 0.808 | 0.148 | 0.838 | 5.650*** |
| ACiii | 0.544 | 0.176 | 0.544 | 3.091*** |
| ***Firm performance (composite reliability = 0.818; AVE = 0.605)*** | | | | |
| PFi | 0.762 | 0.098 | 0.769 | 6.732*** |
| PFii | 0.901 | 0.038 | 0.911 | 15.423*** |
| PFiii | 0.609 | 0.136 | 0.628 | 6.449*** |

***,*** p < 0.05 and p <0.01, respectively*

The discriminant validity was examined by comparing the square roots of the AVEs of the constructs with the correlation coefficient between the constructs [55]. The result of discriminant validity is shown in table IV where by the correlation coefficients are in off-diagonal, and the square root of the AVEs are shown on diagonal. Our model indicated that there is adequate discriminant validity due to the fact that all the square root of the AVEs are greater than the correlation coefficients.

TABLE IV. DISCRIMINANT VALIDITY AND HETERO MONOTRAIT RATIO (HTMT) IN PARENTHESIS

| | AC | FL | PF | UFS |
|---|---|---|---|---|
| AC | 0.686 | | | |
| FL | 0.085 (0.722) | 0.638 | | |
| PF | -0.283 (0.353) | -0.295 (0.572) | 0.778 | |
| UFS | -0.334 (0.562) | -0.294 (0.370) | 0.229 (0.438) | 0.686 |

*Diagonal item are square root of AVE, and non-diagonal elements are the correlation of the elements*

The other approach to confirm the model discriminant validity is to examine the cross loading. In principle indicators are supposed to load well on their target factors and the cross loading not supposed to be significant. The heuristic limit loading for cross loading should be ≤ 0.3 [57]. From table V note that all cross loading are ≤0.3, further endorsing that our model has discriminant validity.

TABLE V. VARIANCE INFLATION FACTOR (VIF) AND CROSS LOADING

| | VIF | AC | FL | PF | UFS |
|---|---|---|---|---|---|
| ACi | 2.208 | 0.753 | -0.234 | 0.154 | -0.145 |
| ACii | 2.217 | 0.838 | 0.288 | 0.179 | -0.207 |
| ACii | 1.011 | 0.544 | -0.303 | 0.145 | -0.302 |
| FA | 1.690 | -0.168 | 0.731 | -0.239 | 0.047 |
| FB | 1.624 | -0.305 | 0.730 | -0.273 | 0.047 |
| FK | 1.022 | -0.302 | 0.607 | -0.223 | 0.200 |
| GFL | 1.089 | -0.154 | 0.438 | -0.144 | 0.127 |
| PFi | 1.533 | 0.185 | -0.273 | 0.769 | -0.272 |
| PFii | 2.533 | 0.246 | -0.324 | 0.911 | -0.231 |
| PFiii | 1.478 | 0.062 | -0.224 | 0.628 | -0.140 |
| UFSi | 2.801 | -0.259 | 0.184 | -0.196 | 0.885 |
| UFSii | 2.134 | -0.264 | 0.102 | -0.161 | 0.720 |
| UFSiii | 2.420 | -0.123 | 0.142 | -0.214 | 0.551 |
| UFSiv | 1.082 | -0.201 | 0.052 | -0.209 | 0.530 |

*3) Common Method Bias*

Because of the self-reported data there is a possibility of existence of common method bias. The common method bias can be caused by a number of factors which may include consistency and social interest [58]. In attempt to find out whether our data was affected by common method bias, we use the approach by [59]. In this procedure we included in the PLS model a common method factor where by this common factor indicators included all principal constructs' indicators and then we calculated the variance 's indicators essentially exemplified by the principal construct and by the method. Our results as shown in table VI. The average variance of the indicators is 0.523, whereas the average method variance is 0.240, from this we obtain the ratio of principal constructs variance to method variance to be about 2.2:1. Our result also shows that all the factor loadings for the method variances are not significant. Given the higher variance of the substantive indicators and the insignificant results on the method variance, then we determine that the common method bias is unlikely to be a serious concern for our study.

Also the common method bias problem was assessed by the adopting the approach used by [60]. The practice is to assess the Variance Inflation Factor (VIF) if they are $\leq 3.3$ for the model which has passed convergent and discriminant validity evaluation. The VIF shown in table V presents all results to be less than 3.3, further confirming that our scrutiny was not affected by common method bias.

TABLE VI. COMMON METHOD BIAS ANALYSIS

| | | Substantive Factor Loading | | Method Factor Loading | |
|---|---|---|---|---|---|
| ***Construct*** | ***Indicator*** | ***(R1)*** | $R1^2$ | ***(R2)*** | $R2^2$ |
| Financial Literacy | GFL | 0.757*** | 0.573 | 0.492 | 0.242 |
| | FK | 0.617** | 0.381 | 0.509 | 0.259 |
| | FA | 0.767*** | 0.588 | 0.443 | 0.196 |
| | FB | | | | |
| Use of financial services | UFSi | 0.907** | 0.823 | 0.560 | 0.314 |
| | UFSii | 0.705** | 0.497 | 0.451 | 0.203 |
| | UFSiii | 0.591** | 0.241 | 0.386 | 0.149 |
| | UFSiv | 0.477** | 0.228 | 0.318 | 0.101 |
| Access to finance | ACi | 0.828** | 0.686 | -0.498 | 0.248 |
| | ACii | 0.880*** | 0.774 | -0.602 | 0.362 |
| | ACiii | 0.429 | 0.184 | 0.403 | 0.162 |
| Firm Performance | PFi | 0.919*** | 0.845 | -0.547 | 0.299 |
| | PFii | 0.729*** | 0.531 | -0.401 | 0.161 |
| | PFiii | 0.667*** | 0.445 | -0.650 | 0.423 |
| Average | | 0.713 | 0.523 | 0.067 | 0.240 |

***,*** p < 0.05 and p <0.01, respectively*

## V. RESULTS AND DISCUSSION

After verification of the model measurement and confirming no worry of the common method biases we carry on examining the structural model. Table VII and fig. 2 provides us with the detailed results of the PLS analysis. Sheet A shows the direct relationship and sheet B shows an indirect relationship.

TABLE VII. PATH COEFFICIENT AND T-STATISTICS RESULTS

| | Sample | Mean | Standard Deviation | T Statistics |
|---|---|---|---|---|
| AC->PF | 0.017 | 0.013 | 0.122 | 0.135 |
| AC->UFS | 0.298 | 0.314 | 0.052 | 2.184* |
| FL->AC | 0.448 | 0.467 | 0.075 | 5.996*** |
| FL->PF | 0.309 | 0.317 | 0.108 | 2.855** |
| FL->UFS | 0.044 | 0.051 | 0.136 | 0.326 |
| UFS->PF | 0.223 | 0.243 | 0.097 | 2.038* |

**, **,*** p < 0.05 and p <0.01, and p <0.001 respectively*

$H_1$ predicted a positive relationship between SMEs managers' financial literacy and use of financial products in developing economies. The PLS structural path on this shows insignificant results. This explains that financially literate doesn't necessarily mean using them effectively. Other factors like the access to such funding sources might explain the relationship between financial literacy and the use of financial services by SMEs managers in developing economies.

$H_2$ predicted the positive association between SMEs managers' financial literacy and firm performance in developing economies. The structural model shows a significant result at ($p<0.01$) in the hypothesized direction. The results suggest that the SMEs managers' financial literacy is such an important driver for the firm performance. Therefore $H_2$ is supported by our model.

$H_3$ Also predicted that SMEs managers' financial literacy has a positive impact on firm financial access in developing. The results from the structural PLS model shows a significant results at ($p<0.01$) in the hypothesized direction. Implying that the more the SMEs managers in developing economies be financially literate the more enhancement of access to finance by the firm. Therefore $H_3$ is supported by our model. The significant results in this hypotheses might explain the reasons for insignificant results in $H_1$.

In $H_4$ we hypothesized that there is a positive relationships between the use of financial services and the firm performance in developing economies. In this we have a significant result also at ($p<0.05$) from the model. This shows that as firms in developing economies use more of formal financial services there will also be improvement in the firm performance. Therefore $H_4$ is also supported by our model.

$H_7$ projected that there is a positive relationship between firm accesses to finance and the utilization of the financial services. The results of our PLS structural model provide that there is significant result at ($p<0.01$) in the predicted direction. Confirming our prediction that as more access to finance firms

has in developing economies they are likely to use more financial services. Hence $H_7$ is supported in our model.

$H_8$ Also predicted appositive association of access to financial services on the performance of the firm. Our PLS structural model shows insignificant results. Therefore H8 is not supported by our model. This results may explain that there might be indirect relationship between access to finance and performance of the firm. We therefore take care of the case by analyzing the mediation role of use of financial services to access to financial services-firm performance relationship.

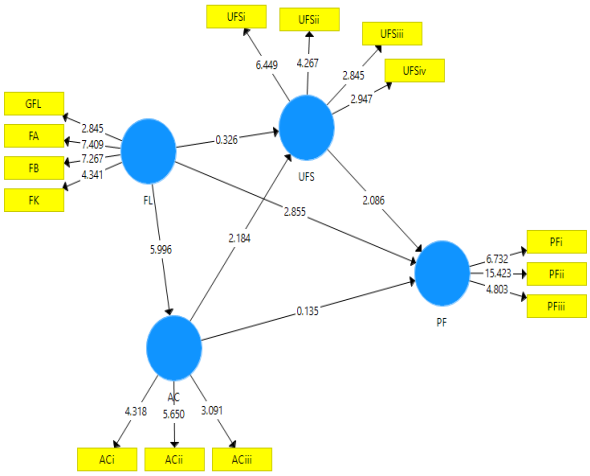



Fig. 2. Results after Bootstrapping

The fact that we had insignificant results in some relationship in our model and exploring further inferences we examine the supposedly mediation role of the firm managers use of financial services into the financial literacy-firm performance and access to finance-firm performance relationship. Hence we have $H_5$ and $H_6$ to do this we use the Sobel test of mediation available online at https://www.danielsoper.com.The Sobel test of mediation uses three formulas to calculate the mediation:

The first formula is the error function

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \qquad (1)$$

The second formula is the normal distribution cumulative distribution function (CDF)

$$F(x; \mu, \delta^2) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\delta\sqrt{2}}\right)\right] \qquad (2)$$

Where μ is the mean, δ is the standard deviation and 'erf' is the error function.

Sobel test statistic:

$$Z = \frac{ab}{\sqrt{b^2 SE_a^2 + a^2 SE_b^2}} \qquad (3)$$

Where 'a' is the regression coefficient for the relationship between the independent variable and the mediator 'b' is the regression coefficient for the relationship between the mediator and the independent variable, $SE_a^2$ is the standard error of the relationship between the independent variable and the mediator and $SE_b^2$ is the standard error of the relationship between the mediator variable and the independent variable.

Using this approach then we test $H_5$ and $H_6$

$H_5$ Predicted that the SMEs firm use of financial services mediate financial literacy-firm performance relationship in developing economies. From our model we have the firm performance (PF) as the dependent variable, use of financial services (UFS) as the mediator and financial literacy (FL) as independent variable. We have the original sample path coefficient FL->UFS is 0.044 and the standard error is 0.136, the original sample path coefficient for UFS->PF is 0.236 and the standard error is 0.097. Therefore we input these data into the online Sobel test and we get Z= 0.32037257, one tailed probability = 0.37434296 and two tailed probability = 0.74868591. Our results shows insignificant mediation role of the use of financial services and on the relationship of financial literacy and the firm performance. Therefore $H_5$ is not supported.

$H_6$ Also projected that the use of financial services mediate the access to finance-firm performance relationship in developing economies. The firm performance is our dependent variable, the use of financial services is the mediating variable and access to financial services is the independent variable. From our results we have the following data:

The original sample path coefficient for AC->UFS is 0.298 and the standard error is 0.052, the path coefficient for UFS->PF is 0.236 and the standard error is 0.097. Inputting the data in the Sobel test we yield Z = 2.13368290, one tailed probability =0.01643437, and the two tailed probability = 0.03286875. The results is significant at $p < 0.05$ indicating that the use of financial services significantly explain the relationship between access to financial services and firm performance. Therefore H6 is supported.

### A. Discussion of the Results

The objective of this study was to examine the SMEs financial literacy, access to financial services, use of financial services influence each other and the eventual impact on firm performance. Our study offer evidence of the significant results between financial literacy and access to finance, financial literacy and financial performance. The results cement evidences of the positive impact of financial literacy on either firm access to financial services and ultimately improves SMEs performance, access to financial service and the use of financial services and financial performance. Our findings is in conformity with prior studies of [24], [61]-[63]

Also our study finds that the use of financial services significantly mediate the relationship between access to financial services and firm performance, such that the effect impact of access to financial services on firm performance is more explained when there is use of financial services by the SMEs. Our results implies that use of financial services significantly justify the causal order of access to financial services on firm performance such that the higher the use of financial services the more clearer we can appreciate the impact of access to financial services on firm performance.

The results are profound contribution to scholarly literatures on how we can understand and review the variables relationships in our model. The literature confirms that there is positive impact of access to financial services to firm performance [64]-[67] but none of these tried to examine the anticipated effect of the mediating role of the actual use of financial services.

Our model also reveal that there is an insignificant between financial literacy to use of financial services, this may speculatively explains that even though awareness of financial resources remain to be an important driver for use such services some other social and economic factors might influence individuals and firm on the actual use, and also the same applies to access to financial services-firm performance relationships which might be insignificant as well.

Our model uncover that there is insignificant mediating role of use of financial services on the financial literacy-firm performance relationship. This implies that use of financial services has insignificant explanation of the awareness of funding sources and firm performance relationships. In other words the causal order between financial literacy and firm performance is immaterially explained by the use of such financial services. The results makes sense because there is significant positive direct relationship of financial literacy-firm performance path.

Many studies have treat access and use of financial services as a single variable, by breaking down these parts our study brings down a more thorough interplay analysis of the financial accessibility and usability issues facing individuals and firms and therefore add to the body of literature.

## VI. CONCLUSIONS

Even though there are ample efforts by the governments, private sectors, regional and multilateral organizations, there is still more to be done when it comes to SMEs performance improvements. SMEs managers at personal level still need to have more knowledge and insight on how they manage the finances of their venture. Need more flexible mindset to match with the pace of the growth of the finance supply side for their firm sustenance. For SMEs in developing countries to thrive there is a need to improve more aggressive efforts need to be implemented in terms of personal initiative training, the funding suppliers which includes the formal banks, microfinances, governments agencies and NGOs need widen their scope of financial literacy not only to their customers but also to the wider community. While providing education it is important to fight the barrier towards access to financial services monitoring the actual use of the financial products by firms and noting the reasons for non uses. Boosting the competitive behavior of the financial institutions would result into ample supply of financial services to SMEs and low-income households at less costs of intermediation.

## REFERENCES

[1] E. Lukács, "The economic role of SMEs in world economy, especially in Europe," European integration studies, vol. 4, pp. 3-12, 2005.

[2] K. Hayes, "SMEs and the Economy," in Business Journalism, ed: Springer, 2014, pp. 175-185.

[3] L. I. Savlovschi and N. R. Robu, "The role of SMEs in modern economy," Economia, Seria Management, vol. 14, pp. 277-281, 2011.

[4] S. Ahmed, "the role of SMEs in developing economy," Abuja, Omotayo and co. ltd, 2006.

[5] G. S. Fields, "Labor market analysis for developing countries," Labour Economics, vol. 18, pp. S16-S22, 2011.

[6] A. F. Gockel and S. K. Akoena, Financial Intermediation for the Poor: Credit Demand by Micro Small and Medium Scale Enterprises in Ghana: a Further Assignment for Financial Sector Policy?: ILO, 2002.

[7] J. Abor and P. Quartey, "Issues in SME development in Ghana and South Africa," International Research Journal of Finance and Economics, vol. 39, pp. 215-228, 2010.

[8] K. Dalitso and Q. Peter, "The policy environment for promoting small and medium-sized enterprises in Ghana and Malawi," University of Manchester, 2000.

[9] G. F. Udell, "SME Access to intermediated credit: What do we know and what don't we know," in Reserve Bank of Australia, Proceedings of Small Business Conditions and Finance Conference, 2015, pp. 61-109.

[10] S. Spinelli and H. Neck, "The Timmons model of the entrepreneurial process," Entrepreneurship: The Engine of Growth, vol. 2, p. 6, 2007.

[11] L. Guiso and T. Jappelli, "Awareness and stock market participation," Review of Finance, vol. 9, pp. 537-567, 2005.

[12] J. Nunoo and F. K. Andoh, "Sustaining Small and Medium Enterprises through Financial Service Utilization: Does Financial Literacy Matter?," in Annual Meeting, 2012.

[13] A. Demirgüç-Kunt, L. F. Klapper, D. Singer, and P. Van Oudheusden, "The global findex database 2014: Measuring financial inclusion around the world," 2015.

[14] N. A. Waithaka J.M, "Effects of Informal Finance on the Performance of Small and Medium Enterprises in Kiambu County," International Journal of Scientific and Research Publications, vol. 5, 2015.

[15] R. Cull and K. Scott, "Measuring household usage of financial services: does it matter how or whom you ask?," the world bank economic review, vol. 24, pp. 199-233, 2010.

[16] A. Grohmann, T. Klühs, and L. Menkhoff, "Does Financial Literacy Improve Financial Inclusion? Cross Country Evidence," 2017.

[17] M. I. Wachira and E. N. Kihiu, "Impact of financial literacy on access to financial services in Kenya," International Journal of Business and Social Science, vol. 3, 2012.

[18] A. Atkinson and F.-A. Messy, "Promoting financial inclusion through financial education," OECD Working Papers on Finance, Insurance and Private Pensions, p. 1, 2013.

[19] H. A. Simon, "Theories of bounded rationality," Decision and organization, vol. 1, pp. 161-176, 1972.

[20] R. Rugimbana and E. Kojo Oseifuah, "Financial literacy and youth entrepreneurship in South Africa," African journal of Economic and management studies, vol. 1, pp. 164-182, 2010.

[21] M. Van Rooij, A. Lusardi, and R. Alessie, "Financial literacy and stock market participation," Journal of Financial Economics, vol. 101, pp. 449-472, 2011.

[22] BBB, "Small Business Finance Market," ed: British Business Bank, 2015.

[23] A. A. Eniola and H. Entebang, "Financial literacy and SME firm performance," International Journal of Research Studies in Management, vol. 5, 2015.

[24] O. O. Fatoki and F. Asah, "The impact of firm and entrepreneurial characteristics on access to debt finance by SMEs in King Williams' Town, South Africa," International Journal of Business and Management, vol. 6, p. 170, 2011.

[25] S. Claessens, "Access to financial services: a review of the issues and public policy objectives," The World Bank Research Observer, vol. 21, pp. 207-240, 2006.

[26] T. Beck, A. Demirguc-Kunt, and M. S. M. Peria, "Reaching out: Access to and use of banking services across countries," Journal of Financial Economics, vol. 85, pp. 234-266, 2007.

[27] J. Talledo, "Access To And Use Of Financial Services: Evidence From Peru," 2015.

[28] T. Beck and A. De la Torre, The basic analytics of access to financial services: The World Bank, 2006.

[29] G. I. Nwanna, "Financial Accessibility And Rural Sector Development/Accès Au Financement Et Developpément Du Secteur Rural," Savings and Development, pp. 453-491, 1995.

[30] A. Hung, J. Yoong, and E. Brown, "Empowering women through financial awareness and education," OECD Working Papers on Finance, Insurance and Private Pensions, p. 1, 2012.

[31] P. Marriott, "An Analysis of First Experience Students' Financial Awareness and Attitude to Debt in a Post-1992 UK University," Higher Education Quarterly, vol. 61, pp. 498-519, 2007.

[32] D. Ramasamy, T. Savilla, S. Anoop, and M. Ramen, "A study of the level of awareness of financial literacy among management undergraduates," Google Scholar, 2013.

[33] A. Demirgüç-Kunt, T. Beck, and P. Honohan, Finance for all?: Policies and Pitfalls in Expanding Access: World bank, 2008.

[34] A. Demirguc-Kunt, T. Beck, and P. Honohan, "Access to Finance and Development: Theory and Measurement," Finance for All? Policies and Pitfalls in Expanding Access. The World Bank, p. 22, 2008.

[35] K. Chakrabarty, "Financial Inclusion in India—Journey So Far and Way Forward," Keynote Address at the Finance Inclusion Conclave. Organized by CNBC TV, vol. 18, 2013.

[36] B. A. Iqbal and S. Sami, "Role of banks in financial inclusion in India," Contaduría y Administración, vol. 62, pp. 644-656, 2017.

[37] T. Arun and R. Kamath, "Financial inclusion: Policies and practices," IIMB Management Review, vol. 27, pp. 267-287, 2015.

[38] A. Sjauw-Koen-Fa and I. Vereijken, "Access to financial services in developing countries," Rabobank Nederland, September, pp. 1-32, 2005.

[39] F.-A. Messy and C. Monticone, "The status of financial education in Africa," 2012.

[40] BASA, "Save for Better Future," ed. Johanmrsburg, South Africa: Banking Association in South Africa 2011.

[41] UN, "The millennium development goals report 2009," E. a. P. information, Ed., ed: United Nations Publications, 2009.

[42] P. Dahmen and E. Rodríguez, "Financial literacy and the success of small businesses: An observation from a small business development center," Numeracy, vol. 7, p. 3, 2014.

[43] A. A. Eniola and H. Entebang, "SME Managers and Financial Literacy," Global Business Review, p. 0972150917692063, 2017.

[44] B. Fowowe, "Access to finance and firm performance: Evidence from African countries," Review of Development Finance, vol. 7, pp. 6-17, 2017.

[45] L. Chauvet and L. Jacolin, "Financial Inclusion, Bank Concentration, and Firm Performance," World Development, vol. 97, pp. 1-13, 2017.

[46] U. r. o. Tanzania, "Small and Medium Enterprises Dvelopment Policy," M. o. i. a. Trade, Ed., ed. Tanzania, 2003.

[47] F. Olawale, "The financial literacy of Micro Enterpreneurs in South Africa," Journal of social sciences, vol. 42, p. 7, 2014.

[48] D. A. Adomako S, "Financial Literacy and firm Performance: the Moderating role of financial Capital availability and Resource Flexibility," International Journal of Management and Organizational studies, vol. 3, 2014.

[49] H. Latan and N. A. Ramli, "The Results of Partial Least Squares-Structural Equation Modelling Analyses (PLS-SEM)," 2013.

[50] J. F. Hair, M. Sarstedt, C. M. Ringle, and J. A. Mena, "An assessment of the use of partial least squares structural equation modeling in marketing research," Journal of the academy of marketing science, vol. 40, pp. 414-433, 2012.

[51] C. Fornell and F. L. Bookstein, "Two structural equation models: LISREL and PLS applied to consumer exit-voice theory," Journal of Marketing research, pp. 440-452, 1982.

[52] R. P. Bagozzi, Y. Yi, and K. D. Nassen, "Representation of measurement error in marketing variables: Review of approaches and extension to three-facet designs," Journal of Econometrics, vol. 89, pp. 393-421, 1998.

[53] J. Hulland, "Use of partial least squares (PLS) in strategic management research: A review of four recent studies," Strategic management journal, pp. 195-204, 1999.

[54] W. Kwong, "Partial Least Square Structural Equation Modelling (PLS-SEM) Techniques Using SmartPLS," Marketing Bulletin, vol. 24, 2013.

[55] C. Fornell and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error," Journal of marketing research, pp. 39-50, 1981.

[56] W. W. Chin, "The partial least squares approach to structural equation modeling," Modern methods for business research, vol. 295, pp. 295-336, 1998.

[57] D. Garson, Partial Least Square: Regression and Structural Equation Models. North Carolina State University: Statistical Associates 2016.

[58] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, "Common method biases in behavioral research: a critical review of the literature and recommended remedies," Journal of applied psychology, vol. 88, p. 879, 2003.

[59] H. Liang, N. Saraf, Q. Hu, and Y. Xue, "Assimilation of enterprise systems: the effect of institutional pressures and the mediating role of top management," MIS quarterly, pp. 59-87, 2007.

[60] N. Kock, "Common method bias in PLS-SEM: A full collinearity assessment approach," International Journal of e-Collaboration (IJeC), vol. 11, pp. 1-10, 2015.

[61] F. Olawale and D. Garwe, "Obstacles to the growth of new SMEs in South Africa: A principal component analysis approach," African journal of Business management, vol. 4, p. 729, 2010.

[62] J. Abor and N. Biekpe, "Small business financing initiatives in Ghana," Problems and Perspectives in Management, vol. 4, pp. 69-77, 2006.

[63] O. O. Fatoki, "The impact of human, social and financial capital on the performance of small and medium-sized enterprises (SMEs) in South Africa," Small, vol. 50, p. R4m, 2011.

[64] A. W. Butler and J. Cornaggia, "Does access to external finance improve productivity? Evidence from a natural experiment," Journal of Financial Economics, vol. 99, pp. 184-203, 2011.

[65] M. M. Rahaman, "Access to financing and firm growth," Journal of Banking & Finance, vol. 35, pp. 709-723, 2011.

[66] T. Beck and A. Demirguc-Kunt, "Small and medium-size enterprises: Access to finance as a growth constraint," Journal of Banking & finance, vol. 30, pp. 2931-2943, 2006.

[67] R. Atieno, Linkages, access to finance and the performance of small-scale enterprises in Kenya: Research paper/UNU-WIDER, 2009.

# Method of Graph Mining based on the Topological Anomaly Matrix and its Application for Discovering the Structural Peculiarities of Complex Networks

Artem Potebnia
Independent Investigator
Korosten, Ukraine
ORCID: 0000-0002-8162-5613

***Abstract*—The article introduces the mathematical concept of the topological anomaly matrix providing the foundation for the qualitative assessment of the topological organization underlying the large-scale complex networks. The basic idea of the proposed concept consists in translating the distributions of the individual vertex-level characteristics (such as the degree, closeness, and betweenness centrality) into the integrative properties of the overall graph. The article analyzes the lower bounds imposed on the items of the topological anomaly matrix and obtains the new fundamental results enriching the graph theory. With a view to improving the interpretability of these results, the article introduces and proves the theorem regarding the smoothness of the closeness centrality distribution over the graph's vertices. By performing the series of experiments, the article illustrates the application of the proposed matrix for evaluating the topology of the real-world power grid network and its post-attack damage.**

***Keywords*—Topological anomaly matrix; complex network; graph topology; closeness centrality; betweenness centrality; power grid**

## I. Introduction

The distinctive feature characterizing the upcoming fourth wave of the industrial revolution lies in the rapid expansion, complication, and integration of the complex networks serving the needs of humanity and world economy [1, 2]. While enabling the development of the more efficient business processes leading to the increase in the produced outcome and quality of service, such tendency makes the entire society extremely vulnerable to the disruptions of the most critical infrastructural networks [3, 4]. Meanwhile, the functionality and reliability of any complex network heavily relies on its topology inspiring the emergent properties that could not be deduced from the separate network's entities and arise only in result of their interaction [5, 6]. For example, the United States of America has suffered from several catastrophic blackouts caused by the cascading failures in the power grid steaming largely from the low redundancy of its topological design [7 – 9]. These observations contribute to the particular reasonableness of assessing the topology of complex networks while making decisions regarding their reliability or need for the additional protection. Remark that this article focuses on considering the complex networks modeled by the undirected simple graphs $G=(V,E)$. In turn, the topology of any graph $G$ could be regarded as the class of all possible graphs that are isomorphic to $G$. The evaluation of such topology is extremely challenging due to its underlying combinatorial nature and serves as a core problem of the emerging Big Data graph mining and analytics [10, 11]. In the prior works, the graph topology is assessed based on applying the quantitative metrics summarized in the review [12]. However, these metrics give a limited insight into the qualitative topological properties such as the concentration of bottlenecks inspiring the non-uniform load on the entities and links of the modeled network, which points to the presence of the research gap. Thereby, ***the objective of this article*** lies in constructing the mathematical object of the topological anomaly matrix providing the qualitative evaluation of the graph topology and its richness in bottlenecks, while satisfying the computational efficiency demands imposed to the instruments of the Big Data analytics.

## II. Related Work: Vertex Importance Metrics

The inhomogeneous topology of graph gives rise to the differentiation in the relative importance of its nodes for ensuring the normal activity of the modeled complex network. However, the vertex importance is difficult for analyzing due to the possibility of its consideration from the radically different conceptual viewpoints. Thereby, in the existing works, the comprehensiveness of assessing the importance of the graph's nodes is ensured through applying a family of the formalized centrality metrics. In particular, the degree $d(v)$ of the vertex $v$ reflects the extent of its local importance and serves as the simplest centrality metric. Nevertheless, the value of degree is incapable of capturing the position of the examined vertex within the entire graph. At the same time, the metrics of the closeness and betweenness centrality [13, 14] provide the formal way for evaluating the global importance of the graph's nodes and are defined in the following way:

**Definition 1.** The *closeness centrality* $c(v)$ of the node $v$ belonging to the vertex set $V$ of the connected graph $G$ represents the inverted value of its average geodesic distance $d(v,\ k)$ to all nodes $k \in V \setminus \{v\}$, i.e.

$$c(v)=\frac{|V|-1}{\sum\limits_{k\in V\setminus\{v\}} d(v,\ k)}.$$

**Definition 2.** The *betweenness centrality* $b(v)$ reflects the likelihood that the examined vertex $v$ appears on the shortest path between a pair of other nodes and is calculated as follows:

$$b(v) = \sum_{\substack{k,l \in V \setminus \{v\} \\ k \neq l}} \frac{\sigma_{kl}(v)}{\sigma_{kl}}.$$

Here $\sigma_{kl}$ denotes the total number of the shortest paths between the vertices $k$ and $l$ that differ in at least one edge, while $\sigma_{kl}(v)$ stands for the number of such paths transiting the vertex $v$.

Intuitively, the closeness centrality could be interpreted as the velocity of the information broadcasting from the examined vertex to all other nodes of the graph. For example, by starting to spread from the nodes with the highest closeness centrality, the computer worms could potentially reduce the time required for infecting all vertices. For its part, the betweenness centrality could be viewed as the extent to which the examined vertex is involved as an intermediate in the communication flows between the other graph's nodes. Moreover, the vertices that ensure gluing together multiple implicit communities take the crucial responsibility for the exchange of information between them and, thereby, are typically characterized by the high betweenness centrality (especially in the case of the strong community structure) [15, 16].

## III. Proposed Concept of the Topological Anomaly Matrix and its Fundamental Properties

The main contribution of this article lies in introducing the following mathematical object embodying the strategy of translating the local vertex-level characteristics into the property of the overall graph $G$:

**Definition 3.** The *topological anomaly matrix* $\mathbf{A}_{\mathbf{\Omega}}(G)$ of the graph $G$ with respect to the *base vector* $\mathbf{\Omega} = [\omega_1 \ \ldots \ \omega_n]$ containing $n$ vertex importance metrics $\omega_i : V \to \mathbf{R}$ is given in the form of the following $n \times n$ array:

$$\mathbf{A}_{\mathbf{\Omega}}(G) = \begin{bmatrix} a_{\omega_1}^{\omega_1}(G) & a_{\omega_2}^{\omega_1}(G) & \ldots & a_{\omega_n}^{\omega_1}(G) \\ a_{\omega_1}^{\omega_2}(G) & a_{\omega_2}^{\omega_2}(G) & \ldots & a_{\omega_n}^{\omega_2}(G) \\ \ldots & \ldots & \ldots & \ldots \\ a_{\omega_1}^{\omega_n}(G) & a_{\omega_2}^{\omega_n}(G) & \ldots & a_{\omega_n}^{\omega_n}(G) \end{bmatrix}.$$

Here the value of each item $a_{\omega_k}^{\omega_i}(G)$ lies within the range $[-1,1]$ and represents the bivariate correlation coefficient over all pairs of the set $DS(G, \omega_i, \omega_k) = \{(\omega_i(v), \omega_k(v)) | v \in V\}$. Note that $a_{\omega_k}^{\omega_i}(G)$ is taken to be undefined if either $\omega_i$ or $\omega_k$ is constant on the entire vertex set $V$ (i.e. if there exists such $x \in \mathbf{R}$ that $\omega_i(V) \to \{x\}$ or $\omega_k(V) \to \{x\}$).

By definition, the matrix $\mathbf{A}_{\mathbf{\Omega}}(G)$ is symmetric, while its undefined components should be organized into the rows and columns crossing at the diagonal entries $a_{\omega_i}^{\omega_i}$ and, thereby, indicating the incapability of the corresponding metrics $\omega_i$ to distinguish the vertices of $G$. In turn, all defined components comprising the main diagonal of $\mathbf{A}_{\mathbf{\Omega}}(G)$ should be equal to one. For convenience, the matrices $\mathbf{A}_{\mathbf{\Omega}}(G)$ deprived of the undefined entries are referred to as *perfect* through this article.

The selection of metrics into the base vector $\mathbf{\Omega}$ is driven by essential need for ensuring the descriptiveness of the constructed matrix $\mathbf{A}_{\mathbf{\Omega}}(G)$ in assessing the topology of $G$ at the optimal utilization of resources involved in the process of its calculation. In particular, the conceptual interpretability and computational efficiency of the metrics discussed in the previous section points to the reasonableness of introducing the *canonical base vector* defined as $\tilde{\mathbf{\Omega}} = [d \quad c \quad b]$. At the same time, the *canonical matrix* $\mathbf{A}_{\tilde{\mathbf{\Omega}}}(G)$ relying on such vector has the size of $3 \times 3$, while its full specification requires values of only three items $a_c^d(G)$, $a_b^d(G)$, and $a_b^c(G)$.

Remark that the matrices $\mathbf{A}_{\tilde{\mathbf{\Omega}}}(R)$ characterizing the purely random (and thereby unstructured) connected graphs $R$ following the binomial distribution of vertex degrees tend to have the close-to-one values of all non-diagonal components. This tendency steams from the fact that, simply by chance, the higher-degree vertices demonstrate a larger probability of being located at the lower average distance to all other nodes and are likely to participate in the larger fraction of the shortest paths between them. In view of these considerations, every low (i.e. close-to-zero or negative) entry of the matrix $\mathbf{A}_{\tilde{\mathbf{\Omega}}}(G)$ clearly points to the significant non-randomness of the graph $G$ and reveals the presence of the unexpected anomaly in its topology. In total, the matrix $\mathbf{A}_{\tilde{\mathbf{\Omega}}}(G)$ could encapsulate three major anomalies originating from the manner of fragmenting the graph $G$ into the cohesive implicit communities.

In particular, the low value of $a_c^d(G)$ indicates that the larger number of the direct neighbors attached to an arbitrary vertex of $G$ does not shrink its farness from the rest nodes of the graph to the statistically significant extent. The main topological property responsible for producing such anomaly consists in differentiating the entire communities of $G$ into the central and peripheral ones (depending on the average distance to the other communities in terms of the inter-community edges). In this context, the high-degree vertices involved in the peripheral communities as well as the low-degree nodes occurring in the central ones serve as the key factors contributing to the reduce in the value of $a_c^d(G)$.

Conversely, the topological anomaly evidenced by the low value of $a_b^d(G)$ implies that the higher-degree vertices do not act as the significantly more preferred intermediates in the

shortest paths of the graph $G$. The topological pattern provoking such effect is characterized by the incidence of many critical inter-community edges to the low-degree nodes along with the presence of the high-degree vertices adjacent exclusively to the members of their own communities. Finally, at the low value of $a_b^c(G)$, the ability of an arbitrary vertex to be involved into the shortest paths in the graph $G$ (and control the corresponding communication flows) is not strongly dependent on its average distance to the other vertices. From the topological viewpoint, the anomalous decrease in $a_b^c(G)$ is driven by the nodes that, while being located in the central communities, are neither directly incident to the inter-community edges nor lie on the shortest path between any pair of vertices equipped with such edges.

In order to provide a fruitful insight into the entries of $\mathbf{A}_{\tilde{\mathbf{\Omega}}}(G)$, let us introduce and prove the following fundamental relationship between the closeness centrality values of the adjacent graph's nodes:

**Theorem 1.** The closeness centrality $c(v)$ of any vertex $v$ in the connected graph $G=(V,E)$ is bounded below by

$$c(v)\ge\frac{|V|-1}{\frac{|V|-1}{c_m(v)}+|V|-2},$$

where $c_m(v)=\max\{c(u)|(v,u)\in E\}$ stands for the highest closeness centrality among all direct neighbors of $v$.

▲ Let us assume that $v$ is adjacent to the node $u$ having the closeness centrality of $c(u)$. This, for its part, implies that every vertex $h\in V\setminus\{v,u\}$ could be reached from $v$ based on the walk (i.e. sequence of edges with allowed repetitions) composed of the edge $(v,u)$ and shortest path from $u$ to $h$. Accordingly, the geodesic distance between $v$ and $h$ is bounded above by the condition $d(v,h)\le d(u,h)+1$, while $d(v,u)=1$. In view of this observation, the entire closeness centrality of $v$ is constrained in the next manner:

$$c(v)\ge\frac{|V|-1}{|V|-2+\theta(v,u)};\quad \theta(v,u)=1+\sum_{h\in V\setminus\{v,u\}}d(u,h).$$

In turn, $\theta(v,u)$ could be expressed based on the closeness centrality of $u$ as $\theta(v,u)=(|V|-1)/c(u)$, which completes deriving the desired relationship. At the same time, the increase in $c(u)$ over the whole allowed range $(0,1]$ leads to the monotonic growth of the imposed bound at any fixed $|V|\ge 3$. This remark clearly points to the largest restrictiveness of the bound produced by the neighbor with the highest closeness centrality. ▼

The most significant implication of the above theorem lies in the smooth nature of distributing the closeness centrality values over the graph's vertices. On the contrary, the values of the betweenness centrality could be distributed in much more rugged manner implying the extreme differences between the adjacent nodes. For example, each leaf vertex $l$, by definition, is associated with zero betweenness centrality $b(l)=0$ regardless the properties of its single neighbor. Conversely, the closeness centrality of $l$ takes the lowest possible value satisfying the bound given in Theorem 1. Remark that such bound demonstrates the close-to-linear behavior at the low values of $c_m(v)$ (since its derivative with respect to $c_m(v)$ approaches one as $c_m(v)\to 0$). This observation clearly shows that the leaf nodes of the sparse large-scale graph $G$ typically tend to have almost the same closeness centrality as their neighbors. In view of such relationship, the leaf vertices appearing in the central communities are characterized by the relatively high closeness centrality compared to the other graph's nodes and, thereby, serve as the most evident contributors to the reduce in the value of $a_b^c(G)$.

## IV. Analysis of the Lower Bounds Imposed on the Entries of the Canonical Topological Anomaly Matrix

Meanwhile, the anomalous effects indicated by the matrix $\mathbf{A}_{\tilde{\mathbf{\Omega}}}(G)$ are not inspired solely by the intentional self-organizing process of the complex network modeled by the graph $G$. Additionally, the values of $a_c^d(G)$, $a_b^d(G)$, and $a_b^c(G)$ are affected by the structural constraint taking the form of the vertex degree multiset $D(G)=\{d(v)|v\in V\}$ containing the degrees of all nodes in $G$. Each multiset $D(G)$, for its part, characterizes the family $\Gamma_{D(G)}$ composed of all non-isomorphic graphs $G'\in\Gamma_{D(G)}$ such that $D(G')=D(G)$. In this sense, the specification of $D(G)$ restricts the possible topologies of $G$ only to ones contained in $\Gamma_{D(G)}$ and imposes the structural bounds on the components of $\mathbf{A}_{\tilde{\mathbf{\Omega}}}(G)$.

Furthermore, such quantitative characteristics of $G$ as the order $|V|$ and density $\varphi(G)=2|E|/(|V|(|V|-1))$ are derived from $D(G)$ and by themselves provide the lower bounds on the items of $\mathbf{A}_{\tilde{\mathbf{\Omega}}}(G)$. For convenience, let us denote the minimum values of $a_c^d(G)$, $a_b^d(G)$, and $a_b^c(G)$ over all graphs $G$ containing $|V|$ nodes and having the density of $\varphi(G)$ respectively by $m_c^d(|V|,\varphi(G))$, $m_b^d(|V|,\varphi(G))$, and $m_b^c(|V|,\varphi(G))$. Notice that all these lower bounds are defined over the domain restricted by $\varphi_{tree}(|V|)\le\varphi(G)<1$, where

$\varphi_{tree}(|V|) = 2(|V|-1)/(|V|(|V|-1))$. Such restriction steams from the impossibility of constructing any connected graph sparser than a tree along with the presence of only undefined items in the matrix $\mathbf{A}_{\tilde{\Omega}}(K)$ of each complete graph $K$ having all possible edges.

With a view to simplifying the discussion of the results given in Fig. 1, let us use the notations $m_c^d(\varphi(G))|_k$, $m_b^d(\varphi(G))|_k$, and $m_b^c(\varphi(G))|_k$ for the dependences of $m_c^d(|V|,\varphi(G))$, $m_b^d(|V|,\varphi(G))$, and $m_c^b(|V|,\varphi(G))$ on $\varphi(G)$ at the value of $|V|$ fixed to $k$ (representing slices of the illustrated surfaces). As evident from Fig. 1a, the dependence $m_c^d(\varphi(G))|_k$ for any considered $k$ exhibits a single minimum located close the lowest allowed density $\varphi_{tree}(k)$. Moreover, such minimum becomes deeper with the increase in $k$, which is directly attributed to the growing number of possible topologies. Another notable feature of the analyzed surface consists in the presence of the wide plateau-like region where $m_c^d(|V|,\varphi(G))$ takes the close-to-one values. While being located at the high density $\varphi(G)$, this region complies with the limited suitability of the dense graphs to the elaboration the high-modular topology underlying the emergence of the structural anomalies. Conversely, the tree graphs could be strongly segregated into the sparse implicit communities, which acts as an explanation for the relatively low values of $m_c^d(|V|,\varphi_{tree}(|V|))$. However, the requirement regarding the sparsity of communities also hinders the formation of the structural anomalies. Accordingly, the minima of all considered dependences $m_c^d(\varphi(G))|_k$ are slightly deviated from $\varphi_{tree}(k)$. For example, Fig. 1a depicts the graph $G_7$ responsible for producing the minimum of $m_c^d(\varphi(G))|_7$. Remark that this graph implies the inclusion of the three-degree vertices into the peripheral communities (represented by cycles) and placement of the two-degree node as the connector between these communities. As a result, such connector is associated with the largest closeness centrality compared to all other vertices. The graphs on six or fewer nodes, in turn, could not contain the lower-degree vertex characterized by the larger closeness centrality than the higher-degree one due to the influence of the structural restrictions.

At the same time, the shape of the surfaces constructed in Figs. 1b and 1c requires the more careful investigation. The distinctive feature expressed by the experimentally registered dependences $m_b^d(\varphi(G))|_k$ and $m_b^c(\varphi(G))|_k$ consists in the presence of multiple local minima whose number grows with the increase in $k$ (one at $k=4$ and $k=5$, two at $k=6$, and three at $k=7$). Remark that for every considered $k$, the local minima of both $m_b^d(\varphi(G))|_k$ and $m_b^c(\varphi(G))|_k$ are exhibited at the identical graph topologies and same values of $\varphi(G)$. Furthermore, the presented results allow noticing that the bounds $m_b^d(|V|,\varphi(G))$ and $m_c^b(|V|,\varphi(G))$ are lower than $m_c^d(|V|,\varphi(G))$ at the intermediate density $\varphi(G)$. These effects are fully attributable to the fact that the betweenness centrality is capable of producing the rugged distributions over the graph's nodes, while the closeness centrality is unavoidably subjected to the smoothing requirement proved in Theorem 1.

The inspection of the callouts in Fig. 1 shows that the topologies underlying the local minima of $m_b^d(\varphi(G))|_k$ and $m_b^c(\varphi(G))|_k$ are characterized by the presence of the densely interconnected group of the highest-degree vertices along with the inclusion of the low-degree nodes into the chain-like substructures. Moreover, the collected results allow discovering that the formation of such topologies is driven by the hidden fundamental rules. In particular, each graph labeled in Fig. 1 as $G_1^k$ for $k \in \{5,6,7\}$ could be obtained based on constructing the diamond graph (i.e. complete graph on four vertices with one removed edge) with the subsequent linking of its two-degree nodes by the path containing $k-3$ edges. Each graph labeled as $G_2^k$ for $k \in \{6,7\}$, in turn, contains such basic substructures as the three-length cycle $C_3$ and star $S_{k-3}$ represented by a tree with $k-4$ leaf vertices. Its formation involves placing all possible edges between the nodes of $C_3$ and $S_{k-3} \setminus \{r\}$, where $r$ denotes the central node of the star $S_{k-3}$. These trends suggest that the additional local minima arising in the dependences $m_b^d(\varphi(G))|_k$ and $m_b^c(\varphi(G))|_k$ with the increase in $k$ are caused by the graph topologies following the new fundamental rules.

## V. Application of the Proposed Matrix for Assessing the Topology of the Power Grid Network and its Post-Attack Damage

The role of this section lies in demonstrating the descriptive potential of the introduced mathematical structure in evaluating the qualitative topological properties of the real-world complex networks. As a sample dataset for investigation, this work uses the benchmark model of the power grid infrastructure of the United States of America available at the open-access network collection [17] and given by the undirected graph $P = (V_P, E_P)$. Notice that this graph is connected and contains 4 941 vertices reflecting the facilities responsible for producing and distributing electricity along with 6 594 edges modeling the high-voltage transmission lines. The canonical matrix $\mathbf{A}_{\tilde{\Omega}}(P)$ calculated for the described graph $P$ is given by $a_c^d(P) = 0.2306$, $a_b^d(P) = 0.2766$, and $a_b^c(P) = 0.3536$. These values indicate the involvement of all considered structural anomalies in the topological organization of $P$, which serves as the natural result for the spatially distributed technological man-made system needing the constant supervision for preserving the desired functionality.

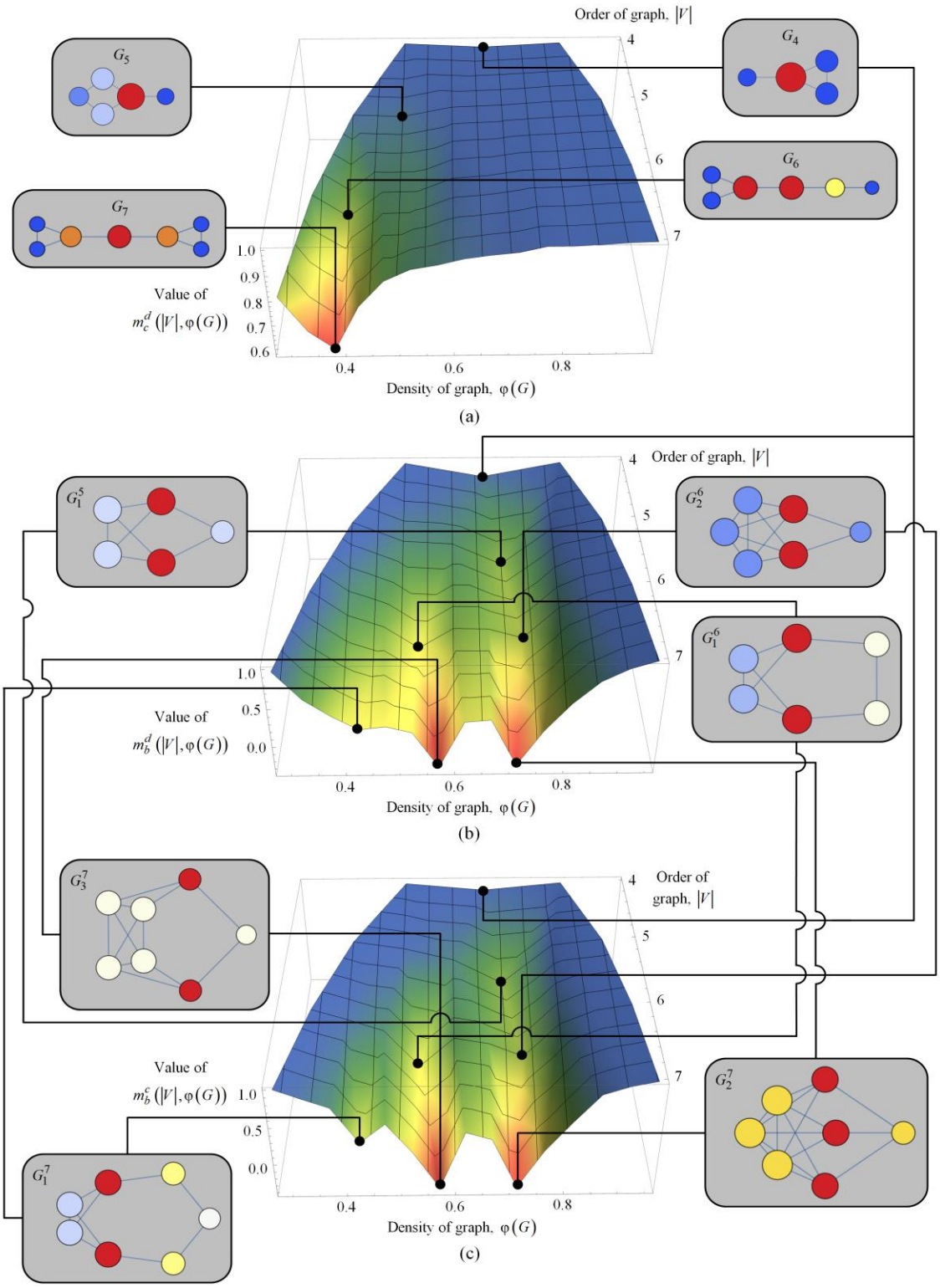



Fig. 1. Three-dimensional plots showing the lower bounds $m_c^d(|V|,\varphi(G))$, $m_b^d(|V|,\varphi(G))$, and $m_b^c(|V|,\varphi(G))$ on entries of the matrix $\mathbf{A}_{\tilde{\Omega}}(G)$ as functions of the graph order $|V|$ and density $\varphi(G)$. All given continuous surfaces were constructed by processing the scattered points representing the experimentally calculated data with the bilinear interpolation scheme. In addition, the surfaces are equipped with the callouts depicting the graph topologies underlying the minima of the considered bounds each fixed $|V|$. The size size and color of vertices in every callout graph encode respectively their closeness and betweenness centrality (the largest size and red color correspond to the highest values of metrics).

The more in-depth analysis shows that the sets of pairs $DS(P,d,c)$ and $DS(P,d,b)$ underlying the calculation of the correlation coefficients $a_c^d(P)$ and $a_b^d(P)$ are organized in line with inverted cone-shaped pattern of heteroscedasticity implying the decrease in the variability of the closeness and betweenness centrality values of nodes with increasing their degree. Remark that the emergence of such phenomenon steams heavily from the vertex degree distribution of $P$ pointing to the presence of fewer nodes accommodating more neighbors. In turn, the set $DS(P,c,b)$ constituting the basis for computing the component $a_b^c(P)$ exhibits the direct cone-shaped form of heteroscedasticity characterized by the tendency of vertices with the higher closeness centrality to demonstrate the larger variability of the betweenness centrality values.

With a view to illustrate the usefulness of applying the proposed matrix as a measure of the topological damage, let us consider the attack on $P$ implying the removal of all its nodes having the degree of at least $t$, i.e. comprising the subset $\Psi(t)=\{v \mid d(v)\geq t\}$. The post-attack graph on the remaining nodes of $V_P^t = V_P \setminus \Psi(t)$ is represented by $P_t=(V_P^t,E_P^t)$, where $E_P^t = E_P \setminus \{(v,v') \mid (v\in\Psi(t))\vee(v'\in\Psi(t))\}$. In turn, let us use the notation $W_t=(V_W^t,E_W^t)$ for the largest connected component of $P_t$. At the high fraction $f(t)=|V_W^t|/|V_P^t|$, $W_t$ is additionally referred to as the giant component of $P_t$, while its topology accumulates the majority of damage that is not related to the connectivity issues [3].

Fig. 2 illustrates the application of the matrix $\mathbf{A}_{\tilde{\Omega}}(W_t)$ for assessing such damage by presenting the experimentally calculated values of its components $a_c^d(W_t)$, $a_b^d(W_t)$, and $a_b^c(W_t)$ as functions of the degree threshold $t$. Remark that with the decrease in $t$, all obtained dependences demonstrate the tendency to fall after the plateau-like region and reach their global minima at the same critical threshold $t_c=6$. Meanwhile, all post-attack graphs $P_t$ for $t<t_c$ are deprived of the giant connected component (as evidenced by the dependence of $f(t)$ on $t$), while their subgraphs $W_t$ are trees. In light of these observations, the sharp growth of $a_c^d(W_t)$, $a_b^d(W_t)$, and $a_b^c(W_t)$ at the subsequent reduce in $t$ is driven by the structural constraints studied in the previous section. For its part, the graph $W_{t_c}$ corresponding to the global minima of the traced dependences is characterized by the most significant anomalies reflecting the accumulation of the largest topological damage. Conceptually, with the decrease in $t$, such damage stimulates the collapse of $W_{t_c}$ into the numerous small connected components. Furthermore, the dependence of $a_c^d(W_t)$ on $t$ demonstrates the deepest global minimum $a_c^d(W_{t_c})=0.0784$. This result allows noting that the topological damage of $W_{t_c}$ is expressed primarily by the more significant differentiation of its communities into the central and peripheral ones, which follows from the degradation of the inter-community relationships.

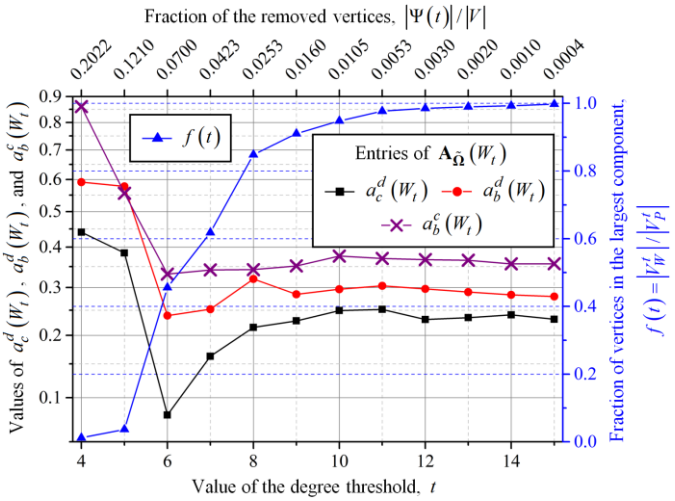



Fig. 2. Entries of the matrix $\mathbf{A}_{\tilde{\Omega}}(W_t)$ associated with the largest connected component $W_t$ of the post-attack graph $P_t$ obtained by deleting all vertices with the degree of at least $t$ from the graph $P$ modeling the power grid network. To trace the significance of the obtained results for the overall graph $P_t$, the plot additionally gives the fraction of its vertices included in $W_t$.

## VI. Conclusions

The findings presented in the preceding sections clearly substantiate the crucial role of the topological anomaly matrix $\mathbf{A}_{\Omega}(G)$ in discovering the unexpected topological patterns of the real-world complex networks and producing the new fundamental results advancing the frontiers of the graph theory. The canonical form of the proposed matrix $\mathbf{A}_{\tilde{\Omega}}(G)$ is recommended for the widespread usage, while the need for performing the more in-depth analysis could be addressed by applying the matrices of larger size relying on the extended base vectors containing additional metrics. Conceptually, at the low values of $a_c^d(G)$ and $a_b^d(G)$, the network modeled by the graph $G$ is characterized by the tendency of the entities accommodating only a few neighbors to act as hubs managing the significant portions of traffic. In turn, the links attached to such entities are subjected to the enhanced risks of overloading and, thereby, play a role of the primary structural bottlenecks. Meanwhile, the low value of $a_b^c(G)$ is caused by the entities that, while being located close to all other nodes, do not use their beneficial geodesic position to support the traffic transmission in the network and, in this sense, contribute to the formation of bottlenecks. In sum, the opportunity of ensuring the balance between the descriptive potential and computational complexity of the matrix $\mathbf{A}_{\Omega}(G)$ (through the

selection of metrics into the base vector **Ω** ) allows its consideration as the promising tool in the Big Data graph mining and analytics. Due to its usefulness in describing the topological damage (as illustrated in the previous section), the topological anomaly matrix could be potentially applied as one of the robustness metrics in assessing the attack tolerance of complex networks. Similarly, the proposed matrix could assist in detecting the differences in the topological organization between the whole network and its important subnetworks (such as the rich-clubs). Another possible application lies in tracing the evolutionary topological transformation of complex networks (by comparing the matrices calculated for the giant connected components of the graph models constructed for the series of the time-indexed network snapshots).

## REFERENCES

[1] M. Khan, X. Wu, X. Xu, and W. Dou, "Big Data challenges and opportunities in the hype of industry 4.0," in *Communications (ICC)*, 2017 IEEE International Conference on, pp. 1 – 6, 2017.

[2] Y. Lu, "Industry 4.0: A survey on technologies, applications and open research issues," *Journal of Industrial Information Integration*, vol. 6, pp. 1 – 10, 2017.

[3] A. Potebnia, "Innovative metrics for assessing the catastrophic collapse of the complex networks under the greedy attacks on their most important vertices and edges," in *Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET)*, 14th International Conference on, pp. 564 – 569, 2018.

[4] A. Potebnia, "Innovative concept of the strict line hypergraph as the basis for specifying the duality relation between the vertex separators and cuts," in *Advances in Intelligent Systems and Computing II. CSIT 2017. Advances in Intelligent Systems and Computing*, vol. 689, Springer International Publishing, pp. 386 – 403, 2018.

[5] Y. Huang, G. Wang, and Y. Tang, "Bottleneck attack strategies on complex communication networks," in *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence. ICIC 2010. Lecture Notes in Computer Science*, vol. 6216. Springer-Verlag Berlin Heidelberg, pp. 418 – 425, 2010.

[6] A. Potebnia, "New method for estimating the tree-likeness of graphs and its application for tracing the robustness of complex networks," in *Computing, Communication and Networking Technologies (ICCCNT)*, 9th International Conference on, 2018.

[7] Y.-K. Wu, S.M. Chang, and Y.-L. Hu, "Literature review of power system blackouts," *Energy Procedia*, vol. 141, pp. 428 – 431, 2017.

[8] S. Soltan, D. Mazauric, and G. Zussman, "Analysis of failures in power grids," *IEEE Transactions on Control of Network Systems*, vol. 4(2), pp. 288 – 300, 2017.

[9] L. Liu, Y. Yin, Z. Zhang, and Y.K. Malaiya, "Redundant design in interdependent networks," *PLoS ONE*, vol. 11(10):e0164777, 2016.

[10] J.A. Miller, L. Ramaswamy, K.J. Kochut, and A. Fard, "Research directions for Big Data graph analytics," in *Big Data (BigData Congress)*, 2015 IEEE International Congress on, pp. 785 – 794, 2015.

[11] M.U. Nisar, A. Fard, and J.A. Miller, "Techniques for graph analytics on Big Data", in *Big Data (BigData Congress)*, 2013 IEEE International Congress on, pp. 255 – 262, 2013.

[12] B. Kantarci and V. Labatut, "Classification of complex networks based on topological properties," in *Cloud and Green Computing (CGC)*, 2013 Third International Conference on, pp. 297 – 304, 2013.

[13] K. Das, S. Samanta, and M. Pal, "Study on centrality measures in social networks: a survey," *Social Network Analysis and Mining*, vol. 8:13, 2018.

[14] N. Matas, S. Martincic-Ipsic, and A. Mestrovic, "Comparing network centrality measures as tools for identifying key concepts in complex networks: A case of Wikipedia," *Journal of Digital Information Management*, vol. 15(4), pp. 203 – 213, 2017.

[15] S. Fortunato and C. Castellano, "Community structure in graphs," in *Computational Complexity*. Springer New York, pp. 490 – 512, 2012.

[16] G. Lin, Z. Di, and Y. Fan, "Cascading failures in complex networks with community structure," *International Journal of Modern Physics C*, vol. 25(5): 1440005, 2014.

[17] US power grid network dataset. KONECT: The Koblenz Network Collection. Available online: http://konect.uni-koblenz.de/networks /opsahl-powergrid. Accessed: April 2018.

# Comparison of Intelligent Methods of SOC Estimation for Battery of Photovoltaic System

Tae-Hyun Cho[1], Hye-Rin Hwang[2], Jong-Hyun Lee[3], In-Soo Lee[4, *]
School of Electronics Engineering
Kyungpook National University
Daegu, Korea

***Abstract*—It is essential to estimate the state of charge (SOC) of lead-acid batteries to improve the stability and reliability of photovoltaic systems. In this paper, we propose SOC estimation methods for a lead-acid battery using a feed-forward neural network (FFNN) and a recurrent neural network (RNN) with a gradient descent (GD), a levenberg–marquardt (LM), and a scaled conjugate gradient (SCG). Additionally, an adaptive neuro-fuzzy inference system (ANFIS) with a hybrid method was proposed. The voltage and current are used as input data of neural networks to estimate the battery SOC. Experimental results show that the RNN with LM has the best performance for the mean squared error, but the ANFIS has the highest convergence speed.**

***Keywords*—Lead-acid battery; SOC; FFNN; RNN; ANFIS; gradient descent; levenberg-marquardt; scaled conjugate gradient**

## I. Introduction

Today, several environmental issues exist, including the depletion of fossil fuels and the dangers of nuclear power generation. For these reasons, the application of renewable energy has increased, and research on solar power has been actively conducted. Photovoltaic (PV) systems can be categorized into grid-connected systems and stand-alone power systems, depending on whether the system is connected to an electrical power-generation system. In particular, in the stand-alone power system, the demand for solar streetlights is urgently increasing, including low-power light-emitting diode (LED) lamps that are used to replace conventional halogen security lighting [1].

The solar streetlight system consists of solar-panel modules that convert solar energy into electricity, a secondary battery that stores the developed power, and a stand-alone (off-grid) system. The stand-alone system has the following advantages: 1) it does not require electric power installation, because the commercial power is not connected, and 2) it has a $CO_2$-reduction effect because it is operated by PV power generation.

The operation time of the solar streetlight only depends on a secondary battery. Days when sunshine time is less than 0.1 hour are considered sunless days, and solar streetlights should be guaranteed to operate for more than three sunless days.

It is essential to estimate the state of charge (SOC) of lead-acid batteries in real time for the following reasons. First, we must avoid reliance on the initial SOC of the battery. This allows more effective control of the power consumption of LED lamps. Second, the lead-acid battery is the most commonly used energy-storage device for PV systems. According to some researchers, lead-acid batteries will survive in the future.

A battery is a device that generates electrical energy through a chemical reaction, and it has nonlinear characteristics in response to parameters such as the ambient temperature, internal resistance, and capacitance. For these reasons, it is very difficult to estimate the SOC of a battery correctly.

There are many methods for estimating battery SOC [2]. The internal-impedance method estimates the SOC by measuring the internal-impedance change according to the charging and discharging of the battery. However, it is difficult to apply in a state where the cell is reacting, because it is very sensitive. The kalman-filter method is difficult to apply because of the complexity of the parameters and algorithms. The current-integration (CI) method [3] involves subtracting the initial SOC value by integrating the actual charge and discharge current. However, this method cannot estimate the initial SOC, and because of the accumulated errors of the leakage current and current sensing over time, accurate SOC estimation is impossible. The open-circuit voltage (OCV) [4] method involves measuring the voltage in the no-load state. However, this method is difficult to apply to real-time systems because it uses the measured OCV at the chemical equilibrium inside the battery.

Neural networks have proven to be a promising paradigm for intelligent systems. They have been trained to perform complex functions in various fields, such as pattern recognition, identification, and classification [5]. Their ability to learn complex nonlinear input/output relationships, their use of sequential training procedures, and their adaptability to data are three outstanding characteristics of neural networks. Some popular modules of neural networks have shown abilities of associative memory and learning [6-8]. To allow the network to perform a specific classification and clustering task efficiently, the learning process comprises updating the network architecture and modifying the weights between the neurons. The neural network can efficiently model a variety of input and output relationships. Compared with procedural models, it has the advantage of a shorter execution time [9, 10].

In this paper, SOC estimation methods for a lead-acid battery using a feed-forward neural network (FFNN), a

*Corresponding Author.

recurrent neural network (RNN), and an adaptive neuro-fuzzy inference system (ANFIS) were proposed. The FFNN and RNN models were applied with three different training methods: gradient descent (GD), scaled conjugate gradient (SCG), and levenberg–marquardt (LM), under the same network configuration. The ANFIS model was applied with a hybrid training method. Also, we compared the performance of the different networks.

In Section II, research on battery SOC estimation is introduced. In Section III, neural network based battery SOC estimation methods were proposed. In Section IV, experimental results for the FFNN, RNN, and ANFIS models were proposed. Finally, the paper is concluded in Section V.

## II. Study on Battery SOC Estimation

### A. Open Circuit Voltage Method

The open circuit voltage (OCV) method is widely known as a battery SOC estimation method. The OCV method is very accurate in estimating the SOC when the lead-acid battery reaches the stabilization phase. In addition, data must be constructed through SOC-OCV relationship experiments. Although estimating the SOC from a battery is one of the most effective methods for estimating the SOC using OCV, this method requires a condition in which the circuit is opened or no current flows, and it takes time to wait for the battery to stabilize internally. For this reason, this method is difficult to use in real-time estimation.

### B. Current Integration Mehtod

The current integration (CI) method is a real-time battery SOC estimation method based on CI. The SOC at time $t$ is shown in (1), and the basic principle of the CI method is to add all the charges flowing into and out of the battery in terms of ampere-hours [11]. The CI method requires determining the initial SOC and performing accurate current measurement. Although the initial SOC can be estimated via the CI method using the OCV-SOC data sheet, the OCV-SOC data sheet becomes inaccurate as the battery ages. Thus, It is needed to renew the data sheet to perform accurate battery SOC estimation. In addition, there is the disadvantage of accumulated errors over time.

$$SOC(t) = SOC(t_0) + \int_{t0}^{t} \frac{I(t)dt}{C_N} \quad (1)$$

Where, $t_0$ is the initial time, $I(t)$ is the battery current, and $C_N$ is the nominal capacity of the battery.

## III. Proposed SOC Estimation Methods based on Neural Network

In this paper, we propose methods for estimating the lead-acid battery SOC using an FFNN, an RNN, and an ANFIS. To perform lead-acid battery SOC estimation through the neural networks, we selected the voltage and the current as input parameters for these models. The reason for using these two parameters is that they can be measured easily, and have important relationships with the battery SOC. A schematic diagram of the lead-acid battery SOC estimation methods based on the proposed neural networks is shown in Fig. 1. Also, SOC at time $t$ is defined as shown in (2) by us, to set the battery SOC range as 0% to 100% [12].

$$SOC(t) = SOC(t_0) + \int_{t0}^{t} \frac{I(t)dt}{C_M} \quad (2)$$

Where, $t_0$ is the initial time, $I(t)$ is the battery current, and $C_M$ is the measured capacity of the battery.

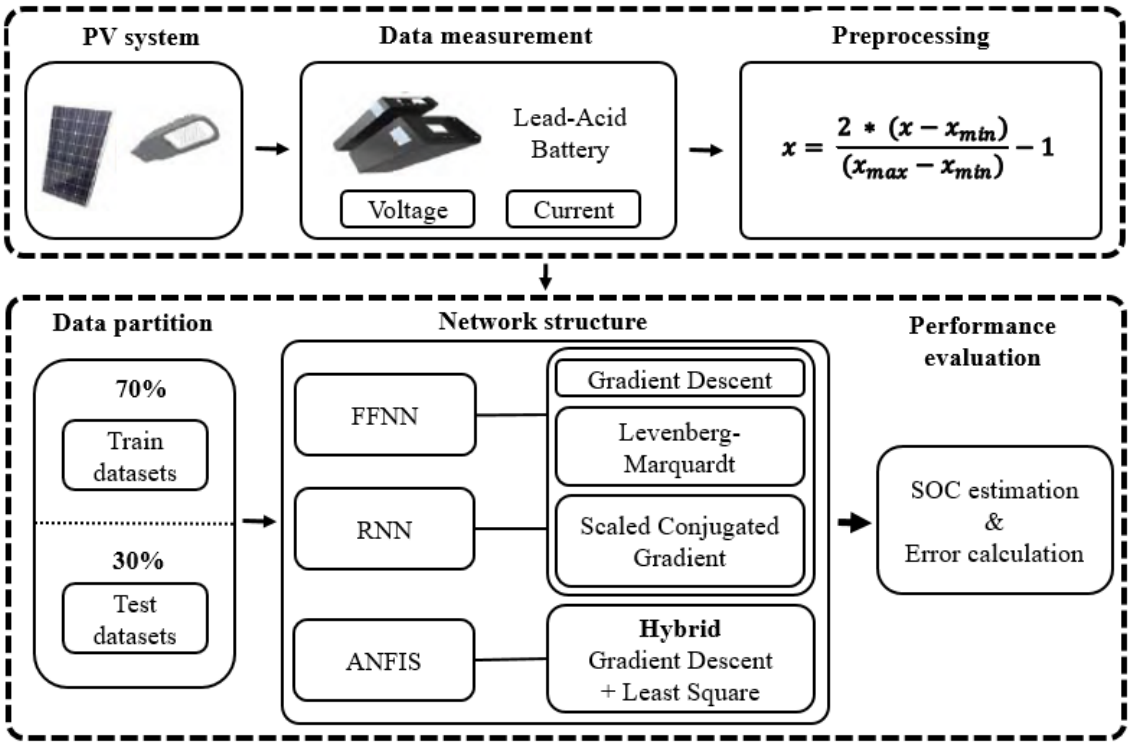



Fig. 1. Schematic Diagram of the Neural Network-based Lead-Acid Battery SOC Estimation.

### A. Feed-Forward Neural Network

The feed-forward neural network (FFNN) is a multilayer neural network with one or more hidden layers. The basic structure is composed of an input layer, a hidden layer, and an output layer, as shown in Fig. 2. Each layer is composed of nodes (neurons) and the weights connected between each neuron.

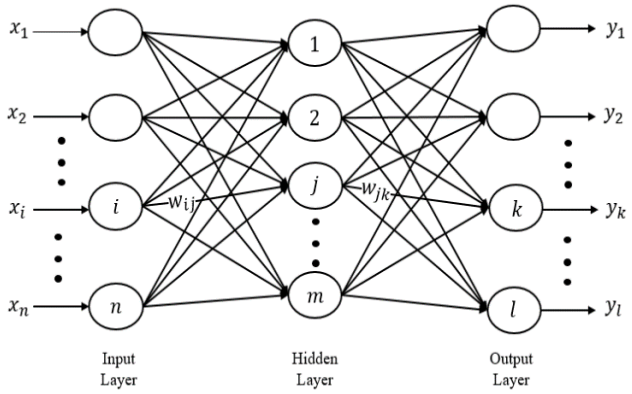



Fig. 2. Structure of the FFNN.

The input pattern is multiplied by the weights between the input layer and the hidden layer, and it passes through the activation function of the hidden layer. The output of node j in the hidden layer at epoch $p$ and output of node $k$ in the output layer at epoch $p$, as given by (3) and (4).

$$y_j(p) = f_j(\sum_{i=1}^{n} x_i(p) w_{ij}(p)) \quad (3)$$

Where, $n$ is the number of input samples, and $f_j$ is the activation function of the hidden node $j$.

$$y_k(p) = f_k(\sum_{j=1}^{m} y_j(p) w_{jk}(p))) \quad (4)$$

Where, $m$ is the number of hidden nodes, $f_k$ is the activation function of the output node $k$.

The error of the output pattern for the target pattern at output node k at epoch $p$ is defined by (5). The error is calculated and the weight is modified again through the backpropagation method. As this process repeats, it is corrected to the optimal weight to obtain the target output pattern using the input pattern. If this model is generalized well, we can obtain the target pattern for un-learned input patterns.

$$e_k(p) = d_k(p) - y_k(p) \quad (5)$$

Where, $d_k(p)$ is the target value, and $y_k(p)$ is the output value at output node $k$ at epoch $p$.

### B. Recurrent Neural Network

The recurrent neural network (RNN) is a representative model for processing sequence data. It differs from the generalization model of the existing FFNN structure, as shown Fig. 3. The most important feature of an RNN is that it has the state of the hidden layer at time $t$, as shown in (6).

$$h(t) = f_h(x(t) W_{ih} + h(t-1) W_{hh}) \quad (6)$$

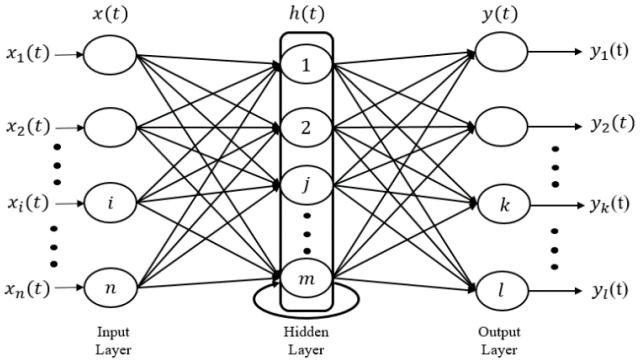



Fig. 3. Structure of the RNN.

Where, $h(t-1)$ represents the activation values of the hidden layer at time $t-1$, which are replicated at time $t$ and accumulate past information, $f_h$ is the activation function of the hidden layer, $x(t)$ is the input at time $t$, $W_{ih}$ is the weight matrix between $x(t)$ and $h(t)$, and $W_{hh}$ is the connection matrix between $h(t-1)$ and $h(t)$. The output at time $t$ as shown in (7).

$$y(t) = f_o(h(t) W_{ho}) \quad (7)$$

Where, $f_o$ is the activation function of the output layer, and $W_{ho}$ is the weight matrix between $h(t)$ and y$(t)$.

### C. Adaptive Neuro Fuzzy Inference System

In the adaptive nuero fuzzy inference system (ANFIS) model [13], the sugeno fuzzy model is a systematic method for generating fuzzy rules from the input data and the output data sets. It uses a hybrid learning method, which combines the least-squares estimator and the GD method. The hybrid learning method has the advantage of converging quickly [14].

As shown in Fig. 4, the ANFIS is divided into six layers: the input layer, fuzzification layer, rule layer, normalization layer, defuzzification layer, and summation neuron layer. Term $x_n$ represents the input values, $A_n$ and $B_n$ are fuzzy sets, $f_n$ is the output function, and $y$ is the output at epoch $p$, as shown in (8).

$$y(p) = k_0(p) + k_1(p) x_1(p) + \cdots + k_m(p) x_m(p) \quad (8)$$

Where, $k_{i0}$, $k_{i1}$, and $k_{i2}$ are the sets of arguments for rule $i$.

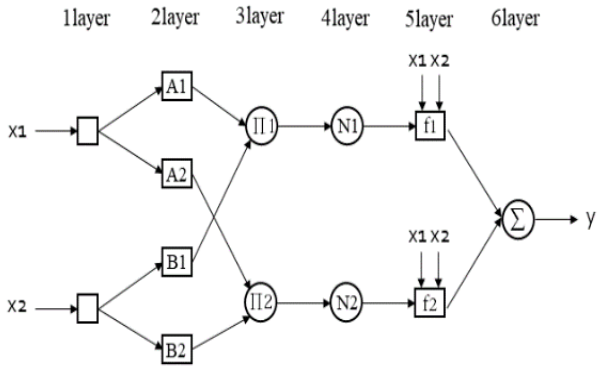



Fig. 4. Structure of the ANFIS.

In the ANFIS learning algorithm, each epoch is composed of a forward calculation and a backward calculation. In the forward calculation, the training set of the input pattern is input to the ANFIS, the neuron output is calculated for each layer, and the rule consequent factor is obtained via least-squares estimation.

*1) GD training method*: The gradient descent (GD) method minimizes the error between the target output and the output value by adjusting weights. The error gradient is transmitted in the reverse direction at the k[th] output node at epoch p and is expressed as (9)

$$\delta_k(p) = \dot{f}_k(p) \cdot e_k(p) \tag{9}$$

Where, $e_k(p) = d_k(p) - y_k(p)$, and $\dot{f}_k(p)$ is the derivative of the activation function of output node $k$ at epoch $p$.

The slope and the weight ($w_{kj}$) of the cost function for the weight between the hidden layer and the output layer are improved at the $p + 1$[th] epoch to (11).

$$\Delta w_{jk}(p) = -\eta \cdot y_j(p) \cdot \delta_k(p) \tag{10}$$

$$w_{jk}(p+1) = w_{kj}(p) + \Delta w_{kj}(p) \tag{11}$$

Where, $\Delta w_{jk}(p)$ is the weight improvement amount between the $j$[th] hidden node and the $k$[th] output node at epoch $p$, and $\eta$ is the learning rate.

Similarly, the error signal at the $j$[th] hidden node and the adjustment of the weight ($w_{ij}$) between the input layer and the hidden layer at the $p + 1$[th] epoch is as follows (12),(13), and (14).

$$\delta_j(p) = \dot{f}_j(p) \cdot \sum_{k=1}^{N} \delta_k(p) \cdot w_{jk}(p) \tag{12}$$

$$\Delta w_{ij}(p) = -\eta \cdot x_i(p) \cdot \delta_j(p \tag{13}$$

$$w_{ij}(p+1) = w_{ij}(p) + \Delta w_{ij}(p) \tag{14}$$

Where, $\Delta w_{ij}(p)$ is the weight improvement amount between the $i$[th] input node and the $j$[th] hidden node, and $\delta_j(p)$ is the error gradient that is reversed from the hidden layer to the input layer at epoch $p$.

*2) SCG training method*: The scaled conjugate gradient (SCG) training method is known to be effective for large problems. It uses the second-order information, without the line-search process. Thus, amount of using memory can be reduced by reducing the amount of computation of gradient information. The final SCG algorithm is detailed below [15].

① Choose the weight vector $\tilde{w}_1$ and scalars $0 < \sigma \leq 10^{-4}$,

$0 < \lambda_1 \leq 10^{-6}, \bar{\lambda}_1 = 0$.

Set $\tilde{p}_1 = \tilde{r}_1 = -E'(\tilde{w}_1), p = 1,$ and $success = true$.

② If *success* = *true*, calculate the second-order information:

$$\sigma_p = \sigma / |\tilde{p}_p|,$$

$$\tilde{s}_p = (E'(\tilde{w}_p + \sigma_p \tilde{p}_p) - E'(\tilde{w}_p))/\sigma_p,$$

$$\delta_p = \tilde{p}_p^T \tilde{s}_p.$$

③ Scale $\delta_p$: $\delta_p = \delta_p + (\lambda_p - \bar{\lambda}_p)|\tilde{p}_p|^2$.

④ If $\delta_p \leq 0$, make the Hessian matrix positive definite:

$$\bar{\lambda}_p = 2(\lambda_p - \delta_p / |\tilde{p}_p|^2),$$

$$\delta_p = -\delta_p + \lambda_p |\tilde{p}_p|^2,$$

$$\lambda_p = \bar{\lambda}_p.$$

⑤ Calculate the step size:

$$\mu_p = \tilde{p}_p^T \tilde{r}_p,$$

$$\alpha_p = \mu_p / \delta_p.$$

⑥ Calculate the comparison parameter:

$$\Delta_p = 2\delta_p [E(\tilde{w}_p) - E(\tilde{w}_p + \alpha_p \tilde{p}_p)] / \tilde{\mu}_p^2.$$

⑦ If $\Delta_k \geq 0$, a successful reduction in the error can be made:

$$\tilde{w}_{p+1} = \tilde{w}_p + \alpha_p \tilde{p}_p,$$

$$\tilde{r}_{p+1} = -E'(\tilde{w}_{p+1}),$$

$$\bar{\lambda}_p = 0, \textit{success} = \textit{true}.$$

If $k$ mod $N = 0$, restart the algorithm:

$$\tilde{p}_{p+1} = \tilde{r}_{p+1}$$

else:

$$\beta_p = (|\tilde{r}_{p+1}|^2 - \tilde{r}_{p+1}^T \tilde{r}_p)/\mu_p,$$

$$\tilde{p}_{p+1} = \tilde{r}_{p+1} + \beta_p \tilde{p}_p.$$

If $\Delta_p \geq 0.75$, reduce the scale parameter:

$$\lambda_p = \tfrac{1}{4} \lambda_p.$$

else:

$$\bar{\lambda}_p = \lambda_p,$$

*success* = *false*.

⑧ If $\Delta_p < 0.25$, increase the scale parameter:

$$\lambda_p = \lambda_p + (\delta_p (1 - \Delta_p) / |\tilde{p}_p|^2).$$

⑨ If the steepest descent direction $\tilde{r}_p \neq \tilde{0}$, set $p = p + 1$ and go to ②; else, terminate and return $\tilde{w}_{p+1}$ as the desired minimum.

*3) LM training method*: The levenberg-marquardt (LM) training method [16] is a deformation of the Newton method. The algorithm has a faster convergence of second orders, fewer iterations, and does not need to compute the Hessian matrix. For some network models with few parameters, the training speed of the algorithm is higher [17]. Let N be the vector of the weight and bias of each layer in the iterative training is given by (15). The delta N is the adjustment

quantity of N, and adjusting N means the regulation of the weights and thresholds of each layer in the network; finally, the goal of training the network is achieved.

$$N = \begin{bmatrix} w_{11} & w_{12\ldots}w_{IH} & \theta J\,(1)\ldots\theta J(H) \\ V_{11} & V_{12\ldots}V_{HO} & \theta I\,(1)\ldots\theta I(O) \end{bmatrix} \quad (15)$$

$$P(N) = \sum_{i=1}^{I} e_i(N)^2 \quad (16)$$

Where, P(N) is expression function, $e_i(N)^2, i = 1 \rightarrow I$ represents the square of error.

$$\nabla P(N) = J^T(N)E(N) \quad (17)$$

$$\nabla^2 P(N) = J^T(N)E(N) + S(N) \quad (18)$$

Where, E(N) is expressed as $e_i(N)$, $i = 1 \rightarrow I$ constituent vectors, $J(N)$ is the Jacobian matrix given by (19), and S(N) is the error function given by (20).

$$J(N) = \begin{bmatrix} \frac{\partial e_1(N)}{\partial N_1} & \frac{\partial e_1(N)}{\partial N_2} & \frac{\partial e_1(N)}{\partial N_i} \\ \frac{\partial e_2(N)}{\partial N_1} & \frac{\partial e_2(N)}{\partial N_2} & \frac{\partial e_2(N)}{\partial N_i} \\ \frac{\partial e_I(N)}{\partial N_1} & \frac{\partial e_I(N)}{\partial N_2} & \frac{\partial e_I(N)}{\partial N_i} \end{bmatrix} \quad (19)$$

$$S(N) = \sum_{i=1}^{I} e_i\,(N)\nabla^2 e_i(N) \quad (20)$$

Because the LM algorithm is an improved form of the Gauss Newton method, ΔN is given as shown in (21).

$$\Delta N = J^T(N)E(N)[J^T(N)J(N) + \mu I]^{-1} \quad (21)$$

Where, $I$ is the unit matrix, and $\mu > 0$ is a constant.

When $\mu$ = 0 is used for the Gauss Newton method, when larger, LM approaches the small-step GD method. For training, the modification factor is changed from $\mu$ to α. If the training fails, $\mu$ is increase or, decrease it. The solution of $J^T(N)E(N)[J^T(N)J(N) + \mu I]^{-1}$always exists, because of it is positive value. Thus, the LM algorithm is superior to the Gauss Newton method. The LM algorithm steps are as follows.

① The allowable values, coefficients, thresholds, weights of error training, and $k$ = 0 ($k$ is the number of iterations) are initialized.

② Calculate the output of network and the expression function P(N) and Jacobian matrix $J$(N).

③ Calculate $\Delta N$.

④ If $P(N) < \varepsilon$, the end; otherwise, use $N + \Delta N$ as the weights and thresholds to recalculate the expression function $P(N)$. When P(N) is less than the P(N) of ②, set $\mu = \mu/\alpha$ and $p = p + 1$, and return to ②; otherwise, set $\mu = \mu \cdot \alpha$, and return to ③.

## IV. Experiment Results

### A. Data-Acquisition Environment

We set up the experimental environment to obtain the real data required for lead-acid battery SOC estimation, as shown in Fig. 5.

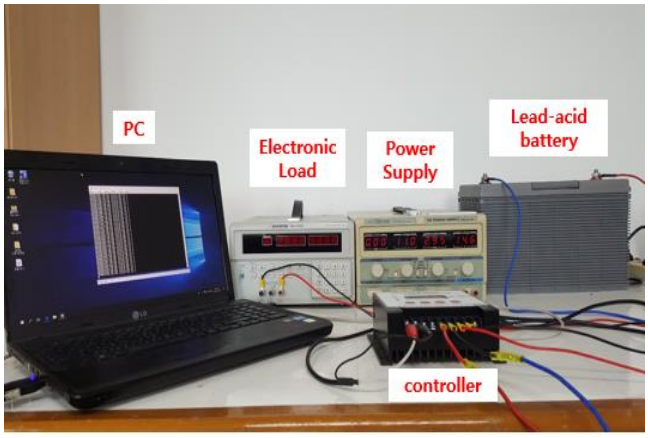



Fig. 5. Experimental Setup.

TABLE I. Specifications of Experimental Equipment

| Item | Specification | Quantity |
|---|---|---|
| Power supply | Model: UP-150DT<br>Max values: 5 A/30 V (DUAL) | 1EA |
| Lead-acid battery | Model: KB100-12<br>Nominal capacity: 12 V, 100 Ah | 1EA |
| Electronic load | Model: PEL-300<br>Power: 1–300 W<br>C.V mode: 3–6 V<br>C.C mode: 6 mA–60 A | 1EA |

The experimental environment composed of a lead-acid battery, an electronic load, a power supply, a battery controller, and a PC. Further, by connecting the battery controller and the PC in series, the real-time battery data are monitored and collected by the PC. The power supply performs battery charging, and the electronic load performs battery discharging. The battery controller includes a protection function for over-charge or over-discharge of the battery and monitoring of the current and voltage of the battery in real time. The detailed specifications of the experimental equipment are shown in TABLE I.

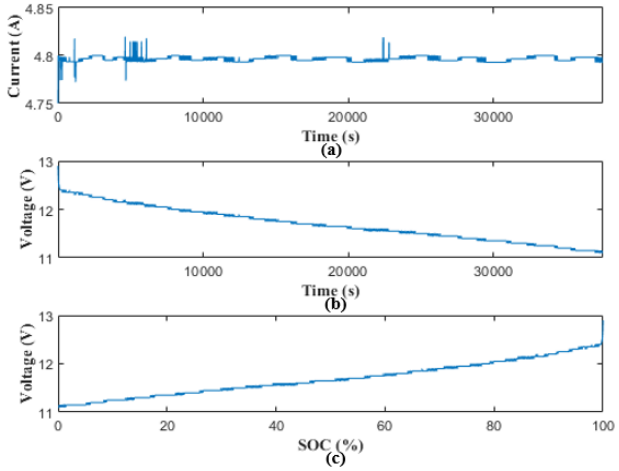



Fig. 6. One Cycle of Discharge: (A) Current, (B) Voltage, and (c) Voltage Against the SOC.

In addition, the collected data are nominalized to a range of between -1 and 1 for the positive effect that the network model has faster convergence in training, as shown in (22).

$$x = \frac{2(x - x_{\min})}{x_{\max} + x_{\min}} - 1 \quad (22)$$

Where, $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of $x$ from the data that we collected.

A solar controller was used to prevent over-charging and over-discharging when the data were acquired. The battery was discharged at a constant load from the fully charged state to the discharge end voltage. The battery voltage range was 11.1 ~ 12.9. A graph of the battery discharge is shown in Fig. 6.

### B. Configuration of Network

The neural network models are implemented and tested by using MATLAB. The configuration of the FFNN and RNN model is as follows: the input nodes are 2, the output node is 1, the hidden nodes are 30, the learning rate is 0.01, and the training error goal is 0.0005. We used the hyperbolic tangent sigmoid function as the activation function of the hidden layer, and the linear transfer function as the activation function of the output layer. The initial weights were random values. The cost function was the mean squared error (MSE) and is given by (23).

$$MSE = \frac{1}{n}\sum_{p=1}^{n}\sum_{k=1}^{l}(d_k(p) - y_k(p))^2 \quad (23)$$

Where, $n$ is the number of training sets, $d_k(p)$ is the $k^{\text{th}}$ target value, and $y_k(p)$ is the output at output node $k$ at epoch $p$.

The configuration of the ANFIS model is as follows: the input nodes are 2, the output node is 1, the membership functions are Gaussian, and the number of rule neurons is 25. Training is performed until the root-mean-square error (RMSE) reaches the error tolerance 0.01, and the RMSE is given by (24).

$$RMSE = \sqrt{\frac{1}{n}\sum_{p=1}^{n}(d(p) - y(p))^2} \quad (24)$$

Where, $n$ is the number of training sets, and $d(p)$ and $y(p)$ are the target value and output value at epoch $p$.

### C. SOC Estimation Results

We used the data obtained from the experimental setup presented in Section 4.1 and compared the proposed methods for estimation of the lead-acid battery SOC. First, the FFNN model was tested by applying three training modes, and the SOC estimation results are shown in Fig. 7, Fig. 8, and Fig. 9.

The FFNN-GD method passed 100,000 epochs to converge to the desired MSE during training, and the test MSE was 2.11. FFNN-SCG recorded 1,401 epochs, and the MSE was 2.03, which was better than that of FFNN-GD. FFNN-LM converged on 35 epochs, and the MSE was 1.78 as the test data. Thus, the MSE and epochs were superior to those of the other two training methods.

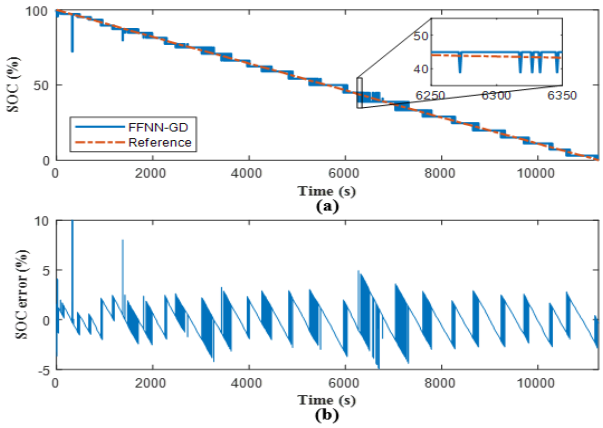



Fig. 7. SOC Estimation using FFNN-GD: (a) SOC and (b) SOC Error.

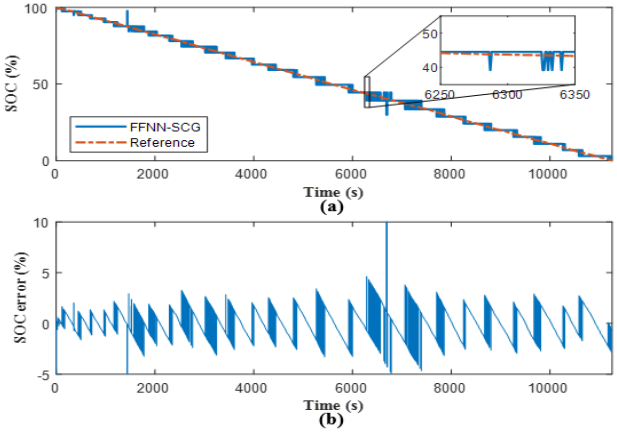



Fig. 8. SOC Estimation using FFNN-SCG: (a) SOC and (b) SOC Error.

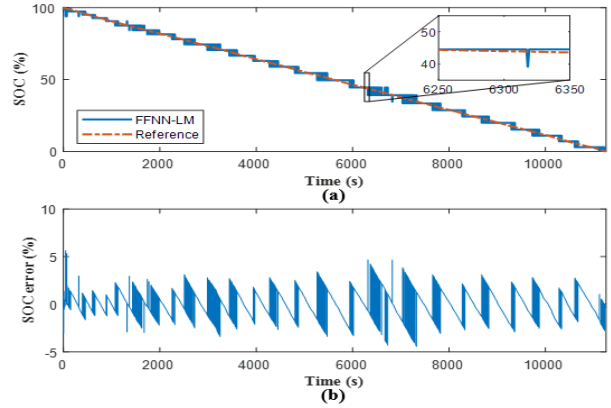



Fig. 9. SOC Estimation using FFNN-LM: (a) SOC and (b) SOC Error.

Second, the RNN model was tested by applying the same three training modes, and the SOC estimation results are shown in Fig. 10, Fig. 11, and Fig. 12. The RNN-GD method passed 100,000 epochs to converge to the desired MSE during training, and the test MSE was 1.67. The RNN-SCG recorded 764 epochs, and the MSE was 1.19, which was better than that of the RNN-GD method. RNN-LM converged to epochs. The MSE of 1.08 was better than those of RNN-GD and RNN-SCG.

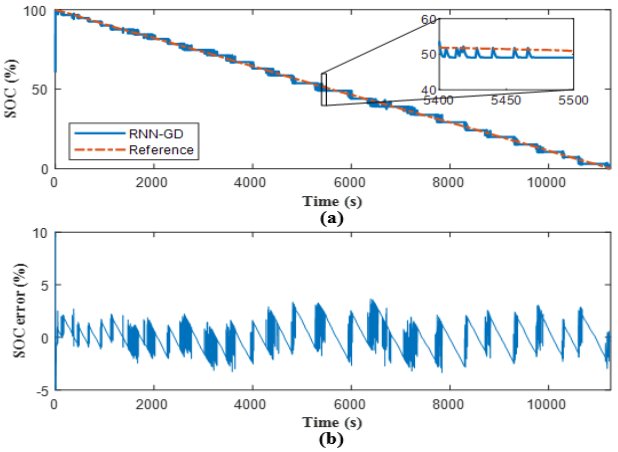



Fig. 10. SOC Estimation using RNN-GD: (a) SOC and (b) SOC Error.

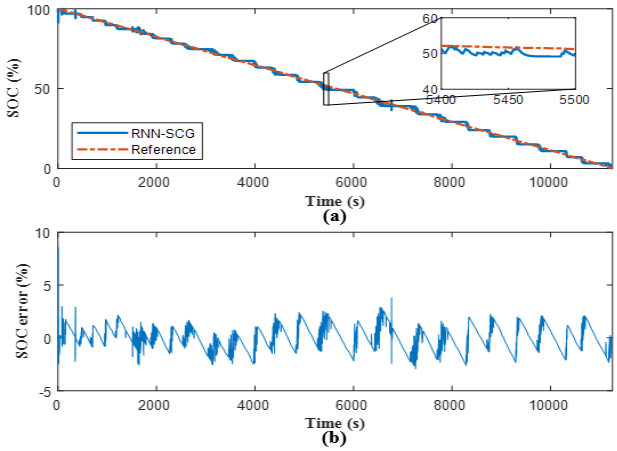



Fig. 11. SOC Estimation using RNN-SCG: (a) SOC and (b) SOC Error.

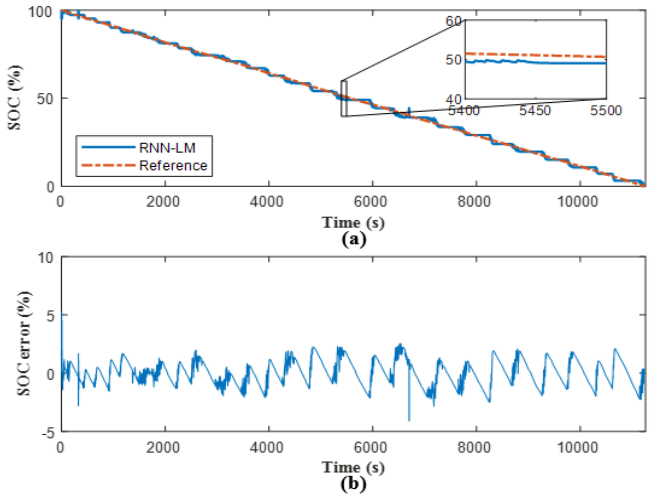



Fig. 12. SOC Estimation using RNN-LM: (a) SOC and (b) SOC Error.

Finally, SOC estimation was performed with ANFIS using a hybrid training method, and the results are shown in Fig. 13. The training was completed in only one epoch, and the MSE was 1.82 as a result of the SOC estimation with test samples. All the experimental results are comprehensively compared in Fig. 14 and TABLE II.

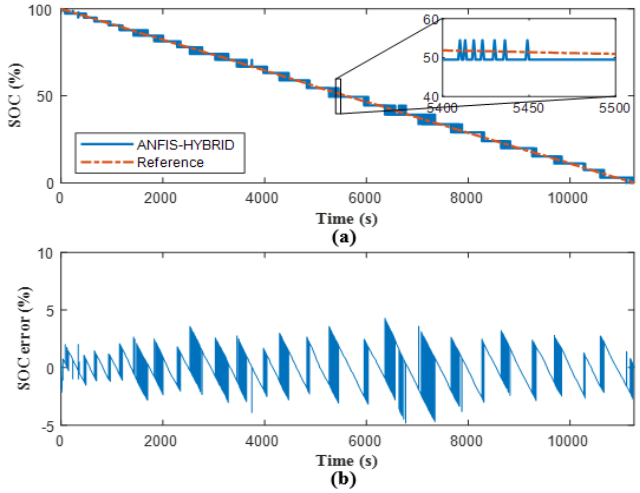



Fig. 13. SOC Estimation using ANFIS-Hybrid: (a) SOC and (b) SOC Error.

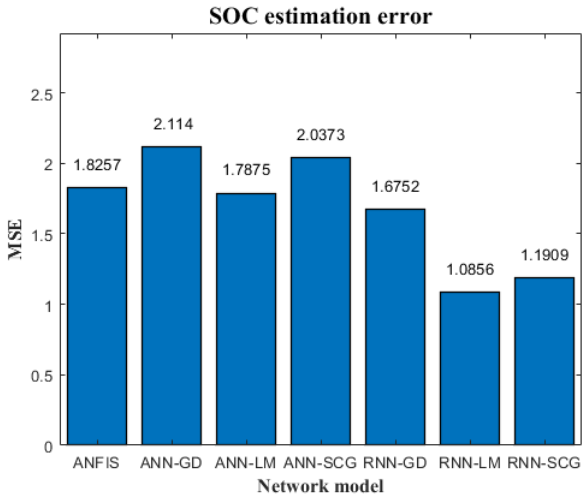



Fig. 14. Comparison of the SOC Estimation Error.

TABLE II. COMPARISON OF TRAINING EPOCHS AND MSE

| Para-meter | Network model | | | | | | |
|---|---|---|---|---|---|---|---|
| | *FFNN* | | | *RNN* | | | *ANFIS* |
| | *GD* | *SCG* | *LM* | *GD* | *SCG* | *LM* | *Hybrid* |
| Epochs | Over 100, 000 | 1,401 | 35 | Over 100, 000 | 767 | 6 | 1 |
| MSE | 2.11 | 2.03 | 1.78 | 1.67 | 1.19 | 1.08 | 1.82 |

## V. EXPERIMENT RESULTS

Three intelligent models (FFNN, RNN, and ANFIS) for estimating the lead-acid battery SOC in PV systems were presented. Additionally, we compared the proposed methods with regard to the MSE and epochs.

The experimental results are as follows. For the FFNN and RNN, we used three training methods and found that FFNN-LM (MSE of 1.78, 35 epochs) and RNN-LM (MSE of 1.08, 6 epochs) demonstrated excellent performance. The ANFIS

used the hybrid learning method with GD, and according to the least-squares method, the MSE was 1.82, with 1 epoch. Moreover, the convergence speed was the highest among all the models. In summary, RNN-LM can learn at a high speed and is the most accurate among all the methods; thus, the RNN-LM method is suitable for estimating the lead-acid battery SOC because it can be generalized to the greatest extent. In a future study, the proposed methods will be applied to a lithium battery.

## REFERENCES

[1] C. K. Park, "Study on the obsolescence forecasting judgement of PV systems adapted micro-inverters", Journal of Korea Multimedia Society, vol. 18, no. 7, pp. 864-872, 2015.

[2] J. H. Jung, J. H. Jung, "New State-of-Charge Polynomial using Hermite Interpolation", Journal of the Institute of Electronics Engineers of Korea, vol.48, no.1, pp. 9 – 17, 2011.

[3] M. Coleman, C. K. Lee, C. Zhu, and W. G Hurley, "State-of-Charge Determination From EMF Voltage Estimation: Using Impedance, Terminal Voltage, and Current for Lead-Acid and Lithium-Ion Batteries", IEEE transactions on industrial electronics, vol. 54, no. 5, Oct, 2007.

[4] K. S. Ng, C.-S. Moo, Y.-P. Chen, and Y.-C. Hsieh, ''Enhanced coulomb counting method for estimating state-of-charge and state-of-health of lithium-ion batteries,'' Appl. Energy, vol. 86, no. 9, pp. 1506–1511, 2009.

[5] SchurmFFNN J, "Pattern classification, a unified view of statistical and neural approaches", john Wiley and Sons, New York, 1996.

[6] Kohonen T, Self-organizing maps, Springer, Berlin, 1997.

[7] Fausett L, Fundamental of neural netwroks, Prentice Hall, 1994.

[8] Timmermans, A.J.M and A.A hulzebosch, "Computer vision system for on-line sorting of pot plants using an artificila neural network classifier,"Computer and Electronics in Agriculture, vol. 15, no. 1, pp. 41-55, May,1996.

[9] Th.F. Elshatter, M.T. Elhagry, E.M. Abou-Elzahab, A.A.T. Elkousy,"Fuzzy modeling of photovoltaic panel equivalent circuit", in: Proceedings of the Photovoltaic Specialists Conference, vol. 15, no. 22,pp. 1656–1659, Sep. 2000.

[10] M. AbdulHadi, A.M. Al-Ibrahim, G.S. Virk, "Neuro-fuzzy-based solar panel model", IEEE Trans. Energy Convers. vol. 19, no. 3, pp. 619–624,Aug. 2004.

[11] M. S. H. Lipu, M. A. Hannan, A. Hussain, and M. H. M. Saad, "Optimal BP neural network algorithm for state of charge estimation of lithium-ion battery using PSO with PCA feature selection," J. Renew. Sustain. Energy, vol. 9, no. 6, p. 64102, Nov. 2017.

[12] K. S. Yu, B. K. Kim, D. J. Kim, M. S. Jang, H. S. Go and H. C. Kim, "A State-of-Charge estimation using extended Kalman filter for battery of electric vehicle" Journal of Korea Academia-Industrial cooperation Society, vol. 18, no. 10, pp. 15-23, 2017.

[13] Michael Negnevitsky, Artificial Intelligence, 2rd ed., Addison Wesley.

[14] S. H. Par, K. J. Lim, S. H. Kang, J. M. Seo, and Y. G. Kim, "Comparison of recognition rates between BP and ANFIS with FCM clustering method on off-line PD diagnosis of defect models of traction motor stator coil," KIEE International Transactions on Electrophysics and Applications, vol. 5-C, no. 4, pp. 138-142, 2005.

[15] M. F. Moller, "A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning", Neural Networks, Vol. 6, pp. 525-533, 1993.

[16] Zeng Min, Liang Xiao, Lin Cao, Hangcheng Yan, "Application of the Neural Network in Diagnosis of Breast Cancer Based on Levenberg-Marquardt Algorithm", International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), p.268 – 272, 2017.

[17] Arun. D. Kulkarni. Fuzzy Neural Network Models for Classification. Applied Intelligence, 2000, 1 2(3): 207-215.

# Novel Mechanism of Classifying the Brain Tumor for Identifying its Critical State

Deepthi Murthy T S
Assistant Professor
School of ECE, REVA University,
Bangalore, India

Sadashivappa G
Professor
Department of Telecommunication Engineering, RV College of Engineering, Bangalore, India

Ravi Shankar D
Assistant Professor
School of ECE, REVA University,
Bangalore, India

***Abstract*—Classification of brain tumor is one of the most challenging tasks in the clinical and radiological research. Upon investigating the existing research contribution, we find that still there is wide open scope of addressing classification problem pertaining to brain tumor. Therefore, this manuscript presents a simple mechanism of classifying the brain tumor in order to categorize its state of criticality. The proposed system applies a multi-level preprocessing to enhance the input image followed by image thresholding for feature extraction and decomposition using wavelet transform. The extracted features are further subjected to process of dimensional reduction that maintains a balance between good number of enriched feature and less size of redundant feature using statistical approach. Further, a supervised learning approach is implemented that further optimizes the classification process. The study outcome is further benchmarked with different process of classification to show the efficient computational environment of proposed system.**

***Keywords*—Brain tumor; classification; categorization; clustering; identification; segmentation; MRI; Brain X-Ray**

## I. Introduction

Brain tumor is considered to be a dreadful clinical condition affecting both patient and doctor to deal with. It is basically an abnormal growth of tissues which could be either lethal (malignant) or non-lethal (benign). However, there are also fair chances of transformation of non-lethal to lethal tumors in many cases. At present, with the advancement of medical imaging, it is now not a big deal to obtain the imaging of the brain tumor [1]. Magnetic Resonance Imaging or popularly known as MRI is one of the best mechanism to obtain multi-dimensional as well as discrete visualization of brain tumor [2]. However still with such advance imaging, there are significant problems associated with both identification of tumors as well as classifying it. The first problem associated with tumor detection is illumination factor [3]. Variable illumination factors e.g. abnormal brightness, contrast, saturation, etc may significant affect the detection rate. Similarly, there are other set of problems associated with classifying what sort of tumor it is. The prime reason behind such classification problem is many. Normally, there are neoplastic tissues in brain that are considered to be non-homogeneous in its natural state when the imaging is taken [4]. Apart from this, such tissues are also found to be overlapped with each other as found in the imaging that poses a significant level of challenge for the physician to actually even understand the type of tumor. The normal tissues are also found to be overlapped with such neoplastic tissues making it further impossible even to identify that is it a tumor or normal tissue. On the other hand, there is another type of tumor called as Glioma that associated with various types of brain cells. It will mean that Gliomas exhibit multiple characteristics in imaging making it more difficult to classify. In order to solve such problem, the medical community has evolved up with a reference model of brain neoplasm that considered analysis based on histopathological images [5]. This process is further confirmed by surgical biopsy. In simpler way, there is no straight forward imaging technique of classification of brain tumor although identification of tumor is somewhat feasible based on skills of radiologist.

In the last few decades, there have been certain serious research attempts toward exploring classification problems [6]-[10]. Majority of these existing research contributions has been focusing more on detection and segmentation rather than on classification problems. A closer look into the problems associated with the existing research work are i) more work towards detection and less towards classification, ii) few studies with benchmarked outcomes, iii) usage of region of interest, iv) consistent usage of machine learning approach with computational complexity. Hence, all these above problems are essential to be mitigated as the present demand of the application is not only accuracy but also computational friendly application. At present, software used for diagnosis are modeled to run on low-power and low-resourceful computing mobile device. Hence, there is a need to evolve up with computational model that can balance both the classification performance as well as computational efficiency. Therefore, the proposed manuscript introduces a simple and yet novel classification process to determine the malignant and benign case of brain tumor. The paper also outlines the simple usage of feature detection, detection, dimensional reduction, and learning involved. Section II highlights existing research works in tumor classification followed by problems associated with existing system are discussed in Section III. Section IV briefs the proposed research methodology and Section V gives the algorithm design while the results analysis described in Section VI. Finally, Section VII summarizes the proposed system.

## II. Related Works

This section discusses about research contribution of existing authors towards addressing the problems associated with the classification of the tumor in medical images.

Bahadure et al. [11] have emphasized on performing segmentation technique to ensure better detection performance of brain tumors using wavelet transform and supervised learning approach. Most recently a classification approach using ensemble classifier is introduced by Cong et al. [12] where multiple clustering techniques has been used for diagnosis cancerous section. Usage of ensemble classifier was also reported in the work carried by Esener et al. [13]. The author have used region-of-interest followed by preprocessing and feature extraction for introducing classification of multistage cancer. Jadoon et al.[14] have adopted deep learning method in in order to perform classification. The author have also used curvelet-based transform as well as adaptive histogram and wavelet transform in the preliminary stages followed by training using convolution neural network and support vector machine. Ravi et al. [15] have emphasized on the segmentation scheme for enhanced detection of abnormal region in brain imaging. The authors have used stochastic approach as well as semantic approach for performing classification. Adoption of deep learning approach was also seen in the work of Song et al. [16] towards solving classification problems associated with radiological images. Vu et al. [17] have presented a dictionary learning technique for effective classification of histopathological images. The scheme assists in generation of rich number of features from brain tumor images. Xi et al. [18] have implemented swarm intelligence and supervised learning technique for enhancing the selection process of feature that directly assisted in solving classification problems. Zhao and Jia [19] have advocated the use of convolution neural network that is claimed to have its superior effect in segmentation performance. The scheme is also said to maintain equal emphasis on both local and global region.

Usage of neural network was also reported in the work carried out by Marcomini et al. [20]. The authors have used self-organizing map along with multiple segmentation process assisting in better classification process. Barbalata and Mattos [21] have used anisotropic filtering mechanism in order to perform detection and classification of various critical lesions on the radiological image. Anitha and Murugavalli [22] have enhanced k-means algorithm to incorporate adaptiveness in order to assist in solving segmentation and classification problems. The authors have also used neural network and wavelet transform. Kaur et al. [23] have introduced the usage of ensemble classifier as well as Hilbert transformation for enhanced representation of radiological image. The implemented algorithm was proven to have maximum accuracy level on brain tumor images. Jui et al. [24] have emphasized on the feature extraction scheme for enhancing the segmentation process as well as classification process all together. Emphasis on feature selection process is also carried out by Huda et al. [25], where the authors have morphometric analysis of radiological images followed by applying ensemble based classifier. Svensson et al. [26] have investigated on supervised classification technique as well as random forest in order to evolve up with a better classification scheme. Work carried out by Sharma and Marikkannu [27] have introduced as unique classification technique. The authors have used Gaussian filter in preliminary state followed by a non-conventional mechanism of extracting features based on distribution of gray level. Further optimization of features were carried out by search algorithm followed by series of learning algorithms e.g. support vector machine, back propagation, k-nearest algorithm etc to obtain better classification performance. Automation process in extraction of significant feature assists in precise identification of tumors. This fact has been proven by Chadded [28] by using Gaussian mixture model. Bukovsky et al. [29] have used neural network in order to perform identification followed by prediction of tumor. This process is a way advanced to classification process using feed-forward neural network. Acevedo et al. [30] have implemented memory-based scheme of bidirectional nature in order to increase the accuracy of classification process. Kothavari et al. [31] have used machine learning approach along with search optimization in order to emphasize on the feature extraction process. Chen and Jian [32] have used regression approach for carrying out classification of tumor. Huang et al. [33] have introduced a projection–based mechanism as well as regression for assisting in solving classification problems. The technique assists in categorizing voxels of clinical importance and thereby emphasizing on the locality factor of the given radiological image. Huml et al. [34] have used genetic programming along with the usage of morphological marker in order to assists in classification of brain tumor. The technique further introduces usage of mining approaches in order to extract more precise clinical information directly assisting in tumor classification. Corso et al. [35] have used Bayesian classification approach for localizing the brain tumor. Saddique et al. [36] have used a contour-based technique in order to identify the symmetry factor followed by segmentation process. Nanthagopal and Sukanesh [37] have used wavelet-based technique for feature extraction and used both support vector machine and neural network for classification of brain tumor. Sun et al. [38] have implemented a learning approach using support vector machine and Eigen value for classification of tumor. The technique also uses independent component analysis. Usage of neural network was reported in the work of Brougham et al. [39] on spectroscopic data for solving its associated classification problems. Usage of matrix factorization is adopted in the work of Zheng et al. [40] for solving similar problems associated with tumor classification. The behavior of tumor is found in Astana and Vashith [41] in which brain MRI is considered for defining the criticality of tumor. In [41], time taken by manual segmentation is minimized by using appropriate quantitative technique which can be useful for futuristic clinical research. The performance analysis of [41] states significant outcomes.The next section briefs up the research problem associated with existing research work.

## III. Problem Identification

From the prior section, it can be seen that there has been various research techniques work on solving classification problems. All the problems offers advantages as well as is also associated with certain unsolved problems e.g. i) existing techniques are more inclined towards machine learning algorithms where the proportion of learning operation is too much complex and more as compared to other lightweight operations. This process although shows good increase in accuracy but they are highly subjective in nature. It will mean

that such algorithms don't offer similar performance when the environmental variables are changed, ii) existing techniques also doesn't emphasizes on pre-processing steps. All the existing techniques apply any one preprocessing without knowing if they are really suitable or not, iii) existing feature extraction techniques using wavelet based transforms are too much complex and involves too many recursive steps. This reduces computational performance although they can offer better accuracy performance, iv) usage of supervised learning algorithms e.g. neural network, support vector machine, etc have been used to ensure better accuracy, but usage of such techniques doesn't emphasize on convergence performance as well as techniques e.g. neural network has higher dependency to perform repetitive training on trained images rather than using it from features. Such process significant cost the execution time of algorithms that has not been much discussed in existing technique. Hence, these are the open issues in existing research techniques that are yet to be solved. The next section introduced methodology to address such problems.

## IV. Proposed Research Methodology

The primary aim of the proposed system is to carry out a simple and effective classification process for assisting in understanding the state of criticality of the tumor condition for a given brain image. The proposed system considers the analytical research methodology in order to perform the implementation of the process flow as exhibited in Fig.1. The proposed system basically acts as a framework that considers the input radiological image of brain, which is further subjected to multi-stage preprocessing e.g. gamma correction, adaptive histogram equalization linear/non-linear filtering process in order to obtain superior enhancement of the input image. The significant level of features are extracted from the enhanced image in the form of binarized threshold image which is further subjected to maximum 3 level of decomposition using wavelet transforms. This process significant assists in obtaining enriched and smaller number of features in the form of decomposed coefficients. Although, the enrichment of the features are good but they are very large in number that could offer significant consumption of processing time. We address this problem by using an algorithm that performs dimension reduction. This algorithm successfully extracts the abstract of statistical data from the feature obtained in prior stage and thereby makes the data much precise with reduced size. We use Principal Component Analysis (PCA) for this purpose. The final step is to apply algorithm of classification where supervised classification technique e.g. Support Vector Machine (SVM) is utilized for categorizing the criticality state of tumor in the form of malignant or benign state of output (Fig.1).

## V. Algorithm Description

This section discusses about the algorithms implemented in order to carry out classification of brain tumor. The proposed system initially takes the input of a brain images and subject it to a multi-level image enhancement process. This process allows performing on-demand preprocessing based on the complexity of the visualization of an image. The proposed system introduces three algorithms that perform sequential operation in order to achieve the research goal. The discussions of the algorithms are as follows: -

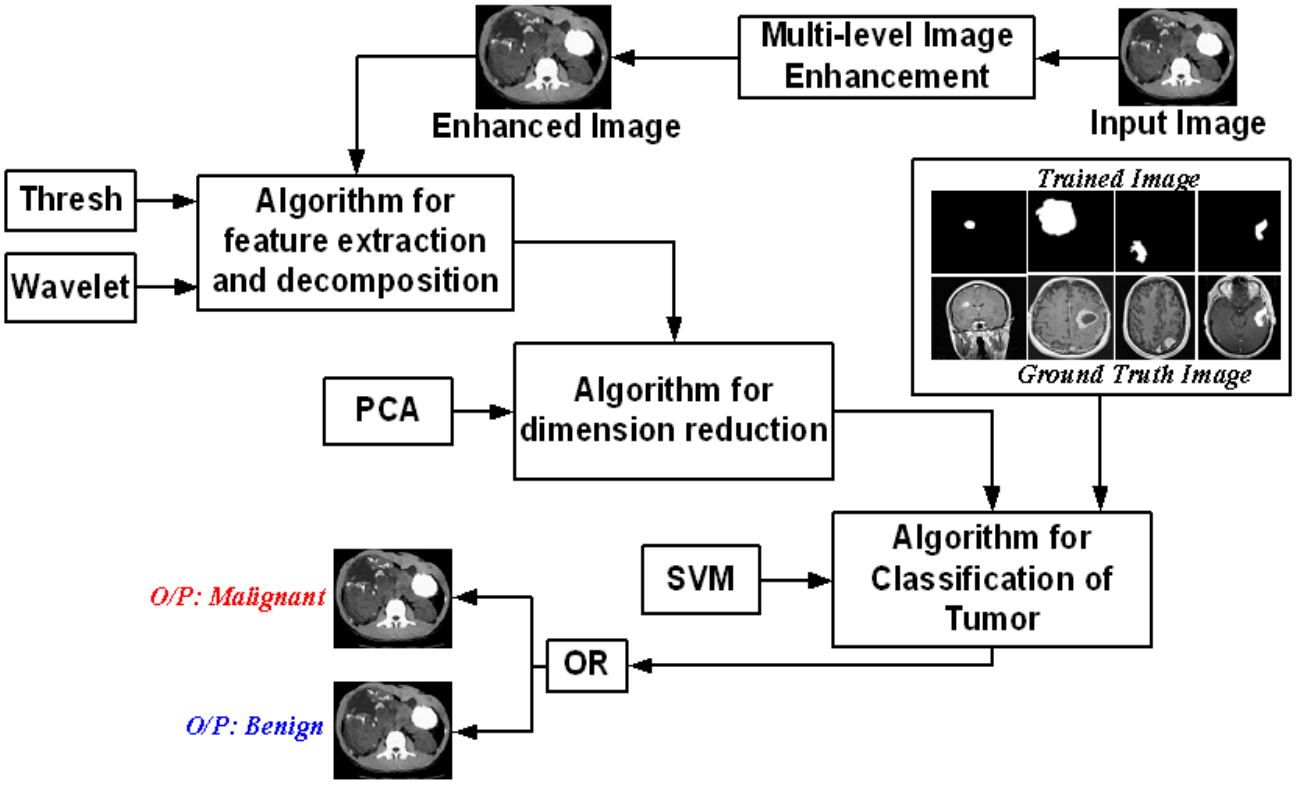



Fig. 1. Process Flow of Proposed System.

*A. Algorithm for Feature Extraction and Decomposition*

This algorithm is responsible for extracting explicit set of preliminary features that will assists in classification process in later phase of implementation. Let L be the number of gray levels for representing a pixel in an image, the gray level ranges from [1, 2 … L].The probability distribution function for the normalized gray level histogram is given by,

$$p_i = \frac{n_i}{N} \quad p_i > 0, \sum_{i=1}^{L} p_i = 1$$

Where, $n_i$→ number of pixels at the $i^{th}$ level

N→ Total number of pixels ($n_1$, $n_{2\ldots}$ $n_L$)

The image pixel is differentiated into pixels of two classes which are identified as foreground and background objects, this is obtained by setting a threshold at level k. The upper bound and the lower bound of the threshold is represented by $C_0$ and $C_1$, where, $C_0$ ranges from 1 to k and $C_1$ ranges from k+1 to L. The probability of occurrence is given as follows,

$$w_0 = P_r(C_0) = \sum_{i=1}^{k} p_i = w(k), \; w_1 = P_r(C_1) = \sum_{i=1 p_i}^{L} = 1 - w(k)$$

The respective mean is computed as follows,

$$\mu_0 = \sum_{i=1}^{k} iP_r(i \mid c_0) = \sum_{i=1}^{k} i\, p_i / w_0 = \mu(k)/w(k)$$

$$\mu_1 = \sum_{k+1=1}^{L} iP_r(i \mid c_1) = \sum_{i=k+1}^{L} i\, p_i / w_1 = \mu_\tau - \mu(k)/1 - w(k)$$

Where,

$$w(k) = \sum_{i=1}^{k} p_i$$

and

$$\mu(k) = \sum_{i=1}^{k\ i} p_i, \qquad \mu_v = \mu(L) = \sum_{i=1}^{L} i_{p_i}$$

Where, $\mu_v$→ total mean level of the original picture. The variances are given by,

$$\sigma_0^{\ 2} = \sum_{i=1}^{k} (i - \mu_0)^2 \Pr(i \mid C_0) = \sum_{i=1}^{k} \frac{(i - \mu_0)^2 p_i}{\omega_0}$$

$$\sigma_1^{\ 2} = \sum_{i=k+1}^{L} (i - \mu_1)^2 \Pr(i \mid C_1) = \sum_{i=k+1}^{L} \frac{(i - \mu_1)^2 p_i}{\omega_1}$$

The following discriminant criterion measure is applied to evaluate the goodness of the threshold value,

$$\lambda = \sigma^{2}{}_B \big/ \sigma^{2}{}_w, k = \sigma^{2}{}_T \big/ \sigma^{2}{}_w, \eta = \sigma^{2}{}_B \big/ \sigma^{2}{}_T$$

Where,

$$\sigma^{2}{}_w = \omega_0 \sigma_0^{\ 2} + \omega_1 \sigma_1 2$$

$$\sigma_w^2 = \omega_0 (\mu_0 - \mu_\gamma) 2 + \omega_1 (\mu_1 - \mu_\gamma)^2$$

$$= w_0 w_1 (\mu_1 - \mu_0)^2$$

The discriminant criteria which maximizes λ, k and η for k are however equivalent to one another, due to the following relation

$$\sigma_w^2 + \sigma_B^2 = \sigma_T^2$$

The optimal threshold $k^*$ which maximizes $\eta$ or $\sigma_B^2$ is selected for sequential search by using a simple cumulative search,

$$\eta(k) = \sigma_B^2(k) \big/ \sigma_T^2$$

where,

$$\sigma_B^2 = \frac{[\mu_{Tw}(k) - \mu(k)]^2}{w(k)[1 - w(k)]}$$

Optimal threshold for $k^*$ is given by,

$$\sigma_B^2(k*) = \max_{1 \le k \le L} \sigma_B^2(k)$$

From the above optimization problem the range of K over which the maximum is sought can be restricted to

$$S^* = \{k; \omega_0 \omega_1 = \omega(k)[1 - \omega(k)] > 0, or\, 0 < \omega(k) < 1\}$$

The above equation is known as the effective range of gray level histogram. Converting the image from rgb to binary format using the above threshold value.The image of RGB format is transformed to binary image by considering the threshold value k defined above, the value at each pixel position is compared with the threshold value. The following condition is applied for the RGB to binary image transformation,

$$I_1(i, j) = \begin{cases} I(i_x, j_y) \le k, then I(i_x, j_y) = 0 \\ I(i_x, j_y) > k, \; then (i_x, j_y) = 1 \end{cases}$$

Where, $i_x$, $i_v$ is current pixel position of the image, k is threshold value, $l_1$(i, j) is resulting Binary image, and l(i,j) is Input RGB image. The next part of the study considers implementing discrete wavelet transformation as decomposition is required Consider an image f(x, y) of size MxN whose forward discrete transform is T(u, v, …) is expressed in terms of general relation as mentioned below.

$$T(u, v) = \sum_{x,y} f(x, y) g_{u,v,} \ldots (x, y)$$

Where, x and y are spatial variables and u, v are transform domain variables. Given T (u, v …), f(x, y) is obtained using generalized inverse discrete transform which is given below,

$$f(x, y) = \sum_{u,v} T(u, v, \ldots) h_{u,v,} \ldots (x, y)$$

Where, $g_{u,v}$ and $h_{u,v}$ in these equations are called forward and inverse transformation kernels respectively. By applying the Discrete Fourier Transform (DFT), in this case

$$h_{u,v}(x, y) = g_{u,v}{}^*(x, y) = \frac{1}{\sqrt{MN}} e^{j2\Pi(ux/M + vy/N)}$$

Where, u → horizontal frequency and v→ vertical frequency.

The kernels are *separable* since

$$h_{u,v}(x,y) = h_u(x)h_v(y)$$

for

$$h_u(x) = \frac{1}{\sqrt{M}} e^{j2\Pi ux/M} \text{ and } h_v(y) = \frac{1}{\sqrt{N}} e^{j2\Pi vy/N}$$

and *orthonormal* because,

$$\langle h_r, h_s \rangle = \delta_{rs} = \begin{cases} 1 & r = s \\ 0 & otherwise \end{cases}$$

The Daubechies wavelet is based on the work proposed by Ingrid Daubechies which belong to a family of orthogonal wavelets. It defines the discrete wavelet transform and is characterized by a maximal numbers of vanishing moments for some given support. The scaling sequence is generally represented by an orthogonal discrete wavelet transform with approximation order A

$$a(z) = 2^{1-A}(1+z)^A p(z)$$

with N = 2A having real coefficients, p(1) = 1 and degree (p) = A-1, the orthogonality condition can be given as

$$a(z)a(z^{-1}) + a(-z)a(-z^{-1}) = 4$$

The above equation can also be represented as,

$$(2-x)^A P(X) + X^A P(2-X) = 2^A$$

The algorithm takes the input of *I* (preprocessed image), *l* (level of decomposition) that after processing leads to generation of *feat* (feature) as an output. The steps of algorithm are as follows:

**Algorithm for feature extraction and decomposition**
**Input**: *I*, *l*
**Output**: *feat*
**Start**
1. init I
2. $I_2 \rightarrow f_{bin}(I)$
3. $sig_1 \rightarrow I_2$
4. **switch***l*
5. case '1'
6. $[C_A, C_H, C_V, C_D] = \psi_2(sig_1, \text{'db4'})$
7. case '2'
8. $[C_{A1}, C_{H1}, C_{V1}, C_{D1}] = \psi_2(sig_1, \text{'db4'})$
9. $[C_A, C_H, C_V, C_D] = \psi_2(C_{A1}, \text{'db4'})$
10. case '3'
11. $[C_{A1}, C_{H1}, C_{V1}, C_{D1}] = \psi_2(sig_1, \text{'db4'})$
12. $[C_{A2}, C_{H2}, C_{V2}, C_{D2}] = \psi_2(sig_1, \text{'db4'})$
13. $[C_A, C_H, C_V, C_D] = \psi_2(C_{A2}, \text{'db4'})$
14. **End**
15. feat→[ $C_A$, $C_H$, $C_V$, $C_D$]
**End**

The algorithm takes the input image and digitizes it (Line-1) followed by applying threshold-based binarization function $f_{bin}$ (Line-2). This mechanism leads to a binarized image $I_2$ that is further subjected to wavelet-based decomposition of 3 level *l* (Line-4). Depending on the selected case of level of decomposition (Line-5, Line-7, and Line-10), the proposed system applying a two-dimensional wavelet function $\psi_2$ on signal $sig_1$ (Line-3) i.e. binarized image. The algorithm implements Daubechies 4 (Line-4 to Line-14) for performing decomposition in order to obtain 4 different coefficients i.e. approximation coefficient ($C_A$), horizontal coefficient ($C_H$), vertical coefficient ($C_V$), and diagonal coefficient ($C_D$). Finally, a matrix is constructed *feat* that stores all the final features upon selection of any specific level of decomposition (Line-15). The outcome is further subjected to algorithm of dimension reduction.

### *B. Algorithm for Dimension Reduction*

This algorithm is responsible for further improving the search towards better and more accurate features. It performs minimization of dimensions of the features obtained. The purpose of implementing the principal component analysis is primarily reducing the dimensionality of the data which consists of large number of inter-correlated variables, while still maintaining the variation present in the dataset. The principal component which are uncorrelated and retains the first few variations present in all of the original variables.To derive the form of the PCs, consider, $\alpha_1'x$where, $\alpha_1$ maximizes $var[\alpha_1'x] = \alpha_1'\sum x$. To maximize $\alpha_1'\sum x$subject to $\alpha_1'\alpha_1 = 1$, the standard approach is to use the technique of Lagrange multipliers. Maximize

$$\alpha_1'\sum \alpha_1 - \lambda(a_1'\alpha_1 - 1)$$

Where λ is a Lagrange multiplier. Differentiation with respect to $\alpha_1$ gives,

$$\sum \alpha 1 - \lambda\alpha_1 = 0 \; or \; (\sum - \lambda I_p)\alpha_1 = 0$$

Where $I_p$ is the (p × p) identity matrix. Thus, λ is an eigenvalue of ∑and $\alpha_1$ is the corresponding eigenvector. To decide which of the p eigenvectors gives $\alpha_1'x$ with maximum variance, note that the quantity to be maximized is

$$\alpha_1'\sum a_1 = \alpha'_1\lambda\alpha_1 = \lambda\alpha_1'\alpha_1 = \lambda$$

Where, λ is very large.

In general, the $k^{th}$ PC of x is $\alpha_k'x$ and $var(\alpha_k'x) = \lambda_k$, where, $\lambda_k \rightarrow k^{th}$ largest eigenvalue of ∑and $\alpha_k$ is the corresponding Eigen vector. The second Principal Component, $\alpha_2'x$, maximizes $\alpha_2'\sum\alpha_2$ subject to being uncorrelated with $\alpha_1'x$ or equivalently subject to $cov[\alpha_1'x, \alpha_2'x] = 0$, where cov(x, y) denotes the covariance between the random variables x and y . But

$$\text{cov}[\alpha_1'x, \alpha'_2 x] = \alpha'_1\sum\alpha_2 = \alpha'_2\sum\alpha_1 = \alpha'_2\lambda_1\alpha'_1 =$$
$$\lambda_1\alpha'2\alpha'_1 = \lambda 1\alpha'_1\alpha_2$$

Thus, any of the equations

$$\alpha'_1 \sum \alpha_2 = 0, \alpha'_2 \sum \alpha_1 = 0$$

$$\alpha'_2 \alpha_2 = 0, \alpha'_2 \alpha_1 = 0$$

could be used to specify zero correlation between $\alpha'_1 x$ and $\alpha'_2 x$. Choosing the last of these (an arbitrary choice), and noting that a normalization constraint is again necessary, the quantity to be maximized is

$$\alpha'_2 \sum \alpha_1 - \lambda(\alpha'_2 \alpha_2 - 1) - \phi \alpha'_2 \alpha_1$$

where λ, φ are Lagrange multipliers. Differentiation with respect to $\alpha_2$ gives

$$\alpha'_1 \sum \alpha_2 - \lambda_{\alpha 2} - \phi \alpha_1 = 0$$

and multiplication of the above equation on the left by $\alpha_1$ gives,

$$\alpha'_1 \sum \alpha_2 - \lambda \alpha_1 \alpha_2 - \phi \alpha'_1 \alpha_1 = 0$$

which, since the first two terms are zero and $\alpha_1\alpha_1=1$, reduces to $\phi=0$. Therefore, $\sum\alpha_2-\lambda\alpha_2=0$, or equivalently $(\sum-\lambda I_p)$ $\alpha_2=0$, so λ is once more an eigenvalue of ∑, and $\alpha_2$ the corresponding eigenvector. Again, $\lambda= \alpha^1{}_2\sum\alpha_2$, so λ is to be as large as possible. Assuming that ∑ does not have repeated eigenvalues. If it did, it follows that $\alpha_2=\alpha_1$, violating the constraint $\alpha'_1\alpha_2=0$. Hence λ is the second largest eigenvalue of ∑, and $\alpha_2$ is the corresponding eigenvector. As stated above, it can be shown that for the third, fourth, ..., $p^{th}$ PCs, the vectors of coefficients $\alpha_3, \alpha_4, \ldots. \alpha_p$ are the eigenvectors of ∑ corresponding to $\lambda_3, \lambda_4, \ldots. \lambda_p$, the third and fourth largest and the smallest eigenvalue, respectively. Furthermore, $var[\alpha'_{kx}]= \lambda_k$ for k = 1, 2 ,… p. The vectors $\alpha_k$ are referred to as principal components. The algorithm takes the input of obtained *feat* (decomposed feature) from first algorithm and gives the output of $feat_{red}$ (features with reduced dimension). The significant steps of the algorithm are as follows:

**Algorithm for dimension reduction**
**Input**: *feat* (decomposed feature)
**Output**: $feat_{red}$ (features with reduced dimension)
**Start**
1. [G]=Ω(feat)
2. g→$f_1$(G)
3. st→$f_2$(g)
4. (con cor eng hom)→*st* (con cor eng hom)
5. obtain M, std, Ent, RMS→G
6. Var→M(G)
7. smoothness→1-(1/(1+∑G)
8. obtain skewness & kurtosis
9. **For** i=1:m
10. for j=1:n
11. t→G(I,j)/(1+(i-j)$^2$)
12.**End**
13. $feat_{red}$=[ con cor eng hom M st dent RMS Var skewness kurtosis t]
**End**

The algorithm implements an orthogonal transformation function of principal component analysis *Ω* on the obtained feature *feat* from first algorithm (Line-1). Applying the function Ω results in a coefficient G (Line-1) that is further used for constructing a new occurances matrix of gray level i.e. by using $f_1$ function (Line-2). This process results in a new variable *g* (Line-2). A consecutive function $f_2$ is applied on obtained variable g for normalizing the obtained matrix (Line-3). This process results in generation of matrix with summation 1 and thereby dimensionality is reduced to greater extent. The obtained statistical information *st* (Line-3) can now be said to represent the properties of the matrix that can further extract contrast *con*, correlation *cor*, energy *eng*, and homogeneity *hom* (Line-4). The next step of the algorithm is to obtain mean *M*, standard deviation *std*, entropy *Ent*, and root mean square RMS (Line-5). More statistical information is obtained e.g. variance *Var*, smoothness, kurtosis, and skewness (Line-8) from the matrix G (Line-6 to Line-8). Considering all the rows and columns of matrix G, a temporary feature *t* is obtained (Line-11). Finally, the outcome of feature with reduced dimensionality is computed from all the above obtained metrics (Line-13). This process reduces the redundant information and retains enriched statistical pixel data that significant enrich the feature obtained from first algorithm.

### C. Algorithm for Classification of Tumor

The proposed system uses supervised learning algorithm in order to perform classification. We use Support Vector Machine for this purpose. The primary purpose of SVM method is to identify and classify complex patterns present in the data. There are two approaches to this problem, the first approach is a computational approach and the second is the statistical approach. The basic principle of classification is that the class of kernel methods implicitly defines class of possible patterns by introducing a notion of similarity between the data. The choice of similarity is considered by the choice of relevant features. General consideration is that kernel methods use information about the inner products between the data items. Hence the kernel functions could be defined as inner product of some feature space which is potentially very complex. Therefore, the necessity of specifying the features becomes lesser when the kernel is given. The first feature of the support vector machines corresponds to that of the duality, i.e. SVM's are known to be linear learning machines which are represented in a dual fashion

$$f(x) = \langle w, x \rangle + b = \sum \alpha_i y_i \langle x_i x \rangle + b$$

Where, $w=\sum\alpha_i y_i x_i$, if $\alpha_i>0$→ linear combination of training points. The data appears only within the dot products (in decision function and training algorithm). The update rule in dual representation is given as follows i.e.

$$yi\left(\sum \alpha_j y_j \langle x_j, x_i \rangle + b\right) \leq 0 \; Then, \alpha_i \leftarrow \alpha_i + \eta$$

It is to be noted that the data appears only inside the dot products in dual representations. Hence, a closer look into the technique adopted here shows that it has an inclusion of regularization attribute that significantly assist in over-fitting problems in the classification of brain tumor. From the computational complexity viewpoint, it was already seen that

existing machine learning approaches suffers from it. However, proposed system addresses no local minima by solving the convex optimization problem. This offers the algorithm not only to obtain the convergence performance in lowest rounds but also in each rounds the algorithm offers significant performance improvement in terms of accurate classification process. This is the final algorithm that perform classification of the given brain tumor image. The algorithm takes the input of $feat_{red}$(feature with reduced dimension) from second algorithm and gives the output of classification of tumor as malignant/benign case. The steps involved in the proposed algorithm are as follows:

**Algorithm for Classification of Tumor**
**Input**: $feat_{red}$
**Output**: classification of malignant/benign case
**Start**
1. load $t_{data}$
2. xd→meas
3. gr→lab
4. $s_{struc}$→$f_3$(xd, gr)
5. sp→$f_4$($s_{struc}$, $feat_{red}$)
6. **If***$f_5$*(sp, malig)
7. disp (MALIGNANT)
8. **Else**
9. disp(BENIGN)
10. **End**
**End**

The preliminary step of this algorithm is to load the trained data $t_{data}$ (Line-1). The algorithm access the elements of the table using the variable *meas* assigned to matrix *xd* (Line-2) as well as it also assign a label *lab* to the group matrix *gr* (Line-3). The algorithm implements a support vector machine in the form of function $f_3$ on *xd* and *gr* that gives the output in the form of structure $s_{truck}$ (Line-4). This new structure $s_{truck}$ consists of all the relevant data associated with supervised trained classifier. The system also uses dot product as the linear kernel function. The next part of the algorithm is to perform classification using function $f_4$ (Line5). This function assists in categorizing the data existing in the row of matrix feature of reduced dimension $feat_{red}$ using the data of trained classifier structure. We use a comparative function $f_5$ in order to check the comparison among all the classifiers i.e. *sp* (Line-6). A true comparison will result in finding the malignant case of tumor (Line-7) or else the benign case is confirmed (Line-9).

It should be known that all the illustrated three algorithms performs sequential operation where the algorithm performs the run-time execution and doesn't store anything on its memory. This process is carried out by constructing a temporary memory of restricted minimal size that stores the run-time process and eliminates them once the results are displaced. This properties offers significant freedom from any form of storage complexity as well as lower dependencies. Hence, the proposed algorithm has better assurance of computational performance apart from accuracy in the classification performance. The next section discusses about the results obtained by implementing algorithm.

## VI. Result Obtained

The algorithms discussed in prior section have been assessed using dataset of brain tumor images. The scripting of the proposed logic was designed using numerical computing MATLAB on normal 32 bit windows machine. The dimensions of the images are approximately in the range of 225x225. Upon taking the input image of brain tumor (Fig.2), the proposed system can undertake multiple forms of preprocessing algorithm in order to perform image enhancement. Fig.3 highlights the enhanced version of the input image.

The proposed algorithm also implements a threshold based scheme for the purpose of performing clustering of image pixels on the gray scale processed image. It significantly minimizes the extent of gray scale that further assist in segmentation process of binary image as shown in Fig.4

Fig. 2. Input Image of Brain Tumor.

Fig. 3. Preprocessed Image.

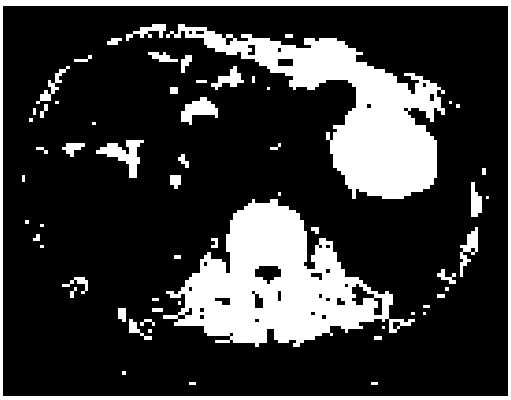
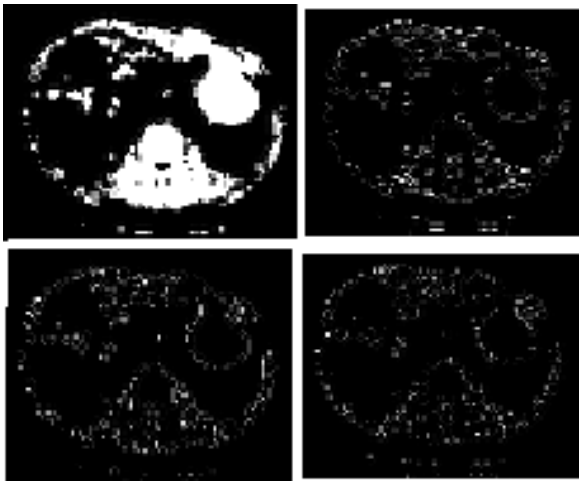
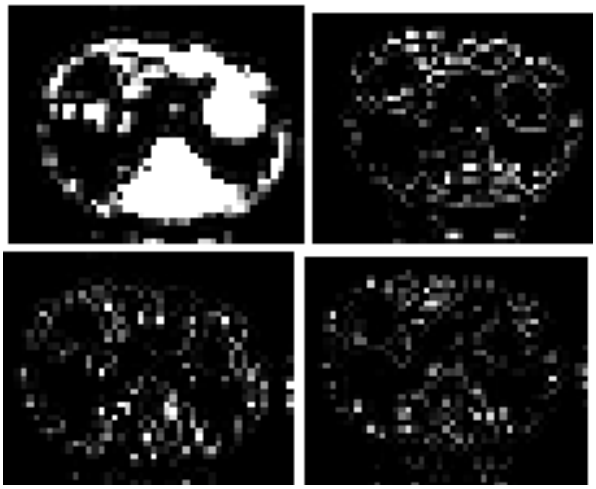


Fig. 4. Threshold Image.

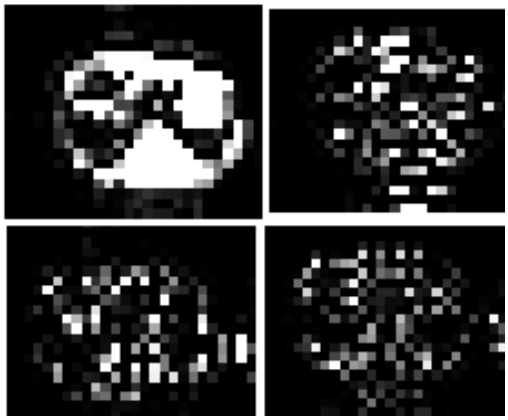
The proposed system uses threshold image information as the preliminary feature for further subjecting to be decomposed. After the thresholding of the image is carried out, the next step was to obtain the decomposition of the image using discrete wavelet transform. The proposed system uses Daubechies approach in order to perform decomposition supportable up to 3 levels. Fig.5 highlights the visuals of all the three levels of decomposition to be considered for further dimensional reduction.



(a) Level-1 Decomposition



(b) Level-2 Decomposition



(c) Level-3 Decomposition

Fig. 5. Decomposed Image.

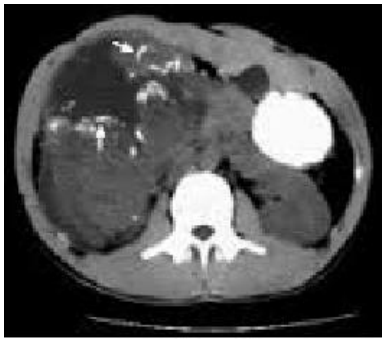


Fig. 6. Final Classified Image.

The decomposed image is further subjected to principal component analysis using algorithm-2, where the prominent intention is to retrieve certain statistical information e.g. mean, standard deviation, entropy, root mean square value, variance, smoothness, kurtosis, skewness, contrast, correlation, energy, and homogeneity. All these statistical data are computed and easy to assess owing to its numerical significance with the probability theory. Finally supervised algorithm-3 is implemented in order to obtain the classified image to be malignant (or else benign otherwise) (Fig.6). The average time taken in order to perform classification is approximately 0.36551 seconds on core i7 processor.

The study outcome of the proposed system is also benchmarked considering various forms of the machine learning techniques that assists in performing classification with respect to kernel e.g. quadrature, neural network, polynomial, linear, and radial basis function. The numerical assessment of the study was carried out with respect to sensitivity, specificity, precision, and accuracy.

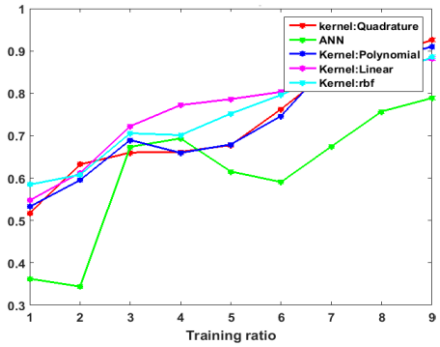



Fig. 7. Analysis of Sensitivity.

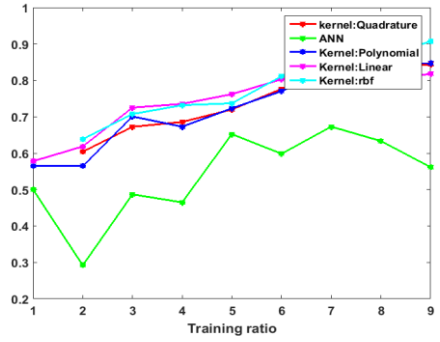



Fig. 8. Analysis of Specificity.

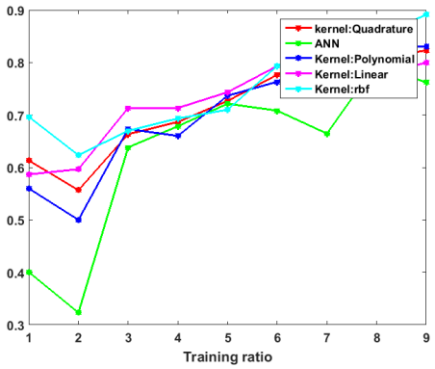



Fig. 9. Analysis of Precision.

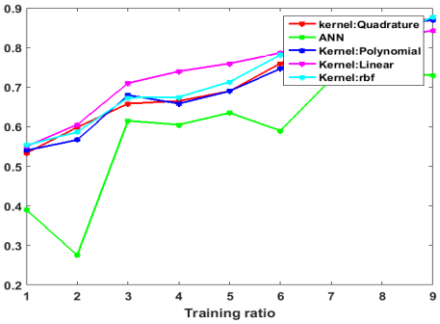



Fig. 10. Analysis of Accuracy.

Fig.7 highlights the sensitivity analysis or the true positive factor for the image queried and trained image within the database. It shows that sensitivity for linear-based kernel shows better performance altogether. This trend is followed by radial-basis function. However, there is no significant difference is found between the quadrature-based and polynomial based machine learning mechanism. The outcome directly infers that better identification as well as classification can be carried out using linear-based kernel system for the machine learning approach. The overall performance of neural network is not optimal and there are various fluctuating points owing to failure in exploring better match of elite result with activation function. Moreover achieving convergence is too slow in neural network for the given scenario and hence we recommend insufficiency of neural network in order to perform classification. Nearly similar performance can be seen in the specificity trends in Fig.8 i.e. true negative factor. Although, the specificity trend of neural network is low, but is not much recommended owing to time consumption in the classification process which is definitely not scalable. Although, the trend of linear-based kernel is found little increasing in beginning, but it is found to lower down the true negatives at the end to a large extent. The performance of polynomial and quadrature is nearly similar just like that of sensitivity. The precision and accuracy curve is shown in Fig.9 and Fig.10 which is also in the agreement with sensitivity and specificity curve. The overall performance of linear-based kernel system is found to offer higher precision and accuracy, while the trend for quadrature and polynomial remains the same. Therefore, the inference of the entire numerical benchmarked outcome is that-Support Vector Machine definitely proves as a best classifier in comparison to the frequently used neural network.Compared to existing ANN approaches, the proposed Kernel approaches have better sensitivity (ranging between 87-93%), specificity (ranges 82-92%), precision (of 80%-85) and accuracy (of 87%-88%) approx. Thus the Kernel approaches are significant than ANN approaches (sensitivity of 79%, specificity of 57%, precision of 75% and accuracy of 72%).It is also now proven that there is significant difference in adoption of 4 different kernel system of support vector machine, where linear-based kernel has proven to offer better classification performance for complex forms of brain tumor images. Also, recommend for usage of polynomial and radial basis function that also offers better significance towards accuracy performance when it comes to classification problems. Therefore, the proposed system offers a simple but yet novel classification techniques using supervised learning approach for efficient binary categorization of malignancy and benign state of brain tumor.

## VII. Conclusion

There are various schemes implemented for enhancing the classification performance towards identifying the malignancy or benign state of brain tumor. However, existing systems are also associated with research problems that lead to start of proposed investigation. We find that it is essential to emphasize on the preprocessing that can potentially lead to minimization of the many effort of accurate classification in the latter stage. We also believe that different preprocessing has different effect on classification performance. However, at present, we generalize this preprocessing and then emphasize on the latter stage of it. The preprocessed image is subjected to binarized thresholding scheme followed by applying of wavelet transform in order to accomplish more number of enriched features. We further apply simple process of dimensional reduction followed by supervised learning. The study outcome shows different forms of performance on different form of kernel system and is found to be much better in comparison to existing technique of neural network.

The future researches can consider this work as base for classifying the malignancy and benign state of brain tumor by using supervised learning approach which may bring higher accuracy in classification of tumor. Also, this work can be considered with diagnosis time, and sensitivity for brain MRI classification.

## Acknowledgement

I thank Dr. Rajshekar M, radiologist who helped me in carrying out research work by giving real time images.

## References

[1] E. Hattingen, U. Pilatus, Brain Tumor Imaging, Springer, 2015

[2] R. R. Colen, Imaging of Brain Tumors, An Issue of Magnetic Resonance Imaging Clinics of North America, E-Book, Elsevier Health Sciences, 2016

[3] M. M Kaila, R. Kaila, Quantum Magnetic Resonance Imaging Diagnostics of Human Brain Disorders, Elsevier, 2010

[4] J. Hodler, G. K. von Schulthess, C. L. Zollikofer, Diseases of the Brain, Head & Neck, Spine, Springer Science & Business Media, 2012

[5] H. Huang, J. P. Owen, P. Mukherjee, Neuroanatomy of Human Brain Development, Frontiers Media, 2017

[6] J. Sachdeva, V. Kumar, I. Gupta, N. Khandelwal and C. K. Ahuja, "Multiclass Brain Tumor Classification Using GA-SVM," 2011 Developments in E-systems Engineering, Dubai, 2011, pp. 182-187.

[7] R. Sonavane and P. Sonar, "Classification and segmentation of brain tumor using Adaboost classifier," 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), Jalgaon, 2016, pp. 396-403

[8] L. Farhi and A. Yusuf, "Comparison of brain tumor MRI classification methods using probabilistic features," 2017 13th IASTED International Conference on Biomedical Engineering (BioMed), Innsbruck, Austria, 2017, pp. 55-62

[9] H. B. Nandpuru, S. S. Salankar and V. R. Bora, "MRI brain cancer classification using Support Vector Machine," Electrical, Electronics and Computer Science (SCEECS), 2014 IEEE Students' Conference on, Bhopal, 2014, pp. 1-6

[10] M. Sornam, M. S. Kavitha and R. Shalini, "Segmentation and classification of brain tumor using wavelet and Zernike based features on MRI," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, 2016, pp. 166-169.

[11] N. B. Bahadure, A.K. Ray and H. P. Thethi, "Research Article Image Analysis for MRI Based Brain Tumor Detection and Feature Extraction Using Biologically Inspired BWT and SVM", Hindawi International Journal of Biomedical Imaging, pp. 12, 2017

[12] J. Cong, B. Wei, Y. He, Y. Yin and Y. Zheng, "A Selective Ensemble Classification Method Combining Mammography Images with Ultrasound Images for Breast Cancer Diagnosis", Computational and Mathematical Methods in Medicine, pp. 7, 2017

[13] I. I. Esener, S. Ergin, and T. Yuksel, "Research Article A New Feature Ensemble with a Multistage Classification Scheme for Breast Cancer Diagnosis", Journal of Healthcare Engineering, pp. 15, 2017

[14] M. M. Jadoon, Q. Zhang, I. Ul Haq, S. Butt, and A. Jadoon, "Research Article Three-Class Mammogram Classification Based on Descriptive CNN Features", BioMed Research International, pp. 11, 2017

[15] D. Ravì, H. Fabelo, G. M. Callic and G. Z. Yang, "Manifold Embedding and Semantic Segmentation for Intraoperative Guidance With Hyperspectral Brain Imaging," in *IEEE Transactions on Medical Imaging*, vol. 36, no. 9, pp. 1845-1857, Sept. 2017.

[16] Q.Z. Song, L. Zhao, X.K. Luo, and X.C.Dou, "Research Article Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images", Journal of Healthcare Engineering, pp. 7, 2017

[17] T. H. Vu, H. S. Mousavi, V. Monga, G. Rao and U. K. A. Rao, "Histopathological Image Classification Using Discriminative Feature-Oriented Dictionary Learning," in *IEEE Transactions on Medical Imaging*, vol. 35, no. 3, pp. 738-751, March 2016.

[18] M. Xi, J. Sun, L. Liu, F. Fan and X. Wu, "Research Article Cancer Feature Selection and Classification Using a Binary Quantum-Behaved Particle Swarm Optimization and Support Vector Machine", Computational and Mathematical Methods in Medicine, pp. 9, 2016

[19] L. Zhao and K. Jia, "Research Article Multiscale CNNs for Brain Tumor Segmentation and Diagnosis", Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine, pp. 7, 2016

[20] K. D.Marcomini, A.A. O. Carneiro and H. Schiabel1, "Research Article Application of Artificial Neural Network Models in Segmentation and Classification of Nodules in Breast Ultrasound Digital Images", Hindawi Publishing Corporation International Journal of Biomedical Imaging, pp. 13, 2016

[21] C. Barbalata and L. S. Mattos, "Laryngeal Tumor Detection and Classification in Endoscopic Video," in *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 1, pp. 322-332, Jan. 2016.

[22] V. Anitha and S. Murugavalli, "Brain tumour classification using two-tier classifier with adaptive segmentation technique," in *IET Computer Vision*, vol. 10, no. 1, pp. 9-17, 2 2016.

[23] T. Kaur, B. S. Saini and S. Gupta, "Quantitative metric for MR brain tumour grade classification using sample space density measure of analytic intrinsic mode function representation," in *IET Image Processing*, vol. 11, no. 8, pp. 620-632, 8 2017.

[24] S. L. Jui *et al*., "Brain MRI Tumor Segmentation with 3D Intracranial Structure Deformation Features," in *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 66-76, Mar.-Apr. 2016.

[25] S. Huda, J. Yearwood, H. F. Jelinek, M. M. Hassan, G. Fortino and M. Buckland, "A Hybrid Feature Selection With Ensemble Classification for Imbalanced Healthcare Data: A Case Study for Brain Tumor Diagnosis," in *IEEE Access*, vol. 4, no. , pp. 9145-9154, 2016.

[26] C-M. Svensson, R. Hübler and M.T. Figge, "Research Article Automated Classification of Circulating Tumor Cells and the Impact of Interobsever Variability on Classifier Training and Performance", Hindawi Publishing Corporation Journal of Immunology Research, pp. 9, 2015

[27] R. R. Sharma and P.Marikkannu, "Research Article Hybrid RGSA and Support Vector Machine Framework for Three-Dimensional Magnetic Resonance Brain Tumor Classification", Hindawi Publishing Corporation e Scientific World Journal, pp. 14, 2015

[28] A. Chaddad, "Research Article Automated Feature Extraction in Brain Tumor by Magnetic Resonance Imaging Using Gaussian Mixture Models", Hindawi Publishing Corporation International Journal of Biomedical Imaging, pp. 11, 2015

[29] I. Bukovsky, N. Homma, K. Ichiji, M. Cejnek, M. Slama, P.M. Benes and J. Bila, "Research Article A Fast Neural Network Approach to Predict Lung Tumor Motion during Respiration for Radiation Therapy Applications", Hindawi Publishing Corporation BioMed Research International, pp. 13, 2015

[30] M. E. Acevedo, M. A. Acevedo and F. Felipe, "Research Article Classification of Cancer Recurrence with Alpha-Beta BAM", Hindawi Publishing Corporation , Mathematical Problems in Engineering, pp. 14, 2009

[31] K. Kothavari, B. Arunadevi, and S. N. Deepa, "Research Article A Hybrid DE-RGSO-ELM for Brain Tumor Tissue Categorization in 3D Magnetic Resonance Images", Hindawi Publishing Corporation Mathematical Problems in Engineering, pp. 14, 2014

[32] X. Chen and C. Jian, "A tumor classification model using least square regression," *2014 10th International Conference on Natural Computation (ICNC)*, Xiamen, 2014, pp. 753-758.

[33] M. Huang, W. Yang, Y. Wu, J. Jiang, W. Chen and Q. Feng, "Brain Tumor Segmentation Based on Local Independent Projection-Based Classification," in *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 10, pp. 2633-2645, Oct. 2014

[34] M. Huml, R. Silye, G. Zauner, S. Hutterer and K. Schilcher, "Research Article Brain Tumor Classification Using AFM in Combination with Data Mining Techniques", Hindawi Publishing Corporation BioMed Research International, pp. 11, 2013

[35] J. J. Corso, E. Sharon, S. Dube, S. El-Saden, U. Sinha and A. Yuille, "Efficient Multilevel Brain Tumor Segmentation With Integrated Bayesian Model Classification," in *IEEE Transactions on Medical Imaging*, vol. 27, no. 5, pp. 629-640, May 2008.

[36] M. Saddique, J. H. Kazmi and K. Qureshi, "Research Article A Hybrid Approach of Using Symmetry Technique for Brain Tumor Segmentation", Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine, pp. 10, 2014

[37] A. P. Nanthagopal and R. Sukanesh, "Wavelet statistical texture features-based segmentation and classification of brain computed tomography images," in *IET Image Processing*, vol. 7, no. 1, pp. 25-32, February 2013.

[38] Z. L. Sun, C. H. Zheng, Q. W. Gao, J. Zhang and D. X. Zhang, "Tumor Classification Using Eigengene-Based Classifier Committee Learning Algorithm," in *IEEE Signal Processing Letters*, vol. 19, no. 8, pp. 455-458, Aug. 2012.

[39] D. F. Brougham, G. Ivanova, M. Gottschalk, D.M. Collins, A. J. Eustace, R. O'Connor and J. Havel5, "Research Article Artificial Neural Networks for Classification inMetabolomic Studies ofWhole Cells Using 1H NuclearMagnetic Resonance", Hindawi Publishing Corporation, Journal of Biomedicine and Biotechnology, pp. 8, 2011

[40] C. H. Zheng, T. Y. Ng, L. Zhang, C. K. Shiu and H. Q. Wang, "Tumor Classification Based on Non-Negative Matrix Factorization Using Gene Expression Data," in *IEEE Transactions on NanoBioscience*, vol. 10, no. 2, pp. 86-93, June 2011.

[41] P. Asthana and S. Vashisth, "Segmentation of brain tumor from MRI images," *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, Gurgaon, 2017, pp. 262-266.

# Quality Flag of GOSAT/FTS Products Taking into Account Estimation Reliability

Kohei Arai[1], Takashi Higuchi[2], Hiroshi Okumura[3]
Department of Information Science
Saga University, Saga City, Japan

Hirofumi Ohyama[4]
NIES: National Institute of Environmental Studies, Tsukuba Japan

Shuji Kawakami[5], Kei Shiomi[6]
JAXA: Japan Aeronautics Exploration Agency, Tsukuba, Japan

***Abstract*—Quality and cloud flags of GOSAT/FTS: Fourier Transform Spectrometer onboard Greenhouse gasses Observation Satellite products taking into account cirrus clouds and thick aerosols are considered and proposed. Influence due to cirrus and thick aerosol on estimation of column $CO_2$ and $CH_4$ with GOSAT/FTS data is clarified. Relatively large estimation errors are observed in column $CO_2$ and $CH_4$ retrievals with FTS data in some atmospheric conditions. In order to find such cases, retrieval results and quality/cloud flags in the GOSAT/FTS data products are checked. Through the investigation, it is found that relatively large error is caused by convergence problem due to cirrus clouds and thick aerosols. In the proposed paper, some of the cases of which relatively large estimation error is occurred at the Saga TCCON (The Total Carbon Column Observing Network) site are investigated. Also, a comparative study is conducted between standard products provided by NASA/JPL and the Levenberg-Marquardt based least square method of column $CO_2$ and $CH_4$ retrieval. It is suggested that some improvements of estimation accuracy of column $CO_2$ and $CH_4$ retrieval with GOSAT/FTS data can be expected.**

***Keywords*—Levenberg-Marquardt; FTS; GOSAT; aerosol; cirrus cloud***

## I. Introduction

GOSAT: Greenhouse gases Observing Satellite is a greenhouse gas observation technology satellite ("IBUKI" in Japanese) 1 jointly developed by the Ministry of the Environment, the National Institute for Environmental Studies (NIES), and the Japan Aerospace Exploration Agency (JAXA). Observe the concentration distribution of greenhouse gases such as carbon dioxide and methane gas, which are considered to be the cause of global warming, from outer space. On January 23, 2009, it was launched from the Tanegashima Space Center at H-IIA Launch Vehicle No. 15. Starting acquiring observation data from February 2009, in May 2009 uncalibrated value analysis results on a global scale have been published.

GOSAT carries TANSO - FTS (TANSO: Thermal And Near Infrared Sensor for Carbon Observation, FTS: Fourier Transform Spectrometer)[2] is a main IBUKI sensor measuring carbon dioxide and methane gas. We observe spectra of solar light (near infrared rays) reflected by the earth surface and spectra of the light radiated from the earth's atmosphere and the ground surface (far infrared rays) by Fourier spectroscopy[3].

Since carbon dioxide and methane present in the atmosphere have the property of absorbing light of a specific wavelength, the amount of carbon dioxide and methane that existed in the path of light, depending on the degree of absorption of light transmitted through the atmosphere, It can be calculated. Absorption band of carbon dioxide around 1.6 μm and around 2.0 μm is important as a wavelength band containing much information near the ground surface. On the other hand, the absorption band around 14 μm is mainly used to obtain altitude information higher than 2 km. The accumulated amount of carbon dioxide gas is measured by short wavelength infrared band 1 to 3 (SWIR: Short Wave Infrared Band 1 - 3), and the vertical concentration distribution of carbon dioxide is measured by thermal infrared band 4 (TIR: Thermal Infrared Band 4). For conditions without error factors such as clouds and aerosols, we aim to measure errors within 1%.

GOSAT also carries TANSO-CAI (CAI: Cloud and Aerosol Imager)[4] is an image sensor used for determination of the presence or absence of clouds and measurement of aerosol (atmospheric particulate matter), which is an error factor when measuring carbon dioxide at TANSO-FTS. It is a secondary sensor of IBUKI. It is used for correction of measured data obtained by TANSO-FTS.

CAI observes the state of the atmosphere and the surface of the earth as an image in the daytime. The presence or absence of clouds in a wide range including the field of view of FTS is judged from the observation data, and in the case of aerosol or thin clouds, the characteristics of the cloud and the amount of aerosol are calculated. These information are used to correct the influence of clouds and aerosols contained in the spectrum obtained from FTS.

Development of transportable Lidar[5] for validation of GOSAT satellite data products is well reported [1]. Also, advanced validation of the GOSAT observed $CO_2$ and $CH_4$ at TCCON (The Total Carbon Column Observing Network)[6] and prioritized observation sites is reported [2]. Meanwhile, Observations of $XCO_2$ and $XCH_4$ (total column $CO_2$ and $CH_4$

[1] http://www.gosat.nies.go.jp/
[2] http://www.eorc.jaxa.jp/GOSAT/instrument_1.html
[3] https://en.wikipedia.org/wiki/Fourier-transform_spectroscopy
[4] http://www.eorc.jaxa.jp/GOSAT/instrument_2_j.html
[5] https://ja.wikipedia.org/wiki/LIDAR
[6] https://tccon-wiki.caltech.edu/

$CO_2$ and $CH_4$) with ground-based high-resolution FTS at Saga, Japan and comparisons with GOSAT products is discussed [3]. On the other hand, evaluation of cirrus cloud detection accuracy of GOSAT/CAI and Landsat-8 OLI[7] with laser radar: lidar and confirmation with CALIPSO/CALIOP [8] data is conducted [4] together with a comparative study on cloud parameter estimation among GOSAT/CAI, MODIS [9], CALIPSO/CALIOP and Landsat-8/OLI with laser radar as truth data [5].

Advanced validation of the GOSAT observed $CO_2$ and $CH_4$ at priotized observation sites is reported [6]. On the other hand, Web based data acquisition and management system for GOSAT validation Lidar data analysis is developed [7]. Meanwhile, observation of aerosol parameters at Saga using GOSAT product validation Lidar is discussed [8]. Meantime, improvement of web-based data acquisition and management system for GOSAT validation Lidar data analysis (2013) is also discussed [9] together with observation of aerosol properties at Saga using GOSAT product validation LiDAR [10].

Impact of aerosol and cirrus clouds on the GOSAT observed $CO_2$ and $CH_4$ inferred from ground based lidar, sky-radiometer [10] and FTS data at prioritized observation sites,(2013) is clarified [11]. Comparison of lower tropospheric ozone column observed by DIAL: Differential Absorption Lidar [11] and GOSAT TANSO-FTS TIR is made [12]. Furthermore, impact of aerosols and cirrus on the GOSAT onboard $CO_2$ and $CH_4$ inferred from ground based Lidar, sky-radiometer and FTS data at prioritized observation sites is discussed [13]. Then, on the validation results of five year GOSAT SWIR $XCO_2$ and $XCH_4$ (total column $CO_2$ and $CH_4$ ) data is well reported [14].

Lidar observations at priotized sites for GOSAT validation is reported [15] together with lidar observation at TCCON site to investigate the influence of particles for GOSAT data [16]. On the other hand, influence due to cirrus and thick aerosol on estimation of column $CO_2$ and $CH_4$ with GOSAT/FTS data is discussed [17].

From the methodological point of view, estimation of $XCO_2$ and $XCH_4$ can be done with FTS data based on least square method. FTS data, however, is affected by clouds and thick aerosols. In order to avoid the cloud influence, clouds are detected with CAI imagery data then the FTS data which suffered from the clouds are identified with cloud flag. It, however, does not work properly. Thin cirrus type of clouds cannot be detected with CAI instrument. Other than this, quality flag can be used for analysis of the FTS data in concern. FTS data seems to also be suffered from thick aerosols. Such these FTS data are not reliable enough for analysis. Therefore, it is desirable to reconsider the method for reliability flag rather than the present quality flag.

Past FTS data are checked from the point of view of reliability. By using TCCON site of calibration site data, it is possible to check the estimation accuracy (discrepancy between the true and the estimated $XCO_2$ and $XCH_4$). Through careful check of the reliability, it is found that there is relation between reliability and clouds and thick aerosol. This paper intends to discuss this matter and suggest some reliability flag to the FTS data in concern.

Due to the fact that GOSAT/FTS does not cover the wavelength of water vapor absorption, it is hard to estimate influence of water vapor and clouds on estimation of $XCO_2$ and $XCH_4$. Therefore, it is required to use other sensor data than the GOSAT/FTS data for estimation of water vapor, aerosol, and clouds influences..

The following section describes current status of the GOSAT/FTS derived $XCO_2$ and $XCH_4$ followed by the relation between discrepancy between the true and the estimated $XCO_2$ and $XCH_4$ as well as clouds and thick aerosols measured with sky-radiometer. Then, some methods which allows to detect clouds and thick aerosols by using other data sources derived from the other mission instruments such as MODIS. After that, conclusion is described together with some discussions and with future research works.

## II. Research Background

### A. Present Status of GOSAT Project

GOSAT satellite is operated in good health condition. In order to calibrate GOSAT TANSO-FTS, TCCON sites are also operated. One of the TCCON sites is situated at Saga University in Saga, Japan. Saga ground based-FTS operation started from July 2011. GOSAT targets at Saga are acquired every 3 days. In conjunction of GOSAT operation at Saga, OCO-2: Orbiting Carbone Observatory-2[12] can target 2 days in 16-day revisit cycle of GOSAT satellite. Fig. 1 shows outlook of the Saga TCCON site.


Fig. 1. Outlook of the Saga TCCON site.

[7] https://en.wikipedia.org/wiki/Landsat_8
[8] https://www-calipso.larc.nasa.gov/
[9] https://modis.gsfc.nasa.gov/
[10] http://www.kippzonen.com/ProductGroup/31/Sky-radiometers
[11] https://www.esrl.noaa.gov/csd/groups/csd3/instruments/lidar/dial.html
[12] https://oco.jpl.nasa.gov/

From the roof of the FTS container situated at Saga University, Saga Japan, solar irradiance comes in the ground based FTS instrument. The FTS measures solar irradiance which is absorbed by atmospheric molecules, aerosol, water vapor. Based on Fourier spectrum analysis, $XCO_2$ and $XCH_4$ can be estimated. On the other hand, sky-radiometer and sky-view camera are situated at the top roof of the building at Saga University nearby FTS container. From the sky-radiometer, aerosol refractive index and size distribution can be estimated while existing visible clouds can be observed with sky-view camera.

Fig. 2 shows an example of GOSAT observation pattern around Japanese vicinity. Along with the GOSAT satellite track, grid observation can be done while target areas can also be observed in GOSAT operation. Fig. 2 also shows the location of Saga TCCON site. From the 2011 to up to now, not only $XCO_2$ and $XCH_4$ but also XCO, $XN_2O$ can be estimated.

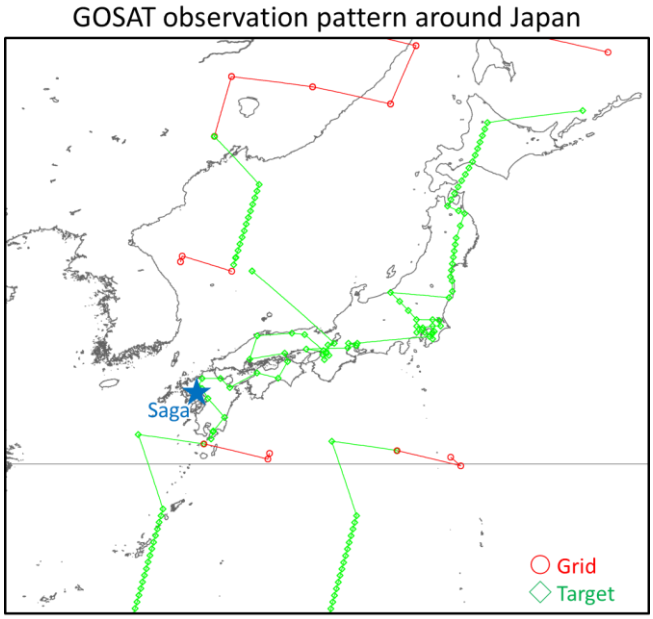



Fig. 2. Example of GOSAT observation pattern around Japanese vicinity.

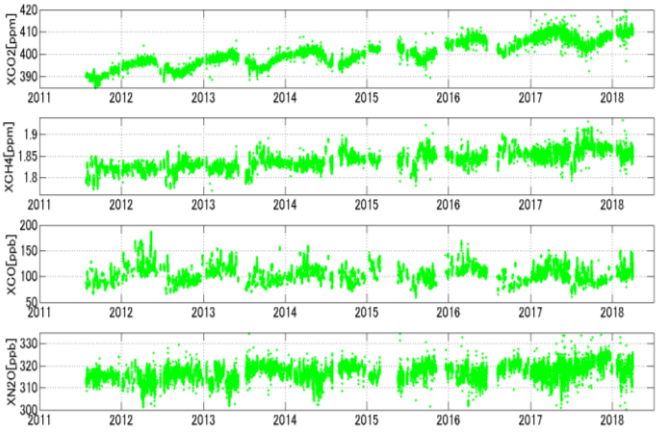



Fig. 3. Trends of the $XCO_2$ (upper), $XCH_4$ (second), XCO (third), $XN_2O$ (bottom)

Fig. 3 shows the trends of the $XCO_2$ (in unit of ppm), $XCH_4$ (in unit of ppm), XCO (in unit of ppb), $XN_2O$ (in unit of ppb) which are estimated with GOSAT FTS data. On the other hand, Fig. 4 shows $XCO_2$ and $XCH_4$ estimated from the ■NIES GOSAT V02.72 observation and from the ground based FTS observation at ●Saga-TCCON site. Overall trends of $XCO_2$ and $XCH_4$ estimated with GOSAT-FTS data and the ground based FTS show almost same.

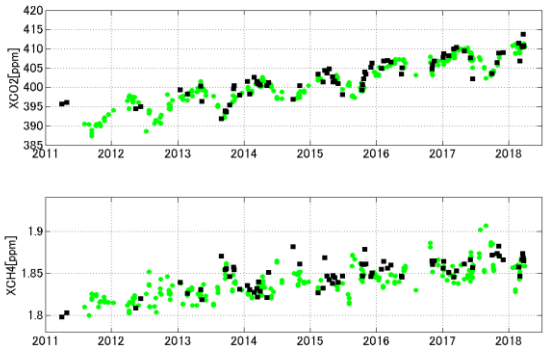
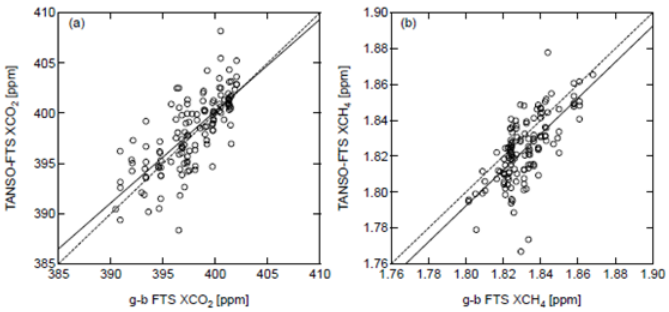



Fig. 4. $XCO_2$ and $XCH_4$ estimated from the ■NIES GOSAT V02.72 observation and from the ground based FTS observation at ●Saga-TCCON site

There are five major datasets, ACOS/GOSAT, AIRS/AQUA, OCO-2/OCO-2, TES/AURA, TCCON (ground based FTS)[13].

Correlation plots of NIES GOSAT v02 $XCO_2$ and $XCH_4$ and ground based FTS data derived $XCO_2$ and $XCH_4$ are shown in Fig. 5. The ground based FTS within ±30min. of GOSAT overpass time are averages. Solid and dashed lines denote linear fit with an intercept of 0 and 1-to-1 line respectively.




(a) $XCO_2$ (b)$XCH_4$

Fig. 5. Correlation plots of NIES GOSAT v02 $XCO_2$ and $XCH_4$

Average differences as (TANSO-FTS minus ground-based FTS) are as follows,

$XCO_2$ : 0.40 +/- 2.51 ppm

$XCH_4$ : -7.6 +/- 13.7 ppb

As shown in Fig. 4 and 5, there are some significant discrepancy between GOSAT/FTS derived and the ground based FTS derived $XCO_2$ and $XCH_4$. In particular, relatively

[13] https://co2.jpl.nasa.gov/#mission=ACOS

large difference between both is observed at the specific days. Therefore, careful check at the specific dates of estimated $XCO_2$ and $XCH_4$ has to be done.

### B. Check the GOSAT Products, Flags

Flag information is stored in the analyzed data. There is a high possibility that the estimation does not go well if the flag $\neq$ 0. For instance, Fig. 6 shows an example of a relation between sky-view camera images and Saga-TCCON flag. As shown in Fig. 6, flag=0 for the clear sky condition (a) while flag=33 for the partly cloudy condition (b).

(a)js20170401saee0a.0020 (b) js20170401saee0a.0062

Flag=0 Flag=33

Fig. 6. Example of a relation between sky-view camera images and Saga-TCCON flag.

Fig. 7 shows a result from the comparison between the changes of solar irradiation data (red) and Saga-TCCON flag (blue).

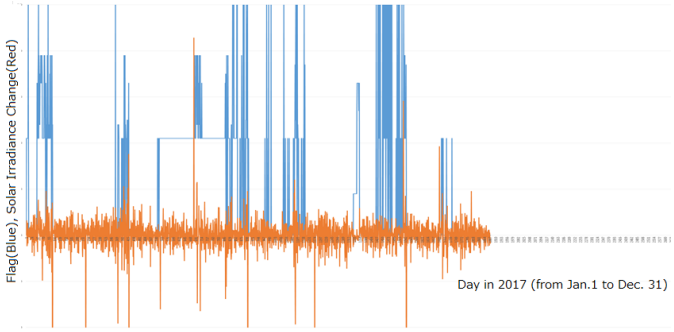



Fig. 7. Result from the comparison between the changes of solar irradiation data (red) and Saga-TCCON flag (blue).

Coincidentally, the flag is large when the solar irradiation changes a lot. When the weather condition is changed from clear sky to cloudy, solar irradiation is also changed a lot. Therefore, both are occurred coincidentally. Table 1 shows the relation between Saga-TCCON flag and Mean and Standard deviation of solar irradiation as well as correlation coefficient between Saga-TCCON flag and solar irradiation.

TABLE I. RELATION BETWEEN SAGA-TCCON FLAG AND MEAN AND STANDARD DEVIATION OF SOLAR IRRADIATION AS WELL AS CORRELATION COEFFICIENT BETWEEN SAGA-TCCON FLAG AND SOLAR IRRADIATION.

| Saga-TCCON flag | Off | On |
|---|---|---|
| Mean of daily changes of solar irradiation | 1.979 | 2.472 |
| Standard deviation | 3.340 | 4.218 |
| Correlation | 0.481 | 0.580 |

Saga-TCCON flag is on when the solar irradiation is relatively large. Therefore, there is a possibility to find estimated $XCO_2$ and $XCH_4$ with relatively poor accuracy by taking into account the changes in solar irradiation.

### C. Comparison of Estimated $XCO_2$ and $XCH_4$ by the Proposed Least Square Method based on the Levenberg-Marquardt with the Conventional Method

Typical cases of estimation accuracy is relatively good (fine weather: clear sky condition) and of estimation error seems to be large comparatively (cloudy condition) are selected. Those are FTS data of April 4, 12, 13, 19, and 24 2018. Fig. 8 shows sky-view camera images (Top left: April 4, Top right: April 12, Middle left: April 13, Middle right: April 19 and Bottom: April 24). The left side image shows cloud condition in the morning while the right side image shows cloud condition in the afternoon. It is clear that it is clear in the morning on April 4, 13 and 24 while there are some clouds in the afternoon on April 4, 12, 24. In accordance with the Saga brunch of the JMA: Japan Meteorological Agency, CC: Cloud Coverage at 12:00 and 15:00 of each day is as follows,

April 4: 2, 7. April 12: 2, 3. April 13: 0, 7. April 19: 3, 7. April 24: 0, 8.

It, however, is not always true. There are some visible cirrus clouds in the morning on April 12 clearly. Therefore, FTS data derived $XCO_2$ and $XCH_4$ on April 12 is compared to that of April 13.

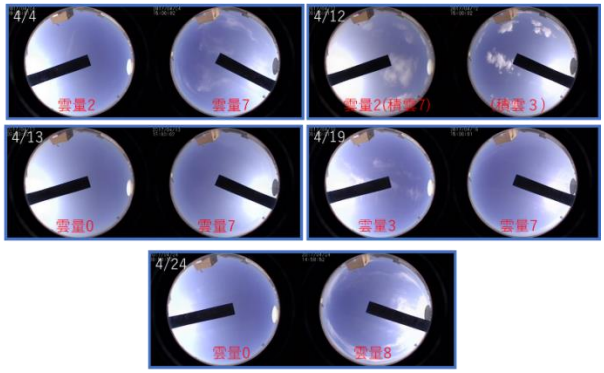



Fig. 8. Sky-view camera images (Top left: April 4, Top right: April 12, Middle left: April 13, Middle right: April 19 and Bottom: April 24)

As mentioned in the previous section, the conventional method for estimation of $XCO_2$ and $XCH_4$ is called GGG algorithm. Table 2 shows a comparison of mean and standard deviation of the estimated $XCO_2$ between GGG and the proposed least square method based on the Levenberg-Marquardt of nonlinear least square method. As shown in Table 2, relatively large discrepancy between the estimated $XCO_2$ is observed during 9 a.m. and 11:30 a.m. on April 12. Those estimated $XCO_2$ is shown in Fig. 9. In the figure, black dots denote $XCO_2$ of GGG while blue dots denote $XCO_2$ of the proposed least square method.

These discrepancy is caused by the difference between algorithms of the proposed least square and GGG. Therefore, there is a possibility to check the reliability of estimated $XCO_2$ by using the discrepancy between both.

TABLE II. COMPARISON OF MEAN AND STANDARD DEVIATION OF THE ESTIMATED $XCO_2$ BETWEEN GGG AND THE PROPOSED LEAST SQUARE METHOD

| | April 4 | April 12 | April 13 | April 19 | April 24 |
|---|---|---|---|---|---|
| Mean (Proposed) | 415.98 | 373.64 | 415.23 | 413.6 | 413.33 |
| St.Dev.(Proposed) | 0.81 | 37.58 | 0.93 | 0.59 | 1.11 |
| Mean (GGG) | 409.45 | 404.21 | 410.69 | 408.66 | 410.47 |
| St.Dev.(GGG) | 0.52 | 5.92 | 1.16 | 1.18 | 1.05 |

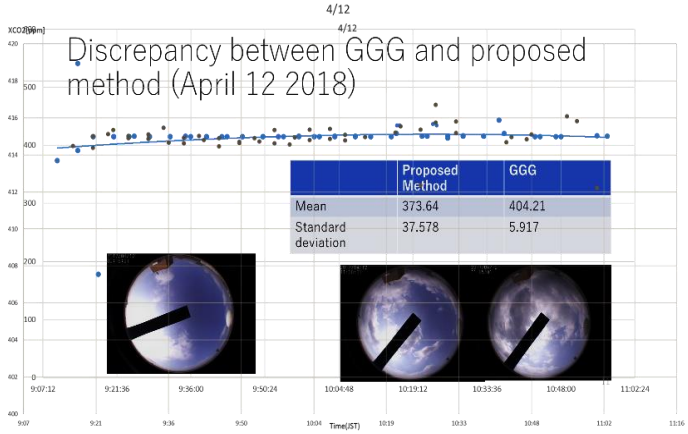



Fig. 9. Estimated $XCO_2$ derived from GGG and from the proposed least square method (Black dots denote $XCO_2$ of GGG while Blue dots denote $XCO_2$ of the proposed least square method.

### D. *Possible Causes of Relatively Large $XCO_2$ Estimation Error*

Cloud condition on April 12 2018 is shown in Fig. 10. Cloud Coverage ratio (CC) at 9:00 is 7 (Cumulus clouds), then CC at 12:00 is 2 (Cirrus clouds), and CC at 15:00 is 3 (Cumulus clouds), respectively. Also, hourly Solar Irradiance (SI) at 9:00 is 1.22, then SI at 12:00 is 1.96, and SI at 15:00 is 2.72. Averaged CC of the day is 6 and total solar irradiance a day is 20.88 $MJ/m^2$.

Other than these, solar direct and diffuse irradiance are measured with sky-radiometer, POM-1 of sun-photometer and aureole-meter which is manufactured by Prede Co. Ltd. in Japan. Using this data, aerosol refractive index, size distribution, single scattering albedo, atmospheric optical depth, etc. can be estimated.

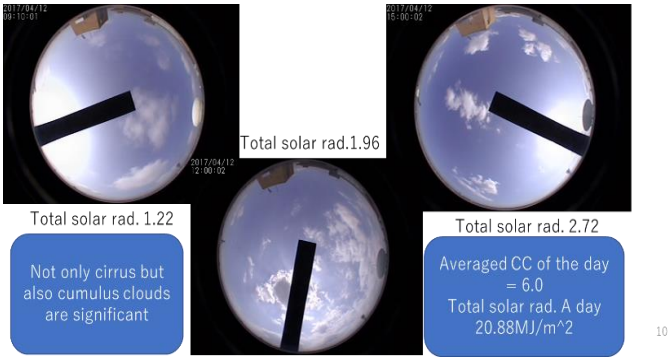



Fig. 10. Cloud condition on April 12 2018

TABLE III. SINGLE SCATTERING ALBEDO AND TOTAL ATMOSPHERIC OPTICAL DEPTH

(A)SINGLE SCATTERING ALBEDO

| Wavelength[μm] | April 4 | April 12 | April 13 | April 19 | April 24 |
|---|---|---|---|---|---|
| 0.38 | 0.841 | 0.826 | 0.846 | 0.787 | 0.808 |
| 0.4 | 0.961 | 0.927 | 0.960 | 0.908 | 0.954 |
| 0.5 | 0.965 | 0.948 | 0.989 | 0.938 | 0.953 |
| 0.675 | 0.962 | 0.948 | 0.990 | 0.932 | 0.952 |
| 0.87 | 0.750 | 0.854 | 0.813 | 0.897 | 0.717 |

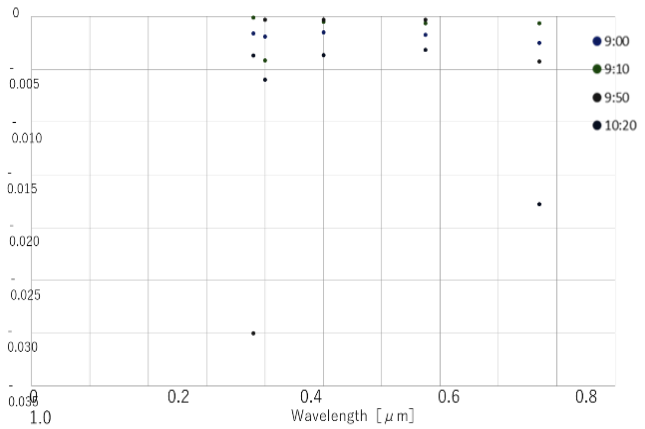
(b)Total Atmospheric Optical Depth

| Wavelength[μm] | April4 | April 12 | April 13 | April 19 | April 24 |
|---|---|---|---|---|---|
| 0.38 | 0.303 | 0.681 | 0.503 | 0.257 | 0.263 |
| 0.4 | 0.271 | 0.622 | 0.456 | 0.238 | 0.226 |
| 0.5 | 0.201 | 0.501 | 0.344 | 0.201 | 0.165 |
| 0.675 | 0.123 | 0.369 | 0.217 | 0.159 | 0.102 |
| 0.87 | 0.0995 | 0.312 | 0.171 | 0.145 | 0.0878 |

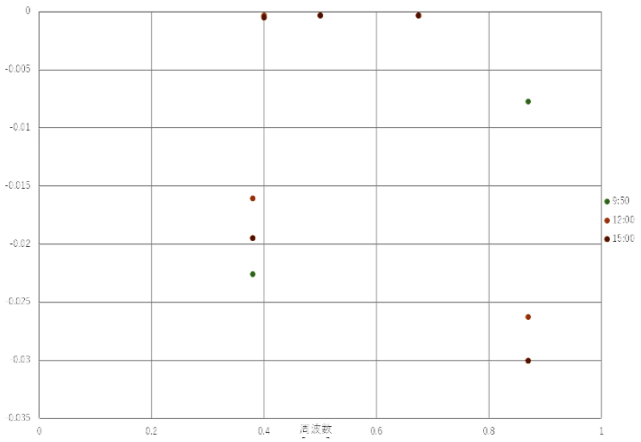
Single scattering albedo of each day is shown in Table 3 (a) while total atmospheric optical depth of each day is shown in Table 3 (b), respectively.

The single scattering albedo on April 12 is relatively small while the total atmospheric optical depth on April 12 is comparatively large. Therefore, absorption in the atmosphere is relatively large while comparatively thick atmosphere on April 12.

Imaginary part of aerosol refractive index estimated for 9:00 to 10:20 a.m. on April 12 is shown in Fig. 11 (a) while that of 9:50 to 15:00 on April 13 is shown in Fig. 11 (b), respectively. Not so significant difference is observed between both of imaginary part of aerosol refractive index estimated for April 12 and April 13. Also, relatively large changes are observed for imaginary part of aerosol refractive index during from 9:00 to 10:20 on April 12 while that for 9:50 to 15:00 on April 13 is very stable.



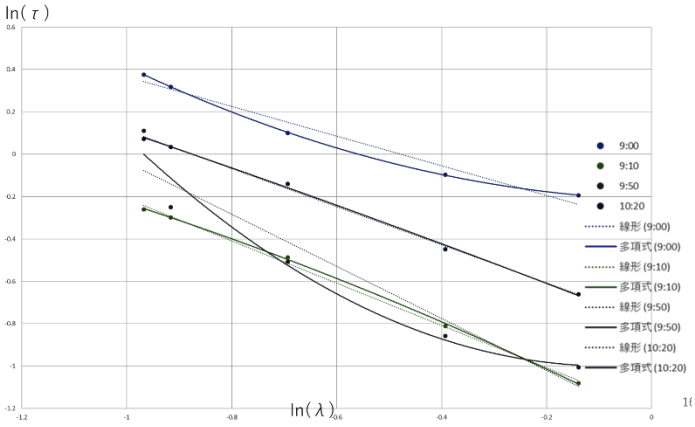

(a)April 12



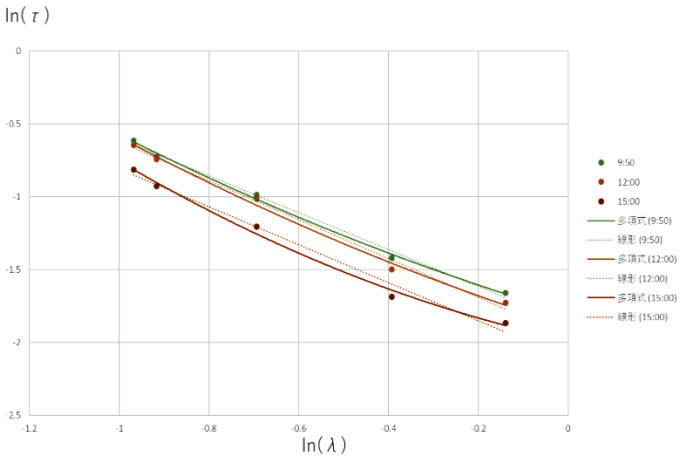

(b)April 13

Fig. 11. Estimated imaginary part of aerosol refractive index

On the other hand, real part of aerosol refractive index which is estimated for 9:00 to 10:20 on April 12 and 9:50 to 15:00 on April 13 is shown in Fig. 12 (a) and (b), respectively.



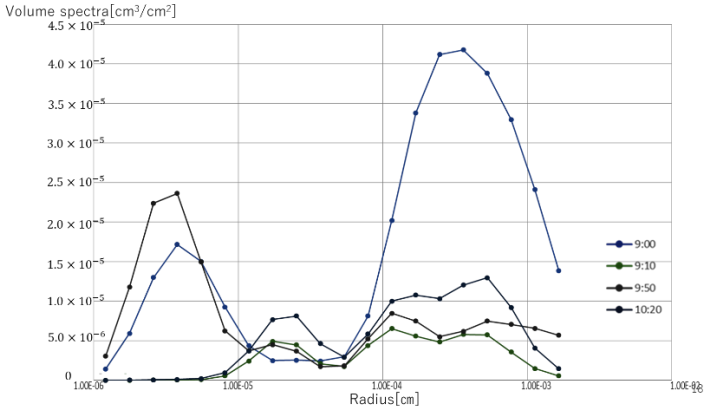

(a)April 12



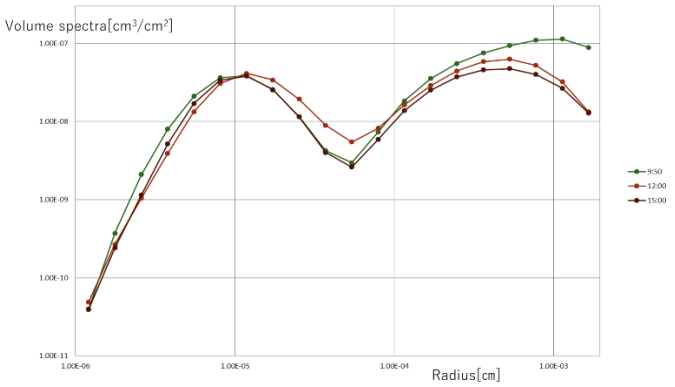

(b)April 13

Fig. 12. Real part of aerosol refractive index which is estimated for April 12 and April 13

Also, relatively large changes are observed for real part of aerosol refractive index during from 9:00 to 10:20 on April 12 while that for 9:50 to 15:00 on April 13 is very stable. Meanwhile, aerosol size distribution which is estimated for 9:00 to 10:20 on April 12 and 9:50 to 15:00 on April 13 is shown in Fig. 13 (a) and (b), respectively. Also, relatively large changes are observed for aerosol size distribution during from 9:00 to 10:20 on April 12 while that for 9:50 to 15:00 on April 13 is very stable. More importantly, aerosol size distribution on April 13 is bi-modal and stable during from 9:50 to 15:00 while that on April 12 is changed dramatically during from 9:10 to 9:50. It is thought type of aerosol particles are changed during the period.




(a)April 12

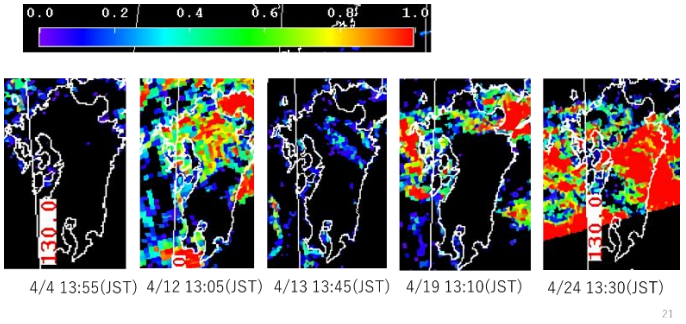



(b)April 13

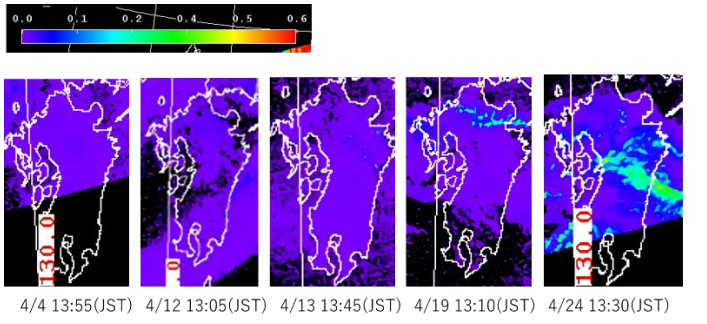
Fig. 13. Aerosol size distribution which is estimated for April 12 and April 13

### *E. Another Possible Method for Identification of Relatively Poor Estimation Accuracy of $XCO_2$*

Another possible method for identification of relatively poor estimation accuracy of $XCO_2$ is investigated. The most reliable method is comparison between ground based FTS data derived $XCO_2$ and $XCH_4$ and GOSAT/TANSO-FTS, obviously. It does work for the data location at TCCON sites. As mentioned in the previous section, most significant causes for poor accuracy of $XCO_2$ and $XCH_4$ estimations are clouds. Therefore, the method which uses sky-view camera data might be useful. It, however, can be used for the FTS data nearby TCCON sites. Also, TANSO-CAI is applicable to find clouds. It, however, does not work for cirrus cloud detection.




(a) Cloud reflectance




(b) Cirrus cloud reflectance

Fig. 14. MODIS data derived cloud and cirrus cloud reflectance

Therefore, the following method utilizing MODIS data derived cloud reflectance as well as cirrus cloud reflectance is proposed. Fig. 14 (a) shows cloud reflectance on April 4, 12, 13, 19, and 24 2018 while Fig. 14 (b) shows cirrus cloud reflectance on the same days, respectively. These reflectance are derived from AQUA/MODIS data. Therefore, it measures in the afternoon because local mean time of the AQUA satellite orbit is afternoon. It, however, shows great possibility of cirrus cloud detection.

## III. Conclusion

Quality and cloud flags of GOSAT/TANSO-FTS: Fourier Transform Spectrometer onboard Greenhouse gasses Observation Satellite products taking into account cirrus clouds and thick aerosols are considered and proposed. Influence due to cirrus and thick aerosol on estimation of column $CO_2$ and $CH_4$ with GOSAT/TANSO-FTS data is clarified. Relatively large estimation errors are observed in column $CO_2$ and $CH_4$ retrievals with FTS data in some atmospheric conditions.

In order to find such cases, retrieval results and quality/cloud flags in the GOSAT/TANSO-FTS data products are checked. Through the investigation, it is found that relatively large error is caused by convergence problem due to cirrus clouds and thick aerosols. In the paper, some of the cases of which relatively large estimation error is occurred at the Saga TCCON (The Total Carbon Column Observing Network) site are investigated. Also, a comparative study is conducted between standard products provided by NASA/JPL and the Levenberg-Marquardt based least square method of column $CO_2$ and $CH_4$ retrieval. It is suggested that some improvements of estimation accuracy of column $CO_2$ and $CH_4$ retrieval with GOSAT/TANSO-FTS data can be expected.

The purpose of this research is to develop a method for identifying poor accuracy of $XCO_2$ and $XCH_4$ estimation with GOSAT/TANSO-FTS data. The conventional method is to use the flag which reflect cloud detection with GOSAT/TANSO-CAI imagery data and quality of FTS data. It, however, GOSAT/TANSO-CAI does not work for cirrus clouds. Therefore, alternative methods are attempted and proposed here. One of those is to use MODIS data derived cloud reflectance and cirrus cloud reflectance. For the FTS data is acquired around TCCON sites, sky-view camera images are useful. Also, solar irradiance is useful for the FTS data is acquired around meteorological stations which allow measurement data of solar direct irradiance.

There are some reasons for relatively large discrepancy between GGG and the proposed $XCO_2$ estimation methods as well as between ground based measurement data and GOSAT based $XCO_2$ measurement data. Therefore, it is possible to create quality flag through comparisons of the estimated $XCO_2$ and $XCH_4$ derived from the different methods, GGG and Levenberg-Marquardt.

Cirrus cloud (cloud type, MODIS derived cloud reflectance including cirrus cloud reflectance) while, thick aerosol (Refractive index, size distribution, single scattering albedo, optical depth, Angstrom exponent, etc.). Therefore, there is a possibility to identify a poor accuracy of $XCO_2$ and $XCH_4$ estimation by using MODIS data derived cloud reflectance as well as cirrus cloud reflection. Cloud flag can be created with MODIS derived cloud and cirrus cloud reflectance as well as sky-radiometer data derived refractive index and size distribution together with sky view camera data

Further study is required for establishment of $XCO_2$ and $XCH_4$ retrieval method with a variety of remote sensing satellite imagery data.

## Acknowledgment

The author would like to tank Dr. Kurtis Thome of NASA/GSFC and Dr. Satoshi Tsuchida of Advanced Institute of Science and Technology for their valuable comments and suggestions as well as Dr. Liang Xing Ming of NOAA/NESDIS for their discussions.

## References

[1] Osamu Uchino, Tetsu Sakai, Tomohiro Nagai, Masahisa Nakasato, Isamu Morino, Tatsuya Yokota, Tsuneo Matsunaga, Nobuo Sugimoto, Kohei Arai, Hiroshi Okumura, Development of transportable Lidar for validation of GOSAT satellite data products, Journal of Remote Sensing Society of Japan, 31, 4, 435-443, 2011

[2] O.Uchino, I.Morino, Y.Yoshida, N.Kikuchi, M.Inoue, K.Nakano, T.Yokota, D.Wunch, P.Wennberg, G.Toon, J.Notholt, V.Sherlock, B.Liley, D.Grifith, S.Kawakami, H.Ohyama, T.Nagai, T.Sakai, Kohei Arai, H.Okumura, Advanced validation of the GOSAT observed CO2 and CH4 at TCCON and prioritized observation sites, Geophysical Research Abstract, Vol. 14, EGU 2012, General Assembly, 2012.

[3] H.Ohyam, S.Kawakami, Tetsu Tanaka, Isamu Morino, O.Uchino, Makoto Inoue, TSakai, T.Nagai, A.Yamazaki, A.Uchiyama, T.Fukamachi, M.Sakashita, T.Kawasaki, T.Akaho, Kohei Arai, H.Okumura, Observations of XCO2 and XCH4 with ground-based high-resolution FTS at Saga, Japan and comparisons with GOSAT products, Journal of Atmospheric Measurement Technics Discussion, 8, 8, 8257-8294, 2015

[4] Kohei Arai, Masanori Sakashita, Evaluation of Cirrus Cloud Detection Accuracy of GOSAT/CAI and Landsat-8 with laser Radar: Lidar and

Confirmation with Calipso Data, International Journal of Advanced Research on Artificial Intelligence, 5, 1, 14-21, 2016.

[5] Kohei Arai, Masanori Sakashita, Hiroshi okumura, Shuichi Kawakami, Kei Shiomi, Hirofumi Ohyama, Comparative Study on Cloud Parameter Estimation AmongGOSAT/CAI, MODIS, CALIPSO/CALIOP and Landsat-8/OLI with Laser Radar as Truth Data, International Journal of Advanced Research on Artificial Intelligence, 5, 5, 21-29, 2016.

[6] I.Morino, O.Uchino, Y.Yoshida, N.Kikuchi, M.Inoue, K.Nakamae, T.Yokota, V.Sherlock, B.Liley, S.Kawakami, H.Ohyama, T.Nagai, T.Sakai, Kohei Arai, H.Okumura, Advanced validation of the GOSAT observed CO2 and CH4 at priotized observation sites, AGU Fall Meeting, A063, 2012.

[7] H. Okumura, S.Takubo, T.Kawsaki, I.N.Abdulah, T.Sakai, T.Maki, Kohei Arai, Web based data acquisition and management system for GOSAT validation Lidar data analysis, Proceedings of the SPIE Vol.8537, Conference 8537: Image and Signal Processing for remote Sensing , Paper #8537-43, system, 2012.

[8] T.Sakai, H. Okumura, T.Kawsaki, I.N.Abdulah, O.Uchino, I.Morino, T.Yokota, T.Nagai, T.Sakai, T.Maki, Kohei Arai, Observation of aerosol parameters at Saga using GOSAT product validation Lidar, Proceedings of the SPIE Vol.8526, Conference 8526: Lidar Remote Sensing for Environmental Monitoring XIII, SPIE Asia-Pacific Remote Sensing, Paper #8295A-50,IP1, 2012.

[9] Hiroshi Okumura, Shoichiro Takubo, Takeru Kawasaki, Indra Nugraha Abdulah, Osamu Uchino, Isamu Morino, Tatsuya Yokota, Tomohiro Nagai, Tetu Sakai, Takashi Maki, Kohei Arai, Improvement of web-based data acquisition and management system for GOSAT validation Lidar data analysis(2013), SPIE Electronic Imaging Conference, 2013.

[10] Hiroshi Okumura, Kohei Arai, Observation of aerosol properties at Saga using GOSAT product validation LiDAR, Proceedings of the Conference on Image and Signal Processing fo Remote Sensing, SPIE #ERS13-RS107-38, 2013

[11] T. Sakai, O. Uchino, I. Morino, T. Nagai, S. Kawakami, H. Ohyama, A. Uchiyama, A. Yamazaki, Kohei Arai, H. Okumura, Y. Takubo, T. Kawasaki, T. Akaho, T. Shibata, T. Nagahama, Y. Yoshida, N. Kikuchi, B. Liley, V. Sharlock, J. Robinson, T. Yokota, Impact of aerosol and cirrus clouds on the GOSAT observed CO2 and CH4 inferred from ground based lidar, sky-radiometer and FTS data at prioritized observation sites,(2013), Proceedings of the 9th International Workshop on Greenhouse Gas measurements from Space, IWGGMS-9, 2013

[12] Osamu Uchino, T.Sakai, T.nagai, I.Morino, H.Ohyama, S.Kawakami, K.Shiomi, T Kawasaki, T.Akaho, H.Okumura, Kohei Arai, T.matsunaga, T.Yokota, Comparison of lower tropospheric ozone column observed by DIAL and GOSAT TANSO-FTS TIR, Proceedings of the AGU Fall Meeting 2013.(2013)

[13] I Morino, T Sakai, T.Nagai, A.Uchiyama, A.Yamazaki, S Kawakami, H.Ohyama, Kohei Arai, H.Okumura, T.Shibata, T.Nagahama, N.Kikuchi, Y.Yoshida, Ben Liley, Vannessa Sherlock, John Robinson, O. Uchino, T.Yokota, Impact of aerosols and cirrus on the GOSAT onboard CO2 and CH4 inferred from ground based Lidar, sky-radiometer and FTS data at prioritized observation sites, Proceedings of the AGU Fall Meeting 2013.(2013)

[14] O.Uchino, I.Morino, M.Inoue, Y.Yoshida, N.Kikuchi, K.Nakane, T.Yokota, Debra Winch, Coleen M. Roahl,Paul U.Wannberg, Geoffrey C.Toon, Nicholas M.Deuscher, Thorsten Weneke, Justus Notholt, David W., T.Griffich, Vennesa Sherlock, John Robinson, Brian Connor, Ralf Sussmann, Markus Rettinger, Petteri Ahonen, Paul Heikkinen, Esko Kiro, Joseph mendonka, Kinbbely Strong, Frank Hase, Susanne Dohe, Skanakami, H.Ohyama, T.Nagahana, T. Machida, Y.Suwa, H.Matsueda, K.Tsuboi, Colun Sunney, Pieter Tans, Arlyn Anderews, Sebastian C.Biraud, Steven C.Wofsy, T.Sakai, T.Nagai, A.Yamazaki, A.Uchiyama, T.Akaho, T.Kawsaki, H.Okumura, Kohei Arai, ban liley, T.Shibata, On the validation results of five year GOSAT SWIR XCO2 and XCH4 data, Proc. of the IWGGMS-10, 2014.

[15] Osamu Uchino, Isamu Morino, Tetsu Sakai, Toshiharu Izumi, Tomohiro Nagai, Takashi Shibata, Hiroshi Okumura, Kohei Arai, Ben Liley, Gerty Bagtasa, Lidar observations at priotized sites for GOSAT validation, Proceedings of the AGU Fall Meeting 2016.

[16] Osamu Uchino, Isamu Morino, Tetsu Sakai, Toshiharu Izumi, Tomohiro Nagai, Takashi Shibata, Hiroshi Okumura, Kohei Arai, Ben Liloy, Dave Pollard, Gerry Bagtasa, Lider observation at TCCON site to investigate the influence of particles for GOSAT data, Proceedings of the IWGGMS 13, at Helshinki, Finland, 2017.

[17] Kohei Arai, Takashi Higuchi, Hiroshi Okumura, Hirofumi Ohyama, Shuji Kawakami, Kei Shiomi, Influence due to cirrus and thick aerosol on estimation of column CO2 and CH4 with GOSAT/FTS data, Proceedings of the COSPAR 2018, A1.1 Space-based and sub-orbital observations of atmospheric physics and chemistry, 2018..

## 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

# An Ensemble approach to Big Data Security (Cyber Security)

Manzoor Ahmed Hashmani[1]
High Performance Cloud Computing Center (HPC3)
Center for Research in Data Science (CeRDaS)
Universiti Teknologi PETRONAS
Sri Iskandar, Malaysia

Syed Muslim Jameel[2], Aidarus M. Ibrahim[3],
Maryam Zaffar[4]
Department of Computer and Information Sciences
Universiti Teknologi PETRONAS
Sri Iskandar, Malaysia

Kamran Raza[5]
IQRA University, Karachi, Pakistan

***Abstract*—In the past, information safety was centered on event correlation designed for observing and spotting previously identified attacks. Due to the dynamic nature of multidimensional cyber-attacks, these models are no more acceptable. Specifically, these attacks use different strategies and procedures to find their way into and out of an organization. Traditional methods have reached their limit and thus new approaches are needed to find a solution for arising issues and challenges for big data security. To understand the current problem, we critically reviewed the literature related to big data security and the solutions proposed by the scientific community. In this paper, an ensemble approach for big data cybersecurity is proposed. To evaluate our approach, the given benchmark data is fed to three different classifiers namely to a k-nearest neighbor (KNN), support vector machine (SVM), multilayer perceptron (MLP) and the output of the single classifiers were compared to ensemble approach of the three classifiers. The reported results show that the ensemble approach for big data cybersecurity performs better than the single classifiers.**

***Keywords*—Big data; cyber security; benign, malicious; ensemble approach; Support Vector Machine (SVM); Receiver Operating Characteristic (ROC); Features (F)***

## I. Introduction

The progress in the current technology has embarked concerns about the risks to data related having weak security issues such as a virus, malware and compromising systems and services [1]. Lack of all aspects of data security may result compromised data in terms of confidentiality, integrity, and availability of data to outsiders [2]. A lot of efforts have been made to deploy cybersecurity monitoring which significantly worked over the last decade, yet these systems face challenges and issues [3], [1]. For example, Host-based security system and intrusion detection system were proposed to provide protection from the attacks, however, these systems failed to capture the new sophisticated attacks having unknown signatures. Moreover, some commercial systems for monitoring were proposed. These systems include Ganglia, Nagios, and Zabbix. They have provided a quick solution to security problems which have impact system performance. Though, subtle attacks were not detected by these systems [4].

To protect the big data in universal resource locator (URL), different methods for web filtering were deployed. For example, proxy servers are another mitigation approach for brows-able space of the internet [4].

Other methods were proposed for malicious URL detection. The popular method to detect malicious URL is the blacklist method and it is extremely fast and very easy to implement [5]. However, this technique suffers from non-trivial false positive and it is difficult for it to maintain an exhaustive list of malicious URLs [6]. Furthermore, Signature-based detection technique (IDS) is capable to identify the malicious pattern from the already defined type of pattern. However, this technique is not capable to identify new type of malicious attack. Heuristic approach detects the possible malicious pattern through the intelligent guesswork, heuristic approach builds rules (rule of thumb) from the experiences instead of any predetermined formula, although due to lack of rules in most cases, it suffers it accuracy to identify correct malicious pattern [6].

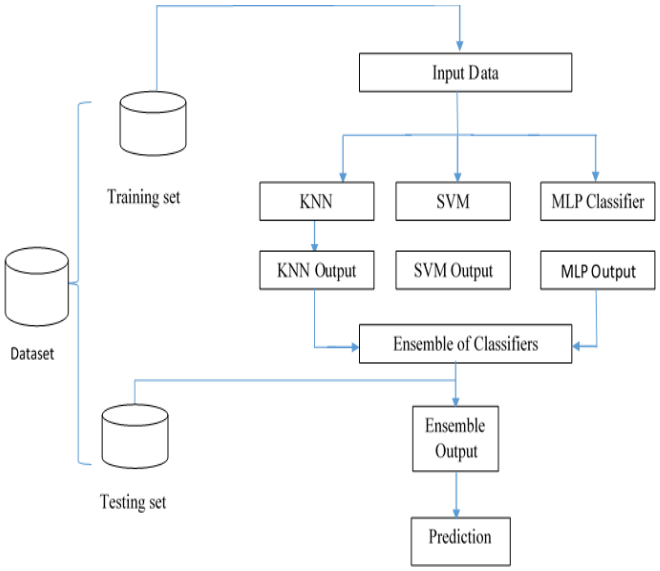



Fig. 1. Proposed Method for Big-Data Cybersecurity.

In this paper, we have proposed an ensemble approach for cybersecurity. In the experimental settings, the data were divided into training and testing sets. The training data were fed to the k-nearest neighbor (KNN), support vector machine, multilayer perceptron (MLP) individually. The output of these single classifiers was combined to form an ensemble approach.

## II. Material and Methods

In this section, a benchmark dataset that consists of about 3.2 million features were used. We used data example from already available dataset. The data set is extracted through feature extraction mechanism from a large mail provider (real-time feed supplies 6000-7000 spam and phishing URL per day). The study [7] provides the complete detail of dataset extraction and preparation. The methods used in this study are explained in detail in the following subsections.

### A. K-Nearest Neighbors

K nearest neighbors is nonparametric classification that stores available data and classify new data based on how similar they are in terms of distance. In the early 1970's, KNN is considered as one of the most prominent nonparametric techniques in statistical estimation and pattern recognition [8],[9].

### B. Support Vector Machine (SVM)

In machine learning, support vector machines supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting).

### C. Multilayer Perceptron (MLP)

Multilayer perceptron (MLP) can be designed by connecting the individual perceptron into neural network-based architecture. MLP is recognize as category of feedforward Artificial Neural Network because all input and intermediate layers provide input to their succeeding layers [9].

### D. Ensemble Approach

Few recent studies in machine learning domain investigate the comparative analysis among single and ensemble classifiers. Through the variety of experimental results, these studies conclude that in most of the cases the ensemble approaches improve the classification performance over single classifier [10]. However, the effect of ensemble approach for the Big Data security is unknown. These integrated approaches based on diverse classification techniques and could attain the unidentical rate of accurate classified individuals, which ultimately results in more reliable, specific and accurate classification outputs than single classifier approach. A study [11] discuss the core parameters which enhance the performance of ensemble approach over single classifier, this study discusses the statistical, representational and computational interpretation and provides justification for better performance in ensemble classifiers. However, in ensemble classifiers, various critical parameters (for example, sum, product, minimum, maximum, average, Byes, dempster Shafer and decision template) tuning are mandatory for significant improvement in classification results. Fig. 1, defines the proposed ensemble-based approach for big-data cybersecurity.

### E. Performance Evaluation

The proposed approach was evaluated using well-known dataset (training and testing. In training part, the parameters of the single classifiers reach to its optimal values (which are near to their target function) from the hypothetical values, and this followed by testing set to validate the performance of the classifiers. This approach reduces the bias and increases the generalization of the reported results. To evaluate and verify the classification performance, a receiver operating characteristics (ROC) were used. A t-test was also applied to see whether benign and malicious URLs are different to validate our results through the statistical analysis of two population mean.

## III. Results and Discussion

The comparative analysis among the single and ensemble approaches are shown in, Table 1. The reports result in terms of accuracy of the classification of the two classes (benign versus malicious) are 0.9867 of KNN, 0.9867 of SVM, 0.9833 of MLP and 0.993 of ensemble approach. The classification performance results clearly show that the proposed ensemble approach is slightly higher than the single classifiers.

TABLE I. Performance f the Single and Ensemble Approach

| No. | Methods | Accuracy (%) |
|---|---|---|
| 1 | KNN | 0.9867 |
| 2 | SVM | 0.9867 |
| 3 | MLP | 0.9833 |
| 3 | Ensemble approach | 0.993 |

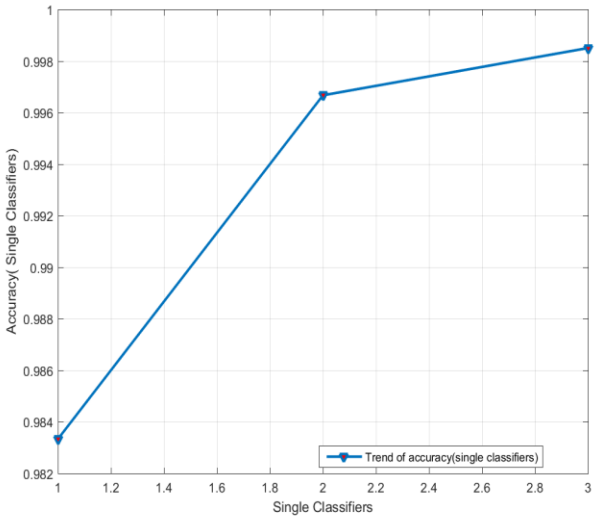



Fig. 2. Classification Performance of Single Classifiers (Benign v/s Malicious).

Fig.2 shows the performance of the single classifiers in terms of accuracy. We have plotted together with the reported accuracy of the single classifiers. The range of the reported

accuracy of the single classifies are in between 0.983 to 0.9867. The line graph starts growing up until it reaches 0.09867 and then turns in to constant and this indicates that SVM and KNN are better in classification performance compared to MLP classifier.

Fig.3. shows, results of the methods those are necessary to combine different classifiers and make the ensemble. The x-axis of the figure represents nine different methods from 1 to 9 (as majority voting, maximum, sum, minimum, average, product, Bayes, decision template and Dempster-Shafer) respectively. The y-axis represents the reported accuracy of the different methods. Our reported results show that Dempster-Shafer is superior in ensemble approach compared to other remaining methods.

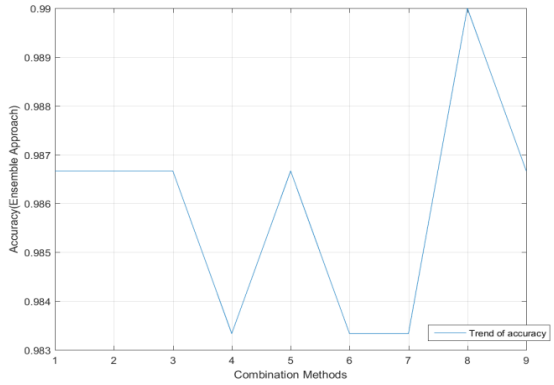



Fig. 3. Results of combination methods in terms of ensemble approach.

Fig. 4. describes the comparison of ROC curve performance of single classifiers and ensemble approach. The reported output shows that ensemble approach is higher than the other methods and this stresses the reported results in Table 1. The same idea was applied to differentiate between benign URLs and malicious URLs. Table 2 indicates the significant difference between benign and malicious features through the mean and standard deviation (having p-value not more than 0.0001). However, all feature (benign and malicious) are not identical.

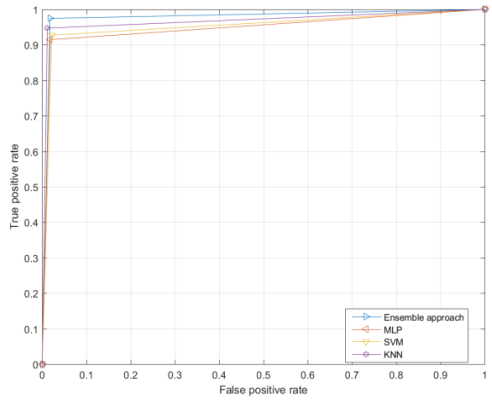



Fig. 4. Comparison of ROC curve performance of single classifiers and ensemble approach.

TABLE II. PERFORMANCE OF THE SINGLE AND ENSEMBLE APPROACH

| F. | Benign | Malicious | T value |
|---|---|---|---|
| F. 1 | 7.087E-02 ± 4.873E-02 | 6.813E-02 ± 3.334E-02 | 0.328 |
| F. 2 | 8.069E-02 ± 3.250E-02 | 8.855E-02 ± 3.812E-02 | 1.11 |
| F. 3 | 0.136 ± 4.670E-02 | 0.118 ± 4.754E-02 | 2.00 |
| F. 4 | 0.455 ± 0.353 | 0.536 ± 0.347 | 1.16 |
| F. 5 | 0.503 ± 0.401 | 0.539 ± 0.392 | 0.460 |
| F. 6 | 0.198 ± 0.250 | 0.333 ± 0.324 | 2.33 |
| F. 7 | 0.123 ± 0.156 | 0.158 ± 0.152 | 1.12 |
| F. 8 | 2.675E-02 ± 4.958E-02 | 3.006E-02 ± 5.465E-02 | 0.317 |
| F. 9 | 1.420E-02 ± 3.466E-02 | 2.319E-02 ± 4.356E-02 | 1.14 |
| F 10 | 3.556E-02 ± 6.019E-02 | 2.444E-02 ± **4.072E − 02** | 1.08 |

## IV. CONCLUSION

Industrial Revolution IR 4.0 Big Data is considered as an opportunity to provide a more reliable and accurate source for business intelligence [12]. However, the versatile characteristics of Big Data possesses the potential to compromise the reliability and integrity of Big Data (which in result may degrades performance accuracy). Big Data security is considered as one of the serious challenges for researchers. Therefore, in this study, we have proposed a more reliable and accurate ensemble-based approach to classify benign and malicious activities to identify and prevent the possible cyber threat. Our proposed approach is highly accurate and able to classify (between benign versus malicious) an accuracy of 0.993. In future, this study will be further investigated to identify the threat pattern in cybersecurity.

## REFERENCES

[1] Y. Ashibani and Q. H. Mahmoud, "Cyber-physical systems security: Analysis, challenges, and solutions," Computers & Security, vol. 68, pp. 81-97, 2017.

[2] C. Everett, "Big data–the future of cyber-security or it's the latest threat?," Computer Fraud & Security, vol. 2015, pp. 14-17, 2015.

[3] A. M. AlMadahkah, "Big Data In computer Cyber Security Systems," International Journal of Computer Science and Network Security (IJCSNS), vol. 16, p. 56, 2016.

[4] M. Mayhew, M. Atighetchi, A. Adler, and R. Greenstadt, "Use of machine learning in big data analytics for insider threat detection," in Military Communications Conference, MILCOM 2015-2015 IEEE, 2015, pp. 915-922.

[5] P. Vinod, R. Jaipur, V. Laxmi, and M. Gaur, "Survey on malware detection methods," in Proceedings of the 3rd Hackers' Workshop on the computer and internet security (IITKHACK'09), 2009, pp. 74-79.

[6] A. Sirageldin, B. B. Baharudin, and L. T. Jung, "Malicious Web Page Detection: A Machine Learning Approach," in Advances in Computer Science and its Applications, ed: Springer, 2014, pp. 217-224.

[7] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Identifying suspicious URLs: an application of large-scale online learning," in Proceedings of the 26th annual international conference on machine learning, 2009, pp. 681-688.

[8] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in Advances in neural information processing systems, 2006, pp. 1473-1480.

[9] L. I. Kuncheva, Combining pattern classifiers: methods and algorithms: John Wiley & Sons, 2004.

[10] T. G. Dietterich, "Ensemble learning," The handbook of brain theory and neural networks, vol. 2, pp. 110-125, 2002.

[11] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[12] S.M. Jameel, M.A. Hashmani, H. Alhussain, and A. Budiman, 2018, June. A Fully Adaptive Image Classification Approach for Industrial Revolution 4.0. In International Conference of Reliable Information and Communication Technology (pp. 311-321). Springer, Cham.

# An Interactive Content Development for Depression Awareness among Tertiary Students

Sarni Suhaila Rahim, Thum Wei Ching
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

***Abstract*—"2D Animation: Depression among Tertiary Students" is a novel interactive content development that gives information of depression to public. It consists of seven modules of depression which are Introduction, Statistics, Types, Symptoms, Causes, Treatments and Video of Depression Information. The objectives of this project are to study the causes and effects of depression among tertiary-level students in Malaysia, to design and develop a 2D animation in raising awareness about depression to the viewers and to investigate the effectiveness of depression animation to the users. The methodology used for this project is Multimedia Production Process which consists of three stages which are pre-production, production and post-production. The testing result shows that the interactive content for depression animation is accepted and effective to the public in order to have a better understanding and awareness on depression among tertiary students.**

***Keywords*—2D animation; awareness; depression; interactive content; tertiary students***

## I. INTRODUCTION

Tertiary students are those pursuing tertiary level of education which includes diploma, bachelor's, masters and doctoral degrees after their secondary level of education. Attending colleges or universities seems challenging and stressful for tertiary students. They have to cope with academic pressure, future worries, independent living from parents, family responsibilities, social life and so on. In order to handle all kind of factors in a perfect manner, tertiary students are having depression. Depression is the most common psychological problems among tertiary students [1]. According to statistics from the 2015 National Health and Morbidity Survey (NHMS), 29.2 percent, or 4.2 million, citizens above the age of 16 years suffer from depression. The study in [2] reviews the number of cases had doubled compared to 2006 (11.2 percent) and 1996 (10.6 percent).

Depression is a common mental disorder, characterized by prolonged sadness, loss of interest, feeling of guilt or worthless, disturbed sleep or appetite, feelings of tiredness, and poor concentration [3]. It is a serious medical illness that negatively affects one's feeling, thinking and behaving. This mental disorder will decrease a person's performance at work and at home [4]. Globally, more than 300 million people of all ages are now living with depression, an increase of depression more than 18% between 2005 and 2015 [5]. Based on statistics, there are 40% of Malaysian suffer from depression [6]. Depression will lead to inclination to suicidal acts. According to World Health Organization, there are nearly 800,000 people die due to suicide every year [7]. Therefore, suicide is the second leading cause of death due to depression among 15-29 years old.

This project aims to create a 2D animation on depression among tertiary students. There is a lack of depression awareness campaign. Information or knowledge on depression is limited among students. In addition, causes and effects of depression needed to be identified. The World Health Organization estimates that by 2020 depression will be the leading cause of disability globally. In addition, Ministry of Health Malaysia provides statistics that show a worsening state of depression among students, that is from one in ten individuals in 2011 to one in five in 2016 [8]. Students are facing independent living challenges as well as educational challenges. According to Kumaraswamy [9], depression among tertiary-level students has become a concerning issue as students are the human capital of society. Students may suffer from Schizophrenia and inclines to suicide if the depression state is prolonged [8]. University students are having cursory knowledge of depression and giving negative attitudes towards depression [10]. According to Nam et al., [11], stigma can be reduced and help-seeking can be increased by improving public understanding of depression and suicide. Therefore, public awareness and understanding have to be raised before the worsening state of depression among tertiary students.

The objectives of this project are to study the causes and effects of depression among tertiary-level students in Malaysia, to design and develop a 2D animation in raising awareness about depression to the viewers and finally to investigate the effectiveness of the depression animation to the users. Target users of this project is university and college students. The modules of this project are Introduction of Depression, Statistics of Depression, Types of Depression, signs and Symptoms of Depression, Causes of Depression, Treatments of Depression and Video of Depression. Methodology used for this project is Multimedia Production Process which consists of three stages which are pre-production, production and post-production. The software used to create this project are Adobe Audition, Adobe Animate, Adobe Photoshop, Adobe Illustrator and Vegas Pro. The outcome of this project is the 2D animation for depression which aims to give public awareness and understanding.

This paper is organized as follows. Section 2 presents previous related work on 2D animation and depression. Section 3 describes in details the overall methodology of this research, while Section 4 presents the developed system testing and

results. The discussion of the research project is presented in Section 5. Finally, Section 6 presents some conclusions and future work.

## II. Literature Review

Animation is used to create multimedia and gaming products by designing, drawing, making layouts and preparation of photographic sequences. An animation is the illusion of movement created by rapid display of images in sequence [12]. Efficiency is one of the 2D animation benefits. Design of 2D animation is less complex and simpler than in 3D animation. Two-dimensional is being chosen for the case that the message is needed to be highlighted. In addition, cost-effectiveness is the other benefit of 2D animation. The simplicity and efficiency of this animation can meet customer needs and also budget constraints, besides artistic freedom. Interactive animation lets the users join on more levels than just-viewing and better than non-interactive animation as it gives better conversion. It can convey the target message efficiently. Furthermore, there is a better engagement as users can decide what they want to watch or to skip. This type of animation is more attractive, compelling and convincing that brings higher responses.

This project will focus on the depression among tertiary students. The use of 2D animation can convey the message and information regarding depression and its treatments to target audiences in order to raise awareness and understanding of depression among them. The signs or symptoms of depression among students can be recognized earlier and proper treatments can be offered before it is too late. An interactive of 2D animation provides interaction with its users by giving control to users on information accessibility through menu which consists of introduction, statistics, types, symptoms, treatments of depression and video.

There are three existing systems found to compare with this project, which are a video named “Depression Motion Graphic Awareness Campaign” and two 2D animation named “Depression & Anxiety Awareness Video” and “What is Depression”. Table I shows the comparison of these existing systems.

Fig. 1 shows the screenshot of the first existing system, “Depression Motion Graphic Awareness Campaign”. It is a 1 minute and 50 seconds video about depression awareness campaign that posted on Youtube. This video portrays information about depression definition, statistics and the way to overcome it. Meanwhile, Fig. 2 shows the screenshot of the second existing system. It is a 3 minutes and 52 seconds 2D animation that posted on Youtube. This animation portrays information about symptoms of depression and its treatments to increase awareness depression and anxiety in urban low income societies in Pakistan. Finally, Fig. 3 shows the screenshot of the “What is Depression” animation. This is a 4 minutes and 28 seconds 2D animation about what is depression that posted on Youtube. This animation portrays information about symptoms of depression and its treatments to help those suffered from depression.


Fig. 1. Screenshot of Depression Motion Graphic Awareness Campaign [13]

Fig. 2. Screenshot of Depression Motion Graphic Awareness Campaign [14]

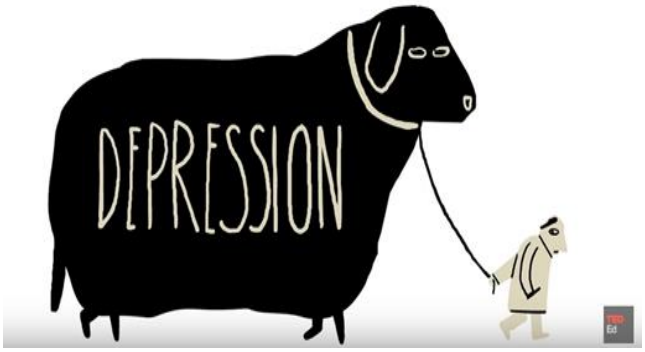



Fig. 3. Screenshot of Depression Motion Graphic Awareness Campaign [15]

Based on Table I, it can be concluded that there is no user interaction in these three existing systems as they are video and 2D animation. Most of these systems are less attractive because they have fewer multimedia elements, which two of the systems consist of four types of multimedia elements only like graphic, audio, video, animation or text. All of these systems are lack of user interaction and interactivity because these systems do not have interactive button for users to control the information or graphics that they want to view or skip. The three existing systems have covered different type of modules such as definition of depression, statistics, symptoms of depression and its treatments. However, these existing systems do not offer an interactive content development for depression. The developed project presents a novel interactive content for depression among tertiary students, besides novel content verification and a new video on depression information by the Counsellor.

TABLE I. COMPARISON OF EXISTING SYSTEMS

| Title / Features | Depression Motion Graphic Awareness Campaign [13] | Depression & Anxiety Awareness Video [14] | What is depression? [15] |
|---|---|---|---|
| Content | Depression definition, statistics and treatments | Symptoms of depression and its treatments | Symptoms of depression and its treatments |
| Attraction Level | Less attractive | Less attractive | Attractive |
| Type | Video | 2D Animation | 2D Animation |
| Format | Mp4 | Mp4 | Mp4 |
| Target User | Public | Public | Public |
| Interactivity | No | No | No |
| User interaction | No | No | No |
| Duration (minutes) | 1:50 | 3:52 | 4:28 |
| Language | English | English | English |
| Interactive button | No | No | No |
| Multimedia Element | Text<br>Graphic<br>Audio<br>Video | Graphic<br>Audio<br>Video<br>Animation | Text<br>Graphic<br>Audio<br>Video<br>Animation |

## III. METHODOLOGY

The methodology used in this project is Multimedia Production Process, consists of three main phases, which are pre-production, production and post-production. There are six phases involved in these three main phases, which are analysis, design, implementation, testing, evaluation and publishing. In this section, all requirements will be further analysed in detail.

### A. Analysis

In order to gather important information which is the content of the project, an interview session is conducted. In addition, online questionnaire is distributed to 80 respondents which are tertiary students in Malaysia in order to get feedback from public about depression.

#### 1) Interview Session with Counsellor

The data are collected through the interview session with the Counsellor at Universiti Teknikal Malaysia Melaka (UTeM) Counselling Unit. Table II shows the details of data collection process.

TABLE II. CONTENT VERIFICATION

| Verification | Subject Matter Expert |
|---|---|
| Contents for six modules | UTeM Counsellor |
| Video on common questions about depression | UTeM Counsellor |
| Justification of five selected types of depression | UTeM Counsellor |

#### 2) Public Online Survey

Besides the interview session, the data are collected through the online survey by using Google Form. The questionnaire is divided into four sections, which are demographic, causes, symptoms and treatments. There are 12 questions included in this questionnaire. The aim of this survey is to study about the prevalence of depression among tertiary students, including symptoms, causes and treatments of depression. Most of the respondents are female aged from 20-25 years old who are undergraduate students. Fig. 4 to Fig. 7 show the findings of the survey. Results are presented and analysed as below.

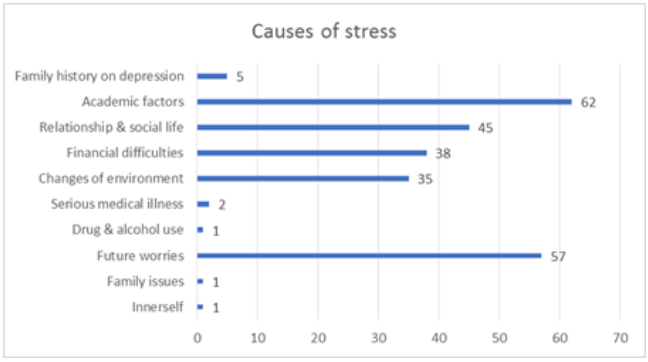



Fig. 4. Causes of stress

Fig. 4 shows academic factors, future worries and relationship and social life are among the factors that bring stress to students representing of 77.5% (62 counts), 71.3% (57 counts) and 56.3% (57 counts) respectively. These followed by 47.5% (38 counts) of financial difficulties and 43.8% (35 counts) of changes of environment factor.

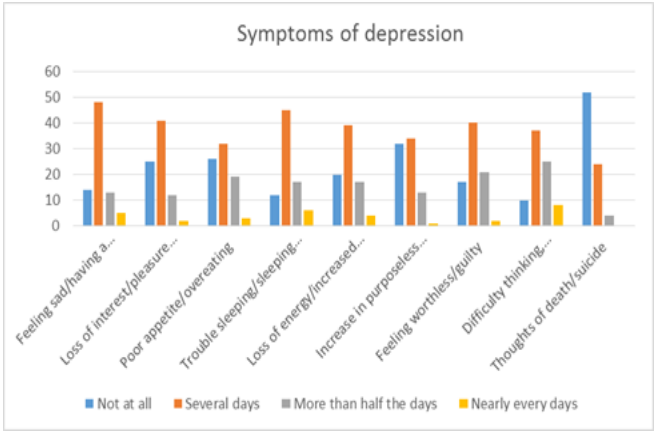



Fig. 5. Symptoms of depression

Fig. 5 shows the result of symptoms of depression faced by the students. Students that choose the option of more than half days and nearly every days might be having depression. Overall, most of the students select option of several days for all symptoms of depression except for thought of suicide. From this result, it reveals that students might just having stress with difficulty in study, relationship factors and so on. In contrast, there are students who choose option of more than half days and nearly every days. For option more than half days, students select difficulty thinking, feeling worthless and poor appetite or overeating the most among other symptoms whereas for option nearly every day, students choose difficult thinking, trouble sleeping and feeling sad the most. By comparing these two options, students are facing difficulty thinking for the symptoms of depression.

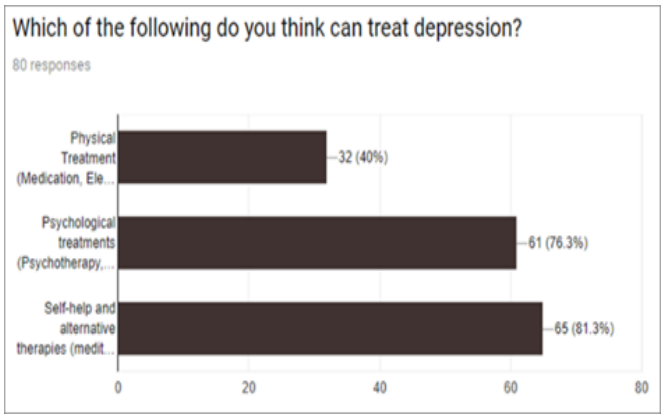



Fig. 6. Treatments of depression

Fig. 6 shows the treatments of depression performed by the students. It presents the highest treatment is self-help and alternative therapies which is 81.3% followed by 76.3% of psychological treatments. The lowest treatment selected is the physical treatment which is 40%.

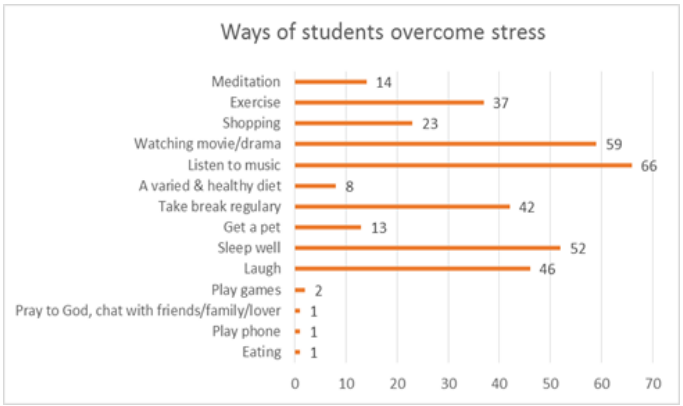



Fig. 7. Ways of students overcome stress

Fig. 7 shows most of the students handle their stress by listening music which is 82.5%. This is followed by watching movie or drama, sleep well, laugh and take break regularly which are 73.8%, 65%, 57.5% and 52.5% respectively.

In conclusion, academic factor, difficulty thinking and self-help and alternative therapies are the main causes, symptoms and treatments of depression of 80 respondents respectively. As most of the students are undergraduate students ranged between 20 to 25 years old, they are facing a lot of challenges and difficulty in their studies. They have to study hard to achieve better results. Therefore, they are facing depression when they cannot cope with their studies which leads to difficulty in thinking nearly everyday. Most of them prefer self-help and alternative therapies like listening to music.

## B. Design

During the design phase, the project is designed to satisfy the requirements that identified in the previous analysis phase. In this phase, it takes time to finish all the design for the 2D animation such as the menu, layout, interactive buttons and interactive elements. During this phase also, the storyboard of the design and character profile are performed. Fig. 8 to Fig. 11 show some of the interfaces of this project.

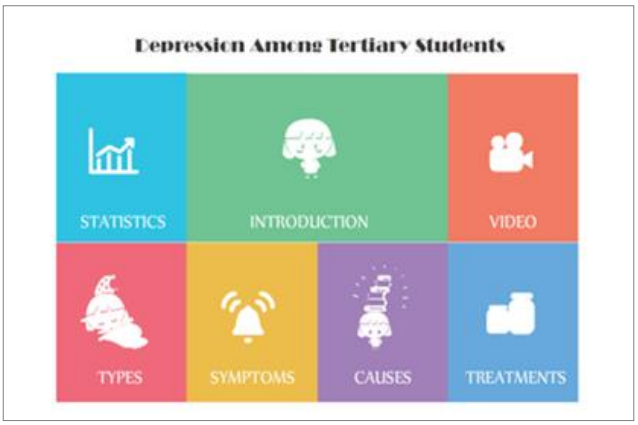



Fig. 8. Interface of main menu


Fig. 9. Interface of Introduction module

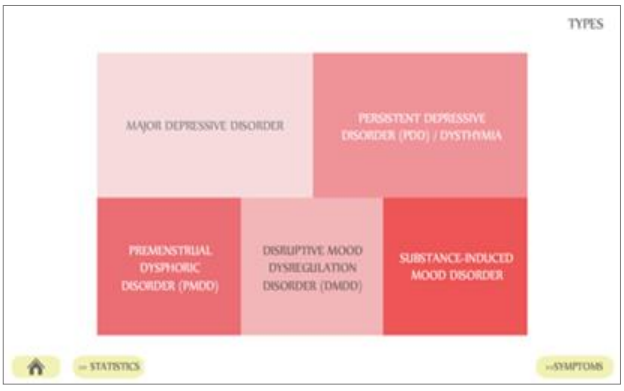



Fig. 10. Interface of Types of Depession module


Fig. 11. Interface of Video Depression Information module

### C. Development

Text is used to give information and explanation of each of the modules. The font of the text used for title of home page and title of menu page are Berlin Sans FB and Broadway respectively. The font of text used for title of menu, subtitle and title of button are Nyala.

The graphics or images used in this project are in Portable Network Graphic (PNG) format and JPEG format. All of the images are vector images in 2D form. Those vector images have been drawn and edited by using Adobe Illustrator CS6. Then, those designed and edited images will imported to Adobe Animate for animate purposes.

The audio is recorded by using voice recording equipment and then cleaned and edited in Adobe Audition. Another source of sound effect is obtained from internet. Royalty free sound effect were acquired from Freesound.

The interview video on depression among tertiary students with the Counsellor has been edited by using Vegas Pro. The video effects used in this video are film dissolve and dither dissolve. The video duration is about 3 minutes and 26 seconds and it is exported in MP4 format. The integration process was done by using Adobe Animate. All the multimedia elements created were arranged according to modules in Adobe Animate. The final product was exported as SWF file. Fig. 12 shows the process of media integration for this project.

### D. Testing

This section discusses the testing on the product of the project, focusing on the aspects proposed in the project objective. The testing aims to achieve the third objective of this project, which is to investigate the effectiveness of the depression animation to the users.

Test plan is used to determine the test user, test environment and the test schedule that will be carried out for this project.

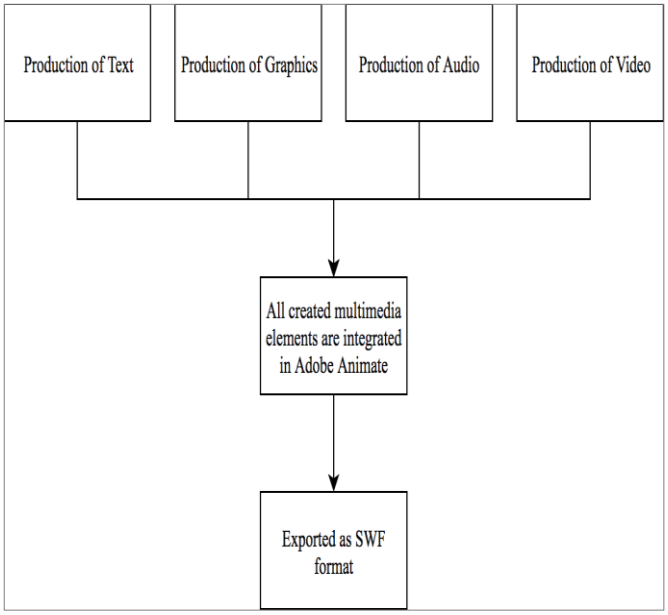



Fig. 12. Process of media creation

#### 3) Test User and Test Environment

Table III shows the details of the users and venue involved.

TABLE III. TEST USER AND ENVIRONMENT

| User | Environment | Total |
|---|---|---|
| Counsellors | UTeM Counselling Unit | 2 |
| Multimedia Experts | FTMK, UTeM | 5 |
| Tertiary Students | Online Testing | 70 |

#### 4) Test Schedule for Target User

Test schedule describes method used for testing, the number of questions, total respondents, date and venue. The method used for testing in this project is questionnaire. Table IV, Table V and Table VI show the schedule of testing among the counsellors, multimedia experts and tertiary students, respectively

TABLE IV. TEST SCHEDULE FOR COUNSELLOR

| **Target User: Counsellors** | | | |
|---|---|---|---|
| **Method: Questionnaire** | **Number of Question** | **Total Respondents** | **Date and Venue** |
| Content | 5 | 2 | 4/7/2018<br>10am – 11am at<br>UTeM<br>Counselling Unit |
| Other | 5 | | |
| Total | 10 | 2 | |

TABLE V. TEST SCHEDULE FOR MULTIMEDIA EXPERTS

| **Target User: Multimedia Experts** | | | |
|---|---|---|---|
| **Method: Questionnaire** | **Number of Question** | **Total Respondents** | **Date and Venue** |
| Functionality Testing | 6 | 5 | 4/7/2018-<br>18/7/2018<br>11am - 12pm at<br>FTMK, UTeM |
| Usability Testing | 6 | | |
| User Interface Testing | 6 | | |
| Total | 18 | 5 | |

TABLE VI. TEST SCHEDULE FOR TERTIARY STUDENTS

| Target User: Tertiary Students | | | |
|---|---|---|---|
| **Method: Questionnaire** | **Number of Question** | **Total Respondents** | **Date and Venue** |
| Content | 10 | 70 | 4/7/2018 - 11/7/2018 via online Google Form |
| Total | 10 | 70 | |

## IV. TEST RESULT AND ANALYSIS

The test result and analysis have been generated into different types of graphical representation for easy analysis of the result.

*1) Test Result for Counsellors at UTeM Counselling Unit*

Table VII shows Question 1 to Question 5 for the Counsellors at UTeM Counselling Unit focusing on the contents of the application. Meanwhile, Fig. 13 shows result of content (Question 1 to Question 5) from two respondents who are Psychology Officer at UTeM Counselling Unit.

TABLE VII. QUESTION FOR COUNSELLOR (CONTENT)

| **Question for Contents** |
|---|
| Question 1: The information provided is valid and correct. |
| Question 2: The information provided is sufficient for public knowledge. |
| Question 3: The explanation of each module is clear and easy to understand. |
| Question 4: The interview video for depression is clear and correct. |
| Question 5: This project can help public know more about depression and raise their awareness towards depression |

TABLE VIII. QUESTION FOR COUNSELLOR (OTHER FEATURES)

| **Question for Other Features** |
|---|
| Question 6: The interface of this animation is pleasant and attractive. |
| Question 7: This project attracts users to know more about depression. |
| Question 8: The flow of 2D animation is simple and easy to understand. |
| Question 9: The narrator's voice pronunciation is clear. |
| Question 10: The navigation of buttons is accurate. |

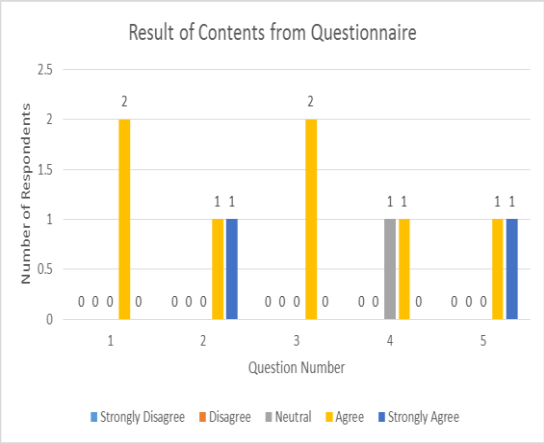



Fig. 13. Result of Contents from Questionnaire

Based on the results presented in Fig. 13, it can be concluded that all of the respondents agree that all of the information provided is valid and correct. For Question 2, one of the respondents strongly agree for this question, while another respondent agree and may has an opinion that more information can be added. Moreover, all of the respondents agree that the explanation of each modules is clear and easy to understand. This is because they think that explanation of each modules can be more clear and easy to understand in order to make this project interesting. For Question 4, one of the respondents agree that interview video for depression is clear and correct, whereas another respondent chooses neutral for this question. Overall, both of the respondents have a thought that the interview video can be improved to be clearer and correct. Finally, for Question 5, one of the respondents strongly agree while another respondent agree that this project can help public know more about depression and raise their awareness towards depression.

There are another five questions, which are Question 6 to Question 10, representing questions on the other features of the application. Table VIII shows the Question 5 to Question 10, while the results from these questions are presented in Fig. 14.

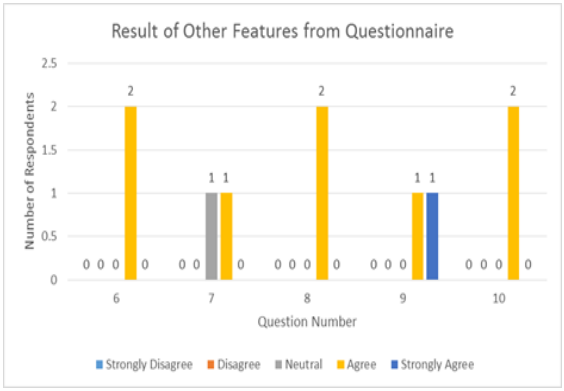



Fig. 14. Result of Other Features from Questionnaire

Fig. 14 shows that all of the respondents agree that the interface of this animation is pleasant and attractive. They think that the interface of this animation is pleasant and attractive, however the interface can be improved to be better.

Meanwhile, for Question 7, one of the respondent agree that this project attracts users to know more about depression whereas another respondent chooses neutral. Based on Fig. 14, both of the respondents have an opinion that this project can be more attractive in order to attract users to know more about depression. For Question 8, it shows that all of respondents agree that the flow of 2D animation is simple and easy to understand. The flow of this project is good and understandable to its users. One respondent strongly agree for Question 9 that the narrator's voice pronunciation is clear, while the other one agree to this question. For Question 10, it shows that all of the respondents agree for this question. They think that the navigation of the buttons are accurate but the navigation of buttons still can be improved.

*2) Test Result for Multimedia Experts at Faculty of Information and Communication Technology (FTMK), UTeM*

The testing for multimedia experts focusing on three parts, which are functionality, usability and user interface. Table IX shows the functionality test questions, while the result of these questios can be referred in Fig. 15.

TABLE IX. FUNCTIONALITY TESTING QUESTION

| Question for Functionality |
|---|
| Question 1: All the buttons are clickable and function well. |
| Question 2: The button displays the right information. |
| Question 3: Animation of this project can be played. |
| Question 4: Audio and background music can be played. |
| Question 5: The interview video for depression can be played. |
| Question 6: The overall system functionality is smooth. |

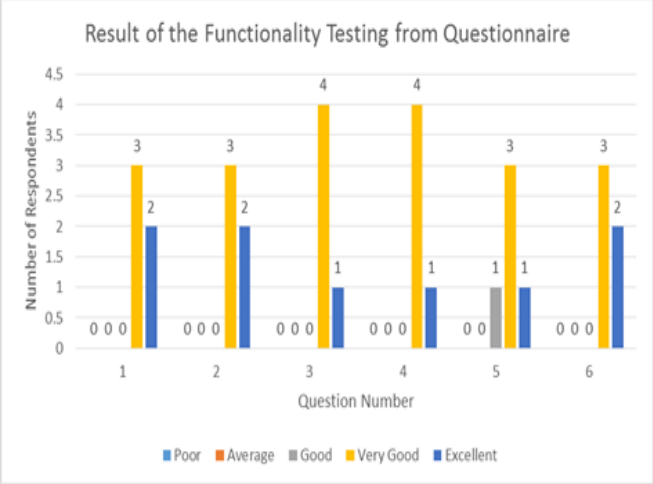



Fig. 15. Result of Functionality Testing from Questionnaire

Fig. 15 shows the result of the functionality testing from questionnaire filled by four multimedia experts. For Question 1, it reveals that all the buttons are clickable and function well

where two out of 5 respondents choose excellent for this question whereas 3 out of 5 respondents choose very good. For Question 2 which is about the button displays the right information, out of 5 respondents, two multimedia experts choose excellent as their answer, whereas 3 out of 5 respondents choose very good. This shows that the button displays the information rightly. For Question 3 which asking that animation of this project can be played, most of the respondents choose very good for this question whereas one of the respondents choose excellent. It shows that the animation of this project can be played smoothly. Question 4 asking about the audio and background music, majority of the respondents choose very good as their answer while only one respondent choose excellent as his or her answer. They agree that audio and background music can be played smoothly and nicely. Most of the respondents choose very good for Question 5 as they think that the interview video for depression can be played. Only two of the respondents choose excellent and neutral, respectively. One of them may have an opinion that the interview video for depression can be played well whereas another respondent may think that improvement for the video can be made. Finally, Question 6 focusing on the overall system functionality. Majority of the respondents choose very good, whereas two of the respondents choose excellent for this question. This shows that the overall system functionality is very smooth.

Table X shows the usability testing questions, comprising of Question 7 to Question 12. Meanwhile, Fig. 16 shows the result of the usability testing from questionnaire filled by four multimedia experts from UTeM.

TABLE X. USABILITY TESTING QUESTION

| Question for Usability |
|---|
| Question 7: The content of all the modules are easy to understand. |
| Question 8: The way of presentation is attractive. |
| Question 9: The explanation on each modules is clear and understandable. |
| Question 10: The information displayed on each modules are organized logically. |
| Question 11: The navigation of the buttons are accurate. |
| Question 12: Users can know more about depression after using this project. |

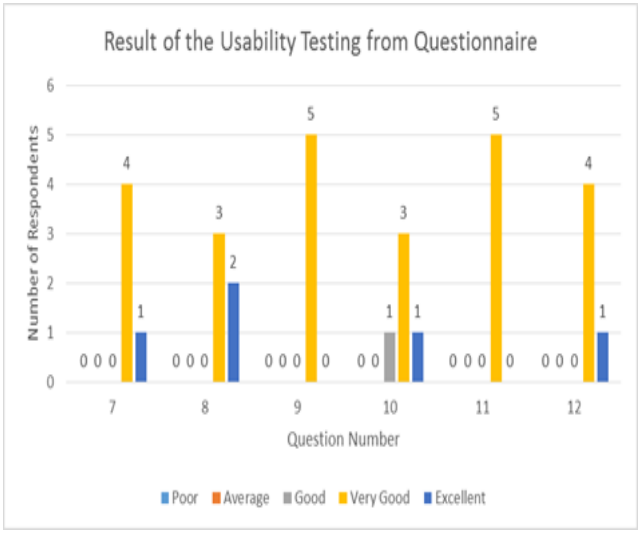



Fig. 16. Result of Usability Testing from Questionnaire

Fig. 16 shows that for Question 7 related to the understanding level of the content, there are four of the respondents choose very good for this question whereas one of the respondents chooses excellent as they think that the content of all modules are very easy to understand. Meanwhile, For Question 8, two out of 5 respondents choose excellent as their answer whereas 3 out of 5 respondents choose very good as their answer to the question about the way of presentation which is very attractive. For the next question, which asking about the explanation on each modules, it shows that all of the respondents choose very good as their option. This is because the explanation on each modules is very clear and understandable for users. Most of the respondents choose very good as their answer whereas one of the respondents chooses excellent as his or her answer for Question 10. They think that the information displayed on each modules are organized logically. Only one of the respondents choose neutral for this question as he or she thinks that the information displayed on each modules can be organized more logically. All of the respondents choose very good for Question 11. They think that the navigation of the buttons are very accurate. Finally, for Question 12, four out of 5 respondents choose very good as

their option for this question, whereas one out of 5 respondents chooses excellent as their option. This shows that users able to know more about depression after using this project.

The testing questions for multimedia experts focusing on the third part, which is the user interface is presented in Table XI. Meanwhile the result of these questios can be referred in Fig. 17.

TABLE XI. USER INTERFACE TESTING QUESTION

| **Question for Usability** |
|---|
| Question 13: Animation of this project attracts users to know more about the information. |
| Question 14: The texts in this 2D animation are readable. |
| Question 15: The colour used in this 2D animation is suitable. |
| Question 16: The video is presented clearly. |
| Question 17: The audio is clear and helps the users understand more on the modules. |
| Question 18: The graphics are clear and suitable. |

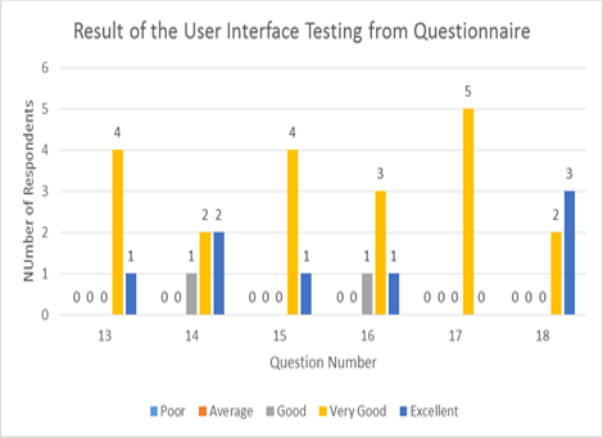



Fig. 17. Result of User Interface Testing from Questionnaire

Based on Fig. 17 which filled by four multimedia experts, it shows that for Question 13 which related to the animation, most of the respondents choose very good as their option, whereas one of the respondents chooses excellent as their option. Overall, the animation of this project attracts users to know more about the information well. Two out of 5 respondents choose very good for Question 14, whereas another 2 out of 5 respondents choose excellent for this question. Only one respondent chooses neutral for this question. For respondents that choose excellent and very good as their option, they agree that the text used in this 2D animation are readable. Only one of the respondents thinks that the text readability can be improved. For Question 15 which asking about the colour used in this animation, majority of the respondents choose very good for this question. Only one respondent chooses excellent for this question. This shows that the colour used in this 2D animation is very suitable. For the next question about the video, three out of 5 respondents choose very good as their option. Meanwhile, two respondents chooses excellent and neutral as their option, respectively. Most of the respondents think that the video is presented clearly, whereas one of the respondents think that the video can be presented more clearly. All of the respondents choose very good for the second last user interface testing question. It is claimed that the audio is very clear and helps users understand more on the modules. For Question 18 which about the graphics used, majority of the respondents choose excellent as their answer whereas minority of the respondents choose very good option. This shows that the graphics used in this project are very clear and suitable.

*3) Test Result for Tertiary Students*

There are 70 respondents that participate in this testing. Most of the respondents are third year undergraduate students aged from 20-25 years old. For the tertiary students testing, it involved an equal percentage of male and female which are 50% (35 respondents) respectively. Meanwhile, majority of the respondents age from 20 to 25 years old which is 73% (51 respondents) whereas minority of respondents aged from 26 to 30 years old which is 4% (3 respondents). There is 23% (16 respondents) of respondents aged from 15 to 20 years old, which is the second highest among other ages. Chinese is the highest ethnicity among other ethnicities which is 57% (40 respondents). The second highest ethnicity is Malay which consists of 37% (26 respondents). Lastly, only 6% (4 respondents) of respondents are Indian, which is the lowest ethnicity.

Majority of the respondents are in their third year of study which is 50% (35 respondents), whereas minority of the respondents are in their second and fifth year of study which are 14% (10 respondents) respectively. In addition, most of the respondents are undergraduate students which consists of 63% (44 respondents) whereas least of the respondents are postgraduate students which consists of 7% (5 respondents) only.

Table XII shows the first five questions prepared for the tertiary students. Fig. 18 shows the test result for tertiary students from Question 1 to Question 5.

TABLE XII. TESTING QUESTION FOR TERTIARY STUDENTS (QUESTION 1-QUESTION 5)

| **Question for Tertiary Students** |
|---|
| Question 1: Do you know more about definition, statistics, types, symptoms, causes and treatments of depression among tertiary students through this 2D animation? |
| Question 2: Do you know more about common causes and way to overcome depression through interview video of this 2D animation? |
| Question 3: Do you understand all the explanation of the seven (7) modules presented from this 2D animation? |
| Question 4: Does this 2D animation raise your awareness towards depression? If No, please specify why. |
| Question 5: Do you think you can apply the knowledge of depression to help those (family, friends, and so on) who suffer from it? If No, please specify why. |

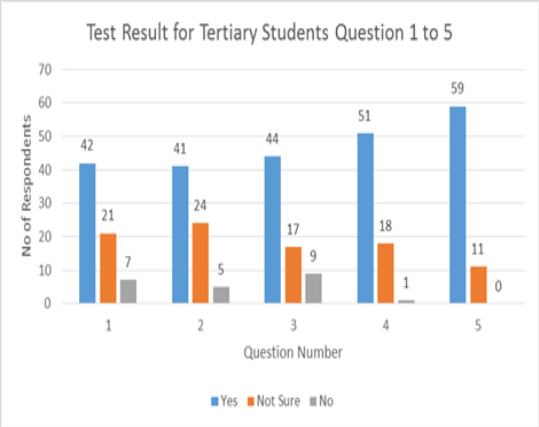



Fig. 18. Test Result for Tertiary Students Question 1 to Question 5

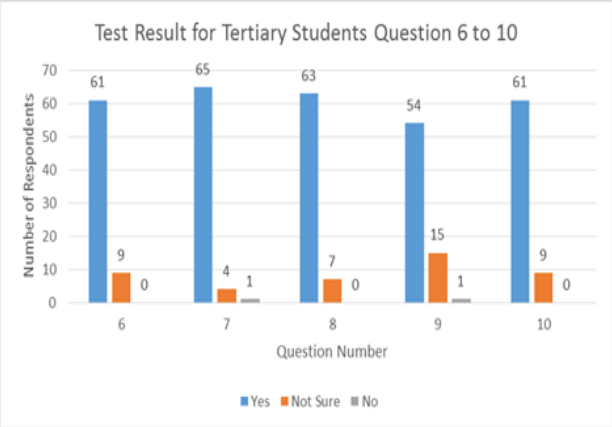



Fig. 19. Test Result for Tertiary Students Question 6 to Question 10

Most of the respondents choose Yes as their answer for Question 1 to Question 5 which are 42 respondents, 41 respondents, 44 respondents, 51 respondents and 59 respondents, respectively. After using this project, majority of the respondents know more about the definition, statistics, types, symptoms, causes and treatments of depression among tertiary students for Question 1 and they also know more about common causes and way to overcome depression through interview video for Question 2. Furthermore, all explanation of seven modules presented in 2D animation are understood by most of the respondents for Question 3 and depression awareness among respondents are raised for Question 4. Knowledge of depression can be applied by most of the respondents in helping those who depressed for Question 5. Only minority of respondents choose No as their answer which are 7 respondents, 5 respondents, 9 respondents and 1 respondent for Question 1, Question 2, Question 3 and Question 4, respectively. They think that there might be improvement to be made for Question 1 to Question 4.

TABLE XIII. TESTING QUESTION FOR TERTIARY STUDENTS (QUESTION 6-QUESTION 10)

| **Question for Tertiary Students** |
|---|
| Question 6: Does the 2D animation is attractive and easy to understand? If No, please specify why. |
| Question 7: The text in this 2D animation is readable. |
| Question 8: The voice of narrator is clear and understandable. |
| Question 9: The interview video about common depression questions is clear and understandable. |
| Question 10: Do you feel enjoyed and satisfied when using this application? |

Table XIII shows the other five questions for the tertiary students and Fig. 19 shows test result answered by 70 respondents who are tertiary students for Question 6 to Question 10.

Based on Fig. 19, it can be seen that 61 out of 70 respondents agree that the 2D animation is attractive and easy to understand, whereas 65 out of 70 respondents agree that the text in this 2D animation is readable. This is because this 2D animation is attractive and understandable and the text is readable for users. For Question 8 to Question 10, majority of the respondents choose Yes as their option which are 63 respondents, 54 respondents and 61 respondents, respectively. This shows that respondents think the voice of narrator and interview video about common depression questions are clear and understandable for Question 8 and Question 9. Moreover, most of respondents feel very enjoy and satisfied when using this application. In contrast, only one respondent chooses No as his or her answer for Question 7 and it same goes to Question 9. They may think that the text and interview video in this project can be improved.

Based on the test result analysis gathered from three categories of respondents, it can be concluded that the 2D animation received a good feedback and suggestions. The 2D animation is strongly accepted to be used as a medium to give the depression information and raise depression awareness to its users.

## V. DISCUSSION

In the development of the project, the strengths and weaknesses were discovered by gathering all the test result analysis. This project is attractive as it consists of many multimedia elements and interactive buttons, which able to attract the user attention. In addition, the characters of the animation are well designed for presenting the depression information. Besides, the flow or movement of the animation are nice and smooth. In addition, the procedure of the project is designed nicely to make it ease of use for the users. The users can easily understand the information presented. The most important thing is the information presented is clear and accurate since it is verified by the Subject Matter Expert which is the counsellor.

However, this project is lack of activity modules. By adding activity modules such as quiz or games, users can interact more with the application and their understanding towards depression can be increased. Moreover, users can easily understand every modules if the explanation of each modules is simple and understandable. Improved quality of the video interview also can give a pleasant and nice view for the users. Therefore, users can easily concentrate on the content of the video. In addition, users can be more enjoyed and pleased by using 3D animation to have a more attractive view that 2D animation.

## VI. CONCLUSION

In conclusion, the 2D animation has developed successfully. The objectives have been achieved where the interactive content development provides the complete depression information to the public. It is envisaged that the developed interactive content would greatly contribute to increase the understanding and awareness to the public especially tertiary students. It is hoped that the developed application will assist the counsellor, parents and community in delivering the depression information efficiently.

This project contributes to public, especially tertiary students by increasing their understanding and awareness on depression. Depression among students can be identified and proper treatments can be provided. Furthermore, help and encouragement can be offered to tertiary students' family or friends who are depressed. This project also contributes to UTeM Counselling Unit and any university and college counselling unit for future use.

## ACKNOWLEDGMENT

This project is part of Final Year Project of Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka. This project won Silver Medal at the e-Learning Carnival & Conference 2018 (eLCC 2018) organized by Centre for Instructional Resources & Technology (PSTP), Universiti Teknikal Malaysia Melaka.

The authors are thankful to the Counsellor from UTeM Counselling Unit for providing the information and involvement in the testing phase. In addition, the deepest gratitude and thanks go to the lecturers from the Department of Interactive Media, Faculty of Information and Communication Technology, UTeM for their expertise during the testing evaluation of the product.

## REFERENCES

[1] D. J. Kirsch, L. A. Doerfler, M.A. Truong, "Mental health issues among college students: who gets referred for psychopharmacology evaluation?", *Journal of American College Health* vol. 63 no. 1, pp. 50–56, 2015.

[2] Institute for Public Health (IPH), "National Health and Morbidity Survey 2015 (NHMS 2015)", *Non-Communicable Diseases, Risk Factors & Other Health Problems*, vol. 2, 2015.

[3] N. Ibrahim, S. Mohd Sidik, C.K. Phang, F. Mukhtar, H. Awang, J. K. Ang, Z. L. Osman and S.F.A Ghaffar, "Prevalence and predictors of depression and suicidal ideation among adolescents attending government secondary schools in Malaysia". *Med Malaysia* vol. 72 no.4, pp. 222-227, 2017.

[4] American Psychiatric Association, 2018, What is Depression?, available online: https://www.psychiatry.org/patients-families/ depression/what-is-depression, last visit:09.03.2018

[5] World Health Organization, 2017, Depression: let's talk. available online:http://www.who.int/mentalhealth/management/depression/en, last visit:06.03.2018

[6] The Star Online, 2017, Depression: 40% of Malaysians will suffer from mental health issues in their lifetime, available online: https://www.thestar.com.my/news/nation/2017/04/02/malaysians-will-suffer-from-mental-health-issues-in-their-lifetime/, last visit:06.03.2018

[7] World Health Organization, 2017, Depression, available online: http://www.who.int/mediacentre/factsheets/fs369/en/, last visit:08.03.2018

[8] The Sun Daily, 2016, Mental health of Malaysian students worrisome, available online: http://www.thesundaily.my/news/1967970, last visit: 10.03.2018

[9] N. Kumaraswamy, "Academic Stress, Anxiety and Depression among College Students- A Brief Review". *International Review of Social Sciences and Humanities* vol. 5, no.1, pp. 135-143, 2013

[10] T. M. Khan, S. A. Sulaiman and M. A. Hassali, "Mental health literacy towards depression among non-medical students at a Malaysian university", *Mental Health in Family Medicine* vol.7, no. 1, pp. 27–35, 2010

[11] Y. Nam, J. M. Bertolote, B. H Chia, T. Maniam, M. R. Phillips, J. Pirkis, and H. Hendin, "Creating Public Awareness in Asia of Depression as Treatable and Suicide as Preventable", *Suicide and Suicide Prrevention in Asia*, pp. 31-38, 2008

[12] India Education, What is Animation, available online: http://www.indiaeducation.net/animation/what-is-animation.html, last visit:11.03.2018

[13] R. Rosdin, Depression motion graphic awareness campaign, available online: https://www.youtube.com/watch?v=9SY hO ED iIVA&t=3s, last visit:14.03.2018

[14] IRDResearch, Depression & Anxiety Awareness video- English, available online: https://www.youtube.com/watch?v= 2s8PYL Mp7RQ, last visit:14.03.2018

[15] TED-Ed, What is depression? - Helen M. Farrell available online:https://www.youtube.com/watch?v=z-IR48Mb3W0, last visit: 14.03.2018

# Model for Predicting Educational Domain Rate based on the Regional Level

Arslan Tariq[1], Iqra Tariq[2], Aneela Abbas[3], Muhammad Aadil Butt[4], Maimoona Shahid[5], Umair Muneer Butt[6]
University of Lahore, Gujrat Campus, Pakistan.

***Abstract*—The geographic information system (GIS) is rapidly becoming the part of current technology trends. GIS can be used to identify the factors that become the reason for an individual to adopt a field or subject. We used GIS as a major tool with the other technologies to identify the key factors. This research has analyzed that mostly people used to migrate to other cities due to unavailability of resources in their own region. Collection of data was done with the help of Survey 123 through which we were able to collect location coordinates of participants. After that, Pilot study approach used to conduct this research. Results show` that mostly user preferred to move to other cities due to unavailability of programs in local institutes. The overall idea can be used to improvement of local institutes and this research can also be used for proper and efficient allocation of facilities and resources in a region, which in turn can save money and time.**

***Keywords*—Geographic information system; education; domain; technology; region**

## I. Introduction

The combination of GIS with online social networking have embedded GIS and location-based services into daily routine of people. This pattern not just gives the GIS people group a remarkable memorable open door for public engagement but additionally brings up some major issues about the importance and role of GIS in education. For the rapid growth of practical development of GIS [8], it is necessary that we begin a genuine exchange on what, why, and how we should instruct and prepare our understudies (and the general population) about GIS and related fields [9, 12]. Numerous have contended that the advancement of spatial knowledge must be given more unmistakable quality in training at all levels, if the up and coming age of clients of geospatial advances, including geospatially empowered online networking, is to make viable and dependable utilization of them.

With GIS and mapping technologies increasingly being used to illustrate issues ranging from earthquake relief and environmental disasters [14] to human rights abuses and the on-going war on terrorism, what extra information and abilities are required [1]? Is GIS training at last about geographic instruction [1]? Assuming this is the case, maybe GIS can take in something valuable from geographers' endeavors to connect with the general population and even conceivably to change the world in significant ways [1]. The GIS people group has a pleased record of drawing in the general population through research on open investment GIS and most as of late through volunteered geographic information (VGI) and online networking [1]. What new aggregate procedure would it be a good idea for us to create in our effort endeavors and open engagement?

GIS can be used to provide a decision-making tool for educational sectors [11] to visualize and monitor the safety, mapping buildings and infrastructure [14, 12], routing transport of campus and recruitment. Students can use GIS to be aware of the contents of different fields and disciplines they are interested in [1, 12]. It can also provide in-demand career pathways and incorporate fieldwork. It also helps the students to explore the relationship among different geographical contents including people, climate, river systems, tree species, historical buildings, water quality, soils, landforms, natural hazards and much more. It helps students to think, understand about the world and explore content knowledge.

The geographic data can be analyzed to discover relationship across time and space [1]. GIS requires thinking critically to study the relationship among various variables including latitude, altitude, and climate and much more for example in the field of cotton production to find a suitable growing area for cotton [15]. Through GIS, spatial analysis becomes necessity for scientists, business persons and epidemiologists to solve daily problems as the world is changing rapidly by meandering rivers [13], erupting volcanoes, and human forces and shifting planets. e.g. With the help of GIS tools and applications, researchers, managers and students can think critically and scientifically to make better decisions for making differences in the changing world.

This study aims to find different resources [13, 14, 15] and fields of study by gathering data geographically. Students will be grouped according to their spatial, educational requirements through an online pool [10]. The interests of students will be critically analyzed and then will be informed with the list of related studies or localities for their future interests to pick related career pathways. This will help students to make better decisions for their future interests and careers ahead.

## II. Related Work

Geographical region based education is an old phenomenon but it is a filed in which research limited to few research articles. In 20th century, students collected their data about their own city and village and send it to other regions of country so that people use it [2]. In 1970, the Foxfire project in Georgia urged a national movement, which aimed, was to investigating, collecting and documenting the regional based educational data across the country. The Foxfire books and magazines were highly famous and successful that time,

students and teachers in other parts of United States used that data in their research and created the similar educational data for others [2, 3, 11]. Some students and teacher also created same type of data other than Foxfire but these never achieved the popularity of Foxfire [3, 10].

The region-based education do not have to be limited to research and studies. Students of Akula School (K-12) [10] have designed a website of their school which contained information related to their school and community. Heartland Community School in Henderson did the same thing [2, 3]. There are different factors that lead many young people to leave their communities [3], is unavailability of resources like higher institutes, unavailability of interested degree program, unavailability of demanded degree program according to market and such others. To overcome these factors students, want to move other regions.

## III. Materials and Methods

This section presents the material and methods. First of all study list of graduated students and their regional background needed to identify the geographical region-based acceptance and availability of certain filed, subject or domain. For this purpose, data was acquired through survey. Survey 123 for ArcGIS used for this purpose to collect the data from different graduated students. Survey123 for ArcGIS is an easy and intuitive form based field data collecting solution [4] that creates making, sharing, and analyzing surveys attainable in 3 easy steps: ask queries, get replies, and make improved conclusions.

Instead of being map oriented, Survey123 works with forms (or surveys), that are often used to collect data for numerous disciplines. Surveys from easy to sophisticated are authored with Survey123 and then shared in ArcGIS [4]. With the help of Survey123 mobile app, teams in the field can collect survey results with their smartphones and tablets. Surveys can also be completed and submitted in a web browser [4].

Data captured is instantly obtainable for analysis in ArcGIS to assist you and your organization make improved adoptions. Survey123 is a multipurpose data gathering solution already popular with users in public health, public works, education, disaster response, resource management and humanitarian agencies [4].

From that online survey, we identified their regional background as well as collected the information and location of their institutes in form of x and y coordinates. After collecting the data through this tool, pilot study approach used to retrieve the refined form of data. Pilot study can be defined as preliminary study that leads toward main study. It is useful to evaluate the time, cost and feasibility [5]. It is convenient for researcher to evaluate different factors in pilot study so that main study become more refine and error free. Pilot study can be defined as primarily study, which is used to investigate whether crucial components of main study are feasible or not [5]. A pilot study is a standard scientific tool for 'soft' research, permitting scientists to conduct an initial examination before committing to a full-blown experiment or study [6].

Collected data had many irrelevant information and some sort of missing information. After removing irrelevant and missed information, data shown and visualize in ArcMap after refinement. Fig 2 shown the detailed methodology for the study. Section 4 presents the results and discussion.

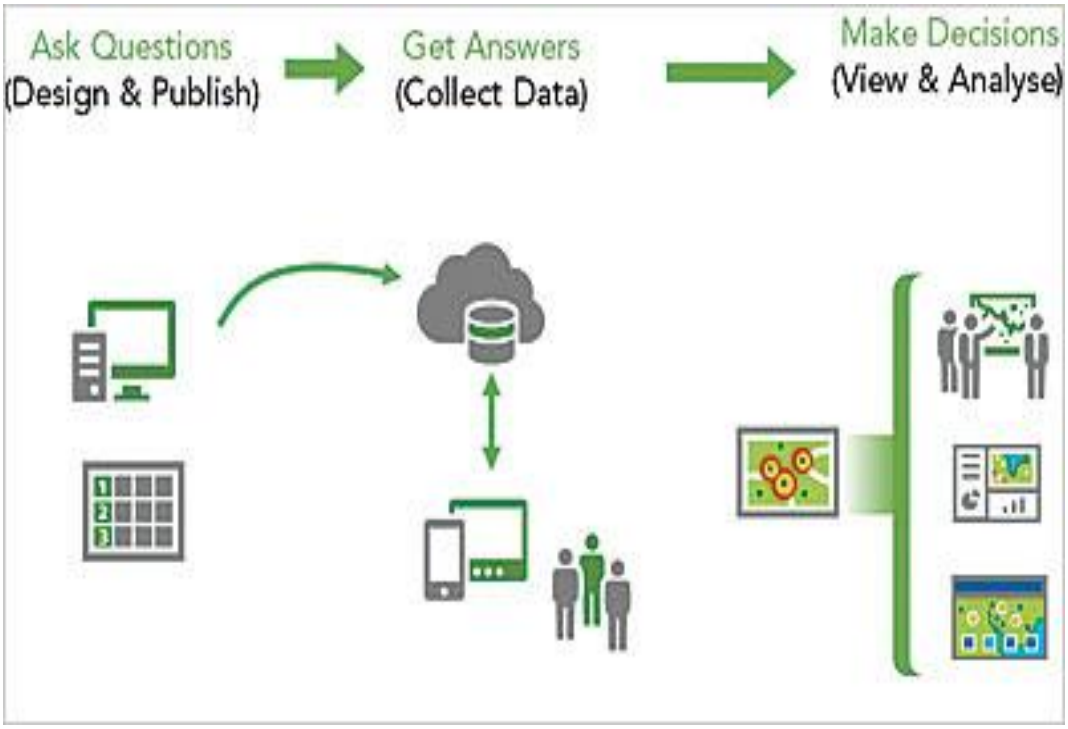



Fig. 1. 1-2-3 Simple Process for Survey 123 for ArcGIS [4].

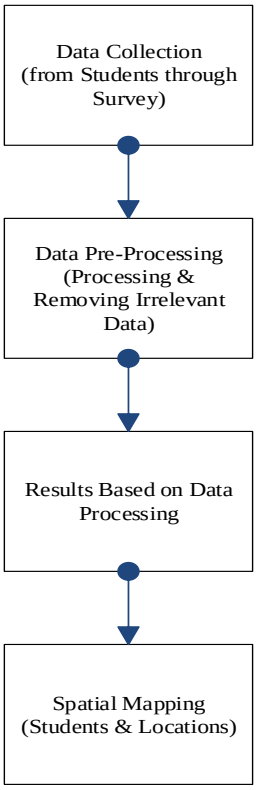



Fig. 2. Detailed Flow Chart of Methodology.

## IV. Results and Discussions

Initially 113 users filled the online survey available online. Data of 61 users was not complete even they did not provide the information of their institutes as well as their coordinates. Some users provided wrong information related to their institutes and coordinates. It was quite difficult to identify wrong information and it was not an easy task. Shape file of Pakistan downloaded from a website [7] and refined data displayed on it in form of .csv. The shape file also have all the coordinates of data that collected through survey. Fig 3 shown the spatial mapping of collected data results.

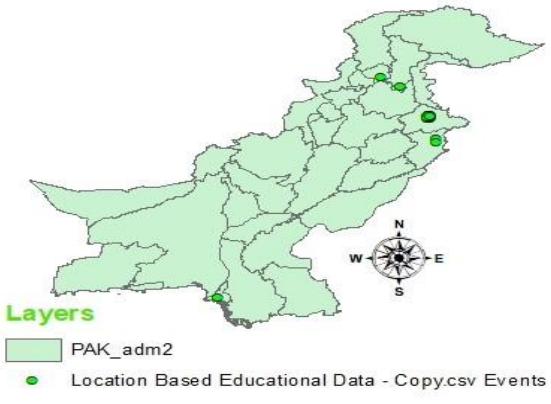



Fig. 3. Spatial Mapping of Collected Data Results.

52 participants filled the survey correctly and provided all he information correctly in all manners. 30 users participated from District Gurjat, 19 from District Gujranwala, 1 from Azad Jamu Kashmir, 1 from Karachi and 1 from Swabi. From the results of survey, we can say that 49 participants participated from Gujranwala Division. There are 21 females and 31 males and there are 7 professionals and 45 student out of 52, filled the survey.

47 participants provided the coordinates of their institutes located in Gujranwala Division, 2 participants marked the location of their institutes in Lahore, 1 participant mark in Islamabad and remaining 2 marked their location of institutes in Mardan and Karachi respectively. Fig 4, 5, 6, 7, and 8 shown the marked location of above-mentioned results.

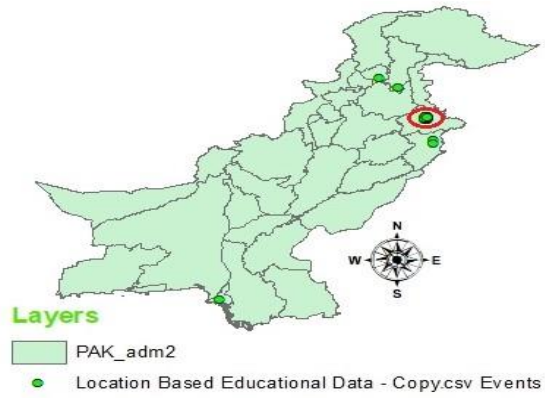



Fig. 4. Marked Location of Institutes Located in Gujranwala Division.

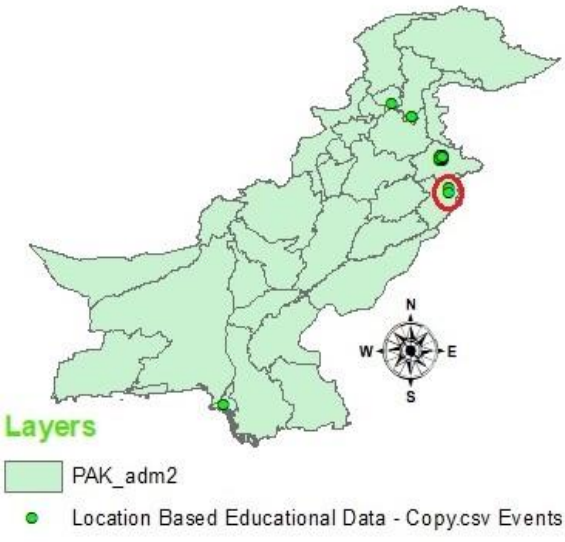



Fig. 5. Marked Location of Institutes Located in Lahore.

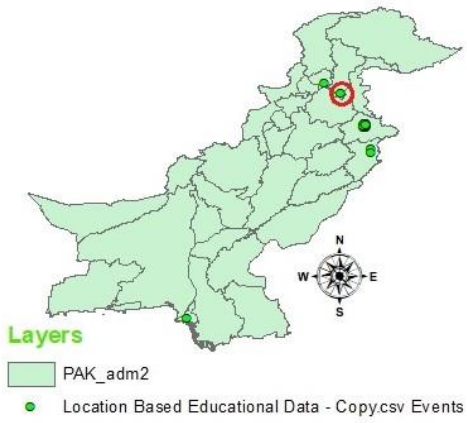



Fig. 6. Marked Location of Institutes Located in Islamabad.

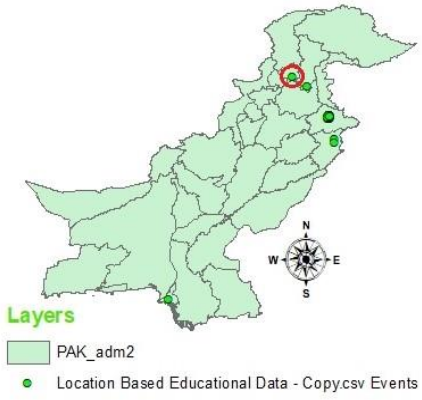



Fig. 7. Marked Location of Institutes Located in Mardan.

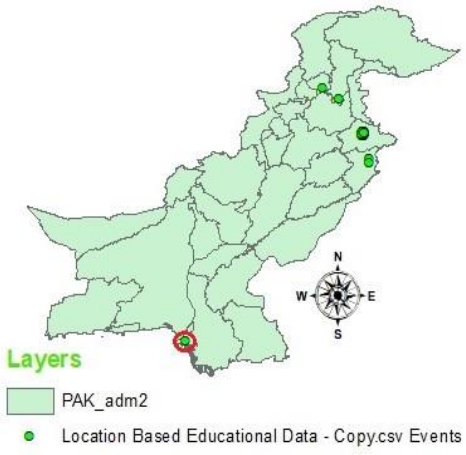



Fig. 8. Marked Location of Institutes Located in Karachi.

The qualification level of 25 participants were graduation, 18 participants had background of intermediate level and rest of 9 had the qualification level of master. Fig 9 shown the qualification level of participants.

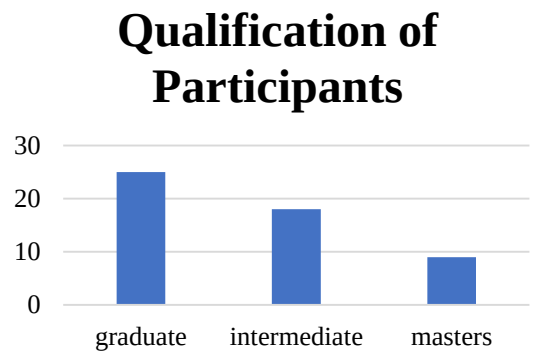



Fig. 9. Qualification Level of Participants.

The major subjects of 20 participants was computing, 15 participants belonged to medical study, 10 participants belonged to engineering study and rest 5 participants and 2 participants belonged to business and arts study respectively. Fig 10 shown the major subjects of participants.

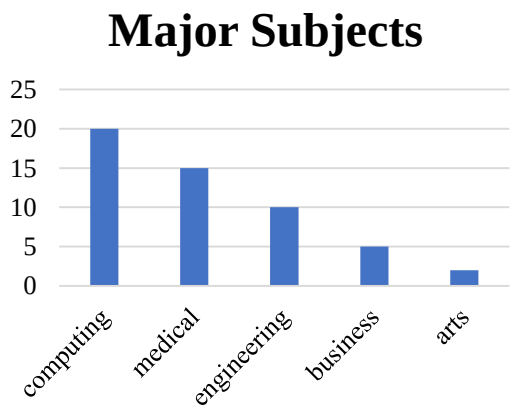



Fig. 10. Major Subjects of Participants.

The 29 participants provided the information that they were interested in the current subjects, 10 participants selected the subjects due to demand of the subjects, 8 participants selected their subjects due to influenced by someone and rest of 5 participants selected their subjects due to growth rate of that subject was increasing. Fig 11 shown the reason for selecting the field of study.

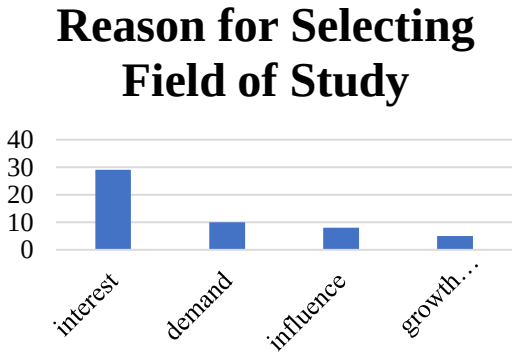



Fig. 11. Reason for Selecting of Field of Study by Participants.

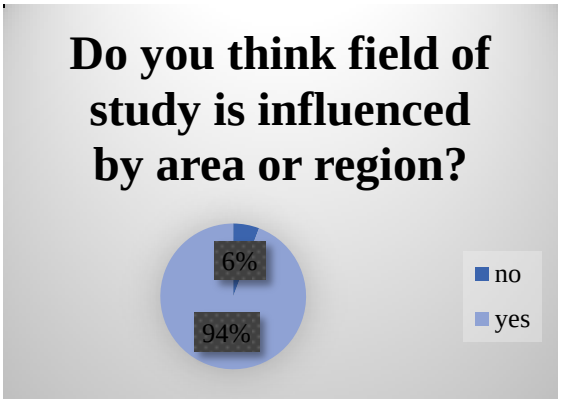



Fig. 12. Field of Study Influenced by Area or Region.

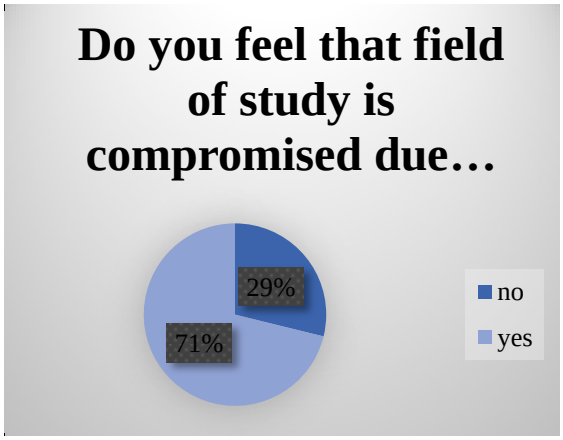



Fig. 13. Unavailability of Degree Programs in Local Institutes.

The question "Do you think field of study is influenced by area or region?" was asked in the survey. The 94% participants answered yes and only 6% answered no of that question. Fig 12 shown the percentage result of question answered by the participants.

The question "Do you feel that field of study is compromised due to unavailability of degree programs in local institutes?" was asked in the survey. The 71% participants answered the question as yes and 29% participants answered no of that question. Fig 13 shown the percentage result of asked question by the participants.

Our research problem was to identify the geographical region-based availability and acceptance of certain domain/subjects/fields. This study guided us that there are some key factors that cause both geographical region and individual to opt or adopt any field". We have mentioned some of those key factors that have influence on individual to adopt any field regarding to their studies.

## V. Conclusion and Discussions

This study proves that the geospatial technologies can be effectively and efficiently used to identify the key factors that cause both geographical region and individual to opt or adopt a specific subject. We had used some simple and effective techniques to collect the data and after processing we were able to find the factors that effects the region based facilities and unavailability of resources. Form the survey, we analyzed that 71% participants compromised on their study due to unavailability of resources. Some of them moved to other cities and some adopted the subject due to demand, growth rate or influenced by someone. Some of them compromised due to unavailability of subject of interest in local institute and select or opt a subject that have better growth rate or demand in the market, so they join local institutes.

## VI. Limitations and Future Work

There are few limitations associated with this study that need to be addressed and improved. We faced some difficulties during conducting of this study. The main problem was incomplete information from participants. Mostly participants did not provide all the asked information and due to missed information we decided to conducted pilot study on remain data. In future, we will conduct main study using this pilot study and will find out more constraints. We will also conduct this study based on some factors like co-education, marks in previous degree and level of satisfaction of users after adapting the subjects or filed.

## References

[1] D. Sui and M. Goodchild, "The convergence of GIS and social media: challenges for GIScience", *International Journal of Geographical Information Science*, vol. 25, no. 11, pp. 1737-1748, 2011.

[2] A. Murphy, "Enhancing Geography's Role in Public Debate∗", *Annals of the Association of American Geographers*, vol. 96, no. 1, pp. 1-13, 2006.

[3] G. Smith, "Place-Based Education: Learning to Be Where We are", *Phi Delta Kappan*, vol. 83, no. 8, pp. 584-594, 2002.

[4] "5 Reasons to Use Survey123 for ArcGIS | ArcUser", *Esri.com*, 2018. [Online]. Available: http://www.esri.com/esri-news/arcuser/spring-2017/5-reasons-to-use-survey123-for-arcgis. [Accessed: 01- Jun- 2018].

[5] "What is a pilot study? - Students 4 Best Evidence", *Students 4 Best Evidence*, 2018. [Online]. Available: https://www.students4bestevidence.net/pilot-studies/. [Accessed: 08- Jun- 2018].

[6] "Pilot Study in Scientific Research - Testing a Study before Launching", *Explorable.com*, 2018. [Online]. Available: https://explorable.com/pilot-study. [Accessed: 08- Jun- 2018].

[7] "Pakistan Shape files | Pakistan GIS", *Pakistan GIS*, 2018. [Online]. Available: http://citypulse.com.pk/pakistangis/tag/pakistan-shape-files/. [Accessed: 08- May- 2018].

[8] R. Johnston, "The point is to change it: geographies of hope and survival in an age of crisis. castree, n., chatterton, p., heynen, n., larner, w. and wright, m. w. (eds)", *Geografiska Annaler: Series B, Human Geography*, vol. 93, no. 4, pp. 338-340, 2011.

[9] A. Murphy, "Enhancing Geography's Role in Public Debate∗", *Annals of the Association of American Geographers*, vol. 96, no. 1, pp. 1-13, 2006.

[10] "Learning to think spatially: GIS as a support system in K-12 education." *National Academies Press*, 2006.

[11] O. Darkwa and F. Mazibuko, "Creating virtual learning communities in Africa: Challenges and prospects", *First Monday*, vol. 5, no. 5, 2000.

[12] W. Kresse and D. Danko, *Springer handbook of geographic information*. Berlin: Springer, 2012.

[13] M. Clark, "Putting water in its place: a perspective on GIS in hydrology and water management", *Hydrological Processes*, vol. 12, no. 6, pp. 823-834, 1998.

[14] W. Craig, T. Harris and D. Weiner, *Community participation and geographic information systems*. New York: Taylor & Francis, 2002, pp. 113-124.

[15] M. McCALL and P. MINANG, "Assessing participatory GIS for community-based natural resource management: claiming community forests in Cameroon", *The Geographical Journal*, vol. 171, no. 4, pp. 340-356, 2005.

# Crypt-Tag Authentication in NFC Implementation for Medicine Data Management

Z. Zainal Abidin[1], N. A. Zakaria[2], N. Harum[3], M. R. Baharon[4], Ee-Song Hong[5]
Information Security Forensics and Computer Networking
Universiti Teknikal Malaysia Melaka (UTeM)
Melaka, Malaysia

Z. Abal Abas[6]
Optimization Modelling Analytic and Simulation
Universiti Teknikal Malaysia Melaka (UTeM)
Melaka, Malaysia

Z. Ayop[7], N. A. Mat Ariff[8]
Information Security Forensics and Computer Networking
Universiti Teknikal Malaysia Melaka (UTeM)
Melaka, Malaysia

***Abstract*—This study focus on the implementation of expiry date detection for medicine using RFID in the health care industry. The motivation for doing this research is the process of searching for the expired medicine is a time consuming and lack of security features included in current NFC implementation. Therefore, the objective of this research is to study the RFID technology used for detecting medicine expiry product and to develop a new system that integrated NFC with authentication feature. Moreover, the problem of current data management for medicine still using manual or barcode system that lead to inconsistency, easy duplication and human error. Here, the NFC is chosen, due to smaller distance of signal coverage, since less interference and the time spending for sniffing activity by the hacker can be reduced. The system is developed using C#, SQLite, Visual Studio, NFC Tag and NFC reader (ACR122U-A9). Experiments have shown that the proposed system has produced medicine expiry date system and only authorized person in charge can monitor the medicine. The impact of the proposed system produces safer, greener and easier environment for better medicine data management. The significance of this study gives a medicine expiry date detection system for health care.**

***Keywords*—Expiry date notification; Radio-Frequency Identification (RFID); Near-Field Communication (NFC); internet of things insider threats; health care**

## I. Introduction

Radio-Frequency Identification (RFID) has been introduced to improve the supply chain process and increase the productivity in various fields. The importance of RFID gives a better solution in terms of lightweight device and remote detection to the industry since it is a small electronic device, that consist of a small chip and an antenna. In fact, RFID increases traceability of product through assembly line or warehouses which provides a unique identifier for an object based on magnetic circuit that must be scanned to get the information. However, current RFID based systems have no data privacy features and there is some limitation to access the RFID devices. Assuming a person has the same radar, they might have access to receive all the data capture and there is a limitation in the frequency length which makes the connection slow. Therefore, the proposed tracking system with security features is developed by using RFID. Nevertheless, a short-range frequency of RFID is used for the prototype development due to its small coverage of data signal that permits data transmission within the respective distance. In fact, the use of short-range frequency features is to prevent the interference in frequencies which another RFID reader captures information from another RFID reader data of a product or values due to its characteristics of long distance range. On the other hand, most of the manufacturer still using manual operations, such as Barcode, QR code [1] and RFID [2]. In the field of Information Security, access control is one of the method to protect or control the systems and infrastructure for physical security implementation such as accessing the building or premises, monitoring and surveillance systems and physical access infrastructure for instance door and window. Thus, in the access control method consists of smart card, biometrics, bar code, QR code and RFID. Each of these technology has its own specialty and limitation. The NFC is a short range type of RFID that provide a smaller distance of frequency coverage and performs basic tasks of scanning and tagging. The theoretical taxonomy of the existing systems is illustrated as in Figure 1 and explained in sections of A, B, C and D.

A Special thanks to Malaysia Research Assessment (MYRA) and Universiti Teknikal Malaysia Melaka (UTeM) for giving a sponsor.

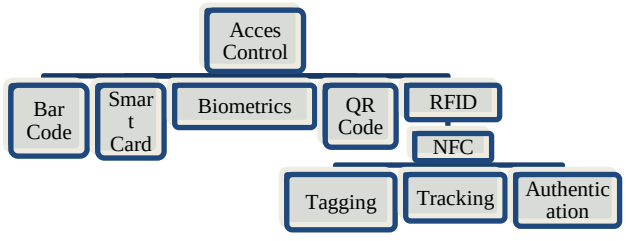



Fig. 1. Taxonomy of Theoretical in Access Control.

### A. Barcode

The barcode system is available in various fields for example healthcare, transportation, manufacturing and retail stores. In fact, the barcode is useful for users to keep track of inventory easily which can make time spent doing inventory checks are minimized and at the same time reduces rates of human error [3].

In medical industry is not the only industry that uses barcodes. Other industry that using barcode includes food, transportation, manufacturing and industrial [4]. A bar code (often seen as a single word, barcode) is the small image of lines (bars) and spaces that is affixed to retail store items, identification cards, and postal mail to identify a particular product number, person, or location. The code uses a sequence of vertical bars and spaces to represent numbers and other symbols. A bar code symbol typically consists of five parts: a quiet zone, a start character, data characters (including an optional check character), a stop character, and another quiet zone as shown in Figure 2.

By using barcode scanners, it helps users to keep track of inventory easily which can shorten the time of work as well as reduces rates of human error. Furthermore, some of the features can allow users to set up alerts for items that are out of stock so orders can be filled promptly [6].

The disadvantages of barcode scanners are it needs a direct line of sight and need to close to the barcode for doing scan or read. Besides that, barcodes have no read or write capabilities. In fact, barcodes have a weak security element as they can be more easily reproduced or forged, it can easily damage which means if a barcode is ripped or damaged there is no way to scan the product. Although barcode scanner is cheaper, but in the long run it requires a lot of money for manually entering information into the system that must be scanned continuously from one product to another till to the end of line.

On the other hand, Quick Response code (QR code) is a two-dimensional barcode that can be read via QR barcode reader or camera and is able to transmit data both in vertical and horizontal track, which is why it is named a 2D barcode. The most significant objective of using QR code is the traceability or monitoring the system [1]. As stated by [4], one of the advantages of the QR Code is that it eliminates the need to type WEB addresses which it is only necessary to launch the application and point the cell phone at a QR Code for the additional content to be displayed in the reader or Web browser.

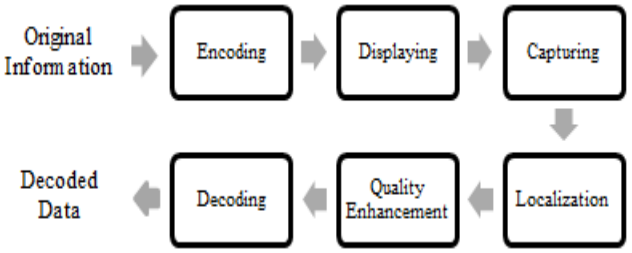



Fig. 2. Barcode Process Diagram [5].


Fig. 3. Left is the Correct Version, Right is the Spammed [8].

However, QR code has a problem that can easily duplicate and manipulate the accurate data, the physical spaces are the most vulnerable to spamming pointing to unsolicited content over the internet [7]. However, the first kind malicious detected by Kaspersky lab which is the attack method used in the QR code was that when a user scans the code, he is directed towards a website and then a malicious file downloads in the user's device without the knowledge of the user [8]. Figure 3 shows how file is modified without being noticed by a user.

### B. Radio-Frequency Identification (FRID)

RFID consist of a tag attached to a product which classifies and tracks the product via radio waves. Besides, tags can carry up to 2000 bytes of records per second. The basic components of an RFID system are a tag, scanner, antenna, writer, control equipment and software [9]. RFID able to capture any added information such as expiry date but RFID contains a limitation space to store information especially about the manufacturer and product.

However, RFID is highly used instead of barcode because it cannot be easily duplicated based on circuit based chip, which represents unique identification number and address [2]. For example, an Italian Food Company Barilla had launched an RFID-enabled product which exploring the use of both passive and active radio frequency identification technology to help track ingredients at the same time maintaining the quality and food safety [10], which the topology of RFID block diagram is illustrates in Figure 4. In fact, RFID technology can avoid or decrease sources of errors, decrease of labor costs and the saving of inventory inaccuracies. [11]. Moreover, the use of RFID prevents sniffing or Eavesdropping, Spoofing, Cloning, Replay, Relay and Denial of Service Attacks are some of the security threats [12].

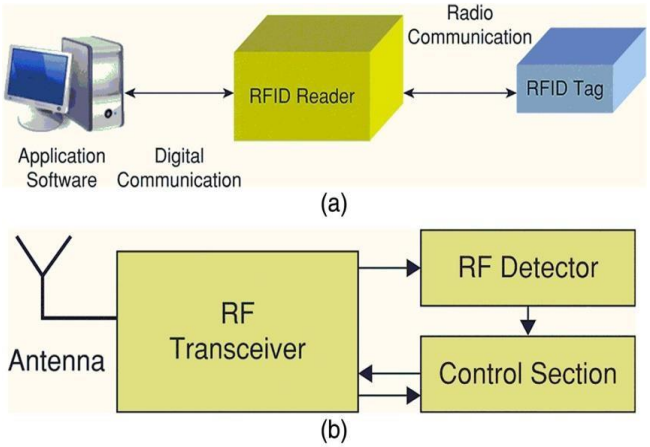



Fig. 4. RFID Reader Block Diagram [10].

However, most organizations control the supplies movement, effective inventory management and maintain productivity, demand higher safety and security checks monitoring [13]. Since the messages are transmitted by using radio waves, RFID systems and protocols are still directly using the real tags' identity, thus, for these reasons, RFID users can be affected by broadcasting the content of the RFID tags because any malicious RFID reader can track the location or obtain the identification and private information [14].

Based on previous studies [15] discussed on the security as the main problem in RFID system since the communication between RFID components is wireless which is still demand a better solution for authentication and detection. Furthermore, RFID is lack of authentication feature of security [16] and vulnerabilities have been found in RFID tagging phase [17], which lead to Man-in-the-middle attack.

A major criticism work in [18] explains that most companies are protecting critical products or items and supplies moving, manage inventory level effectively, maintain productivity, improve safety and security, in order to confirm or agreed requirements and keep emergency transport costs in check. This inconsistency may be due to improve inventory organization at the retail store and together with the supply chain. However, Sun [19] points out that the prime asset for an inventory system is an automatic identification technology system. One of the best examples is the Auto-ID based on RFID technology. This asset is on behalf of two reasons. First, the perceptibility provided by this technology allows a precise information of inventory level by removing the difference between inventory record and physical inventory. Second, RFID technology can avoid or decrease sources of errors. A reasonable approach to tackle this issue could be to decrease of labor costs, the popularization of business processes and the saving of inventory inaccuracies.

It seems that a change of high social concern and high contribution services such as services with banks as opposite to groceries stores [20] have more impact on user's privacy issues when they use QR codes to access the business web sites. Moreover, results suggest that privacy issues when using QR code and [21] describes sniffing or Eavesdropping, Spoofing, Cloning, Replay, Relay and Denial of Service Attacks are some of the security threats to RFID technology. One possible methods being used as countermeasures to the security threats presented against RFID technology.

Park and Lee [22] holds the view that there are many kind of authentication technologies that are developed to protect personal information. However, if the NFC-based services are being used widely, the efficiency and payment information protection of these technologies must be ensured. Initial observations suggest that there may be a link between protection of a user and service provider in NFC-based. Alqarni finds an accurate authentication protocols between the tag and the back-end server is the crucial issue [23]. Since the messages are transmitted by using radio waves through the air in RFID systems and those protocols are still directly using the real tags' identity such as tag identifier and secret key in the authentication phase. For these reasons, RFID users can be affected by broadcasting the content of their RFID tags because any malicious RFID reader can track their location or obtain their identification and private information. The data reported here appear to support the assumption that solving RFID authentication issues that go a long way to persuade people that using these tags are not expose their secret data.

Recent research [24] has suggest that there are many types of applications that can be run using NFC technology. The fact that the lower layer of NFC includes no communication security primitives makes this technology exposed to a wide range of vulnerabilities and attacks. A possible explanation for this might be that NFC security issues.

However, previous method suffers from serious drawbacks and time taken for the valuable information is extracted from the RFID data that is too long [25]. Meanwhile, the object moves within the distance of a RFID reader, it reads the tag attached on that object. It seems possible that these results are due to reading the same tag so the duplicated data or information has been generated. In fact, based on survey conducted [26], respondents agree that human interactions with smart devices makes people feel more confident as the technology becomes an inseparable part of human life.

### *C. Near Field Communication (NFC)*

The technology involved is deceptively simple: evolved from radio frequency identification (RFID), an NFC chip operates as one part of a wireless link. Once the NFC chip is activated by another chip, small amounts of data between the two devices can be transferred when held a few centimeters from each other during recent years. Mobile phones with NFC enabled has become widely used and most common itinerant computing devices, playing an important role socially, emotionally and recreationally. The innovations in communication networks particularly in mobile phones technology have made it prone for a broad range of applications. Thus, Near Field Communication (NFC) service, as one of the most recent technologies in telecommunication area, is going to be developed around the world through transformation from initial testing to full scale deployment. In fact, Near Field Communication (NFC) is a type of passive 13.56 MHz RFID technology that enables short-range wireless data transmissions at 4 centimeters (1.6 inches) or less [27] and it lets consumers use NFC enabled mobile devices to interact with RFID tags or other NFC-enabled devices and products.

The number of applications in which NFC technology is widely used including application which is new secure system should be proposed for managing security in complex mobile and variable conditions such as secure payment tools, access management and retailing industry among others. Keeping smartphone secure is a very fundamental nonetheless, to ensure security for the valuable data that may be found stored in the hand phone is another issue. Users would dislike to waste even a few more seconds of their time to unlock a phone, however, the entire process produce a lot easier with the introduction of technologies like NFC, which faster data transmission and better accessibility with the security feature and based on the standards of communication.

The focus of this research is general security specifications, which should serve as guidelines that cover most of the basic

security requirements within the system [28]. Starting with the initial connection to the system, which is done via NFC, the user is required to provide an identity for optimal protection. As a user taps his or her smartphone next to the tag, then the reader captures that tag's unique ID number [29] which is prompting the phone to access webpage. Although this technology is increasingly becoming main stream, there are issues that need to be addressed mainly regarding on security concerns with Secure Element (SE) personalization, management, ownership and architecture that can be useable by attackers to interval the alteration of NFC within societies. Figure 5 illustrates how the attacker captured data from access control reader through the use of the card emulator in a specific distance. On the hand, hacker injects the malicious codes to the NFC reader to obtain crucial information at the receiver or sender sides yet depending on the frequency range of the NFC device and features available in the hardware and firmware.

In order to protect the NFC device, a lightweight authentication method and a secure way need to be developed in response to these attacks. Thus, adding an authentication element to the NFC provide a better security solution to the user.

The motivation to implement authentication in NFC also inspired by the study done by Jung [30] that stated that with the combination of authentication feature in the NFC contributes a better protection for data management. Figure 6 illustrated Jung case diagram of solution.

Therefore, the proposed study is to develop a system of NFC implementation with authentication for medicine data management using cryptography, named as Crypt-Tag Medicine Data Management. The methodology of the proposed system is explained in section II. Moreover, section III illustrates the results and discussion and section IV concludes the study.

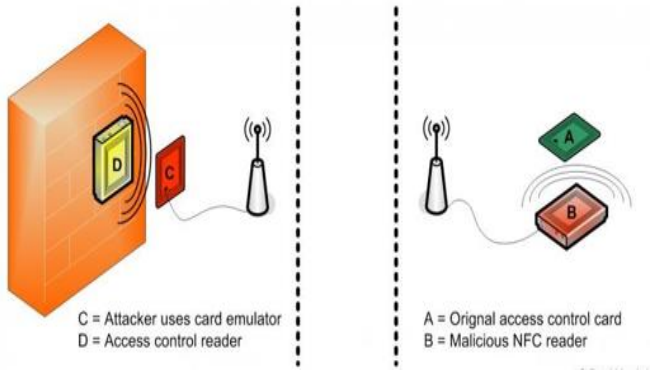



Fig. 5. NFC Attack.

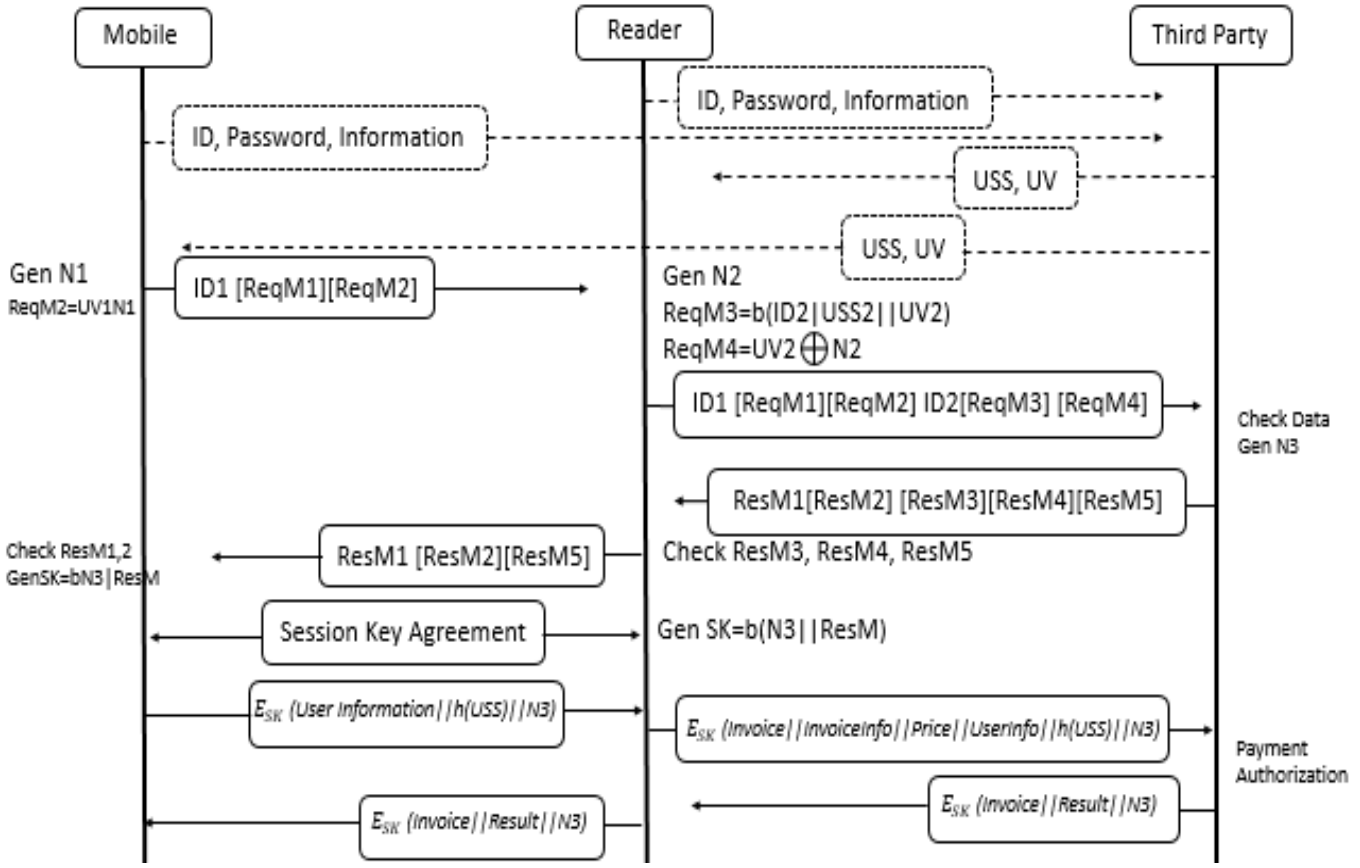



Fig. 6. Case Diagram for Authentication in NFC [30].

## II. METHODOLOGY

The implementation of the Crypt-Tag Medicine Data Management system is using waterfall model as shown in Figure 7 that consists of six phases.

### A. Project Planning and Feasibility Study

In phase 1, the purpose of the project planning and feasibility study needs to gather all information and requirement such as to determine the objectives, to identify the current technology used in health care and to find the suitable software and hardware to be used in the execution of the project. In this phase, research needs to include the implementation idea that involves NFC, how it works and the system requirement to fulfil the study needs.

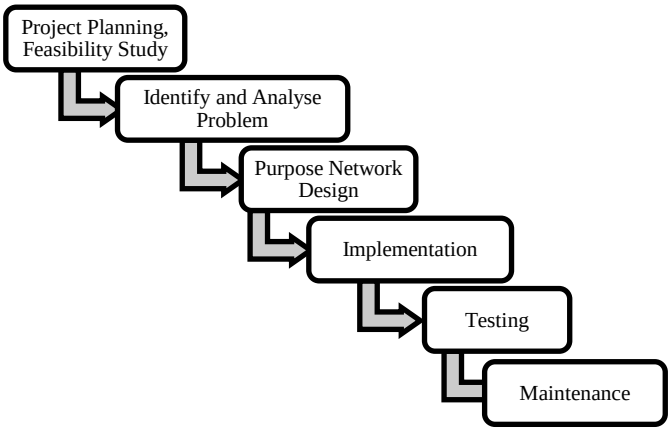



Fig. 7. Waterfall Model Life Cycle.

### B. Identify and Analyze Problem

In phase 2, identify and analyze problem is a process to collect data, to identify the problem and to recommend suggestions in improving the existing system. This phase involves gathering data, finding solution for overcome the limitation of the current system and identify the target users in developing NFC technology. The main objective of this phase is to find out the solution on what, who, when and how the system been and there is a certain technique used to gather data about this study.

### C. Propose Network Design

In phase 3, a new system for the proposed network is designed based on the requirement and analysis such as architecture design, software and hardware selection. Moreover, the user interface for authentication is designed at the beginning of the system using AES cryptography and integrates with the NFC Tag development that called as Crypt-Tag. The AES algorithm is encoded into the authentication system using C# and generates the password into encrypted version. The encrypted password is represented in a series of hexadecimal numbers and alphabets which hacker or intruder is unable to read the real password. The outcome of the proposed system shows that Justin's password is bac363ad53ee1 as shown in Figure 14, which is not his real password.

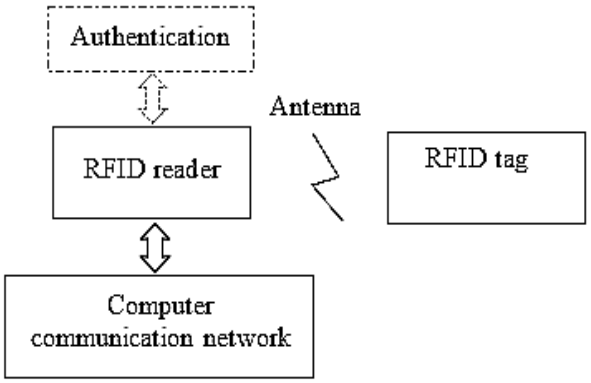



Fig. 8. Proposed Network Design Architecture of Proposed System.

This new design shows the solution to the current network design problem that the existing system of NFC do not has the authentication and cryptography encoded with. All information regarding NFC and detection system are needed to be analyzed in this phase which includes interfaces, NFC reader and NFC tag. The architecture of the network design is created as in Figure 8.

### D. Implementation

Phase 4 shows the installation and configuration of the system development using NFC conducted in the real environment. The hardware and software are required to fully execute the system; and the database with security is setup. This system is using C# for interfaces and SQLite for the database. Sample of system development is provided as in Figure 9.

```
private void staff_id_TextChanged(object sender, TextChangedEventArgs e)
{
    SQLiteConnection sqliteCon = new SQLiteConnection(dbConnectionString);

    try
    {
        sqliteCon.Open();
        string Query = "select * from staff where staff_id = '" + staff_id.Text + "'";
        SQLiteCommand createCommand = new SQLiteCommand(Query, sqliteCon);
        createCommand.ExecuteNonQuery();
        SQLiteDataReader dr = createCommand.ExecuteReader();

        while (dr.Read())
        {


            string sstaffname = dr.GetString(1);
            string sphonenumber = dr.GetString(2);
            string semail = dr.GetString(3);
            string ssusername = dr.GetString(4);
            string sspassword = dr.GetString(5);

            staffname.Text = sstaffname;
            phonenumber.Text = sphonenumber;
            email.Text = semail;
            susername.Text = ssusername;
            spassword.Password = sspassword;
        }

        sqliteCon.Close();
    }

    catch (Exception ex)
    {
        MessageBox.Show(ex.Message);
    }
```

Fig. 9. Data Retrieval from Database.

### E. Testing

In phase 5, unit testing, integration testing, system testing and user acceptance testing are conducted to ensure the system working well. All units developed in the implementation phase are integrated into a system after being tested by each unit to ensure the system is working completely.

### F. Maintenance

In phase 6, researchers need to ensure the requirement statements are fulfilled. This phase dealing with any changes that needed to be done. All recorded data are collected and with the collected results, to determine the effectiveness of the research application. The hardware and software is set up as shown in Figure 10.

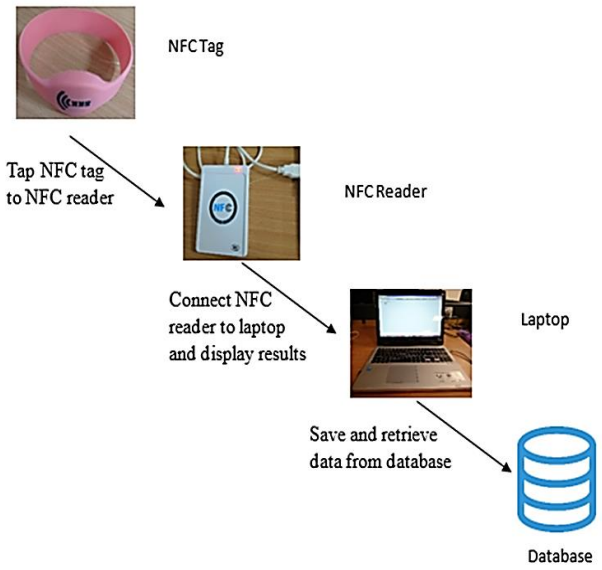



Fig. 10. Hardware and Software Setup of the Proposed System.

## III. Results and Discussion

In this paper, two experiments were conducted during the proposed system development, which is the (a) authentication and (b) detection of medical staff who in charge of the medicine data and medicine records. The first experiment is to develop the authentication feature with AES algorithm, meanwhile the second experiment is to integrate the NFC-Tag Staff Data Management System with authentication.

Figure 11 illustrates the introduction page and Figure 12 shows the authentication using NFC Tag that is scanned by the NFC Reader to access to the proposed system. For this study, user need to click at the "Log in with ID" button for the number from the NFC Tag to display the ID number of the NFC Tag. A user need to key in username and password at the log in page. If the authentication is successful and access is granted, as shown in Figure 13, then only the user able to get to the Crypt-Tag Data Management Page, shown in the main page at Figure 14. If the user is an administrator, then the administrator needs to register to the medical staff who is currently working in the department to avoid misconduct of theft activities in the organization.

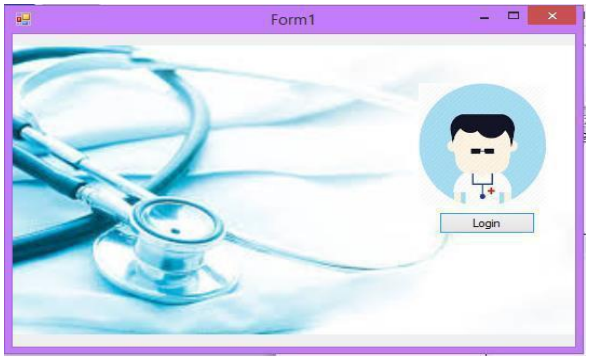



Fig. 11. Introduction to Log in Page.

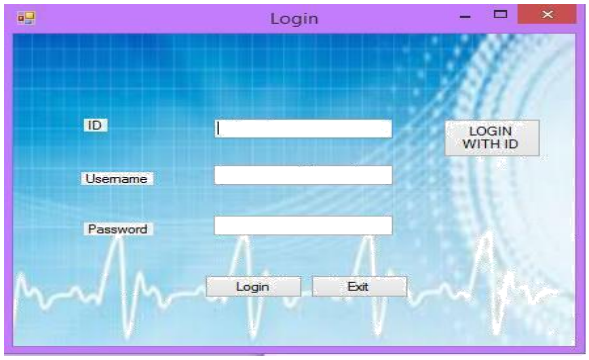



Fig. 12. Authentication using Crypt-Tag for Admin and User.

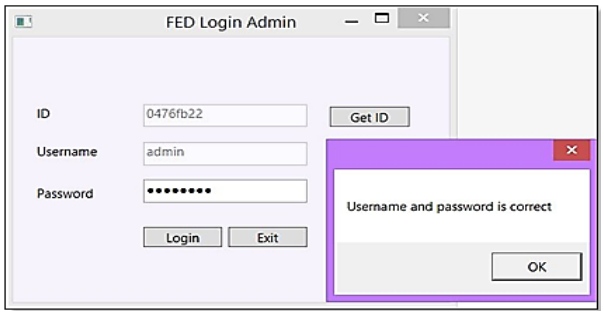



Fig. 13. Administrator Authentication.

The registration of medical staff is done by tapping the staff's identification card (ID) at the NFC Reader and key in the details of the employee; and click the button "Save" to store information into the database as shown in Figure 15. The admin could click the "Exit" button to end the session.


Fig. 14. Crypt-Tag Data Management Page.

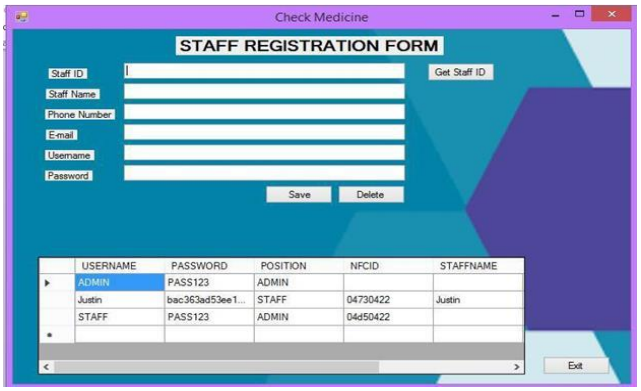



Fig. 15. Integration of NFC Tag Staff Information with Database.

The result indicates that, the staff who has logged-in to the proposed system using his ID (by tapping his ID at the NFC Reader) was Justin. Then, as the NFC Tag of the product is tapped at the NFC Reader, the expiry date and remaining days of the medicine is listed.

Based on our findings, staff can print expired or not expired data, which this process does not only display the item information but also the name of the staff who access the expiry date about the medicine from the proposed system. For future development, the name of the staff no need to be displayed only the medicine name is important.

Another finding that has been encountered is the user interface of this proposed system displayed the process data from receiving and retrieving, which are the list of items tag information with time remaining and current medicine status. This can be improved with an automatically read and save data in database or cloud-based database, which allow the system to be faster and convenience system.

The current system tends to get data leak from several sources such as human being and social media. For example, intruder finds the vulnerability to achieve his or her goal to do malicious activities such as explore or public personnel confidential data or information.

The unauthenticated person easily gain access to an operational process in an organization. For example, attacks will use social engineering such as tailgating or piggybacked to retrieve certain data.

In addition, the current devices used were no security features implemented on it. In fact, there is a lack of security element especially to identify the person who obtain medicine information due to the current devise that not apply encryption technology. Thus, the data or information an easily to achieve by an attacker.

The advantage of this proposed system is it integrates authentication to the proposed system. The authentication features reduce the time taken for sniffing activity by the hacker and able to delay the hacking process before the administrator able to be notified by the system.

## IV. Conclusion

In this paper, the implementation of an expiry date detection using RFID that is NFC (short range) in health care industry. The combination of NFC with medicine data management, helps the health care industry to be more effective and efficient in management. The experimental results have shown the development of authentication and the detection of medicine expiry date. Nevertheless, the proposed system needs to be improved to have a smooth execution, which the information from the NFC Tag appeared instantly without clicking at the "Get Staff ID" button. In future, our proposed system should be able to execute better and able to be applied in other industries that are aligned with the Industrial Revolution.

The future implementation is to automatically get UID tag and save to database. The NFC reader should be able to automatically read NFC tag once the tag is tap to the reader. Suggestion for future work, once the NFC tag is detected by NFC reader, the data will be automatically read and save in the database. For future work, once staff login, they only need to tap medical tag to get data and their name are automatically saved in report. Every user can change their own password. In this project, staff is given a default password which based on their UID and they cannot change it. Suggestion of future work, all staffs can be able to change their own password which make the system more interactive. The system should be able to use cloud or wireless which make it easier to be used by all the user.

## Acknowledgment

High appreciation to the Center of Applied Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi (FTMK) and Universiti Teknikal Malaysia Melaka (UTeM).

## References

[1] I. Tzoulis, and Z. Andreopoulou, "Emerging Traceability Technologies as a Tool for Quality Wood Trade", Procedia Technology, vol. 8, pp. 606–611, 2013.

[2] A. Yewatkar, F. Inamdar, R. Singh, R. Ayushya and A. Bandal "Smart Cart with Automatic Billing, Product Information, Product Recommendation Using RFID and Zigbee with Anti-Theft", Procedia Computer Science, vol. 79, pp. 793–800, 2016.

[3] N.M.Z. Hashim, N.A. Ibrahim, N.M. Saad, F. Sakaguchi and Z. Zakaria, "Barcode Recognition System", International Journal of Emerging Trends & Technology in Computer Science vol. 2 no. 4, pp. 278–283, 2013.

[4] Hashim, N. M. Z., Ibrahim, N. a, Saad, N. M., Sakaguchi, F., & Zakaria, Z. (2013). Barcode Recognition System. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2(4), pp. 278–283.

[5] Changsheng Chena, Alex C. Kota, Huijuan Yang, "A two-stage quality measure for mobile phone captured 2D barcode images", Pattern Recognition, vol 46, pp. 2588-2598, 2013.

[6] R.F. Dos Santos and F.A.S Marins, "Integrated model for reverse logistics management of electronic products and components", Procedia Computer Science, vol. 55, pp. 575–585, 2015.

[7] F. Razzak, "Spamming the Internet of Things: A possibility and its probable solution.", Procedia Computer Science, vol. 10, pp. 658–665, 2012.

[8] A.S. Narayanan, "QR Codes and Security Solutions", International Journal of Computer Science and Telecommunications, vol. 3, no. 7, pp. 1–4, 2012.

[9] E. Ginters, and J. Martin-Gutierrez, "Low cost augmented reality and RFID application for logistics items visualization", Procedia Computer Science, vol. 26, pp. 3–13, 2013.

[10] M. Forouzandeh anf N.C. Karmakar, "Chipless RFID tags and sensors: a review on time-domain techniques", Wireless Power Transfer, vol 2, no. 2, pp. 62-77, 2013.

[11] C. Swedberg. (2016). NFC Tags Take to the Slopes on Ski Jackets NFC Tags Take to the Slopes on Ski Jackets [Online]. Available FTP: http://www.rfidjournal.com/articles/view?15408, C. Sun. (2012). Application of RFID Technology for Logistics on Internet of Things. [Online]. Available FTP: https://doi.org/10.1016/j.aasri.2012.06.019

[12] A. Williamson, L-S. Tsay, I.A. Kateeb and L. Burton, "Solutions for RFID Smart Tagged Card Security Vulnerabilities", AASRI Procedia, vol. 4, pp. 282–287, 2013.

[13] M. Attaran, "Critical success factors and challenges of implementing RFID in supply chain management", Journal of Supply Chain and Operations Management, vol. 10, no. 1, pp. 144–167, 2012.

[14] S.A. Joseph, and N.J. Joby, "Analyzing RFID Tags in a Distributed Environment", Procedia Technology, vol. 24, pp. 1483–1490, 2016.

[15] A. Zavvari, M. Shakiba, M.T. Islam, E. Sundararajan and M.J. Singh, "Computational Cost Analysis on Securing RFID Protocols Conforming to EPC Class-1 Generation-2 Standard", Procedia Technology, vol. 11, pp. 778–784, 2013.

[16] A. Alqarni, M. Alabdulhafith and S. Sampalli, "A Proposed RFID Authentication Protocol based on Two Stages of Authentication", Procedia Computer Science, vol. 37, pp. 503–510, 2014.

[17] T. Feng, M. Hwang and L. Syu, "An Authentication Protocol for Lightweight NFC Mobile Sensors Payment", INFORMATICA, vol. 27, no. 4, pp. 723–732, 2016.

[18] M. Attaran. Critical success factors and challenges of implementing RFID in supply chain management. Journal of Supply Chain and Operations Management, vol 10, no. 1, pp. 144–167, 2012.

[19] Sun, C. Application of RFID Technology for Logistics on Internet of Things. AASRI Procedia, vol. 1, pp. 106–111, 2012.

[20] A. Wilson, T. Hainey, and T. Connolly, Evaluation of Computer Games Developed by Primary School Children to Gauge Understanding of Programming Concepts. Proceedings of the European Conference on Games Based Learning, 2013, pp. 549–558.

[21] Williamson, "Avery, Tsay, L.-S., Kateeb, I. A., & Burton", L. (2013). "Solutions for RFID Smart Tagged Card Security Vulnerabilities." AASRI Procedia, 4, 282–287.

[22] Park, S., & Lee, K. (2014). Advanced Approach to Information Security Management System Model for Industrial Control System. The Scientific World Journal, 2014.

[23] Alqarni, A., Alabdulhafith, M., & Sampalli, S. (2014). A Proposed RFID Authentication Protocol based on Two Stages of Authentication. Procedia Computer Science, 37, 503–510.

[24] Badra, M., & Badra, R. B. (2016). A Lightweight Security Protocol for NFC-based Mobile Payments. Procedia Computer Science, 83(Ant), 705–711.

[25] Joseph, S. A., & Joby, N. J. (2016). Analyzing RFID Tags in a Distributed Environment. Procedia Technology, 24, 1483–1490

[26] Priporas, C.-V., Stylos, N., & Fotiadis, A. K. (2017). Generation Z consumers' expectations of interactions in smart retailing: A future agenda. Computers in Human Behavior.

[27] Mary, B., & Connor, C. O. (2015). New NFC Security Feature Tightens Controls on Tag Data New NFC Security Feature Tightens Controls on Tag Data, 1–2.

[28] Persson, M., & Håkansson, A. (2015). A communication protocol for different communication technologies in cyber-physical systems. Procedia Computer Science, 60(1), 1697–1706

[29] Swedberg, C. (2016). NFC Tags Take to the Slopes on Ski Jackets NFC Tags Take to the Slopes on Ski Jackets, 1–2.

[30] Jung, M. S. (2015). A study on electronic-money technology using near field communication. Symmetry, 7(1), 1–14.

# Thinging Machine applied to Information Leakage

Sabah S. Al-Fedaghi
Computer Engineering Department
Kuwait University
Kuwait

Mahmoud BehBehani
Information Technology Department
Boubyan Bank
Kuwait

***Abstract*—This paper introduces a case study that involves data leakage in a bank applying the so-called Thinging Machine (TM) model. The aim is twofold: (1) Presenting a systematic conceptual framework for the leakage problem that provides a foundation for the description and design of a data leakage system. (2) The aim in (1) is developed in the context of experimentation with the TM as a new methodology in modeling. The TM model is based on slicing the domain of interest (a part of the world) to reveal data leakage. The bank case study concentrates on leakage during internal operations of the bank. The leakage spots are exposed through surveying data territory throughout the bank. All streams of information flow are identified, thus points of possible leakage can be traced with appropriate evidence. The modeling of flow may uncover possible hidden points of leakage and provide a base for a comprehensive information flow policy. We conclude that a TM based on the Heideggerian notion of thinging can serve as a foundation for early stages of software development and as an alternative approach to the dominant object-orientation paradigm.**

***Keywords—Thinging; bank system; abstract machine; software development cycle; heidegger***

## I. INTRODUCTION

In software engineering, models have a central role in achieving a high level of reliability in the design, development, and deployment of systems. Specifically, in this context, we are interested in utilizing a conceptual model, the Thinging Machine (TM); reviewed in the next section), for the specification of the early phase of the life cycle of development in software systems. Without loss of generality, we focus on the problem of leakage of data. Accordingly, our aim is twofold:

*1)* Presenting a systematic conceptual framework for the leakage problem that provides a foundation for the description and design of a data leakage system. The framework is applied to an actual bank system.

*2)* The aim in (1) is developed in the context of experimentation with TM as a new methodology in modeling side by side with other methodologies such as object orientation. We will discuss the justification for pursuing such a venture later in this section.

### *A. Background about Information Leakage*

Information or data (the terms are used interchangeably) leakage has a major impact on the business of many organizations today because valuable data are at risk of loss and possible exposure. The risks include loss of revenue and loss of credibility with customers, shareholders, or society. "Data leakage poses serious threats to organizations, including significant reputational damage and financial losses" [1]. The volume of data leakage has surpassed all expectations, mostly due to processing approaches where enterprises now centrally collect data, instead of keeping data details in various branches, thus maximizing big data benefits. [2].

According to [3], information leakage represents one of "the most common, but misunderstood, security risks faced by business and government alike." Firewalls, intrusion detection tools, and intrusion prevention mechanisms are deployed—"yet, the perception of the secure perimeter may be at odds with reality" [3]. "Despite a plethora of research efforts on safeguarding sensitive information from being leaked, it remains an active research problem" [1]. According to Lachniet [4], it is difficult to identify the requirements for data loss prevention, as well as "to whom and how they apply, and how to address them in a cost effective manner."

The 2017 Global Data Leakage Report [5], which is based on public information, includes the following data:

- The number of leaks increased by 37%.
- 60.5% of intruders were internal.
- 50.3% of violators were employees.

Information leakage is also a major concern to asset managers. In a recent survey, 35% of respondents claimed that information leakage represents the majority of their transaction costs. "Unfortunately, information leakage is hard to measure and harder to attribute to specific venues and behaviors" [6]. It is reported that 63% of the grayware (potentially unwanted programs that are not malicious and not viruses) applications in 2017 leaked phone numbers and 37% revealed the physical locations of phones [7].

Information leakage is a type of system vulnerability where sensitive data is released and such data can be useful for attackers to breach system security. A sample case of the problem is shown in the following scenario.

In many cases, the broker receiving the order is not the same broker who goes on to execute it on the chosen venue(s). As the originating broker, you give up control as soon as you pass an order on to another broker for execution. You do not necessarily know the route your client's order is taking, or how many other parties might get sight of that order—and how much information about that order leaks out—before it eventually hits the market. [8]

Another sample of information leakage involves printing devices. It was found that when a new printer was installed with large internal hard drives, accessible via IP, it retained information after printing jobs were completed, but it was not at all secure [3]. Leakage of information may not necessarily involve "loss" in the sense of *depriving* a victim from such a resource.

A third example is a 2009 case reported in the newspapers as follows. A member of parliament in Kuwait claimed he was in possession of hard evidence of financial irregularities by the prime minister. The lawmaker produced a $700,000 check signed by the prime minister favoring a former MP. The MP demanded to know the reason for handing the check to the lawmaker, insinuating possibilities of corruption. The prime minister's lawyer said he would file lawsuits against the MP for breaching bank confidentiality laws. The bank has also said it would file lawsuits against the MP and any employee who was involved in giving him a copy of the check.

Data leakage can happen because of internal and external breaches, either intentionally or inadvertently. It is reported that internal employees account for 43% of corporate data leakage, and half of these leaks are accidental [9]. "Accidental leaks mainly result from unintentional activities due to poor business process such as failure to apply appropriate preventative technologies and security policies, or employee oversight" [1].

Data leak prevention is the process of monitoring sensitive information, enforcing data handling policies, and assessing incidents of leakage. It is a strategy to ensure that such information does not reach the wrong hands during internal operations of an enterprise, either during communication outside it. The case study in this paper focuses on the former type of leakage.

Data leak prevention also refers to the use of technology products that assist in controlling the transferred data. According to Lachniet [4], "we must be concerned about controlling our sensitive data throughout its entire life-cycle (from creation to destruction)." Current approaches to data leak prevention systems are designed as risk reduction tools for specific hardware/software systems. Hardware/software platforms are installed on network links to analyze traffic for unauthorized transmissions, and they run on end-user servers that monitor data flow between users.

Many technical methods are used for leakage in enterprises [10]. For example, in relational databases, access behaviors are modelled in order to identify intrusions and detect data breaches [11]. Security data policies and traffic inspection can be utilized to protect sensitive information in communication and storage [12]. A typical tool used to handle data leakage is watermarking, where a unique code is implanted in the information container. Watermarks may require some alteration of the data and can sometimes destroy data. Moreover, the distributer (original owner) may have partners or may outsource where the data requires being shared [13].

### *B. Aim and Approach with Regard to Information Leakage*

We study the prospect that data has leaked along several points of the flow path, and we propose a flow-based model that facilitates the identification of leakages. The aim is to identify and monitor unintentional or deliberate disclosure of information in order to take appropriate steps to prevent any leak in enterprise environment.

In our proposed system, all streams of information flow are identified, thus pointing to potential leakages that can be traced with appropriate evidence. This modeling of flow may uncover possible hidden points of leakage and provide a base for a comprehensive information flow policy. For example, it can be used to draw the specification of the privileges of administrators and employees and the internal information flow among them.

### *C. Aim with Regard to Exploring a New Modeling Methodology*

As mentioned previously, presenting a systematic conceptual framework for the data leakage problem is developed in the context of experimentation with a TM as a new methodology in modeling, alongside other methodologies such as object orientation. Many researchers have extended the use of object-oriented software design languages such as UML in order to apply them at the conceptual level (e.g., [14]). Although the huge development efforts and time that have been invested in UML and object-orientation-based studies, tools and mechanisms are marvelous achievements, this ought not be considered as the final word and should not discourage new research such as TM that points in other directions or may enrich the object-oriented paradigm itself.

The TM model is a diagrammatic language that is founded on slicing the domain of interest (a part of the world) to "bring out" things so that we can perceive them (nearness [15]) through thinging (presencing [15]) and describes how these things behave. According to Malafouris [16], humans evolve by creating new things, which in turn transform the ways we sense the world. "This applies to the modern forager of digital information as it applies to the Paleolithic hunter-gatherer and tool-maker" [16].

The notion of a thing and thinging in general plays an important role in modeling, contending with the salience of the widely acclaimed significance of the word *object*, the term currently in vogue among most software engineers. Heidegger [15] analyzed what makes a thing different from an object; a thing is self-sustained, self-supporting, or independent—something that stands on its own. The condition of being self-supporting transpires by means of *producing* the thing. On the other hand, objects are things locked into their final forms, closed in upon themselves: "It is as though they had turned their backs on us" [17]. Ingold [17] described the difference:

Using a square of paper, matchstick bamboo, ribbon, tape, glue and twine, it is easy to make a kite. Indoors, we were assembling an object. [In] a field outside, they suddenly leaped into action, twirling, spinning, nose-diving, and—just occasionally—flying. The kite that had lain lifeless on the table indoors had become a kite-in-the-air. It was no longer an object, if indeed it ever was, but a thing. As the thing exists in its thinging, so the kite-in-the-air exists in its flying. [17]

TM takes thinging as a basic conceptualization notion. TM modeling consists of an arrangement of machines, wherein

each thing has its unique stream of flow. TM modeling puts together all of the things/machines required to assemble a system (a grand machine). Accordingly, an additional aim of this paper is to explore the TM model capabilities in developing the notion of information leakage.

In the next section, we present a review of the TM model (also called the Flow thing model) as it is introduced in several publications [18-24]. The example in the section is a new contribution. Section 3 focuses on our case study of a *bank as a thing*. Applying TM to data leakage is the topic of section 4.

## II. THINGING MACHINE

According to Richard [25], diagramming is a thinking tool that transforms abstract issues into intelligible and actionable forms. TM modelling utilizes an abstract thinging machine (hereafter, machine) with five stages of thinging as shown diagrammatically in Fig. 1. A thing things; that is, a thing creates, processes, receives, releases, and transfers things. A machine that handles things is itself a thing that is handled by other machines, as illustrated in Fig. 2 (left). Fig. 2 (right) shows the snake as a machine that processes a frog and simultaneously as a thing that flows to an owl. The TM model is a grand thing/machine that forms the thinging of a system. Thinging here refers to the creation, processing, receiving, releasing, and/or transferring of the system (grand machine) or any of its submachines.

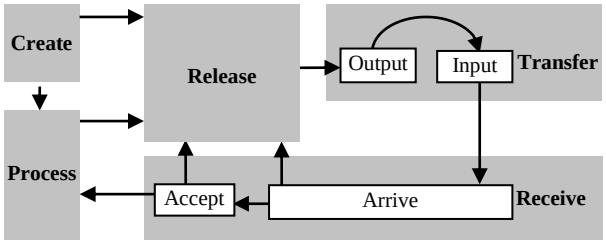



Fig. 1. Thinging Machine

Accordingly, a thing is a machine that manifests itself in the stages of creation, processing, receiving, releasing, and/or transferring, as shown in Fig. 1. The stages in the machine can be briefly described as follows.

**Arrive:** A thing flows to a new machine (e.g., packets arrive at a buffer in a router).

**Accept**: A thing enters a flow machine; for simplification purposes, we assume that all arriving things are accepted; hence, we can combine arrive and accept as the **receiving** stage.

**Release**: A thing is marked as ready to be transferred outside the machine (e.g., in an airport, passengers wait to board after passport clearance).

**Process** (change): A thing changes its form, but not its "identity" (e.g., a number changes from binary to hexadecimal).

**Create**: A new thing is born in a machine (e.g., a logic deduction system deduces a conclusion).

**Transfer**: A thing is input or output in/out of a machine.

TM includes one additional notation—triggering (denoted by dashed arrow)—that initiates a flow from one machine to another.

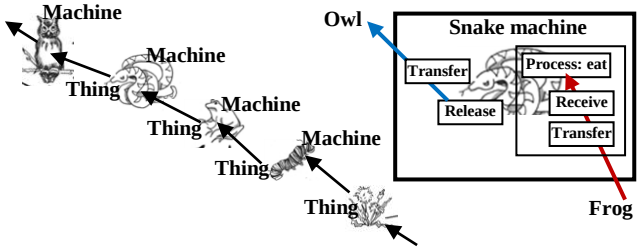



Fig. 2. Illustration of things that are machines and vice versa

**Example**: In contrast to object modeling, a *state* in TM is a thing with its own machine. Consider the classical object-oriented modeling of a coffee mug [26]. In object-oriented modeling, a coffee mug is an object with two states: *empty* and *filled*. TM takes a less abstracted view and considers a state as a submachine of a machine, as shown in Fig. 3. A mug (circle 1 in the figure) is a machine that involves the flow of coffee (2) that triggers the creation of the two states (3 and 4).

In such a scenario, we can identify four mutually exclusive *events*, as shown in Fig. 4. An event is a machine that is defined in terms of a time submachine and a region submachine (in addition to other machines). This notion of time as a thing/machine is not far from the Platonic view of time as a moveable image of eternity. Accordingly, the relevant events in the example are as follows:

Event a ($E_a$): Coffee is poured into the mug.
Event b ($E_b$): The mug is filled.
Event c ($E_c$): Coffee is poured out of the mug.
Event d ($E_d$): The mug is empty.

Accordingly, the behavior of the coffee/mug system is described as shown in Fig. 4. Any of the four events can be taken as the initial event. In the figure, time flow (transfer → receive → process [takes its course] → release → transfer) is not shown.

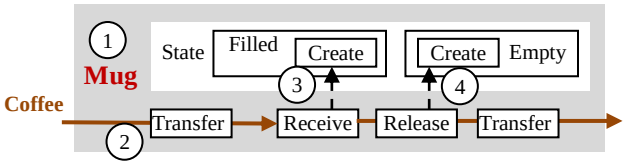



Fig. 3. The diagram of the system that involves filled and empty mug

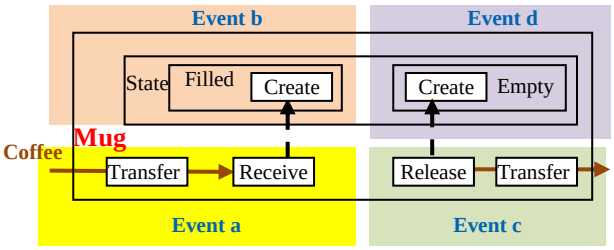



Fig. 4. Events

## III. BANK AS A THING

In this section, our study focuses on a bank as a thing. The bank banks through customers, other financial institutions, government agencies, and so forth. As these bank things encounter one another, the involved flows of different things/machines are interwoven and bundled together in a meshwork. The bank, in this picture, is a certain gathering together of the threads of the business world. It is modeled as a machine with many streams of flows that gather together the threads of banking.

In such a view, the bank switches from the usual perception of being an object to a "liveable" [17] thing that creates, processes, receives, releases, and/or transfers things. Remember, to objectify is to break a thing down into increasingly smaller parts instead of taking it holistically as it is. It is the "lifeless" kite object, as mentioned previously, that is viewed as a square of paper, matchstick bamboo, ribbon, tape, glue, and twine. Indoors, we were assembling an object. Additionally, the bank thing components do the same as exemplified by data leakage, the focus of this paper. In reality, any current bank as an object is a "livable" system to a certain degree, but this, by necessity, is an implicit result of its functions, and it is a partial "livability".

For example, leakage in the bank thing is a gathering of (sub)things and is viewed as one of the inhabitants of the bank, analogous to an octopus in an ecosystem with long arms that extend everywhere: employees, computers, desks, and cabinets, and so forth. This octopus is hiding until it is "brought out" by the bank's thinging. It can cause harm if not dealt with holistically. The main result of our case study in the next section is exposure of this leakage thing through mapping its territories (octopus arms) throughout the bank.

Such a perspective uncovers many hidden things as bank's dwellers. Our task is identifying these hidden occupiers of the bank in its model (diagram), as will be demonstrated by recognizing the information leakage thing. This is analogous to Wittgenstein's [27] work about the differences between "seeing" and "interpreting" (e.g., Wittgenstein's duck-rabbit figure), where in our study, assuming that our focus is on information leakage, we develop the bank's TM model, then we cut off the leakage machines inside it. This exposition of internally hidden machines is used for such purposes as constructing preventive measures and conducting forensics.

The bank as a thing in a modern society encounters ever-present vulnerability to threats. As a real thing, it is "a complicated machine in which every day something breaks down" [28]. A real bank is a gathering place that continuously calls for an unremitting effort to shore it up in the face of the comings and goings of its human inhabitants and nonhuman residents, not to mention the focus on security matters [17]. Much has been tried through developing a bank system that matches the expectations of well-ordered things within its outer boundaries; nevertheless, its function depends on the continual flow of things across these boundaries. The thingness of a bank becomes visible when an interruption or malfunctioning related to these flows appears.

In this paper, we focus on specific control efforts to counter the act of making information available without authorization. According to Lachniet [4], "Many controls are best done internally, such as creating a formal IT security management framework, or identifying the type of data you need to protect."

As used in this paper, leakage includes spilling, which refers to the unintended disclosure of information to unauthorized environments, organizations, or people [29]. In our study, we will exclude the situation of misconfigured systems that permit access to unprotected resources or are made available by hackers.

From the TM perspective, a leakage thing (e.g., information leakage) is a flow that spills out of the grand TM machine. This implies that a submachine has malfunctioned in the bank. Fig. 5 shows four possible types of submachine that leak flow from (1) received, (2) processed, (3) created, and (4) released information.

In the next section, we will identify all possible malfunctioned submachines in the bank used in our case study after developing a TM description of certain operations (e.g., consumer e-purchases) in the bank.

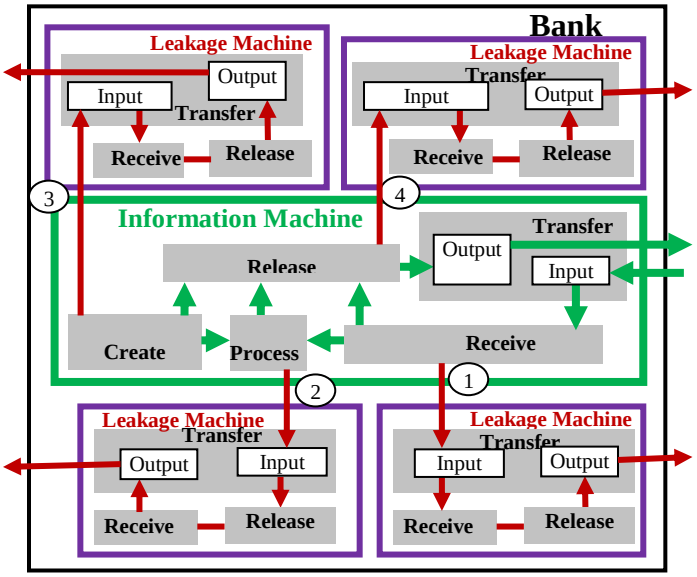



Fig. 5. Leak flows to leakage machines

## IV. BANK AS A CASE STUDY: CONSUMER E-PURCHASES

In the existing system of the bank in our case study (and in all banks, to the best of our knowledge), no explicit focus is given to the issue of information leakage. Such an issue is discreetly handled by the security team that deals with matters such as detecting hacking, collecting evidence, and the use of security tools such as encryption. In such a context, in specific leakage cases a possibility exists that the inability to progress legally due to a lack of a predesigned amount of evidence means the chances of escape for leakers are great.

To demonstrate the application of FM modeling in the area of a leakage of data, we took the following steps:

- First, we developed complete static and dynamic descriptions of the bank TM by focusing on the sample

application of e-purchases because of the paper size limitation.

- Then, we exposed leakage machines inside the bank TM description.
-

### A. Consumer E-purchase

As shown in Fig. 6, a customer (upper left corner) clicks on an icon on the screen that creates a signal (1) that flows to the electronic device software system to be processed (2) and triggers the processing (e.g., filling with relevant data) of a purchase request (3). Note that the request data is provided by the customer selection (click), and the (blank) purchase request is already stored in the device.

The request flows to the merchant server to be validated (4), and this triggers the generation of a formatted message (5) that flows to the payment gateway server (6). There, it is processed (7) to trigger the release of a processed payment page (8) according to the given data of the request. The payment page flows to the merchant server (9), then flows to the customer's electronic device where it is displayed (10). The customer inputs the payment details to trigger the creation of a transaction (11), which includes the card data, PIN number, and bank ID.

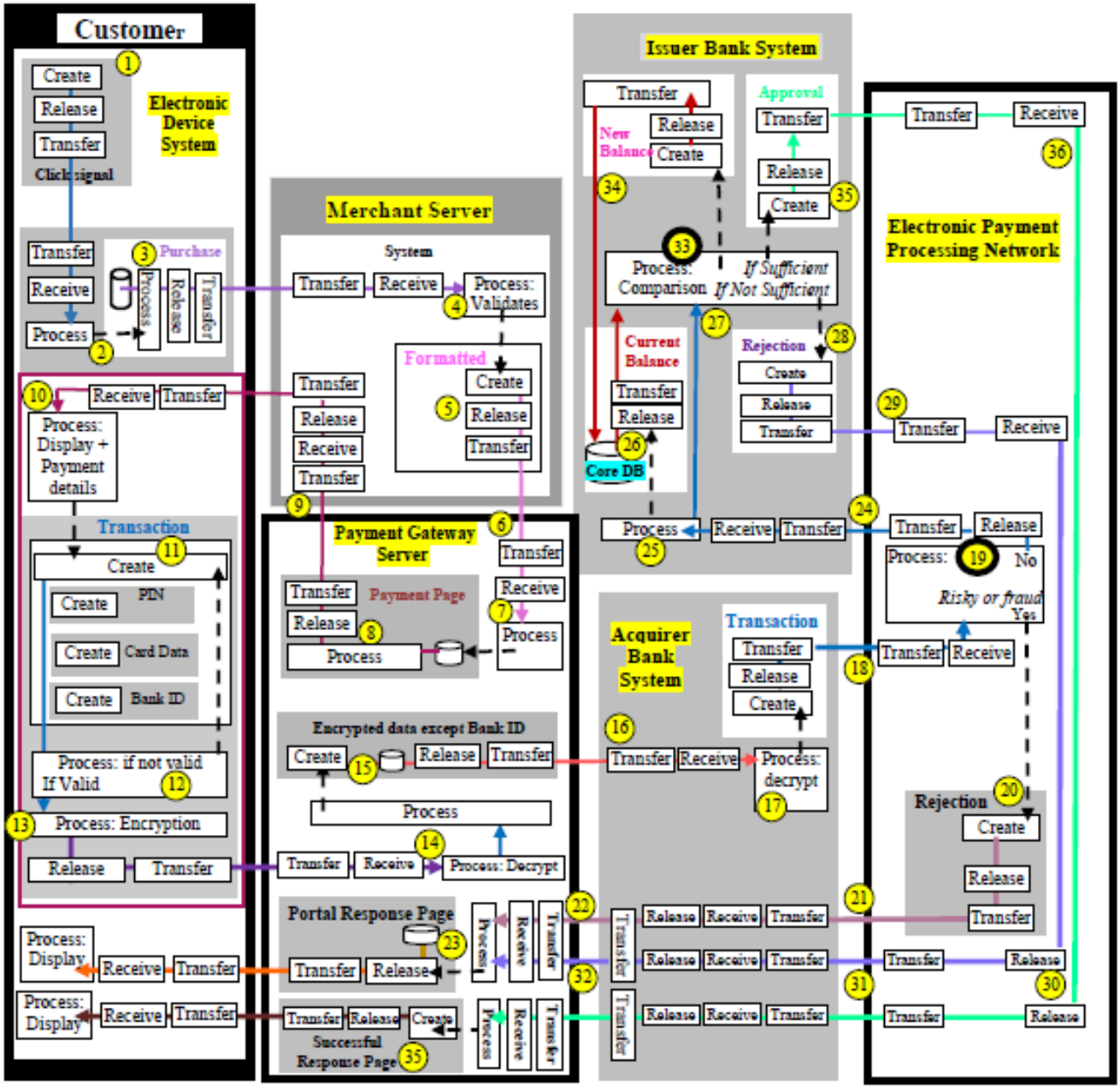



Fig. 6. The static description of a consumer e-purchase

To give a further description of how to create the transaction machine, Fig. 7 shows a sample of constructing it from clicks on the screen, a physical bank card, and a stored value in the inputting device. This thing-oriented depiction contrasts with the typical object-oriented specification (e.g., in UML), which has the mere structure of a class and its attributes.

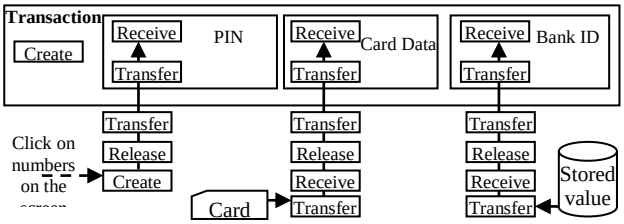



Fig. 7. Sample description of input

Continuing with Fig. 6, the transaction (11) is validated (processed) in the device browser such that

- If the input data is invalid (12), then the transaction is constructed again, or
- If the data is valid, then the transaction is encrypted (13) and flows to the payment gateway server (14).

In the payment gateway server, the encrypted transaction is received and decrypted (14). Then, the original transaction is processed to generate encrypted data without encrypting the bank ID (15). The encrypted data is stored into a database, and then released to the acquirer bank (16).

In the acquirer bank, the transaction data is processed into decrypted data (17). Then, the transaction flows to the payment processing network (18) where it is processed (19) and the following can happen:

- If the transaction is risky or possibly fraudulent, then a rejection response is generated (20) and sent to the acquirer bank (21) and then to the payment gateway server to be processed (22) to release a pre-stored portal page response from the database (23). The portal page response is transferred to the customer's device browser to be displayed.
- Returning back to the processing at (19), if the transaction in the electronic payment network is not considered a risk or possible fraud, then the transaction flows to the issuer bank (24) to be processed (25). This processing triggers a release of the customer's current balance from the centralized banking database (26) to be compared with the transaction amount that is received by the acquirer bank (27).
- If the balance is not sufficient, then a rejection response is created (28) and transferred to the electronic payment processing network (29), and then it flows to the acquirer bank (30), to the payment gateway server (31), and to the payment gateway server (32) to trigger a release of a pre-stored portal page response that is transferred to the customer's device browser.
- Going back to the comparison of the balance with the transaction cost, if the balance is sufficient, then a new balance is calculated (33) and stored in the core banking database (34). Additionally, an approval response is created (35) and flows across servers to create a success page response (36) as described before.

Fig. 6 gives a static description of a consumer e-purchase. To describe its dynamic behavior, we give the following events, as illustrated in Fig. 8, which is a copy of Fig. 6 marked with regions of events.

**Event 1 ($E_1$)**: The customer clicks on his/her electronic device browser, which is processed by the device.
**Event 2 ($E_2$)**: A purchase request is sent to the merchant where it is validated.
**Event 3 ($E_3$)**: The merchant sends a formatted message to the payment gateway server to be processed.
**Event 4 ($E_4$)**: The payment gateway server processes a stored page and sends it to the customer's browser through the merchant server.
**Event 5 ($E_5$)**: The portal payment page instructs the customer to insert his/her payment details.
**Event 6 ($E_6$)**: The payment details are inputted and validated in the electronic device browser.
**Event 7 ($E_7$)**: In case it is invalid, the browser requests that the customer re-input the correct payment information.
**Event 8 ($E_8$)**: The electronic device browser encrypts the transaction.
**Event 9 ($E_9$)**: The electronic device browser sends the transaction to the payment gateway server where it is decrypted.
**Event 10 ($E_{10}$)**: The payment gateway server processes the decrypted transaction by separating the bank ID.
**Event 11 ($E_{11}$)**: The payment gateway server encrypts the transaction, except for the bank ID, then stores it into the database.
**Event 12 ($E_{12}$)**: The encrypted data, with the exception of the bank ID, is transferred from the payment gateway server to the acquirer bank system, where it is decrypted.
**Event 13 ($E_{13}$)**: The decrypted transaction is generated in the acquirer bank and transferred to the electronic payment processing network, where is processed for possible fraud.
**Event 14 ($E_{14}$)**: If the transaction is fraudulent, then a rejection response is sent to the acquirer bank.
**Event 15 ($E_{15}$)**: The payment gateway server received the message from the acquirer bank and processes it.
**Event 16 ($E_{16}$)**: The electronic device browser displays the rejection message.
**Event 17 ($E_{17}$)**: The electronic payment processing network sends the transaction to the issuer bank to be processed.
**Event 18 ($E_{18}$)**: The issuer bank gets the customer's current balance and processes it against the transaction received.
**Event 19 ($E_{19}$)**: If the current balance is not sufficient, then the issuer bank generates a rejection response that flows to the electronic payment processing network then to the acquirer bank.
**Event 20 ($E_{20}$)**: If the current balance is sufficient, then the issuer bank deducts the requested amount from the customer balance.
**Event 21 ($E_{21}$)**: The issuer bank generates an approval message that flows to the electronic payment processing network then to the acquirer bank.
**Event 22 ($E_{22}$)**: The electronic device browser displays the portal payment success page.

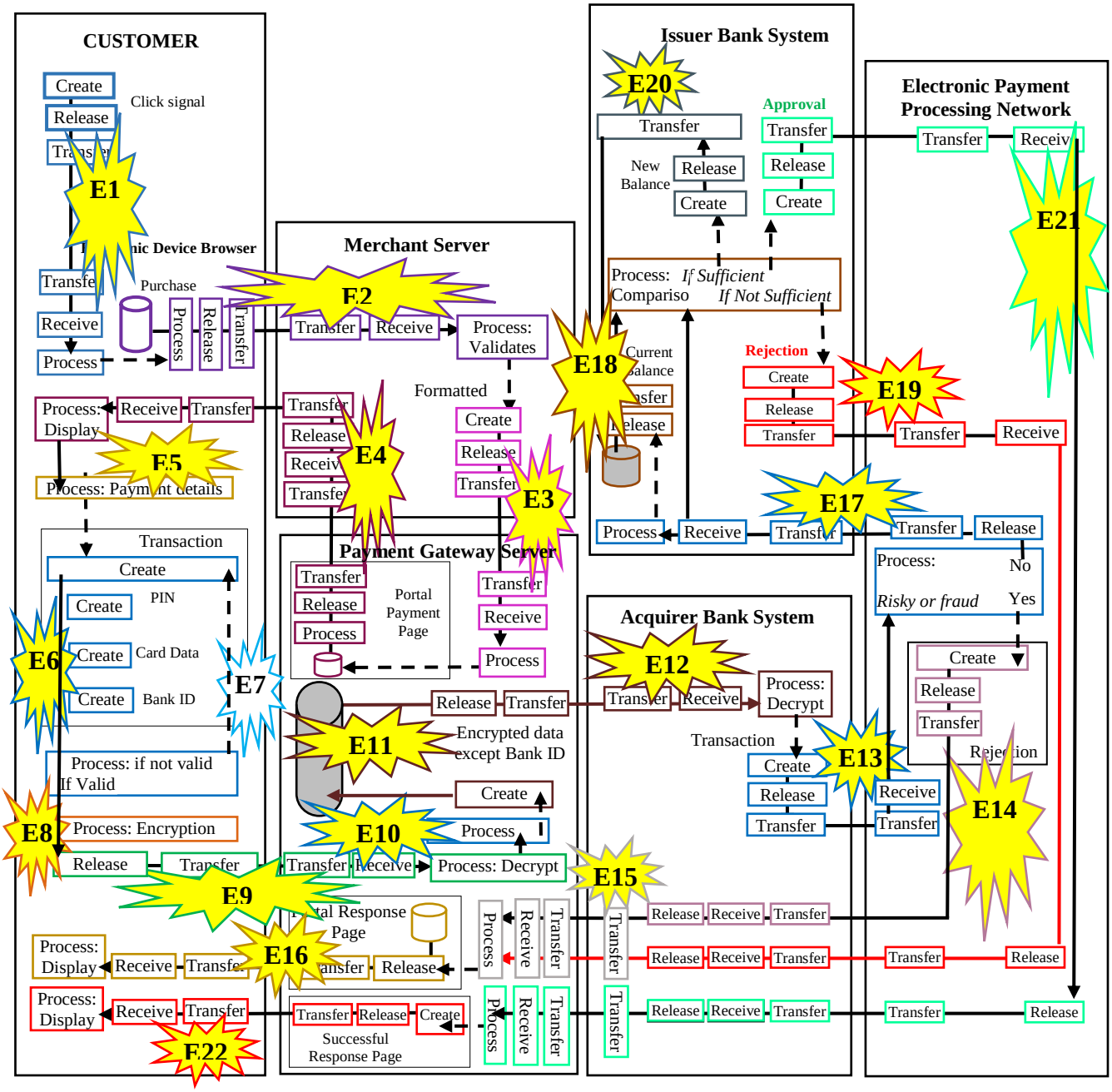



Fig. 8. The events of the consumer e-purchase

## V. Leakage Submachines

In this section, we will identify all possible leakage submachines in the bank. As an example of such identification, we focus on the issuer bank, keeping the numbered circles of Fig. 6. Fig. 9 shows the selected area for analyzing a data leakage. The aim is to model the entire issuing bank as a physical environment of the information system that handles the transaction data during its life cycle. The analysis can be generalized to different areas of consumer e-purchases.

Accordingly, Fig. 10 shows this expanded representation of the issuing bank. The top part of Fig. 10 shows the switch server where the transaction data flows from the electronic payment network to the issuer bank (24) to be processed (25). This processing triggers a release of a customer's current balance from the core database (26) to be compared with the transaction amount that is received by the acquirer bank (27).

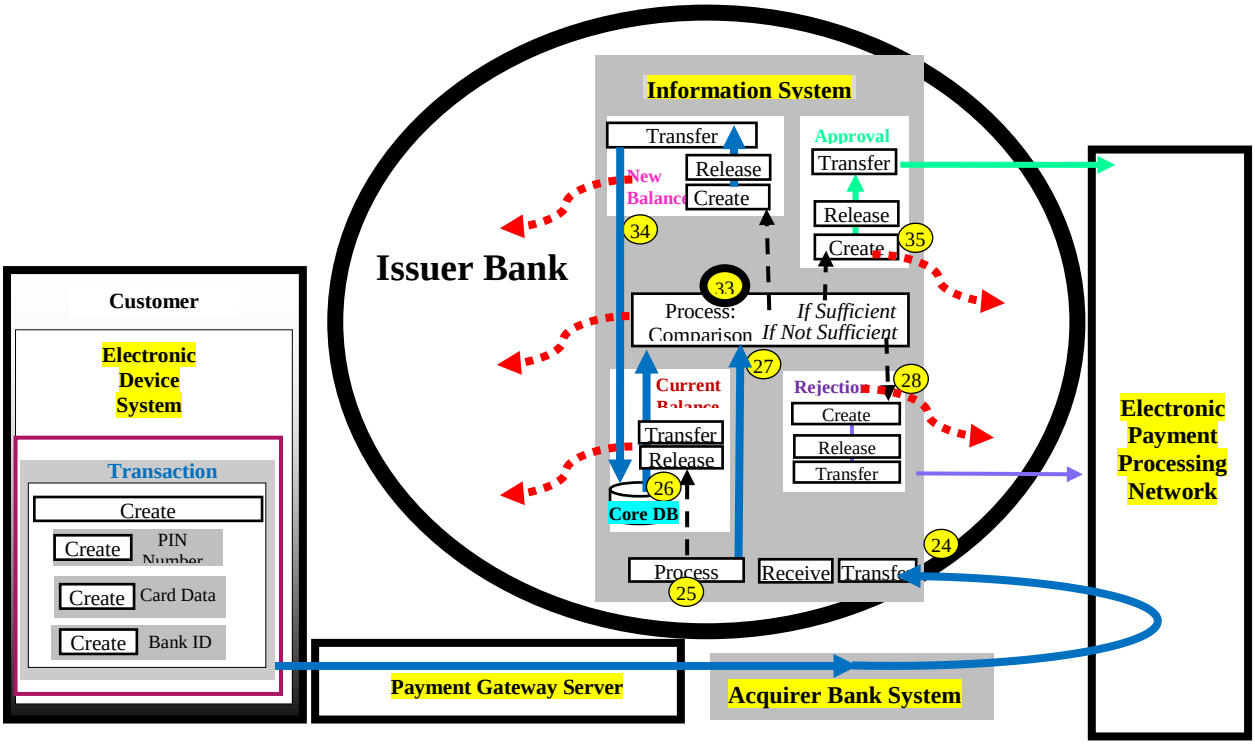



Fig. 9. The selected area for analyzing leakage

In addition to this main flow that involves the switch server and the core database, the server room includes the console, the core server, and the local e-mail server.

In Fig. 10, circles A and B in the server room point to possible leakage because the console screen can be captured by an employee (e.g., cleaning staff member).

The IT employee (C) monitors the system activity through his/her PC. He/she can access data on the switch server by creating an access request that flows to the core server. The currently processed transaction is copied and sent as an attachment (F) of an e-mail to the IT employee.

In the employee's PC (G), it is displayed, flows to the email server (F) and then printed (I). Here, there is an opportunity for leakage (e.g., by taking a picture or using flash drive). The hard copy is sent via a messenger employee to an employee who files it. Accordingly, we can describe all possible leakage machines, identifying their locations and who activates them. An employee can walk out of the bank carrying the data on a flash drive or as a hard copy. He/she can use the regular bank mail to send it out. These examples illustrate the method of identifying all possible leakage machines.

To summarize, Fig. 11 shows a general picture of different flows. The blue arrows in the figure show legitimate flows of the data whereas the red ones indicate leakage. The red flows originate from a leakage machines as follows.

Leakage machine 1 (circle 1): Capturing the console screen by an employee who has access to the server room.

Leakage machine 2 (circle 2): Capturing data from a PC by an IT employee using a camera, flash drive, etc.

Leakage machine 3 (circle 3): Copying data (hard copy) by an IT employee.

Leakage machine 4 (circle 4): Capturing data from a PC by a non-IT employee using a camera, flash drive, etc.

Leakage machine 5 (circle 5): Copying data by a record-keeping employee.

Leakage machine 6 (circle 6): Copying data by a messenger.

Leakage machine 7 (circle 7): An employee obtains a hard copy in an unauthorized way.

This thinging approach means that the leakage machine stands apart from its bank grand machine and is treated as a unified whole. A machine of interest (e.g. leakage) is exposed out of the bank thing with further thinging. It would appear as a subdiagram, in the forefront, clearly contrasted against the ground. This thinging of leakage is an act of "creation" of a machine which is already "exits" in reality even though we only perceive it when it becomes alive.

Such a comprehensive picture of data leakage provides the basis for planners and security personnel to focus on aspects that are suitable for the required prevention level. Additionally, it furnishes a foundation for any forensic investigation.

## VI. Conclusion

This paper sought to accomplish two aims: present a systematic conceptual framework for the leakage and to develop that in the context of experimentation with TM as a new methodology in modeling. The TM model of the bank demonstrates the viability of the TM model.

The TM diagrams may look complex; however, they can be simplified by lumping the details together or omitting stages according to requirements. Many issues remain to be clarified; however, this paper demonstrates the potential feasibility of this approach.

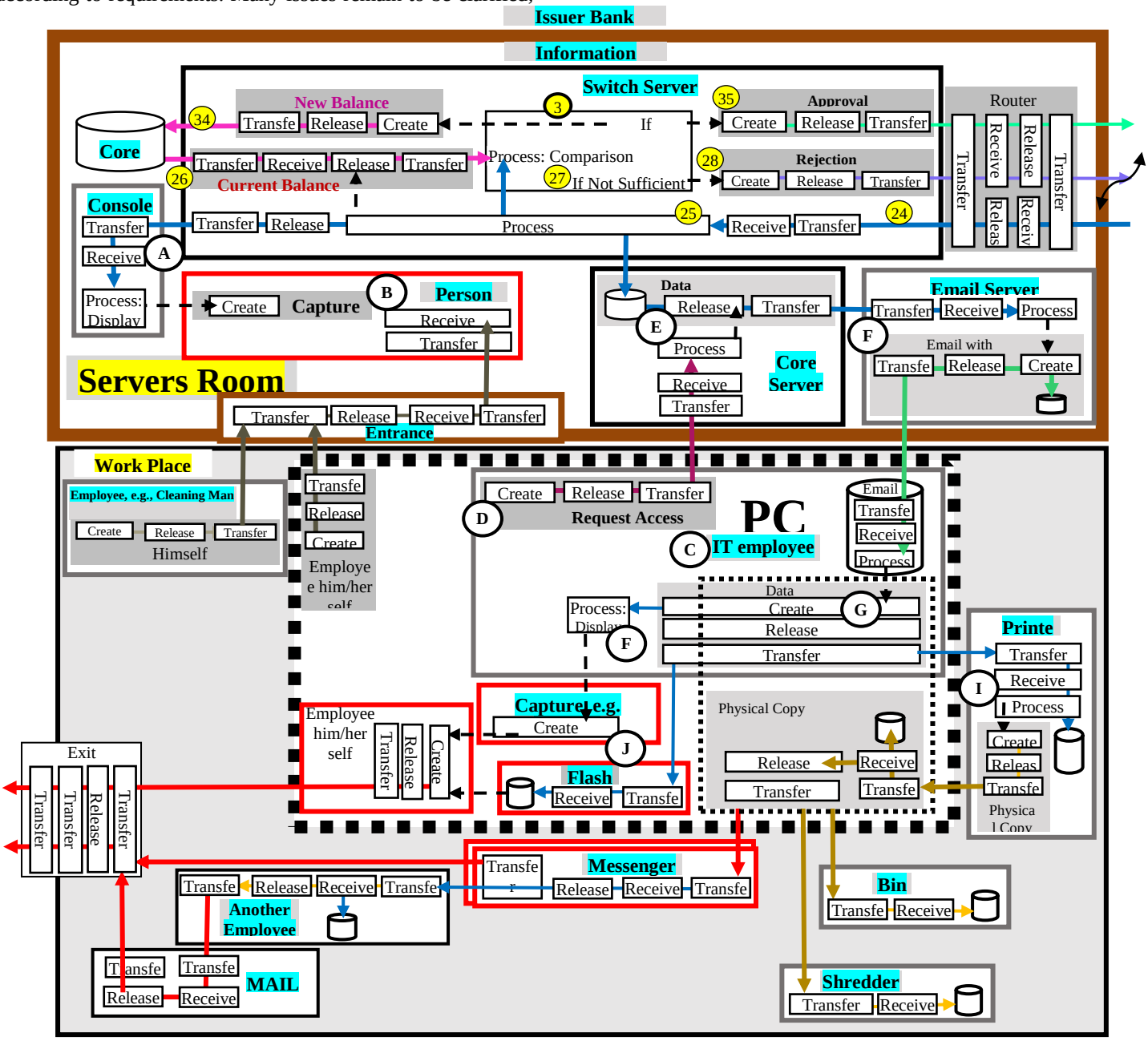



Fig. 10. Description of the flow of data and its physical environment

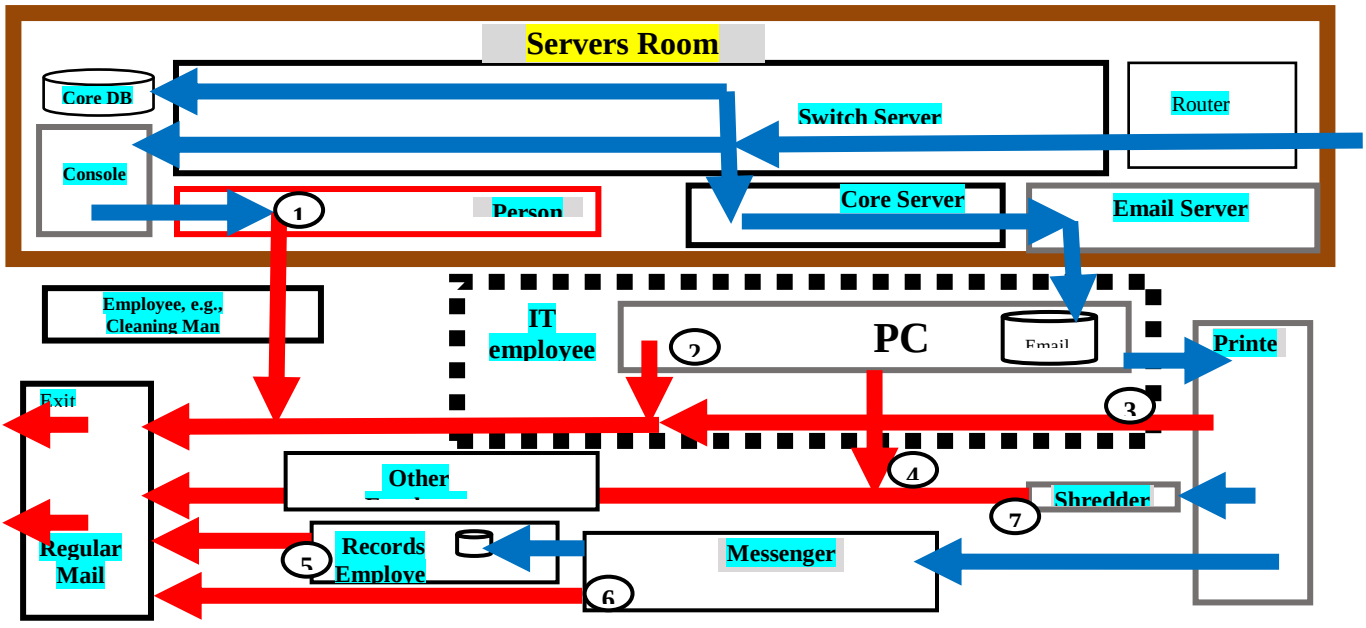



Fig. 11. A general picture of possible leakage in the example

## Reference


[1] L. Cheng, F. Liu, and D. Yao, "Enterprise data breach: causes, challenges, prevention, and future directions," WIREs Data Mining Knowl. Discov., Vol. 7, pp. 1211, September/October 2017. doi:10.1002/widm.1211.

[2] Security: Middle East & Africa, "Data leakage has gone up by four times," April 25, 2018. http://www.securitymea.com/2018/04/25/data-leakage-has-gone-up-by-four-times/

[3] J. Walker, "Information leakage: The most misunderstood security risk," ISACA Now Blog, 30 July 2018 (accessed).

[4] M. Lachniet, "Data Loss – Prevention and controls, slides," July 15th, 2008. https://slideplayer.com/slide/8683051/

[5] InfoWatch, "The 2017 global data leakage report." https://infowatch.com/report2017

[6] B. Polidore, F. Li, and Z. Chen, "Put a lid on it: Controlled measurement of information leakage in dark pools," Investment Technology Group, Inc., August 2017. https://www.itg.com/assets/ITG_PutALidOnIt_20170808.pdf

[7] Symantec Corporation, "Internet security threat report," vol. 23, April 2018.

[8] Itarle, "Preventing information leakage in broker order flow," (Blog) October 2016. https://itarle.com/blog/preventing-information-leakage-broker-order-flow.html

[9] McAfee, "Data exfiltration study: Actors, tactics, and detection," 2017. https://www.mcafee.com/us/resources/reports/rp-data-exfiltration.pdf (accessed March 2017).

[10] S. Alneyadi, E. Sithirasenan, and V. Muthukkumarasamy, "A survey on data leakage prevention systems," J. Netw. Comput. Appl., vol. 62(C), pp. 137–152, 2016.

[11] A. Kamra, E. Terzi, and E. Bertino, "Detecting anomalous access patterns in relational databases," VLDB Journal, vol. 17: pp. 1063–1077, 2008.

[12] X. Shu, D. Yao, and E. Bertino, "Privacy-preserving detection of sensitive data exposure," IEEE Trans. Inform. Forensic Secur, vol. 10, pp. 1092–1103, 2015.

[13] P. Papadimitriou and H. Garcia-Molina, "Data leakage detection," IEEE Trans. Knowl. Data Eng., vol. 23, no. 1, pp. 51-63, January 2011.

[14] [14] J. Evermann, "Thinking ontologically: Conceptual versus design models in UML," in Ontologies and business analysis, M. Rosemann and P. Green, Eds., Idea Group Publishing, 2005.

[15] M. Heidegger, "The thing," in Poetry, Language, Thought, A. Hofstadter, Trans. New York: Harper & Row, 1975, pp. 161–184.

[16] L. Malafouris, "The feeling of and for clay," Pragmatics & Cognition, vol. 22, no. 1, pp. 140–158, January 2014. doi:10.1075/pc.22.1.08mal

[17] T. Ingold, "Bringing things to life: Creative entanglements in a world of materials," Realities working papers #15: Bringing things to life, Department of Anthropology, University of Aberdeen, Scotland, UK, July 2010.

[18] S. Al-Fedaghi, "Privacy things: Systematic approach to privacy and personal identifiable information," Int. J. Comp. Sci. and Inf. Sec., vol. 16, no. 2, February 2018.

[19] S. Al-Fedaghi and H. Aljenfawi, "A small company as a thinging machine," 10th International Conference on Information Management and Engineering (ICIME 2018), University of Salford, Manchester, United Kingdom, September 22-24, 2018.

[20] S. Al-Fedaghi and M. Alsharah, "Security processes as machines: A case study," Eighth International Conference on Innovative Computing Technology (INTECH 2018), August 15-17, 2018, London, UK.

[21] S. Al-Fedaghi and R. Al-Azmi, "Control of Waste Water Treatment as a Flow Machine: A Case Study," 24th IEEE International Conference on Automation and Computing (ICAC 2018), 6-7 September 2018, Newcastle University, Newcastle upon Tyne, UK.

[22] S. Al-Fedaghi and M. BehBehani, "Modeling banking processes," 2018 International Conference on Information and Computer Technologies (ICICT 2018), DeKalb, IL, USA, March 23-25, 2018.

[23] S. Al-Fedaghi and M. Alsulaimi, "Re-conceptualization of IT services in banking industry architecture network," Seventh IEEE International Conference on Industrial Technology and Management (ICITM 2018), Oxford University, Oxford, United Kingdom, March 7-9, 2018.

[24] S. Al-Fedaghi, "Thinging for software engineers," Int. J. Comp. Sci. Inf. Sec., vol. 16, no. 7, July 2018.

[25] A. Richard, Diagrammatic: Beyond Inscription? (Conference), Centre for Research in the Arts, Social Sciences and Humanities, 2 December 2016–3 December 2016.

[26] L.A. Maciaszek, Requirements Analysis and Systems Design (2nd ed.), Harlow: Pearson Education Limited, 2004.

[27] L. Wittgenstein, Remarks on the philosophy of psychology (vol. I). University of Chicago Press, 1980.

[28] A. Siza, Architecture writings, A. Angelillo, Ed. Milan: Skira Editore, 1997

[29] U. M. Wilder, "Why spy now? The psychology of espionage and leaking in the digital age," Studies in Intel., vol. 61, no. 2, June 2017.

# Statistical-Based Trustful Access Control Framework for Smart Campuses

AHMAD B. ALKHODRE
Faculty of computer and information systems
Islamic University of Medina, KSA

***Abstract*—The vision of the Internet of Things (IoT) is based on the idea of offering connectivity to every physical object (e.g., thermometers, banknotes, smart TVs, bicycles, etc.). This connectivity ensures that immediate information about these objects and their surroundings can be obtained and therefore decisions can be taken based on real-time information. This allows increased productivity and efficiency. One of the most important implementations of the IoT is the smart (or digital) cities where the information collected from the connected devices is used in, for instance, configuring energy systems, enhancing the traffic, controlling pollution or ensuring security. However, there is no guarantee that all objects will provide information because, for example, some may be out of service or have lost connectivity bearing in mind that many objects in an IoT network are characterized by their limited resources (e.g., battery life, computing, and connection capacity). Moreover, the decision in an IoT network is mostly based on the information provided by a subset of the objects rather than all of them. In addition, the obtained information can be contradictory for many reasons, such as a defect in the object or malicious interference either in the object itself or during the communication process. Therefore, it is necessary to provide a measure that reflects to what extent the decision in an IoT network is trustful. In this paper, an approach based on statistical science is proposed to measure the trustworthiness of information collected from heat sensors. An architecture and algorithm, based on the confidence interval measurement to reduce the time taken to verify and check the trustworthiness of network sensors or any other type of IoT device.**

***Keywords*—Internet of things; trust management; confidence interval; confidentiality; smart cities***

## I. INTRODUCTION

According to the International Telecommunication Unit (ITU) Report 2005 ([1]), the Internet of Things (IoT) has a vision of providing all objects and devices (e.g., thermometers, banknotes, surveillance cameras, control access badges, etc.) with the ability to connect to a network which is named the IoT. This connectivity ensures that immediate information about these objects and their surroundings can be obtained and decisions can, therefore, be taken based on real-time information. This allows increased productivity and efficiency. One of the important implementations of the IoT is the smart (or digital) cities [2] where information collected from the connected devices is used in, for instance, configuring energy systems, enhancing traffic control, controlling pollution or ensuring security. Hence, the needs of the city inhabitants are efficiently met [3]. However, the connected devices are vulnerable to manipulation and/or physical damage. In practice, there is no guarantee that all the connected devices will provide their information because some of them may be out of service or have lost connectivity. Moreover, the obtained information can be contrary because of a malicious interference.

The Internet of Things (IoT) is an old term. It first became popular due to industrial demand and then secondly because all the pre-requisites that were needed for IoT had never before been available as they are today. The vision of IoT is that all electronic devices will be able to communicate with each other. Although previously this capability existed among a very few devices, ultimately the IoT will generate a massive amount of information which can be used to gather further meaningful data. The IoT concept was presented two decades ago by a technologist, Kevin Ashton. He stated that any two devices can communicate with each other or can connect to the internet with the help of sensors.

The IoT is an interesting concept since it will allow the automation of objects without human intervention through the advancement of technology and other new developments in daily life. However, on the flip side, other issues arise and, of those problems, the big challenges of security and trust need to be given the utmost attention [4]. The security and development of trust among the communication of IoT-based devices as well as the sharing of privacy-preserved data for analysis are the subjects of current critical research.

The IoT devices can be divided into three layers [5] where strong bonds relating to the security and trust of the user's data are required. These are the physical layer, the network, and the application layer. If there is any malfunction in the device or the vulnerability of data, this will create an unreliable environment. In this research, the physical layer of security and data privacy are discussed. In the first place, a plenty of data is received from the physical sensors of the device which needs to be reliable. Secondly, even if the data is correct, it should remain secure and all the concerns regarding the privacy of the user's personal data should be addressed. Furthermore, the data should not be utilized without the user's permission as per the terms and conditions agreed upon.

As discussed, there are three layers in an IoT system and we can find many areas in those layers where trust management is required to promote the worldwide adoptability of the IoT system. These areas have been thoroughly discussed in the various literature. However, before moving on to the contribution of our work to ensure trust in the IoT systems in a specific area, these areas will briefly discuss them. Trust

relationships and decision making are the first objectives to be achieved by establishing a strong IoT based system. For example, trust in data perception, privacy preservation, data fusion and mining, data transmission and communication, and system security are considered.

In this work, an approach based on statistical science is proposed to measure the confidence of the collected measurements. This approach helps to reduce the time taken to verify and check the trustworthiness of network sensors or any other type of IoT devices. The structure of this paper will be as follows. Following this introduction, this paper will look at the security challenges of IoT, then the related work will be presented. Before presenting our approach, section 4 will outline the theoretical background of the proposed approach. Finally, the article will close with a discussion and conclusion.

## II. BACKGROUND

### A. The IoT Security Challenge

Ensuring the security, reliability, resilience, and stability of internet applications and services is critical to promoting trust and the use of the internet. As users of the internet, it needs to have a high degree of trust that the internet, its applications and the devices linked to it are secure enough to perform the kinds of activities we want to do online in relation to the risk tolerance associated with those activities. The Internet of Things is no different in this respect, and security in the IoT is fundamentally linked to the ability of users to trust their environment. If people do not believe that their connected devices and their information are reasonably secure against misuse or harm, the resulting erosion of trust causes a reluctance to use the internet. This has global consequences to electronic commerce, technical innovation, free speech and practically every other aspect of online activities. Indeed, ensuring security in IoT products and services should be considered as a top priority for the sector. As we increasingly connect devices to the internet, new opportunities to exploit potential security vulnerabilities grow. Poorly secured IoT devices could serve as entry points for cyber-attacks by allowing malicious individuals to re-program a device or cause it to malfunction. Poorly designed devices can expose user data to theft by leaving data streams inadequately protected. Failing or malfunctioning devices can also create security vulnerabilities. These problems are just as large, or even larger, for the small, cheap and ubiquitous smart devices in the Internet of Things as they have traditionally been for computers [6].

### B. Confidentiality measuring

Through literary studies, mechanisms have been extensively discussed to determine trust and reputation but there is little research into trust management for the IoT nor into the study of confidence in the field of big data, taking into account the privacy of users and data [7][8].

In digital data processing, the confidence process involves verifying that the collected data is reliable and trustworthy. Through the research presented in [9], trust is calculated via social confidence and the QoS of data metrics through direct and indirect observations and recommendations to update trust metrics.

Indeed, trust has three properties. Honesty, cooperativeness and community interest are considered in the trust evaluation of IoT nodes ("things"). The honesty trust refers to the node as being honest or not, while the cooperativeness trust presents the social cooperation between trustee and trustor [10].

This paper focuses on data perception trust that works on the physical layer of an IoT based system. Data perception trust means that data should be reliably collected and recognised on the physical layer.

The services provided by the IoT vision will demand more data from the user's devices. However, that data needs to be trustworthy and secure from malicious attacks. Chend et al. [11] provided a trust and communication mechanism to securely communicate devices on the Internet of Things network. They presented the trust management model for IoT based on Fuzzy logic in a wireless sensor network environment.

Nitti et al. [12] presented the idea of trustful communication among the social Internet of Things network. In this concept, objects can create their social objects' network. They focused on how the objects were going to share information among nodes and how to evaluate the data received by another social object. The system evaluates the trust level of its peer/friend object based on its personal experience and takes the opinion of friends who are in common with all its objects.

Furthermore, Ruan et al. [13] also worked on the trust management for IoT agents and provided a framework for that. They also observed that, by using a trust-aware IoT network, the error can be reduced. They also enhanced their research findings by applying two different kinds of attacks and detected them via their trust management technique. In addition, they also provided an interface so that an end user can evaluate the whole communication process between agents.

Wand et al. [14] worked on Mobile AdHoc network trust management for a service-oriented approach. This is an ongoing research idea and they have shared the experiences they have learned whilst also providing details of their future direction. Gallahar et al. [15] studied the confidentiality measurement in a health-care environment.

It is obvious that most of the earlier work has focused on obtaining confidence in the IoT (or the connected system in general) by building a system based on reputation. In such a type of system, every time the device/node participates in an operation (e.g., communication, providing information, etc.) its credit will be increased (or respectively decreased) if this participation was correct or honest (or respectively incorrect). However, how does the reputation system decide whether the participation is correct or not? Furthermore, there is no measure that specifies to what extent the services or the operations by an IoT network are trustworthy. This research proposes a statistical-based measure of trust in order that a measure of confidence for the IoT service is provided.

### C. Statistical-based Confidentiality Measure

Statistics is one of the major branches of mathematics and has wide applications. Statistics are concerned with collecting,

summarizing, representing and drawing conclusions from the available data set, trying to overcome problems such as data heterogeneity and divergence [16][17].

In statistical terminology, the population is the entire group of individuals for which statistics are made. In our case, the population is composed of sensor and thermometer measures.

In order to optimize the available resources (money, time and other types of resources), all the individuals in a large population cannot be taken into consideration. Furthermore, it is practically impossible to reach all the individuals. Thus, the decision was taken to base the study on a subset of the individuals called Sample which was selected randomly, i.e., population individuals had an equal chance of being selected. Note that it is not possible that a selected sample will represent 100% of the population and this leads to what is referred to as sample error. In order to evaluate the correctness of a sample result (i.e., to understand how far this result is from the result that could be obtained when all the individuals are used), the confidence interval can be used.

The confidence interval [18] is a concept used to measure the correctness of a sample. It reflects the percentage of error in the sample. In other words, it is seen as a measure of correctness around a sample. The confidence interval reveals the variety of values of the population within a level of correctness. The larger the confidence level, the higher the confidence in the result.

The confidence interval of the control panel decision is proposed to calculate. If this interval is heterogeneous (some values in this interval confirm the violation, while others deny it), then the control panel needs to continue collecting more information from other sensors before making a decision. If the interval is homogeny (i.e. the value of interval confirms a violation in the access control system or denies it), then the control panel reaches a trustworthy result and can make the corresponding decision. The advantages of this system include, firstly, making a decision as soon as sufficient information is gathered rather than waiting for all the information to be gathered. Secondly, this system provides a measure of confidence in the control panel decisions and, finally, it identifies the sensors and thermometers that provided the incorrect information and thus they can be checked.

The proposed solution aims to first design a smart control access system based on data collected from the sensors and the devices in our system. This system also allows making the decision, for example, to evacuate during an emergency. The second objective is to provide a confidentiality measure for the decisions taken by the smart control system.

*1) Statistics and Confidence Interval*

The population size could be very large therefore it is not possible to collect the data from every individual which would be costly in terms of time and resources. Statistics find a solution for such a problem by proposing to obtain the answer based on a group of individuals chosen from the population called Sample [18]. For the most part, the choice is made randomly, however, other methods of selection do exist. Since the Sample does not usually accurately reflect the whole population, an inconsistency exists, a so-called Sample Error, between the sample-based answers and the population-based answers. In practice, the sample-based answer will not be a specific value (so-called point estimate) but rather a range of values (so-called interval estimate) in which the true answer will exist. Remember that the true answer is the answer that we would get if all the individuals in the population were asked. The confidence interval is calculated based on the confidence degree. This latter is a percentage that relates to what degree we are sure that the calculated interval contains the true answer. For better understanding, suppose that a survey claims that a candidate will get between 50% and 60% of the votes and the certitude of this estimation is 95%. The 95% is the confidence degree and thus, under repeated random sampling in identical conditions, the interval [50 – 60]% will contain our sample result 95% of the time.

Our proposal is based on using the confidence interval as a reliability measure for the decision calculated, thanks to the information obtained from the sensors. For instance, if we calculate the 95% confidence interval of the decision, we will be 95% certain that the correct decision belongs to the estimated interval. Thus, if the estimated interval is homogenous (respectively, heterogeneous), we can accept (respectively, refuse) the decision.

*D. Confidence Interval Calculation*

In statistical science, the mean is the most used population parameter [18]. In our context, the calculation of the confidence interval is similar to the calculation of other population parameters such as median or proportion in other disciplines.

Suppose the mean of the sample is $\mu_s$ (i.e., the estimated point) and the error of the sampling is $\lambda$ (standard error) which is calculated according to a specified confidence level (CL), the population mean $\mu$ is determined using the following equation

$$\mu_s - \lambda < \mu < \mu_s + \lambda \qquad (1)$$

The percentage of the confidence interval of the population mean (CL%) is given by the interval $[\mu_s - \lambda, \mu_s + \lambda]$.

The error of the sample for the mean is given by equation (2). Obviously, the key element during the estimation of the confidence interval is the error of sampling of the mean.

In order to calculate the error of sampling, the distribution of the sample should be determined. Equation (2) can be applied only when the distribution follows a normal distribution whereas other equations are used to calculate the sampling error for the distribution which does not follow the normal rule [18].

$$\lambda = z_{\alpha/2}\frac{\sigma_s}{\sqrt{n}} \qquad (2)$$

Whereas the standard deviation $\sigma_s$ of a sample for the normal distribution is calculated as follows:

$$\sigma_s = \sqrt{\frac{\sum_{i=0}^{n} \mu_s - x_i}{n-1}} \qquad (3)$$

Here the factor $z_{\alpha/2}$ correlates with the confidence interval and the distribution of the sampling and $x_i$ represents the sample of the individuals. Consider the bell distribution of data, this factor represents the zones in the tail of the bell. For example, if the confidence interval is 95% then $\alpha/2 = 0.025$. Next, according to the table, 1 z-score is z0:025= 1:96. Table (1), contains the z-scores for the most commonly used confidence levels.

TABLE I. Z-SCORES FOR THE MOST COMMONLY USED CONFIDENCE LEVELS

| Confidence level | z-score |
|---|---|
| 80% | 1.28 |
| 90% | 1.65 |
| 95% | 1.96 |
| 99% | 2.58 |
| 99.9% | 3.29 |

Briefly, the following algorithm characterizes the estimation of the confidence interval of the population:

*1) Select a sample which represents the population.*

*2) Compute the mean of the sample and its standard deviation.*

*3) Compute $\alpha/2$ the zone outside the confidence interval.*

*4) Compute the sampling error using equation 3.*

*5) Compute the estimation of CL as $[\mu_s - \lambda, \mu_s + \lambda]$.*

## III. SCAS: SMART CONTROL ACCESS SYSTEM

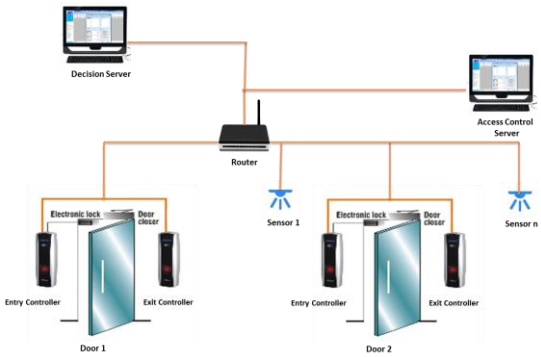



Fig. 1. Smart Control Access System

As shown in figure 1, our IoT system is composed of sensors which can sense fire, movement, pollution, heat, etc. They send their measurements to the decision server. The router (or any type of network connector device) that connects many smart heat sensors together links via the internet to a server. The access control server receives entry/exit requests from the access controllers and sends back the proper response. The decision server is a workstation used for collecting data and analyzing it in order to determine whether the data is trusted or not. Each sensor belongs to a zone that measures the temperature in its zone. The behavior of each sensor is demonstrated in figure 2. This state diagram shows six states representing the behavior of a sensor. When a sensor starts working, it will be in the idle state. When the measurement time arrives, the sensor will read/measure the temperature and store it. The state, sending data, is the state where the sensor is interrogated to give the stored data. Finally, the trust value of the sensor is updated (update trustValue) whenever it is sent by the monitoring system. If the trust value is below the required limit, the sensor receives a deny signal that causes the end state.

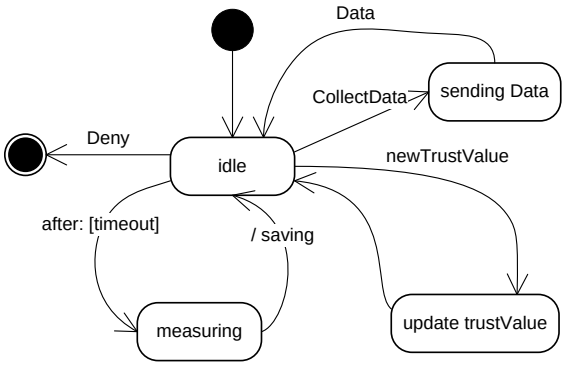



Fig. 2. the states of a smart heat sensor

The sequence diagram (figure 3) represents the diagnostic system. For the first step, a doubtful node is selected. A diagnostic procedure will take place to verify whether the node is sending correct data or not. The second step is to determine the neighbors of the node in focus to form the proof. Next, a subset (sample) is selected from the data sent by those nodes.

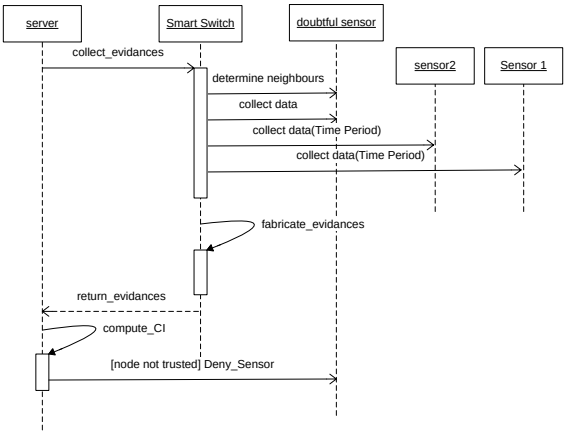



Fig. 3. the sequence diagram of evidences collection

The subset of the data is selected randomly from the neighbors of the doubtful node. In order to apply Formula 4, the weights (wi) are assigned according to the neighborhood. Next, the error of the sample and the CL confidence interval are calculated using formulas 1 to 4. The objective of calculating CL is to limit the number of proofs gathered during the doubtful node's checkup and to measure the extent of using CL in the diagnosis, especially when inconsistent proofs are gathered. For optimization of the resource and for power consumption purposes, the gathering of more proofs (possibly redundant) will not change the result of the diagnosis.

For instance, suppose a doubtful node uses the equation (4)

$$D = \sum_{i=1}^{n} w_i p_i \quad (4)$$

Where wi is a factor, this gives the trust of the neighborhood node where its value is positive or negative and pi is the proof of node i where the value is 1 (the collected data is correct), -1 means the collected data is not correct and 0 means the interrogated sensor has not responded.

After using formula 1 in calculating the sampling error, if the confidence interval is homogenous, this means that the doubtful node cannot be trusted. If the confidence interval is heterogeneous, then the true diagnostic result can either confirm or deny the doubtful node (figure 4).

For instance, suppose a doubtful node and D is calculated according to formula 4.

This is calculated based on proofs p1, p2… pn. and λ is the error of the sampling of D and CL is 95%. Then, the 95% of the confidence interval of the checkup result is [D-λ ;D+ λ]. where the result of the diagnosis is [-1,+1]

If the D+ λ <= 0 (positive values) and CL= 95% this means 95% of the proofs give a negative result and the data set confirms that the doubtful node cannot be trusted.

Similarly, if the D- λ > 0 (negative values) and CL= 95% this means 95% of the proofs give a positive result and the data set confirms that the doubtful node is trusted.

In the case where the results are negatives and positives at the same time, the latter characterizes the doubtful device as trusted but not all the time.

This can be interpreted that the system is not mature enough and more proofs must be collected to continue the diagnostic process.

In conclusion, the correct diagnostic result must be within the estimated confidence interval. Therefore, the result of a sample is accepted only if the estimated interval is homogeneous, that is, all its values are rejected. If the confidence interval is heterogeneous, then the true diagnostic result can confirm the correctness of a device or not. This means that there are still doubts about the diagnostic result obtained and, therefore, more proof must be collected before the diagnosis is complete

**Algorithm** 1 Procedure IoT_node_check ( )

**Input** a selected node
**Output** new trust values

1. Determine node zone and neighbors
2. Collect evidences form a subset from node's neighbor
3. Compute confidence interval (CI)
4. **if** CI is heterogonous **then**
5. **if** CI < 0 then
6. Change the trust value of the selected node
7. Update trust value for all node in the zone
8. **end if**
9. **else goto** 2
10. **end if**

Fig. 4. Integrating the confidence interval into the diagnosis

### A. Reliability of IoT Network Decisions

The collected measurements from the sensors (IoT objects in general) could be incorrect for two reasons:

- Tampered sensors that are controlled by an attacker and provide wrong measurements in order to poison the IoT network
- Damaged or biased sensors because of non-malicious events. For instance, a temperature sensor that is installed next to a light or a warming device will provide biased temperature measurements.

Regardless of the intention behind the incorrect measurements/information, they lead to contradictory information and hence prevent the decision server from making a decision or leads to a wrong decision. Thus, there is a need for a reliability indicator that helps in accepting or refusing the decision taken based on the collected measurements

### B. Experiment and Proof of Concept

#### 1) Confidence Interval as a Reliability Measure

The goal of this experiment is to prove that using the confidence interval helps in avoiding wrong decisions even in the presence of incorrect measurements because of either tampered with or damaged sensors. The measurements of 7 heat sensors are simulated and the different percentage of tampered sensors is discussed using MATLAB (figure 5). This shows the measurements of 7 sensors. The summits indicate the result of an abnormal or damaged sensor.

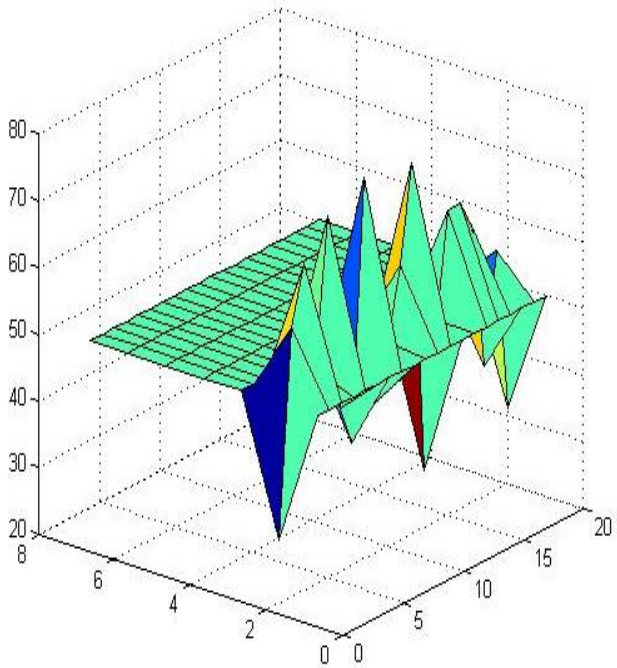



Fig. 5. the measurements of 7 smart sensors

As shown in figure 6, sensor 2 gives abnormal values compared to the other sensors, which indicates that this sensor has abnormal behavior and its measurement data must be verified (the sensor may be to blame).

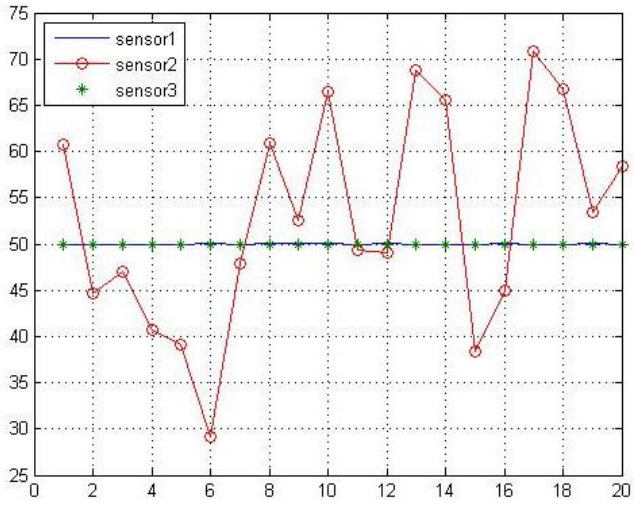



Fig. 6. the measurements of well-behaved and misbehaving sensors

In the simulation, the following Gaussian noise is used to generator erroring signal with SNR (signal-to-noise)

"awgn(V(i),10,'measured');"

Figure 7 shows the updating of trust values for the two different sensors (trusted and accused sensors).

**Algorithm** 2 Procedure update trust_values ( )

**Input** matrix of evidences (E), a matrix of trustvalue of participated node (W)

**Output** matrix of trustvalue of participated node (W)

1. **for** i=1 to **size_of**(W)
2. **if** $e_i$ == -1 **then**
3. $w_i = w_i + 0.2$;
4. **else**
5. $w_i = w_i - 0.2$;
6. **end if**
7. **end for**

Fig. 7. MATLAB code for updating trust value.

The first sensor is trusted so the value of its trustvalue is increased and the second sensor's trustvalue is decreased according to the calculation in figure 8.

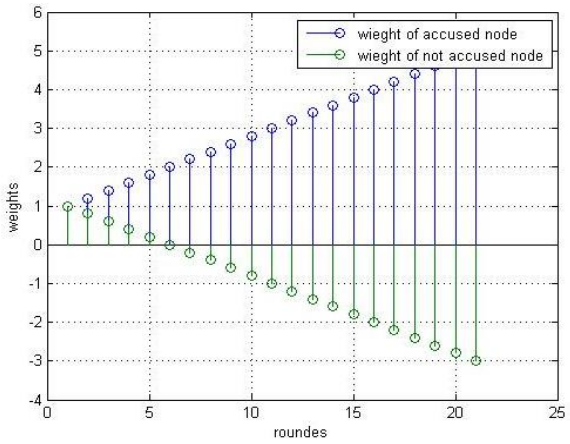



Fig. 8. the trust weights of well-behaved and misbehaving sensors

In the beginning, the 7 sensors start working and measure the temperature periodically. This task is achieved by using a timer that initiates the sensors at certain periods of time.

After the measurement is completed, the data is stored in a matrix in order to start the diagnostic process. Then, a sensor is selected to test its reliability by calculating the confidence interval. In step one, all the collected data is considered as a sample from which we will make the calculations using equation number 4. Next, the system starts to collect proofs. The proofs will be either -1, which means the measurement is not confirmed, or 1, which means the measurement is confirmed and correct. So, from these proofs, in order to prove the validity of the study, a partial data set is defined through the use of the function data-sample of MATLAB functions (figure 9). Then, the confidence interval is calculated for the chosen sample and update the trust value of all the sensors involved in the assessment process.

```
sampleError=1.96*(std2(E2)/sqrt(numel(E2)))
```

```
sample=datasample(E2,5,2);
for i=1:5
for j=1:5
 D(2,count) =D(2,count) + sample(j,i)*W1(j);
end
end
```

Fig. 9. MATLAB code for calculating the sum of the trust values of all sensors.

If the sensors gave a correct data, the trust value increases otherwise the trust values are reduced for all sensors which sent negative results. The previous operation is done for more than one round and, for each round, a different sample is taken and the previous calculations are redone. Figure 10 shows the change in the confidence area values calculated according to equation numbers 3 and 4. The upper curve shows that the values are constant positive, which gives an indication that the sensor is reliable (well-behaving) while the second curve shows that its value is increased by negative values, which means that the sensor cannot be trusted and must be removed (misbehaving).

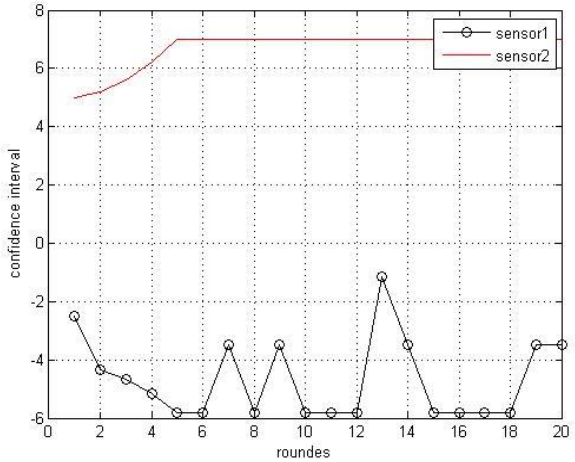



Fig. 10. the confidence interval of well-behaved and misbehaving sensors

## IV. Discussion and Conclusion

This article addressed the three key challenges faced in our system. During the error diagnosis, the sensor that does not behave well may provide incorrect proof. Therefore, the diagnosis can blame a legitimate sensor. In order to make a robust diagnosis against such false proofs, an entropy-based trust model may be proposed and integrated into our approach. In fact, a proof is first weighted by the reliability of its source before being used in the diagnosis. Thus, proof from distrustful (or reliable) nodes has less (or more) impact on the outcome of the diagnosis. The reliability of a node is increased resp. decreased each time it provides correct (or incorrect) diagnostic proof. Therefore, the more the node behaves poorly and provides incorrect proof, the lower its reliability, and therefore the less detrimental impact on the diagnosis.

During the estimation of the reliability of a sensor, its recent participation in the measurement of diagnoses is privileged over older ones. This privilege helps to avoid the effects of intoxication that occur when: (i) a legitimate sensor with a high trust value is compromised and begins to provide incorrect proof, or (ii) a malicious node attempts to gain the trust of the others by providing correct proof for a while before it begins to participate maliciously in the diagnosis.

The reliability of a node is also associated with its role in sending correct data. The risk of incorrect data increases as it evolves and nears its ultimate goals and may lead to more incorrect measurements. Overall, the evaluation of the confidence-based error detection in sensors shows that the defect impact along with the evolution of the trust relationships between the nodes is reduced.

## References

[1] The internet of thing overview, international Telecommunication Unit, 2005

[2] A. Cocchia, "Smart and Digital City: A Systematic Literature Review," in Smart City: How to Create Public and Economic Value with High Technology in Urban Space, Springer International Publishing, 2014, pp. 13-43.

[3] J. G. S. M. a. M. P. J. Jin, "An Information Framework for Creating a Smart City Through Internet of Things," IEEE Internet of Things Journal, vol. 1, 2014, pp. 112-121.

[4] S Sicari, A Rizzardi, LA Grieco, and A Coen-Porisini. Security, privacy, and trust in the internet of things: The road ahead. Computer Networks, vol. 76, 2015, pp.146–164.

[5] M. Wu, T.-J. Lu, F.-Y. Ling, J. Sun, and H.-Y. Du, “Research on the architecture of internet of things,” in Proceedings of the 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE '10), IEEE, Chengdu, China, vol. 5, pp. V5-484–V5-487, August 2010.

[6] K. Rose, S. Eldridge, L. Chapin, "The internet of things: overview, understanding the issues and the challenges of a More connected world", The Internet Society (ISOC), 2015.

[7] Zheng Yan, Peng Zhang, Athanasios V. Vasilakos, "A survey on trust management for the Internet of Things", Journal of Network and Computer Applications, Volume 42, pp. 120-134, June 2014

[8] Zheng Yan, Jun Liu, Athanasios V. Vasilakos, Laurence T. Yang, "Trustworthy data fusion and mining in Internet of Things, Future Generation" Computer Systems, V. 49, August 2015, pp 45-46.

[9] Bao F., Chen I. "Trust management for the Internet of Things and its application to service composition", Proceedings of the IEEE International Symposium on World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp. 1–6, 2012.

[10] Daly EM, Haahr M. "Social network analysis for information flow in disconnected delay-tolerant MANETs". IEEE Trans Mob Comput 2009; vol. 8, no5 pp 606–21, 2009.

[11] D. Chen, G. Chang, D. Sun, J. Li, J. Jia, X. Wang, "TRM-IoT: a trust management model based on fuzzy reputation for Internet of Things", ComSIS, vol. 8, no. 4, pp. 1207–1228, 2011.

[12] Nitti M, Girau R, Atzori L, Iera A, Morabito G. "A subjective model for trustworthiness evaluation in the social Internet of Things", Proceedings of the IEEE 23rd international symposium on Personal Indoor and Mobile Radio Communications (PIMRC), p. 18–23, 2012.

[13] Ruan, Yefeng, Arjan Durresi, and Lina Alfantoukh. "Trust Management Framework for Internet of Things." 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), 2016.

[14] Wang, Yating, Ray Chen, and Jin-Hee Cho. "Trust-based Service Management of Mobile Devices in Ad Hoc Networks", The Eighth International Conference on Dependability, Venice, Italy, August 2015

[15] Shawn Michael Gallaher, "An Approach For Measuring The Confidentiality Of Data Assured By The Confidentiality Of Information Security Systems In Healthcare Organizations", Doctoral Dissertation, University of Central Florida, 2012.

[16] M. A. a. D. Kantarelis., "Essentials of Inferential Statistics", University Press of America, Fourth Edition, 2004.

[17] F. G. a. L. Wallnau, "Essentials of statistics for the behavioral sciences"., Cengage Learning; 8 edition, January 1, 2013.

[18] R. Newcombe, "Confidence Intervals for Proportions and Related Measures of Effect Size", CRC Press, 2012.

# Singkat: A Keyword-Based URL Shortener and Click Tracker Package for Django Web Application

Gottfried Prasetyadi, Utomo Tri Hantoro, Achmad Benny Mutiara
Faculty of Computer Science and Information Technology
Gunadarma University
Depok, Indonesia

***Abstract*—In recent years, Python has been gaining popularity as web scripting/programming language. In this research we propose Singkat, an open source uniform resource locator (URL) shortener package with web user interface built using Python-based Django web framework. It can be deployed and customized in any web project based on Django framework. This makes sure that administrators can gain control over data in their own environment. Users can create shortened links using base62 values to generate pseudo random keyword. To minimize phishing and other abuses, only registered users can create shortened link using their chosen keyword, and it is possible to preview a link before accessing it. Registered users can also monitor each click and get useful information. We also ran some tests to measure Singkat's performance and functionality.**

***Keywords*—Url shortener; click tracker; python; django framework; open source**

## I. Introduction

A web address is a string used to identify a web resource. It includes uniform resource locator (URL) and other strings of characters, which can be used to identify web resources when processed appropriately [1]. Internet today is full of content and each web page needs its own unique URL. To make a URL identifiable by human, some organization must be applied to make it more structured e.g. by using slashes, date of content creation, slugs and keywords from article and blog posts [2]. However, this makes URL hard to memorize, type manually, or distribute (especially via non-digital media), and can only be copied-pasted for reliability.

URL shortener can shorten long URL into shortened one. This kind of web service has been available for at least 16 years [3]. Tinyurl.com and bit.ly are the examples of popular closed source, non-free URL shortener services for public consumption. There are also open source web modules and packages such as Polr, YOURLS, and Shlink, all of which can be used as extension in a web application within a domain. Some are available to Python-based web application, such as microurl, djanurl, and lilliput.

Python is one of the fastest-growing major programming language. It has risen in the ranks, surpassing both C# and PHP in 2018 [4, 5]. There are many web framework based on Python, such as Django, Web2py, and Flask microframework. According to [4], 13% of programmers worldwide use Django framework in 2018. It ranked sixth in the most commonly used frameworks and libraries. Django employs model-view-template (MVT) architectural pattern, which enables code reuse and modularity of features and functions [6].

Singkat is aimed to be a highly customizable URL shortener module alternative for web applications built using Django, with complete user features for a basic deployable package, such as web user interface, shortening using pseudo random and user-chosen keyword, and clicks statistics. The name 'Singkat' is taken from Indonesian, which means short, brief, or concise. This research centered on algorithm and performance of Singkat, especially its model object, harnessing Python scripting language.

## II. Research Methods

### A. Previous Works

There are several URL shortener modules for Django web framework similar to Singkat. Compared to them, Singkat has distinct differences. Microurl (pypi.org/project/microurl) is Python library for URL minification, which requires external services for authentication: Google or bitly. Djanurl (github.com/ericls/djanurl) is a URL shortener module written in Python version 3. Lilliput (github.com/Feverup/lilliput) is a URL shortener based on older version of Django. It treats links as model objects; each has integer identifier converted to base62 characters, with a fixed length of four characters. All of them are open source and can be found on online code repositories.

### B. Models and Relationships

We defined four core models: 'singkat', 'clicker', 'click', and 'clickdetail'. As in any other MVT framework, a model is a table in a database.

Singkat model is used to represent shortened URLs. Clicker model is used to store data of clients that accessed a singkat object. Click model is used as intermediate model to facilitate many-to-many relationship of singkat model and clicker model. Clickdetail model is used to store timestamp of every clicks. The Django built-in user model is also used to represent registered users of the system.

Table 1 to 4 show the detail of attributes of each core model: the type, length, and other options, e.g., a model should have one unique attribute as primary key. Fig. 1 shows the model relationship; it can be concluded that a 'click' model is used to resolve the many-to-many relationship between 'singkat' and 'clicker' models, which has one-to-many relationship with 'clickdetail' model.

### C. Algorithm for Creating Singkat Object

For both pseudo random and user-chosen keywords, we used 'requests' [7], a Python HTTP library, to check whether the input long URL is accessible at that time, preventing link to invalid URL [8], in contrast to Djanurl. By default, maximum user-chosen keyword length is 100 characters, while maximum long URL input length is 2000 characters to meet the specifications of RFC 7230 [9] and popular web browsers.

#### 1) Pseudo Random Keyword

The pseudocode for creating singkat object with pseudo random keyword is as follows:

```
1. If target_url is not valid: raise error 'Cannot
   access target URL'
2. If target_url is singkat: raise error 'Already a
   Singkat URL'
3. new_singkat = Singkat(target_url,
   keyword=generate_random_keyword())
```

Every new Singkat object is instantiated from Singkat class, which takes two main parameters: the target URL and the keyword, which in this case is generated automatically using generate_random_keyword() function.

We used pseudo random keyword identifier (represented as integer), which is designed to be available globally in the system, and converted it to base62 alphabet to generate pseudo random keyword, so that the newly generated keyword would always be available and not already in use. The alphabet consists of 0-9, A-Z, and a-z, so there are $\sum_{n=1}^{100} 62^n$ keyword possibilities, which is different from Lilliput. The generate_random_keyword() function pseudocode is as follows:

```
1. while true:
     a. n = get_new_random_id()
     b. digits = []
     c. while n:
          i. digits.append(int(n%62))
          ii. n //= 62
     d. digits = digits[::-1]
     e. for idx, val in enumerate(digits):
          i. digits[idx] =
             ALPHABET[val]
     f. keyword = "".join(digits)
     g. if keyword is unique: break
2. return keyword
```

The while-loop in the pseudocode above should take only one loop and a unique keyword is then returned.

#### 2) User-chosen Keyword

The pseudocode for creating singkat object with user-chosen keyword is as follows:

```
1. If target_url is not valid: raise error 'Cannot
   access target URL'
2. If target_url is singkat: raise error 'Already a
   Singkat URL'
3. If keyword is not valid: raise error 'Keyword is
   not valid'
4. If keyword is not unique: raise error 'Keyword
   is already taken'
5. new_singkat = Singkat(target_url,
   keyword=remove_invalid_url_char(keyword)).
```

The remove_invalid_url_char() function will remove any invalid character in the keyword for it will be used as a URL segment. Total 71 characters are considered valid, based on RFC 1738 (without apostrophe and plus characters) [10], so there are $\sum_{n=1}^{100} 71^n$ keyword possibilities. By default, only registered users can create singkat object with chosen keyword. It is a way to minimize phishing and other misuses, such as discussed in [11].

### D. Algorithm for Accessing Singkat Object

Singkat acts as intermediary to gather data when a singkat object is accessed by client (clicker), so the owner of that singkat object can monitor the click data, which includes ip address, geolocation, number of total and unique access, and timestamps. By default, appending '+' character to singkat URL will preview it, allowing the client to know where he/she will be redirected.

The algorithm is as follows:

```
1. If clicker is new:
     a. new_clicker = Clicker(ip,
        geolocation_data)
     b. new_click = Click(clicker, singkat)
     c. new_click_detail =
        ClickDetail(new_click, timestamp)
2. Else:
     a. If click is new:
          i. new_click = Click(clicker,
             singkat)
          ii. new_click_detail =
              ClickDetail(new_click,
              timestamp)
     b. Else:
          i. click = Click(clicker, singkat)
          ii. click.number_of_access += 1
          iii. new_click_detail =
               ClickDetail(click, timestamp)
```

The system will check whether the client is new by looking up the IP address in database. If it is not found, then a new clicker object is instantiated from Clicker class, which requires two main parameters: IP address and geolocation data. If the clicker already exists, then the system will simply update the statistics (total access by that particular client and timestamp). The simplicity of Singkat keyword, especially the pseudo random one, makes it prone to brute-force search [12], which may lead to security and privacy vulnerabilities.

TABLE I. DETAIL OF SINGKAT MODEL

| Field | Type | Options |
|---|---|---|
| id | int | Primary key |
| keyword | char | unique, not null, length = 100 |
| target | char (url) | length = 2000 |
| title | char | blank, length = 100 |
| owner | int | Foreign key, null |
| created_at | datetime | |

TABLE II. DETAIL OF CLICKER MODEL

| Field | Type | Options |
|---|---|---|
| id | int | Primary key |
| ip | char (ip address) | unique |
| city | char | blank, length = 100 |
| country | char | blank, length = 100 |
| continent | char | blank, length = 100 |
| latitude | float | default = 0 |
| longitude | float | default = 0 |
| created_at | datetime | |

TABLE III. DETAIL OF CLICK MODEL

| Field | Type | Options |
|---|---|---|
| id | int | Primary key |
| ip | int | Foreign key |
| singkat | int | Foreign key |
| times | int | default = 1 |

TABLE IV. DETAIL OF CLICKDETAIL MODEL

| Field | Type | Options |
|---|---|---|
| id | int | Primary key |
| click | int | Foreign key |
| time | datetime | |

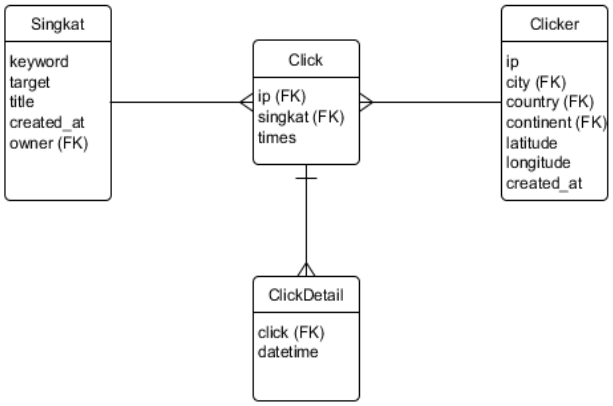



Fig. 1. Singkat model relationship diagram.

## E. Implementation

The Django version used to build the package is 2.0.4 (released on April 2018). The package consists of one file containing view and helper functions, one file containing codes to define all models, one file containing URL patterns, several templates for web interface, additional helpers, and hooks. The total size of the package is 116.3 kB.

For showcase and testing purpose, we temporarily deployed the Singkat package in a Django web project located at http://praz.pythonanywhere.com. However, this hosting has whitelist policy so not every URL can be shortened. We chose to use SQLite as database management system and ipstack API as public IP-to-location service in the testing environment.

Fig. 2 and 3 show the web user interface, which allows users to interact with the package to do operations such as URL shortening, previewing Singkat URL, and accessing campaign statistics. It is currently styled using Bootstrap front-end framework for typography, forms, navigation, and other interface components.

A user who wants to create a new shortened URL will have to enter the original long URL and the keyword that he/she chooses in the web form, and then click 'Generate'. The user can also opt for automatically generated (pseudo random) keyword.

## F. Testing

Three different tests are conducted to ensure the performance and functionality of Singkat:

### 1) Efficiency Testing

This test is to measure how much time it will take for Singkat algorithm to generate pseudo random keyword. As many users may try to shorten URLs at once, Singkat has to be able to generate keywords fast. Our target time is under 100 microseconds.

### 2) Behavioral Testing

This test is to make sure that there is no bug and unexpected behavior in Singkat by listing all possible behaviors and comparing them with actual results.

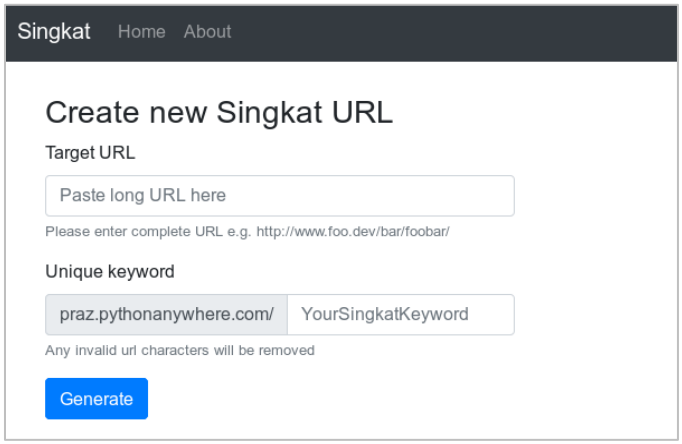



Fig. 2. Singkat object creation form with user-chosen keyword.

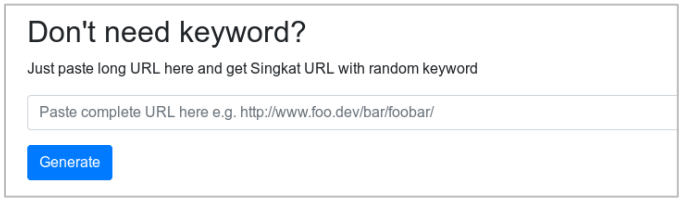



Fig. 3. Singkat object creation form with automatically generated keyword.

*3) Online Campaign*

This campaign is to mimic the real usage of Singkat URL shortener, by distributing a singkat object on various online media over a period, and then analyze whether all the click data have complete attributes, especially the geolocation and click count, so the owner of a singkat object may know whether his/her campaign has reached targeted people.

## III. Results and Analysis

We measured how much time it took to generate a singkat object with pseudo random keyword using Python built-in datetime module. The long URLs as input were from various web page on the internet. We did not take into account the time it took for the 'requests' library to check the validity of long URL. We adjusted the value of current available pseudo random keyword identifier so the generated keyword from base62 conversion was reasonably short and contained multiple characters. We ran the test ten times, each to generate a new singkat objects. We found out that the average time needed to create ten new singkat objects was 0.0000436 second (43.6 microsecond). Because this test was conducted in a free hosting service, the performance should be better in fully dedicated server.

We ran a behavioral testing to ensure acceptable user experience and correct internal functionality based on the proposed algorithms. For example, Singkat package should not allow user to create a Singkat object that links to invalid URL. We excluded some non-core functions such as user registration and login. Table 5 shows expected and actual results. Valid means that the result is as expected, invalid otherwise.

As seen in Table 5, Singkat cannot identify singkat objects created by another Singkat package in different domain. This is a commonly accepted inconvenience, as many URL shortener services can still shorten already shortened URL from other service.

We also created and distributed a singkat object located at http://praz.pythonanywhere.com/TerminalFrost+ on various social media and online forums for research purpose. The experiment began on May 24, 2018 and ended on May 28, 2018. We then extracted all the click data from client area page (ten of them summarized in Table 6, with timestamp field omitted), and then we analyzed them. We managed to gather 156 unique click data; 30 (19.23%) of them had partially complete attributes which was commonly caused by clients accessing using cellular network, or could not be identified by public IP-to-location service. All of the data had complete click count and timestamp attributes, so time series graph could be made, which enabled the owner of the singkat object to track its popularity.

TABLE V. Behavioral Testing Result

| Expected behavior | Actual result |
|---|---|
| User can create singkat object with pseudo random keyword | Valid |
| Registered user can create singkat object with user-chosen keyword | Valid |
| Registered user can access the detail of each of his/her singkat object | Valid |
| The system prevents creating singkat object that links to dead, unreachable, or invalid URL, and shows error notification to the user | Valid |
| The system shows error notification when a registered user tries to create a singkat object with already taken user-chosen keyword | Valid |
| The system automatically removes invalid URL characters in user-chosen keyword when creating singkat object | Valid |
| The system prevents creating singkat object that links to another singkat object, within the same domain | Valid |
| Same as above, but in different domain | Invalid |
| Client can access singkat object and be redirected to destination URL | Valid |
| Client can preview a singkat object before accessing it | Valid |

TABLE VI. Sample Data from Research Campaign

| IP Address | City | Country | Continent | Latitude | Longitude | Number of click |
|---|---|---|---|---|---|---|
| 110.138.149.13 | Bekasi | Indonesia | Asia | -62.349 | 1.069.896 | 1 |
| 61.247.32.164 | Jakarta | Indonesia | Asia | -61.744 | 1.068.294 | 1 |
| 118.137.2.4 | Bandung | Indonesia | Asia | -69.039 | 1.076.186 | 1 |
| 114.124.133.62 | Jakarta | Indonesia | Asia | -61.744 | 1.068.294 | 1 |
| 52.42.129.76 | Boardman | United States | North America | 457.788 | -119.529 | 1 |
| 52.40.20.219 | Boardman | United States | North America | 457.788 | -119.529 | 1 |
| 114.124.234.63 | Jakarta | Indonesia | Asia | -61.744 | 1.068.294 | 1 |
| 120.188.95.134 | Jakarta | Indonesia | Asia | -61.744 | 1.068.294 | 1 |
| 114.124.179.38 | Jakarta | Indonesia | Asia | -61.744 | 1.068.294 | 1 |
| 114.124.176.38 | Jakarta | Indonesia | Asia | -61.744 | 1.068.294 | 1 |

## IV. Conclusion

Performance and all core functions of Singkat had been tested using various method, with satisfying result. Power users can deploy Singkat package by importing it and setting up some configurations, e.g., it is up to the system administrator to choose and set up the database management system powerful enough to prevent locking, which could be caused by the

increasing value of pseudo random keyword identifier and concurrent data access. In Django, this can be configured easily.

The Singkat package can be obtained at github.com/GottfriedCP/Singkat-URL-Shortener. It is licensed under free MIT License.

In the future, RESTful API feature should be developed to make the package more customizable and extensible. Integrated geolocation module should be developed, instead of relying on external IP-to-location service. Algorithms to create both user chosen and pseudo random keyword could also be improved to handle heavy, concurrent usage, and ensuring every keyword will always be unique.

## ACKNOWLEDGMENT

We would like to thank Gunadarma Foundation for their support.

## REFERENCES

[1] Connolly D, Sperberg-McQueen CM. Web addresses in HTML 5. May 2009. [Online]. Available: https://www.w3.org/html/wg/href/draft#url. [Accessed 1 May 2018].

[2] Azar E, Alebicto ME. Swift Data Structure and Algorithms. Birmingham:Packt Publishing. 2016: 236.

[3] prolific. We want 'em shorter. July 2001. [Online]. Available: https://www.metafilter.com/8916/We-want-em-shorter. [Accessed 1 May 2018].

[4] Stack Overflow. Developer Survey Results 2018. [Online]. Available: https://insights.stackoverflow.com/survey/2018/#technology. [Accessed 1 May 2018].

[5] Robinson D. The Incredible Growth of Python. September 2017. [Online]. Available: https://stackoverflow.blog/2017/09/06/incredible-growth-python/?_ga=2.28880508.1369638144.1525141647-1574360211.1503055200. [Accessed 1 May 2018].

[6] George N. Mastering Django Core. Birmingham: Packt Publishing. 2016.

[7] Reitz K, Schlusser T. The Hitchhiker's Guide to Python: Best Practices for Development. Sebastopol:O'Reilly Media. 2016.

[8] Koehler W. A longitudinal study of Web pages continued: a report after six years. *Information Research*. 2004; 9(2): paper 174.

[9] Fielding R, Reschke J. RFC 7230 - Hypertext Transfer Protocol (HTTP/1.1): Message Syntax and Routing. June 2014. [Online]. Available: https://tools.ietf.org/html/rfc7230#section-3.1.1. [Accessed 2 May 2018].

[10] Berners-Lee T, Masinter L, McCahill M. RFC 1738 - Uniform Resource Locators (URL). December 1994. [Online]. Available: https://www.ietf.org/rfc/rfc1738.txt. [Accessed 2 May 2018].

[11] Klien F, Strohmaier M. *Short links under attack: geographical analysis of spam in a URL shortener network*. Proceedings of the 23rd ACM conference on Hypertext and social media. 2012; 83-88.

[12] Georgiev M, Shmatikov V. Gone in six characters: Short urls considered harmful for cloud services. *arXiv*. 2016; arXiv:1604.02734.

# A Quantum based Evolutionary Algorithm for Stock Index and Bitcoin Price Forecasting

Usman Amjad, Tahseen Ahmed Jilani, Humera Tariq
Department of Computer Science
University of Karachi, Karachi, Pakistan

Amir Hussain
School of Natural Science, University of Sterling
United Kingdom

***Abstract*—Quantum computing has emerged as a new dimension with various applications in different fields like robotic, cryptography, uncertainty modeling etc. On the other hand, nature inspired techniques are playing vital role in solving complex problems through evolutionary approach. While evolutionary approaches are good to solve stochastic problems in unbounded search space, predicting uncertain and ambiguous problems in real life is of immense importance. With improved forecasting accuracy many unforeseen events can be managed well. In this paper a novel algorithm for Fuzzy Time Series (FTS) prediction by using Quantum concepts is proposed in this paper. Quantum Evolutionary Algorithm (QEA) is used along with fuzzy logic for prediction of time series data. QEA is applied on interval lengths for finding out optimized lengths of intervals producing best forecasting accuracy. The algorithm is applied for forecasting Taiwan Futures Exchange (TIAFEX) index as well as for Bitcoin crypto currency time series data as a new approach. Model results were compared with many preceding algorithms.**

***Keywords*—Quantum evolutionary algorithm; fuzzy time series; nature inspired computing; fuzzy logic; crypto currency; bitcoin***

## I. INTRODUCTION

Predicting unseen situations and future events hold immense importance and is a need of humanity. Forecasting is important in financial sector as the financial loss in investments can create hurdles for capitalists. Potential areas for forecasting include weather prediction, stock market forecasting, funds management; commodities price calculation, crime rate prediction, floods, droughts and earthquake prediction etc. A lot of research has been done in areas of forecasting and quantum computing. Zadeh [1] introduced fuzzy set theory and worked on its applications later on. Lee et al. [2] proposed algorithms for temperature and stock index prediction on the basis of fuzzy logical relation groups (FLRGs) and genetic algorithm (GA). Jilani and Burney [3], [4] and Jilani, Burney and Ardil [5], [6] proposed new fuzzy metrics for higher-order multivariate FTS forecasting for accidents data of Belgium. Huang et al. [7] formulated a forecasting algorithm on the basis of FTS and Particle Swarm Optimization (PSO), and applied it on enrolments data. Kuo et al. [8] presented a hybrid forecasting algorithm for TAIFEX index forecasting based on FTS and PSO. Jilani, Amjad and Mastorakis [9] presented a hybrid algorithm using GA and PSO for fuzzy time series prediction for TAIFEX index and KSE-100 index and got better results. Amjad, Jilani and Yasmeen [10] formulated a two phase fuzzy time series (FTS) forecasting model using GA and PSO.

Quantum Computing is new computing paradigm based on quantum-mechanical phenomenon like q-bits, superposition and entanglement etc. [11]. Emergence of quantum computing techniques has given immense power to computations and thus these concepts are being applied to different domains and problems of computer science [12]. Concept of Quantum mechanical computers was introduced in early 1980s [11] and the explanation of quantum mechanical computers was enacted in late years of 1980s [12]. Enormous efforts are being made for evolution of quantum computers since inception of Quantum concepts in computing, as these computers are considered more powerful as compared to classical computers for variety of complex problems. Few renowned quantum algorithms include Shor's quantum factoring algorithm [13], [14] presented in 1998, and Grover's database search algorithm [15], [16] presented in mid 1990s. Efforts were made to combine evolutionary algorithms along with quantum computing concepts, which gave birth to quantum-inspired evolutionary computing for a classical computer, an area of Evolutionary Computing (EC) combined with principles of quantum mechanics such as standing waves, interference, coherence, etc. A. Narayanan and M. Moore [17], [18], presented the concept of interference included in a crossover operator.

Concept of crypto currency brought together principles of economics, cryptography and computer science by using encryption technologies for issuance of monetary units. Funds are generated and transferred by using encryption methods without intervention and monitoring of some central governing authority. In 2009, Bitcoin was released as open-source software by Satoshi Nakamoto. Bitcoin is the first peer to peer electronic cash system or crypto currency which revolutionized the concept of digital payment systems [19]. For the prediction of price, studies by Garcia & Schweitzer [20] explain auto regression techniques. They identify two positive feedback loops leading to price bubbles. Similarly, Amjad & Shah [21] propose a theoretical framework and develop a real-time algorithm in order to achieve significant return value on the investment of Bitcoin. Moreover, research study by Kondor et al. [22] explains the Principal Component Analysis related to the block chain network data in order to identify correlation between changes in price exchange and principal variables. Crypto currency web data is used by Kim et al. [23] to predict price fluctuation of Bitcoin. Guo & Antulov-Fantulin [24] construct a temporal mixture model which is a more efficient time-series statistical model than the traditional approaches.

Jang & Lee [25] conducted a comparative study regarding the fluctuation of pricing of Bitcoin.

## II. Review of Key Terms

### A. Quantum Bits (Qubits)

The tiniest entity in which a two-state Quantum computer stores information is called a quantum bit or Qubit [26].

Qubit is atomic unit of information storage in Quantum computer and can be in state "0" or state "1" or any superposition of the two. Bloch sphere representation of Quantum bits is given in Figure 1. The state of quantum bit can be characterized as

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \quad (1)$$

where α and β are two complex values representing the probability amplitudes of respective states. $\alpha^2$ represents the probability of qubit being in state "0" while $\beta^2$ represents probability of being in state "1". Thus the sum of probabilities must be equal to 1

$$|\alpha|^2 + |\beta|^2 = 1 \quad (2)$$

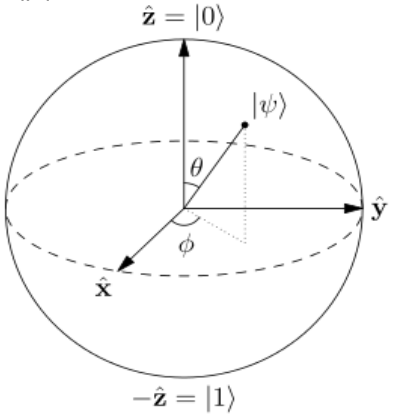



Fig. 1. A Quantum bit represented as a Bloch Sphere

At any given time a Qubit may either be in state "0" or state "1" or a superposition of both the states. Qubit can be represented using a pair of complex values (α,β) like $\begin{bmatrix}\alpha\\ \beta\end{bmatrix}$

### B. Qubit Individual

A Qubit individual is a sequence of qubits with length m represented as

$$q = \begin{bmatrix}\alpha_1 & \alpha_2 & \cdots & \alpha_m\\ \beta_1 & \beta_2 & \cdots & \beta_m\end{bmatrix} \quad (3)$$

where $\alpha_i^2 + \beta_i^2 = 1$ for i=1,2,…,m.

Initially, $\alpha_i$ and $\beta_i$ values are generally set to $\frac{1}{\sqrt{2}}$ for making equal probability of all qubits.

$$q = \begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \cdots & \frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \cdots & \frac{1}{\sqrt{2}}\end{bmatrix} \quad (4)$$

Qubit representation can be used to represent superposition of states. So, a three qubit individual can represent total eight different states, which are $\langle 000|, \langle 001|, \langle 010|, \langle 011|, \langle 100|, \langle 101|, \langle 110|, \langle 111|$

### C. Qubit Population

Qubit population consists of n different Qubit individuals each having m Qubits in it. So, Qubit population can be represented in form of a set of m Qubits like,

$$Q(t) = \{q_{1,}^t q_{2,}^t \cdots, q_{n,}^t\} \quad (5)$$

where $q_i^t$ is ith Qubit individual at time t.

### D. Q-gate

Variation is produced in Qubit individuals using a rotation operator called Q-gate. Rotation is applied to each Qubit and their positions are updated keeping in view the constraint of normalization of probabilities i.e. $|\alpha|^2 + |\beta|^2 = 1$. The rotation operator used as Q-gate usually is

$$U(\Delta\theta_i) = \begin{bmatrix}\cos(\Delta\theta_i) & -\sin(\Delta\theta_i)\\ \sin(\Delta\theta_i) & \cos(\Delta\theta_i)\end{bmatrix} \quad (6)$$

where $\theta_i$ , i=1,2, … , m is rotation angle for each Qubit. Q-gate is applied individually to each Qubit of every individual of population, which moves the Qubit either towards 0 or 1. Furthermore, there are operators other than rotation, which can be used as Q-gates such as NOT gate, controlled NOT gate, Hadmard gate etc.

### E. Quantum Evolutionary Algorithm (QEA)

QEA was proposed by Kuk-Hyun Han in his Ph.D thesis titled "Quantum-inspired Evolutionary Algorithm" in 2003. Concepts of Quantum computing and evolutionary computing were combined in this algorithm. The proposed algorithm is given in Figure 2.

### F. Fuzzy Time Series (FTS)

To deal with uncertain situations, Fuzzy logic and fuzzy set theory was coined by Zadeh [27]. Later on Song and Chisom [28] introduced a time series to tackle fuzziness, called Fuzzy Time Series (FTS). A number of variants of FTS were produced and many researchers used FTS in different areas of application. Few basic concepts of FTS include fuzzy operations, fuzzy relations and fuzzy logical relationship groups [3], [4], [5].

## III. Proposed Algorithm for FTS Forecasting using QEA

**Step 1:** Define the universe of discourse

Define universe of discourse for primary-factor X like U = [$D_{min}$-$D_1$, $D_{max}$+$D_2$], where $D_{min}$ is minimum index value and $D_{max}$ is maximum index value from the historical data under consideration, and $D_1$ and $D_2$ are positive real numbers for extending universe of discourse[3] [4].

Similarly, define the universe of discourse V of the second-factor(Y) V= [$E_{min}$-$E_1$, $E_{max}$+$E_2$], where $E_{min}$ is the minimum value of secondary factor and $E_{max}$ is the maximum value of secondary factor from historical data under consideration, respectively, and E1 and E2 are two positive numbers used for extending universe of discourse for secondary factor. Also find out Range for both main and secondary factors [5].

Higher Education Commission (HEC) Pakistan

**Step 2:** Initialize QEA population

```
Procedure QEA
begin
        t ← o
i)      initialize Q(t)
ii)     make P(t) by observing the states of Q(t)
iii)    evaluate P(t)
iv)     store the best solutions among P(t) into B(t)
v)      while (not termination criteria) do
        begin
            t ← t + 1
vi)         make P(t) by observing the state of Q(t - 1)
vii)        evaluate P(t)
viii)       update Q(t) using Q-gates
ix)         store the best solution among B(t - 1) and P(t) into B(t)
x)          store the best solution b among B(t)
xi)         if (global migration condition)
            then migrate b to B(t) globally
xii)        else if (local migration condition)
            then migrate b_j^i in B(t) to B(t) locally
        end
end
```

Fig. 2. Pseudo code of Quantum Evolutionary Algorithm

QEA population comprises of 15 Qubit individuals and each Qubit individual has 11 Qubit. Number of Qubit individuals is because we will divide Universe of Discourse into 16 intervals, making use of Dmin and Dmax, while 11 Qubits in each individual is because of range of main factor which is 1380 and can be represented using 11 binary numbers.

$$2^n = 1380$$

To find out value of n, we take log base 2 on both sides.

$$\log_2 2^n = \log_2 1380$$

$$n = 10.4308$$

Bits can only be in whole numbers so applying ceiling function on value of n gives,

$$n = 11$$

All Qubit values in population of individuals are initialized with random values of α between 0 and 1, whereas β is initialized with values of $\sqrt{1-\alpha^2}$ thus maintaining normality constraint $\alpha_i^2 + \beta_i^2 = 1$

**Step 3:** Get intervals from QEA individuals

There are 15 individuals in QEA population. For each individual get binary strings using Qubits. Binary strings from Qubits are formed on the principle that if the value of $\alpha^2$ in a Qubit is greater than 0.5, then append '1' to the string otherwise append '0'. In this way a string containing 11 binary bits will be obtained from Qubits individuals. Then convert that binary string into decimal number and add that number into minimum value of time series i.e. Dmin, to get a value in interval. If that value is exceeding the maximum value of time series i.e. Dmax then subtract a random number from it to bring it in universe of discourse. Finally sort all the interval values in ascending order.

TABLE I. INITIAL POPULATION OF QUBIT INDIVIDUALS

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **α** | 0.3151 | 0.4529 | 0.5350 | 0.1194 | 0.5840 | 0.5981 | 0.0389 | 0.8542 | 0.8792 | 0.9267 | 0.9249 |
| **β** | 0.9490 | 0.8916 | 0.8448 | 0.9929 | 0.8118 | 0.8014 | 0.9992 | 0.5199 | 0.4764 | 0.3758 | 0.3803 |
| **α** | 0.0642 | 0.3396 | 0.0561 | 0.7679 | 0.7999 | 0.7621 | 0.9873 | 0.9856 | 0.5291 | 0.6893 | 0.3898 |
| **β** | 0.9979 | 0.9406 | 0.9984 | 0.6406 | 0.6002 | 0.6474 | 0.1589 | 0.1690 | 0.8486 | 0.7245 | 0.9209 |
| **α** | 0.4547 | 0.5315 | 0.3933 | 0.0014 | 0.7920 | 0.9438 | 0.7668 | 0.0901 | 0.1996 | 0.5518 | 0.0671 |
| **β** | 0.8907 | 0.8471 | 0.9194 | 1.0000 | 0.6105 | 0.3305 | 0.6419 | 0.9959 | 0.9799 | 0.8340 | 0.9977 |
| **α** | 0.3674 | 0.0169 | 0.4296 | 0.4483 | 0.2356 | 0.9994 | 0.9944 | 0.6962 | 0.9937 | 0.7832 | 0.4812 |
| **β** | 0.9300 | 0.9999 | 0.9030 | 0.8939 | 0.9719 | 0.0354 | 0.1059 | 0.7179 | 0.1120 | 0.6218 | 0.8766 |
| **α** | 0.7578 | 0.2088 | 0.7668 | 0.6817 | 0.2277 | 0.1810 | 0.7739 | 0.8006 | 0.6641 | 0.6457 | 0.1584 |
| **β** | 0.6525 | 0.9780 | 0.6418 | 0.7316 | 0.9737 | 0.9835 | 0.6333 | 0.5992 | 0.7476 | 0.7636 | 0.9874 |
| **α** | 0.6706 | 0.6942 | 0.8031 | 0.1286 | 0.6713 | 0.2366 | 0.0015 | 0.4067 | 0.4583 | 0.8770 | 0.5725 |
| **β** | 0.7418 | 0.7198 | 0.5959 | 0.9917 | 0.7412 | 0.9716 | 1.0000 | 0.9136 | 0.8888 | 0.4804 | 0.8199 |
| **α** | 0.4198 | 0.5809 | 0.3241 | 0.7772 | 0.8872 | 0.4006 | 0.9499 | 0.5381 | 0.1082 | 0.6396 | 0.0374 |
| **β** | 0.9076 | 0.8140 | 0.9460 | 0.6293 | 0.4614 | 0.9162 | 0.3126 | 0.8429 | 0.9941 | 0.7687 | 0.9993 |
| **α** | 0.8101 | 0.7728 | 0.6614 | 0.0106 | 0.8793 | 0.5823 | 0.7294 | 0.6425 | 0.7786 | 0.5022 | 0.7147 |
| **β** | 0.5862 | 0.6346 | 0.7500 | 0.9999 | 0.4762 | 0.8130 | 0.6841 | 0.7663 | 0.6275 | 0.8648 | 0.6994 |
| **α** | 0.7229 | 0.2582 | 0.6976 | 0.4575 | 0.3229 | 0.6379 | 0.9570 | 0.2487 | 0.5728 | 0.7335 | 0.1288 |
| **β** | 0.6909 | 0.9661 | 0.7165 | 0.8892 | 0.9464 | 0.7702 | 0.2902 | 0.9686 | 0.8197 | 0.6797 | 0.9917 |
| **α** | 0.4721 | 0.1448 | 0.2187 | 0.1061 | 0.5388 | 0.8019 | 0.9054 | 0.3800 | 0.2227 | 0.4961 | 0.5937 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **β** | 0.8815 | 0.9895 | 0.9758 | 0.9944 | 0.8424 | 0.5975 | 0.4246 | 0.9250 | 0.9749 | 0.8683 | 0.8047 |
| **α** | 0.8625 | 0.3368 | 0.5560 | 0.3396 | 0.5309 | 0.9835 | 0.6849 | 0.4845 | 0.8931 | 0.3587 | 0.2710 |
| **β** | 0.5060 | 0.9416 | 0.8312 | 0.9406 | 0.8474 | 0.1806 | 0.7286 | 0.8748 | 0.4498 | 0.9335 | 0.9626 |
| **α** | 0.1341 | 0.5169 | 0.7760 | 0.2016 | 0.1983 | 0.0215 | 0.0814 | 0.1175 | 0.6667 | 0.4900 | 0.4728 |
| **β** | 0.9909 | 0.8561 | 0.6307 | 0.9795 | 0.9801 | 0.9998 | 0.9967 | 0.9931 | 0.7453 | 0.8717 | 0.8812 |
| **α** | 0.5244 | 0.7088 | 0.1133 | 0.4350 | 0.1904 | 0.2031 | 0.8609 | 0.2220 | 0.3372 | 0.3526 | 0.1500 |
| **β** | 0.8514 | 0.7054 | 0.9936 | 0.9004 | 0.9817 | 0.9792 | 0.5088 | 0.9751 | 0.9414 | 0.9358 | 0.9887 |
| **α** | 0.4372 | 0.1942 | 0.1495 | 0.8819 | 0.6340 | 0.2587 | 0.0885 | 0.8281 | 0.1313 | 0.5839 | 0.5641 |
| **β** | 0.8993 | 0.9810 | 0.9888 | 0.4714 | 0.7733 | 0.9660 | 0.9961 | 0.5606 | 0.9913 | 0.8118 | 0.8257 |
| **α** | 0.8276 | 0.3861 | 0.4868 | 0.1154 | 0.6261 | 0.4404 | 0.8680 | 0.9325 | 0.8018 | 0.4465 | 0.2414 |
| **β** | 0.5613 | 0.9224 | 0.8735 | 0.9933 | 0.7797 | 0.8978 | 0.4966 | 0.3611 | 0.5976 | 0.8948 | 0.9704 |

**Step 4:** Find forecast and store best values

State the linguistic value $A_i$ for main factor, where i=1, 2,…, n, and n is the total interval count in universe of discourse for primary factor [3], [4], [5].

$$A_1 = {}^{1}/_{u_1} + {}^{0.5}/_{u_2} + {}^{0}/_{u_3} + \cdots\cdots + {}^{0}/_{u_{n-2}} + {}^{0}/_{u_{n-1}} + {}^{0}/_{u_n}$$

$$A_2 = {}^{0.5}/_{u_1} + {}^{1}/_{u_2} + {}^{0.5}/_{u_3} + \cdots\cdots + {}^{0}/_{u_{n-2}} + {}^{0}/_{u_{n-1}} + {}^{0}/_{u_n}$$

$$\vdots \qquad (7)$$

$$A_n = {}^{0}/_{u_1} + {}^{0}/_{u_2} + {}^{0}/_{u_3} + \cdots\cdots + {}^{0}/_{u_{n-2}} + {}^{0.5}/_{u_{n-1}} + {}^{1}/_{u_n}$$

Similarly, state the linguistic term Bj for secondary factor where j=1, 2,…, m, and m is the total number of intervals present in universe of discourse V of secondary factor:

$$B_1 = {}^{1}/_{v_1} + {}^{0.5}/_{v_2} + {}^{0}/_{v} + \cdots\cdots + {}^{0}/_{v_{n-2}} + {}^{0}/_{v_{n-1}} + {}^{0}/_{v_n}$$

$$B_2 = {}^{0.5}/_{v_1} + {}^{1}/_{v_2} + {}^{0.5}/_{v_3} + \cdots\cdots + {}^{0}/_{v_{n-2}} + {}^{0}/_{v_{n-1}} + {}^{0}/_{v_n}$$

$$\vdots \qquad (8)$$

$$B_n = {}^{0}/_{v_1} + {}^{0}/_{v_2} + {}^{0}/_{v_3} + \cdots\cdots + {}^{0}/_{v_{n-2}} + {}^{0.5}/_{v_{n-1}} + {}^{1}/_{v_n}$$

Historical data is fuzzified as described below.

Determine the interval $u_i$, 1≤ i≤ n, to which the value of the main-factor belongs.

*a)* If the value belongs to interval u1, then it is fuzzified into $1/A_1+0.5/A_2$, denoted by $X_1$.

*b)* If the value belongs to ui, 2≤ i≤ n-1, then it is fuzzified into 0.5/Ai-1+1/Ai+0.5/Ai+1, denoted by Xi.

*c)* If the value belongs to $u_n$, then it is fuzzified into 0.5/An-1+1/An, denoted by Xn.

Determine the interval $v_j$, 1≤ j≤ m, to which the value of the second-factor belongs.

*a)* If value belongs to v1, then it is fuzzified into 1/B1+0.5/B2, denoted by Y1.

*b)* If value belongs to vj, 2≤ j≤ m-1, then it is fuzzified into 0.5/Bj-1+1/Bj+0.5/Bj+1, denoted by Yj.

*c)* If value belongs to vm, then it is fuzzified into 0.5/Bm-1+1/Bm, denoted by Ym.

Build two factor nth-order FLRs based on primary and second factors from the fuzzified historical data obtained in above step. If the fuzzified data of primary factor for day i is $X_i$, then create two factor kth-order FLRs from day i-k to day i as $((X_{ik}, Y_{ik}), \ldots, (X_{i2}, Y_{i2}), (X_{i1}, Y_{i1})) \rightarrow X_i$ where 2≤ k≤ n and $X_{ik}, \cdots\cdots X_{i2}, X_{i1}$ represent the fuzzified data values for primary-factor for days i-k,…, i-2, i-1 respectively; $Y_{ik}, \cdots\cdots Y_{i2}, Y_{i1}$ denote the fuzzified data values of secondary factor for days i-k,…, i-2, i-1, respectively. After that FLRs are divided into fuzzy logical relationship groups (FLRGs) based on the present states of FLRs [3].

For the determined FLRGs forecasted value is calculated using formula given below:

$$F_k = m_{k-1} + \frac{2 * (m_k - m_{k-1}) + 1.5 * (m_{k-1} - m_{k-2}) + 0.5 * (m_{k-2} - m_{k-3})}{4} \qquad (9)$$

Where $F_k$ is forecasted value at time k where as $m_k, m_{k-1}, m_{k-2}, m_{k-3}$ are mid points of intervals in which value at time k lies in.

After finding the forecasted value determine then fitness value of each individual chromosome of the population. We have used Average forecasting error rate (AFER) and mean square error (MSE) as fitness values of each individual for our forecasting algorithm.

Based on fitness value, if fitness value of current lot of individuals is better than previous individuals than store these Qubit individuals separately, otherwise replace these individuals by previously best individuals.

**Step 5:** Apply rotation on individuals suing Q-gate

Apply Quantum gate on individuals with angle θ. Initially value of θ is set to 0, but with each iteration, an increment of 0.01 is made. Q-gate is applied on all individuals and in this way a new population of individuals evolves out.

**Step 6:** Check stopping criteria or repeat

Check if the stopping criteria of algorithm are met or not. Stopping criteria may be the value of fitness function to reach to a specific point or the number of repetitions. If stopping criteria is not met then go to step 3 and repeat all further steps.

## IV. Experimental Results

We have used TAIFEX index and Bitcoin closing price historical data for experiment using our proposed algorithm. We have applied the suggested algorithm to the TAIFEX dataset for the duration of August 3, 1998 to September 30, 1998 where the TAIFEX index values are taken as the primary-factor and the TAIEX index values are taken as the second-factor. We have executed proposed method and compared obtained results for TAIFEX forecasting with existing method's results. Results obtained from our Novel Algorithm and the error criteria AFER and MSE are shown in Table 4. Table 2 shows best Qubits obtained after evolution for TAIFEX forecasting. Table 3 shows the best individual obtained after evolution for TAIFEX forecasting.

We have also applied the proposed algorithm for forecasting of Bitcoin closing prices. Data for Bitcoin prices has been obtained from coinmarketcap, an online source of crypto currency prices and their market capitalization. Obtained dataset comprise of historical data of 3 months for daily closing prices of Bitcoin from 22nd February 2018 till 22nd May 2018. . We have applied 2000 generations of evolution for Bitcoin forecasting. Forecasting results for Bitcoin forecasting are shown in Table 5. Table 6 shows best individual evolved after applying proposed algorithm for Bitcoin forecasting. Figure 5 shows comparison of actual and forecasted time series data for Bitcoin forecasting.

TABLE II. Best Qubits

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **α** | -0.895 | 0.859 | 0.807 | -0.067 | -0.851 | -0.600 | -0.232 | 0.057 | 0.977 | -0.871 | 0.002 |
| **β** | 0.447 | 0.513 | 0.590 | 0.998 | 0.526 | 0.800 | 0.973 | 0.998 | 0.212 | 0.491 | 1.000 |
| **α** | -0.873 | 0.130 | -0.209 | 0.787 | -0.954 | -0.157 | -0.939 | 0.715 | -0.869 | 0.235 | -0.085 |
| **β** | 0.488 | 0.992 | 0.978 | 0.617 | 0.301 | 0.988 | 0.344 | 0.699 | 0.495 | 0.972 | 0.996 |
| **α** | -0.921 | -0.817 | -0.061 | -0.431 | -0.611 | 0.159 | -0.146 | 0.453 | -0.112 | -0.915 | -0.640 |
| **β** | 0.390 | 0.576 | 0.998 | 0.902 | 0.791 | 0.987 | 0.989 | 0.892 | 0.994 | 0.404 | 0.769 |
| **α** | -0.915 | 0.072 | 0.016 | -0.486 | -0.197 | -0.020 | -0.200 | 0.455 | -0.827 | 0.544 | 0.949 |
| **β** | 0.403 | 0.997 | 1.000 | 0.874 | 0.980 | 1.000 | 0.980 | 0.891 | 0.563 | 0.839 | 0.315 |
| **α** | 0.969 | -0.518 | -0.550 | 0.624 | 0.372 | -0.093 | -0.043 | -0.168 | 0.907 | 0.167 | -0.708 |
| **β** | 0.248 | 0.855 | 0.835 | 0.782 | 0.928 | 0.996 | 0.999 | 0.986 | 0.422 | 0.986 | 0.706 |
| **α** | -0.863 | -0.065 | -0.922 | -0.246 | 0.021 | 0.238 | 0.061 | 0.882 | -0.054 | -0.137 | -0.297 |
| **β** | 0.505 | 0.998 | 0.387 | 0.969 | 1.000 | 0.971 | 0.998 | 0.471 | 0.999 | 0.991 | 0.955 |
| **α** | 0.438 | -0.417 | -0.764 | -0.949 | -0.767 | 0.276 | 0.972 | -0.915 | 0.935 | 0.323 | -0.260 |
| **β** | 0.899 | 0.909 | 0.645 | 0.315 | 0.642 | 0.961 | 0.235 | 0.405 | 0.355 | 0.946 | 0.966 |
| **α** | 0.368 | -0.868 | 0.966 | 0.549 | 0.480 | 0.024 | 0.792 | -0.042 | -0.996 | -0.705 | 0.444 |
| **β** | 0.930 | 0.496 | 0.260 | 0.836 | 0.877 | 1.000 | 0.610 | 0.999 | 0.085 | 0.710 | 0.896 |
| **α** | 0.359 | -0.893 | 0.314 | -0.856 | -0.970 | -0.871 | 0.900 | -0.120 | -0.430 | -0.715 | 0.824 |
| **β** | 0.933 | 0.450 | 0.949 | 0.517 | 0.243 | 0.492 | 0.435 | 0.993 | 0.903 | 0.699 | 0.566 |
| **α** | -0.177 | -0.983 | 0.474 | 0.214 | -0.973 | -0.008 | 0.777 | -0.281 | -0.178 | -0.647 | -0.477 |
| **β** | 0.984 | 0.182 | 0.880 | 0.977 | 0.232 | 1.000 | 0.629 | 0.960 | 0.984 | 0.762 | 0.879 |
| **α** | -0.843 | -0.176 | 0.852 | 0.045 | -0.486 | 0.067 | -0.125 | -0.777 | -0.016 | 0.720 | -0.433 |
| **β** | 0.538 | 0.984 | 0.524 | 0.999 | 0.874 | 0.998 | 0.992 | 0.629 | 1.000 | 0.693 | 0.901 |
| **α** | -0.268 | -0.981 | 0.080 | 0.435 | -0.557 | -0.892 | -0.443 | 0.899 | -0.422 | -0.442 | -0.709 |
| **β** | 0.963 | 0.194 | 0.997 | 0.901 | 0.830 | 0.452 | 0.897 | 0.438 | 0.907 | 0.897 | 0.705 |
| **α** | -0.528 | 0.570 | -0.203 | 0.204 | -0.715 | -0.251 | -0.075 | -0.920 | 0.084 | -0.055 | -0.885 |
| **β** | 0.849 | 0.821 | 0.979 | 0.979 | 0.699 | 0.968 | 0.997 | 0.392 | 0.996 | 0.998 | 0.465 |
| **α** | 0.405 | 0.494 | -0.701 | -0.022 | 0.103 | -0.238 | -0.936 | -0.240 | -0.065 | -0.642 | 0.265 |
| **β** | 0.914 | 0.869 | 0.713 | 1.000 | 0.995 | 0.971 | 0.351 | 0.971 | 0.998 | 0.767 | 0.964 |
| **α** | -0.967 | -0.206 | 0.389 | 0.606 | -0.434 | -0.146 | 0.129 | -0.829 | 0.753 | 0.848 | -0.391 |
| **β** | 0.254 | 0.979 | 0.921 | 0.795 | 0.901 | 0.989 | 0.992 | 0.559 | 0.658 | 0.530 | 0.920 |

TABLE III. Best Intervals

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6206 | 6263 | 6666 | 6743 | 6782 | 6945 | 6978 | 7219 | 7219 | 7228 | 7434 | 7478 | 7480 | 7511 | 7511 |

TABLE IV. FORECASTED TAIFEX VALUES

| Date | TAIFEX index | TAIEX index | Forecasted TAIFEX | Date | TAIFEX index | TAIEX index | Forecasted TAIFEX |
|---|---|---|---|---|---|---|---|
| 8/3/1998 | 7552 | 7599 | - | 9/2/1998 | 6430 | 6472 | 6377.5 |
| 8/4/1998 | 7560 | 7593 | - | 9/3/1998 | 6200 | 6251 | 6298.25 |
| 8/5/1998 | 7487 | 7500 | - | 9/4/1998 | 6403.2 | 6463 | 6232.938 |
| 8/6/1998 | 7462 | 7472 | 7409.75 | 9/5/1998 | 6697.5 | 6756 | 6595.075 |
| 8/7/1998 | 7515 | 7530 | 7482 | 9/7/1998 | 6722.3 | 6801 | 6824.938 |
| 8/10/1998 | 7365 | 7372 | 7485.5 | 9/8/1998 | 6859.4 | 6942 | 6833.3 |
| 8/11/1998 | 7360 | 7384 | 7331.5 | 9/9/1998 | 6769.6 | 6895 | 6878.775 |
| 8/12/1998 | 7330 | 7352 | 7343.25 | 9/10/1998 | 6709.75 | 6804 | 6720.038 |
| 8/13/1998 | 7291 | 7363 | 7284 | 9/11/1998 | 6726.5 | 6842 | 6670.25 |
| 8/14/1998 | 7320 | 7348 | 7302.5 | 9/14/1998 | 6774.55 | 6860 | 6755.938 |
| 8/15/1998 | 7300 | 7372 | 7343 | 9/15/1998 | 6762 | 6858 | 6764.737 |
| 8/17/1998 | 7219 | 7274 | 7245.75 | 9/16/1998 | 6952.75 | 6973 | 6862.125 |
| 8/18/1998 | 7220 | 7182 | 7169.688 | 9/17/1998 | 6906 | 7001 | 6992.5 |
| 8/19/1998 | 7285 | 7293 | 7203.563 | 9/18/1998 | 6842 | 6962 | 6905.75 |
| 8/20/1998 | 7274 | 7271 | 7285 | 9/19/1998 | 7039 | 7150 | 6944.938 |
| 8/21/1998 | 7225 | 7213 | 7274 | 9/21/1998 | 6861 | 7029 | 7010.75 |
| 8/24/1998 | 6955 | 6958 | 7075.75 | 9/22/1998 | 6926 | 7034 | 6804.5 |
| 8/25/1998 | 6949 | 6908 | 6843.063 | 9/23/1998 | 6852 | 6962 | 6897.75 |
| 8/26/1998 | 6790 | 6814 | 6831.188 | 9/24/1998 | 6890 | 6980 | 6852 |
| 8/27/1998 | 6835 | 6813 | 6769.875 | 9/25/1998 | 6871 | 6980 | 6890 |
| 8/28/1998 | 6695 | 6724 | 6765.563 | 9/28/1998 | 6840 | 6911 | 6871 |
| 8/29/1998 | 6728 | 6736 | 6645.438 | 9/29/1998 | 6806 | 6885 | 6799.75 |
| 8/31/1998 | 6566 | 6550 | 6582.125 | 9/30/1998 | 6787 | 6834 | 6775.813 |
| 9/1/1998 | 6409 | 6335 | 6471.5 | | | | |
| | | | | | | **AFER** | **0.8478** |

TABLE V. FORECASTED VALUES OF BITCOIN CLOSING PRICE

| Date | Close Price | Forecast | Date | Close Price | Forecast |
|---|---|---|---|---|---|
| 22-Feb-18 | 10,005.00 | - | 8-Apr-18 | 7,023.52 | 6931.557 |
| 23-Feb-18 | 10,301.10 | - | 9-Apr-18 | 6,770.73 | 6931.557 |
| 24-Feb-18 | 9,813.07 | - | 10-Apr-18 | 6,834.76 | 6931.557 |
| 25-Feb-18 | 9,664.73 | 9703.155 | 11-Apr-18 | 6,968.32 | 6931.557 |
| 26-Feb-18 | 10,366.70 | 10420.53 | 12-Apr-18 | 7,889.25 | 7897.048 |
| 27-Feb-18 | 10,725.60 | 10897.8 | 13-Apr-18 | 7,895.96 | 7897.048 |
| 28-Feb-18 | 10,397.90 | 10420.53 | 14-Apr-18 | 7,986.24 | 7897.048 |
| 1-Mar-18 | 10,951.00 | 10897.8 | 15-Apr-18 | 8,329.11 | 8315.837 |
| 2-Mar-18 | 11,086.40 | 10897.8 | 16-Apr-18 | 8,058.67 | 7897.048 |
| 3-Mar-18 | 11,489.70 | 11480.03 | 17-Apr-18 | 7,902.09 | 7897.048 |
| 4-Mar-18 | 11,512.60 | 11480.03 | 18-Apr-18 | 8,163.42 | 8315.837 |
| 5-Mar-18 | 11,573.30 | 11560.43 | 19-Apr-18 | 8,294.31 | 8315.837 |
| 6-Mar-18 | 10,779.90 | 10897.8 | 20-Apr-18 | 8,845.83 | 8873.62 |
| 7-Mar-18 | 9,965.57 | 9703.155 | 21-Apr-18 | 8,895.58 | 8873.62 |
| 8-Mar-18 | 9,395.01 | 9428.164 | 22-Apr-18 | 8,802.46 | 8873.62 |
| 9-Mar-18 | 9,337.55 | 9428.164 | 23-Apr-18 | 8,930.88 | 8873.62 |
| 10-Mar-18 | 8,866.00 | 8873.62 | 24-Apr-18 | 9,697.50 | 9703.155 |
| 11-Mar-18 | 9,578.63 | 9703.155 | 25-Apr-18 | 8,845.74 | 8873.62 |
| 12-Mar-18 | 9,205.12 | 9116.029 | 26-Apr-18 | 9,281.51 | 9116.029 |
| 13-Mar-18 | 9,194.85 | 9116.029 | 27-Apr-18 | 8,987.05 | 9116.029 |
| 14-Mar-18 | 8,269.81 | 8315.837 | 28-Apr-18 | 9,348.48 | 9428.164 |
| 15-Mar-18 | 8,300.86 | 8315.837 | 29-Apr-18 | 9,419.08 | 9428.164 |
| 16-Mar-18 | 8,338.35 | 8315.837 | 30-Apr-18 | 9,240.55 | 9116.029 |
| 17-Mar-18 | 7,916.88 | 7897.048 | 1-May-18 | 9,119.01 | 9116.029 |
| 18-Mar-18 | 8,223.68 | 8315.837 | 2-May-18 | 9,235.92 | 9116.029 |
| 19-Mar-18 | 8,630.65 | 8629.421 | 3-May-18 | 9,743.86 | 9703.155 |
| 20-Mar-18 | 8,913.47 | 8873.62 | 4-May-18 | 9,700.76 | 9703.155 |
| 21-Mar-18 | 8,929.28 | 8873.62 | 5-May-18 | 9,858.15 | 9703.155 |
| 22-Mar-18 | 8,728.47 | 8629.421 | 6-May-18 | 9,654.80 | 9703.155 |
| 23-Mar-18 | 8,879.62 | 8873.62 | 7-May-18 | 9,373.01 | 9428.164 |
| 24-Mar-18 | 8,668.12 | 8629.421 | 8-May-18 | 9,234.82 | 9116.029 |
| 25-Mar-18 | 8,495.78 | 8315.837 | 9-May-18 | 9,325.18 | 9428.164 |

| | | | | | |
|---|---|---|---|---|---|
| 26-Mar-18 | 8,209.40 | 8315.837 | 10-May-18 | 9,043.94 | 9116.029 |
| 27-Mar-18 | 7,833.04 | 7897.048 | 11-May-18 | 8,441.49 | 8315.837 |
| 28-Mar-18 | 7,954.48 | 7897.048 | 12-May-18 | 8,504.89 | 8315.837 |
| 29-Mar-18 | 7,165.70 | 7456.64 | 13-May-18 | 8,723.94 | 8629.421 |
| 30-Mar-18 | 6,890.52 | 6931.557 | 14-May-18 | 8,716.79 | 8629.421 |
| 31-Mar-18 | 6,973.53 | 6931.557 | 15-May-18 | 8,510.38 | 8315.837 |
| 1-Apr-18 | 6,844.23 | 6931.557 | 16-May-18 | 8,368.83 | 8315.837 |
| 2-Apr-18 | 7,083.80 | 7110.363 | 17-May-18 | 8,094.32 | 7897.048 |
| 3-Apr-18 | 7,456.11 | 7456.64 | 18-May-18 | 8,250.97 | 8315.837 |
| 4-Apr-18 | 6,853.84 | 6931.557 | 19-May-18 | 8,247.18 | 8315.837 |
| 5-Apr-18 | 6,811.47 | 6931.557 | 20-May-18 | 8,513.25 | 8315.837 |
| 6-Apr-18 | 6,636.32 | 6931.557 | 21-May-18 | 8,418.99 | 8315.837 |
| 7-Apr-18 | 6,911.09 | 6931.557 | 22-May-18 | 8,041.78 | 7897.048 |
| | | | **MSE** | | **11049.89** |
| | | | **AFER** | | **0.96%** |

TABLE VI. BEST INTERVAL FOR BITCOIN FORECASTING

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7078 | 7114 | 7726 | 8151 | 8542 | 8786 | 8932 | 9317 | 9424 | 10333 | 10617 | 11194 | 11485 | 11549 | 11549 |

## V. COMPARISON WITH OTHER MODELS

Our proposed model is unique in time series forecasting as it uses concepts of Quantum computing and evolutionary algorithms along with fuzzy logic. The proposed algorithm was compared to other models and we observed that it is producing better results than many predecessors. Forecast from proposed model is having AFER of 0.84% and MSE of 5252.28. Comparisons of forecasting results obtained from our algorithm for TAIFEX forecasting are shown in Table 7 and Table 8. Figure 3 shows comparison of actual time series data and forecasted data from our algorithm. Table 8 shows comparison of our forecasting result with various other algorithms' results also shown in Figure 4.

TABLE VII. AFERS OF DIFFERENT FORECASTING MODELS

| Model | AFER |
|---|---|
| Lee et al.'s method (Lee et al., 2008) α=0.5 | 0.84% |
| Lee et al.'s method (Lee et al., 2007) First Order | 1.24% |
| Huarng's Algorithm (2001a) | 1.03% |
| Huarng's Algorithm (2001b) | 0.89% |
| Proposed method | 0.84% |

TABLE VIII. FORECASTS FROM DIFFERENT FORECASTING MODELS AND PROPOSED MODELS

| | Actual TAIFEX Index | Chen's Algorithm [29] | Huarng Algorithm 2001 [30] | Huarng Algorithm 2001 [31] | Proposed Algorithm |
|---|---|---|---|---|---|
| 8/3/1998 | 7552 | -- | -- | -- | -- |
| 8/4/1998 | 7560 | 7450 | 7450 | 7450 | -- |
| 8/5/1998 | 7487 | 7450 | 7450 | 7450 | -- |
| 8/6/1998 | 7462 | 7500 | 7450 | 7500 | 7409.75 |
| 8/7/1998 | 7515 | 7500 | 7500 | 7500 | 7482 |
| 8/10/1998 | 7365 | 7450 | 7450 | 7450 | 7485.5 |
| 8/11/1998 | 7360 | 7300 | 7350 | 7300 | 7331.5 |
| 8/12/1998 | 7330 | 7300 | 7300 | 7300 | 7343.25 |
| 8/13/1998 | 7291 | 7300 | 7350 | 7300 | 7284 |
| 8/14/1998 | 7320 | 7183.33 | 7100 | 7188.33 | 7302.5 |
| 8/15/1998 | 7300 | 7300 | 7350 | 7300 | 7343 |
| 8/17/1998 | 7219 | 7300 | 7300 | 7300 | 7245.75 |
| 8/18/1998 | 7220 | 7183.33 | 7100 | 7100 | 7169.688 |
| 8/19/1998 | 7285 | 7183.33 | 7300 | 7300 | 7203.563 |
| 8/20/1998 | 7274 | 7183.33 | 7100 | 7188.33 | 7285 |
| 8/21/1998 | 7225 | 7183.33 | 7100 | 7100 | 7274 |
| 8/24/1998 | 6955 | 7183.33 | 7100 | 7100 | 7075.75 |
| 8/25/1998 | 6949 | 6850 | 6850 | 6850 | 6843.063 |
| 8/26/1998 | 6790 | 6850 | 6850 | 6850 | 6831.188 |
| 8/27/1998 | 6835 | 6775 | 6650 | 6775 | 6769.875 |
| 8/28/1998 | 6695 | 6850 | 6750 | 6750 | 6765.563 |
| 8/29/1998 | 6728 | 6750 | 6750 | 6750 | 6645.438 |
| 8/31/1998 | 6566 | 6775 | 6650 | 6650 | 6582.125 |
| 9/1/1998 | 6409 | 6450 | 6450 | 6450 | 6471.5 |

| | | | | | |
|---|---|---|---|---|---|
| 9/2/1998 | 6430 | 6450 | 6550 | 6550 | 6377.5 |
| 9/3/1998 | 6200 | 6450 | 6350 | 6350 | 6298.25 |
| 9/4/1998 | 6403.2 | 6450 | 6450 | 6450 | 6232.938 |
| 9/5/1998 | 6697.5 | 6450 | 6550 | 6550 | 6595.075 |
| 9/7/1998 | 6722.3 | 6750 | 6750 | 6750 | 6824.938 |
| 9/8/1998 | 6859.4 | 6775 | 6850 | 6850 | 6833.3 |
| 9/9/1998 | 6769.6 | 6850 | 6750 | 6750 | 6878.775 |
| 9/10/1998 | 6709.75 | 6775 | 6650 | 6650 | 6720.038 |
| 9/11/1998 | 6726.5 | 6775 | 6850 | 6775 | 6670.25 |
| 9/14/1998 | 6774.55 | 6775 | 6850 | 6775 | 6755.938 |
| 9/15/1998 | 6762 | 6775 | 6650 | 6775 | 6764.737 |
| 9/16/1998 | 6952.75 | 6775 | 6850 | 6850 | 6862.125 |
| 9/17/1998 | 6906 | 6850 | 6950 | 6850 | 6992.5 |
| 9/18/1998 | 6842 | 6850 | 6850 | 6850 | 6905.75 |
| 9/19/1998 | 7039 | 6850 | 6950 | 6850 | 6944.938 |
| 9/21/1998 | 6861 | 6850 | 6850 | 6850 | 7010.75 |
| 9/22/1998 | 6926 | 6850 | 6950 | 6850 | 6804.5 |
| 9/23/1998 | 6852 | 6850 | 6850 | 6850 | 6897.75 |
| 9/24/1998 | 6890 | 6850 | 6950 | 6850 | 6852 |
| 9/25/1998 | 6871 | 6850 | 6850 | 6850 | 6890 |
| 9/28/1998 | 6840 | 6850 | 6750 | 6750 | 6871 |
| 9/29/1998 | 6806 | 6850 | 6750 | 6850 | 6799.75 |
| 9/30/1998 | 6787 | 6850 | 6750 | 6750 | 6775.813 |
| **MSE** | | 9668.94 | 7856.5 | 5437.58 | 5252.28 |
| **AFER** | | 1.05% | 1.03% | 0.89% | 0.84% |

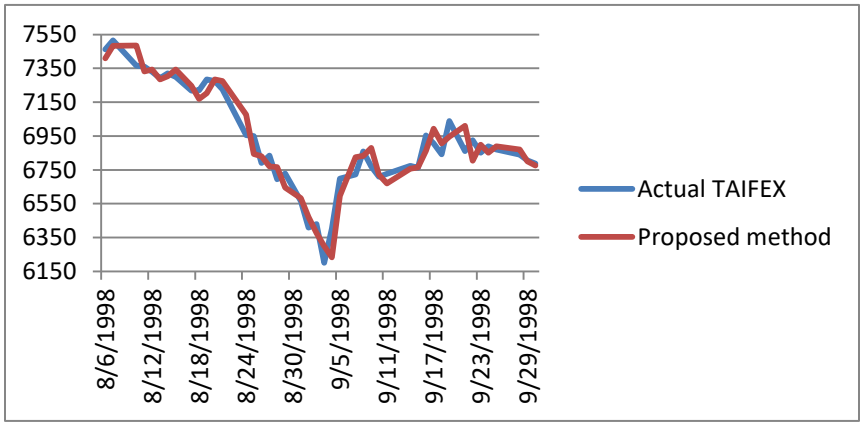



Fig. 3. Graph of actual TAIFEX index and forecasted values

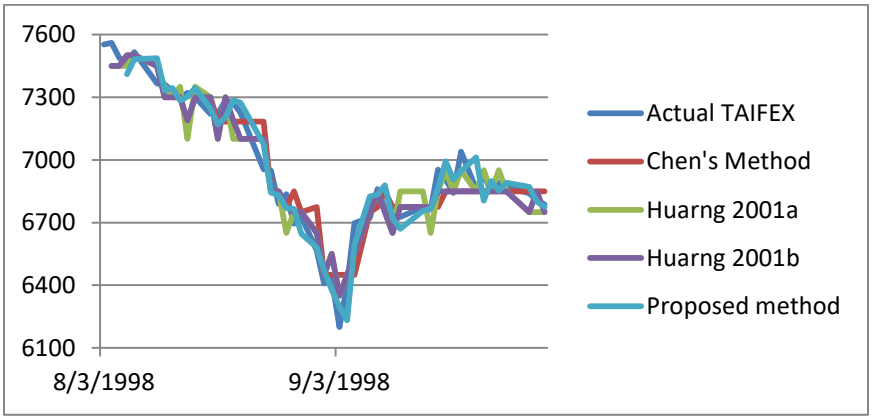



Fig. 4. Graph of forecasts from different models

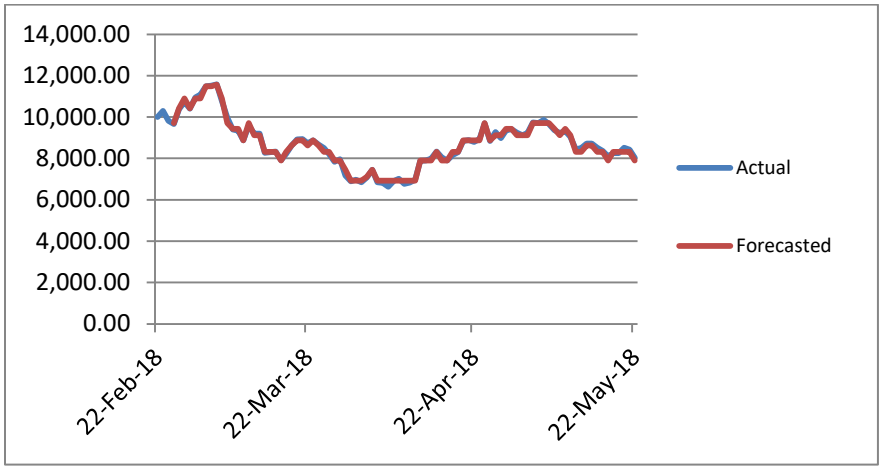



Fig. 5. Actual and forecasted Bitcoin time series

## VI. Conclusion

In this research we have formulated a novel approach for FTS forecasting using Quantum concepts and Evolutionary Algorithms. The Quantum Evolutionary Algorithm is used in combination with FTS forecasting and used to adjust interval lengths for better forecasting values. The unique algorithm was applied on TAIFEX index forecasting and Bitcoin price prediction. TAIFEX index forecasting produced AFER of 0.8478 while Bitcoin prediction produced AFER of 0.96%. AFER of the proposed model for TAIFEX forecasting is better than many of its predecessors. The proposed method is unique in the sense that Quantum computing along with Genetic Algorithms and Fuzzy Logic has never been developed before. The methods provide a new dimension for economic and financial modeling where optimization in various models can be improved with respect to accuracy and computational complexity.

In future, we aim to apply QEA to various algorithms of forecasting and portfolio optimization for improving model efficiency. Moreover, it is aimed at applying Quantum computing concepts on different nature inspired algorithms for designing new and powerful approaches to optimization techniques.

## Acknowledgment

This research is supported in part by the Higher Education Commission (HEC) of Pakistan under HEC Indigenous PhD Fellowship Program. We are thankful to Department of Computer Science, University of Karachi for providing computing facilities for the proposed research.

## References

[1] L.A. Zadeh, Fuzzy sets, Information and Control 8 (1965), pp. 338–353.

[2] [2] Lee L.-W., Wang L.-H., Chen S.-M., 2007. Temperature prediction and TAIFEX forecasting based on fuzzy logical relationships and genetic algorithms, Expert Systems with Applications 33, 2007, pp. 539-550.

[3] T.A. Jilani, S.M.A. Burney, Multivariate stochastic fuzzy forecasting models, Expert Systems with Applications 35 (2008a), pp. 691–700.

[4] T.A. Jilani, S.M.A. Burney, A refined fuzzy time series model for stock market forecasting, Physica-A 387 (2008b), pp. 2857-2862.

[5] T.A. Jilani, S.M.A. Burney, C. Ardil, Multivariate High Order Fuzzy Time Series Forecasting for Car Road Accidents, International Journal of Computational Intelligence 4 (1), (2008c), pp. 15-20.

[6] T.A. Jilani, S.M.A. Burney, C. Ardil, A New Quantile Based Fuzzy Time Series Forecasting Model, International Journal of Intelligent Systems and Technologies 3 (4), (2008d), pp. 201-207.

[7] Y.-L. Huang, S.-J. Horng, M. He, P. Fan, T.-W. Kao, M. K. Khan, J.-L. Lai, I-H. Kuo, A hybrid forecasting model for enrollments based on aggregated fuzzy time series and particle swarm optimization, Expert Systems with Applications 7(38), (2011), pp. 8014–8023.

[8] I-H. Kuo, S.-J. Horng, Y.-H. Chen, R.-S. Run, T.-W. Kao, R.-J. C., J.-L. Lai, T.-L. Lin, Forecasting TAIFEX based on fuzzy time series and particle swarm optimization, Expert Systems with Applications 2(37), (2010), pp. 1494–1502

[9] T.A. Jilani, U. Amjad, N. Mastorakis, A Hybrid Genetic Algorithm and Particle Swarm Optimization based Fuzzy Times Series Model for TAIFEX and KSE-100 Forecasting, Recent Researches in Applied Information Science, ISBN: 978-1-61804-089-3, (2012), pp. 212-218.

[10] U. Amjad, T.A. Jilani, F. Yasmeen, A Two Phase Algorithm for Fuzzy Time Series Forecasting using Genetic Algorithm and Particle Swarm Optimization Techniques, International Journal of Computer Applications (0975 – 8887) 16(55), (2012), pp. 34-40.

[11] P. Benioff, "The computer as a physical system: A microscopic quantum mechanical Hamiltonian model of computers as represented by Turing machines," J. Statist. Phys., vol. 22, pp. 563–591, 1980.

[12] D. Deutsch, "Quantum theory, the Church–Turing principle and the universal quantum computer," Proc. Roy. Soc. London, vol. A 400, pp. 97–117, 1985.

[13] P. W. Shor, "Quantum computing," Doc. Mathematica, vol. Extra Volume ICM, pp. 467–486, [Online]. Available: http://east.camel.math.ca/ EMIS/ journals/ DMJDMV/ xvol-icm/ 00/Shor.MAN.html, 1998.

[14] "Algorithms for quantum computation: Discrete logarithms and factoring," in Proc. 35th Annu. Symp. Foundations of Computer Science. Piscataway, NJ: IEEE Press, Nov. 1994, pp. 124–134.

[15] L. K. Grover, "A fast quantum mechanical algorithm for database search," in Proc. 28th ACM Symp. Theory of Computing, 1996, pp. 212–219.

[16] "Quantum mechanical searching," in Proc. 1999 Congress on Evolutionary Computation. Piscataway, NJ: IEEE Press, July 1999, vol. 3, pp. 2255–2261.

[17] A. Narayanan and M. Moore, "Quantum-inspired genetic algorithms," in Proc. IEEE Int. Conf. Evolutionary Computation. Piscataway, NJ: IEEE Press, May 1996, pp. 61–66.

[18] A. Narayanan, "Quantum computing for beginners," in Proc. 1999 Congress on Evolutionary Computation. Piscataway, NJ: IEEE Press, July 1999, vol. 3, pp. 2231–2238.

[19] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," bitcoin.org, 2008. [Online]. Available: https://bitcoin.org/bitcoin.pdf.

[20] G. Garcia,, and F. Schweitzer, "Social signals and algorithmic trading of Bitcoin," Royal Society Open Science, vol. 2, pp. 5408-5418, Sep, 2015. doi: 10.1098/rsos.150288

[21] M. Amjad and D. Shah, "Trading Bitcoin and Online Time Series Prediction," Proceedings of the Time Series Workshop at NIPS 2016, PMLR, vol. 55, pp. 1–15, 2017.

[22] D. Kondor, M. Posfai, I. Casabai, and G. Vattay, "Do the Rich Get Richer? An empirical analysis of the Bitcoin transaction network," PLOS ONE, vol. 9, no. 2, pp. 1–10, March, 2014.

[23] Y.B. Kim, J.G. Kim, W. Kim, J. H. Im, and C.H. Kim, "Predicting fluctuations in cryptocurrency transactions based on user comments and replies," PLOS ONE, vol. 11, no. 8, e0161197, Aug, 2016. doi.org/10.1371/journal.pone.0161197

[24] T. Guo, and N. Antulov-Fantulin, "Predicting short-term Bitcoin price fluctuations fom buy and sell orders," Feb, 2018.

[25] H. Jang, and J. Lee, "An Empirical Study on Modeling and Prediction of Bitcoin Prices with Bayesian neural networks based on Blockchain information," IEEE Access, vol. 6, pp. 5427- 5437, Feb 2018. doi: 10.1109/ACCESS.2017.2779181

[26] T. Hey, "Quantum computing: An introduction," in Computing & Control Engineering Journal. Piscataway, NJ: IEEE Press, June 1999, vol. 10, no. 3, pp. 105–112.

[27] L. A. Zadeh, "Fuzzy sets," Information and Control, vol. 8, 1996, pp. 338-353.

[28] Q. Song, B. S. Chissom, Fuzzy time series and its models, Fuzzy Sets and Systems, Vol. 54, pp. 269-277, 1993.

[29] S. M. Chen, "Forecasting enrollments based on fuzzy time series," Fuzzy Sets and Systems, vol. 81, 1996, pp. 311-319

[30] K. Huarng, "Heuristic models of fuzzy time series for forecasting," Fuzzy Sets and Systems, vol. 123, issue 3, pp. 369-386, 2001a

[31] K. Huarng, "Effective lengths of intervals to improve forecasting in fuzzy time series," Fuzzy Sets and Systems, vol. 123, issue 3, pp. 387-394, 2001b

# Detection of Railroad Networks in SAR Images

Şafak Altay Açar[1]
Department of Computer Technologies
Yenice Vocational School, Karabük University
Karabük, Turkey

Şafak Bayır[2]
Department of Computer Engineering
Faculty of Engineering, Karabük University
Karabük, Turkey

***Abstract*—In this study, a railroad networks detection method for synthetic aperture radar (SAR) images is proposed. Proposed method consists of three steps. Firstly, railroad segments are detected. An existing line detector is modified by describing some rules for this process. Then segments are connected by utilizing perceptual grouping. Finally, a new line analysis algorithm is applied to determine real parts of railroad networks. A software is developed to achieve and evaluate proposed method. Completeness and correctness values which are obtained after different steps are computed to evaluate proposed method. Two different TerraSAR-X images are used in experiments and obtained results are discussed in detail.**

***Keywords*—Remote sensing; synthetic aperture radar; railroad networks detection; perceptual grouping**

## I. Introduction

Synthetic aperture radar (SAR) is a system which can remote sensing and imaging. SAR could be set on satellites and aerial vehicles. SAR works effectively all day and night, regardless of all weather conditions [1]. These properties increase usability of SAR in civilian and military studies. Majority of passenger and freight transportations is done by roads and railroads. Also, location information of roads and railroads has strategic importance when a sensitive case occurs such as a civilian security. Because of these situations, studies which are about road and railroad detection have high importance but railroads are not taken into consideration as much as roads in studies.

There are many academic studies which extract roads from SAR images [2-13] but there are a few studies which determine railroads using SAR's qualities. Hellwich and Streck [14] present a study which searches visibility of linear structures like roads, rivers and railroads in SAR coherence data. Franceschetti et al. [15] propose an approach which extracts features from SAR raw data. They apply the proposed method to detect railroads. Pigeon et al. [16] present a research which extracts roads, railroads, rivers and energy transmission lines from SAR image. They use rules which are improved by human experts. Okman et al. [17] propose an approach for despeckling SAR images and present a semi-automated railroad detection algorithm to evaluate the performance of proposed despeckling method. A study is presented for polarimetric SAR image classification by Liu et al. [18]. They classify the terrain areas as road, railroad etc. Jin et al. [19] develop a constant false alarm line detector for polarimetric SAR images. They use Wilks' test statistics to develop the detector. The detector determines bright and dark structures such as railroads and roads. In experimental study, they detect buildings and roads.

In addition to SAR imaging studies, many studies are made on railroad determination using various remote sensing and aerial images. Niu et al. [20] present an algorithm to extract linear features from remote sensing images. This algorithm which is based on beamlet transform and wedgelet decomposition is used for railroad detection. Pacifici et. al. [21] present a neural network study for urban area classification. They classify urban land-use as road, railroad, soil, building, tower, tree etc. using extracted textural features from satellite images. Beger et al. [22] propose an approach for railroad track and centre line detection from aerial images and airborne laser scanning data. They use a feature level data fusion to increase accuracy and completeness. Karaman et al. [23] present a fourier based study for railroad extraction. They develop a fourier based feature descriptor to detect railroad objects in aerial images.

Furthermore, there are various studies which use perceptual grouping. Some of them are about road extraction from SAR images [5,8,11]. Others use perceptual grouping in different areas. Hirogaki et al. [24] present a study which carries out complete three dimensional retrieval of dental cast shape. They use perceptual grouping to reconstruct whole shape. Qi et al. [25] propose an approach which draws sketch automatically. A single natural image is utilized for this process. Perceptual grouping is used to create contour segment groups in this approach. Wang et al. [26] present a research which extracts buildings from optical satellite images by utilizing distinctive image primitives. They use a graph search-based perceptual grouping to connect detected line segments. Qi et al. [27] propose a perceptual grouping framework that uses multiple Gestalt principles. They study to increase usefulness of image edges on various researches. A study is presented for segmentation of three dimensional outdoor scenes by Xu et al. [28]. They use hierarchical clustering structure and perceptual grouping laws.

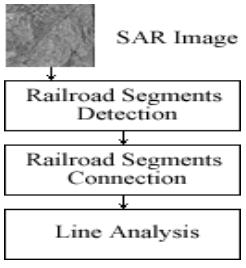



Fig. 1. Workflow of method.

This work was supported by the Scientific Research Coordination Unit of Ankara Yıldırım Beyazıt University as a preliminary research project - 631.

When the previous studies that we can obtain are taken into consideration, it is seen that there are not many studies on railroad detection on SAR images and also the studies are not carried out especially to detect railroads. In this study, which differs from the previous studies, railroad networks detection is achieved and evaluated individually. Improved method is fully automatic and extracts railroads from SAR images directly. We use cross-correlation line detector [2] to detect railroad segments. Then segments are connected by utilizing perceptual grouping. Finally, line analysis is applied to delete non-railroad pixels. Workflow of method is shown in Fig. 1.

The rest of the paper is organized as follows: Section II presents railroad segments detection. In section III, railroad segments connection is explained. Section IV presents line analysis process. In Sections V and VI, experimental results and conclusions are evaluated.

## II. Railroad Segments Detection

In SAR images, railroads appear as bright structures. Furthermore, railroads are narrower than roads. According to these differences, detector model which is proposed in [2] is redesigned for railroad. These model is shown in Fig. 2. There are three regions in this model. Centre region is brighter than adjacent regions. Also, widths of regions are different. Some rules which are related to this model are described. Variables which are used in these rules are similar to the ones used in [29] for roads. Cross-correlation line detector [2] is performed for only pixels which provide rules. By means of these rules, areas which have high probability of being only railroads are investigated. Therefore, more accurate results are obtained and the time is saved. Table I reports rules. In Table I, p(x,y) denotes related pixel and $\mu_1$, $\mu_2$ and $\mu_3$ denote mean values of regions.

Firstly, 3x3 Gaussian filter is applied to smooth SAR image. Then, according to detector model, railroad segments are detected. Different region sizes are used for the model. There are two kinds of situations for detector model. Table II reports them. In Table II, $\ell$ denotes length of regions and $w_1$, $w_2$, $w_3$ denote width of regions.

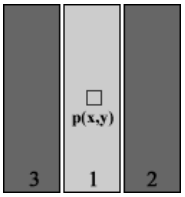


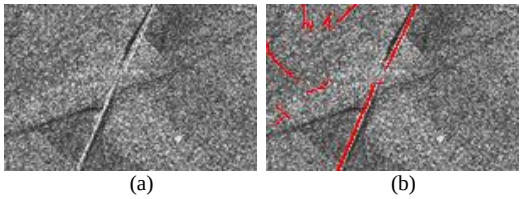

Fig. 2. Model of Railroad Detector.

TABLE I. Described Rules

| |
|---|
| colour ofp(x,y) > 100 |
| $\mu_1 > 100$ |
| $\mu_1 - \mu_2 > 5$ |
| $\mu_1 - \mu_3 > 5$ |

TABLE II. Size of Regions

| $\ell$(pixel) | $w_3$(pixel) | $w_1$(pixel) | $w_2$(pixel) |
|---|---|---|---|
| 11 | 2 | 1 | 2 |
| 11 | 2 | 2 | 2 |



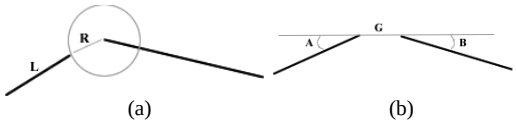
(a) (b)

Fig. 3. Railroad Segments Detection Sample (a) Original Image (B) Detected Railroad Segments.

Detection process is applied for each situation, then results are combined. Eight different directions are tested and the best one is accepted as a result. Decision threshold is defined as 0.4 in the experiments. A railroad segments detection sample is shown in Fig. 3. Red regions denote detected railroad segments.

## III. Railroad Segments Connection

Detected railroad segments are connected by utilizing perceptual grouping so that railroad networks are generated. Proximity and cocurvilinearity which are denoted in [5] are used as perceptual grouping factor. They are shown in Fig. 4.




(a) (b)

Fig. 4. Perceptual Grouping Factors (A) Proximity (B) Cocurvilinearity [5].

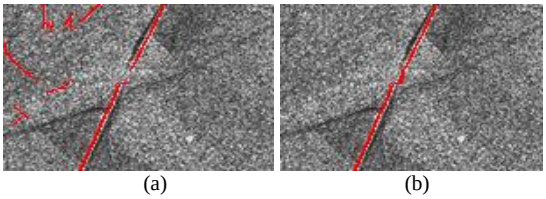
Proximity is formulated by (1). In the equation, L is the minimum length of two segments and R is the minimum distance between two segments at their endpoints. Cocurvilinearity is formulated by (2). In the equation, A and B are tangent angles of segments at joined endpoints. Coefficients $\alpha$ and $\beta$ control the deviation from collinearity and the sensitivity of G respectively [5]. In the experimental studies, we use them as $\alpha$=10 and $\beta$=0.1. Equations (1) and (2) are same as the ones used in [5,8,11].

$$P = \frac{L^2}{2\pi R^2} \tag{1}$$

$$C = \frac{1}{(A^2 + B^2) + (\alpha + \beta G)} \tag{2}$$

Before connection process, segments whose sizes are equal or less than 20 pixels are deleted in order not to waste time with unnecessary segments. Then, segments which are longer than predefined threshold ($th_c$) are determined. In the experiments, two different $th_c$ values (30 and 100) are used and results are evaluated separately. Regions having 41x41 pixels which are around relevant segment's endpoints are searched. If there is any segment and tangent angles (A and B) which are less than 50°, then the value V defined in (3) is calculated. Segments which have the highest V values are grouped with each other. P and C values are multiplied by weighting factors as utilized in the studies [5, 8]. In (3), $\gamma$ and $\theta$ are weighting factors. They are set to 0.5 and $10^3$ respectively in the experiments.

$$V = \gamma P + \theta C \tag{3}$$



(a) (b)

Fig. 5. Railroad Segments Connection Sample (A) Detected Railroad Segments (B) Image which is Obtained after Connection Process.

After connection process, segments whose sizes are equal or less than 100 pixels are deleted. A railroad segments connection sample is shown in Fig. 5. In Fig. 5(a), red regions denote detected railroad segments. Image which is obtained after connection process is shown in Fig. 5(b).

## IV. Line Analysis

A line analysis is applied to determine real parts of railroad networks. A new algorithm is proposed for this process.Adjacent pixels of relevant pixel which is a part of a detected railroad segment are searched. Kernel which is improved for searching is shown in Fig. 6. It consists of three regions as seen in Fig. 6. Red curve denotes a detected railroad segment and $p_r(x,y)$ denotes pixel which is a part of detected railroad segment.

Kernel is placed on the $p_r(x,y)$ for different directions and each time, numbers of detected pixels as railroad which are intersect with centre region and adjacent regions are calculated. If the total number of pixels which intersect with centre region are higher than predefined threshold ($th_1$) and if the total number of pixels which intersect with adjacent regions are less than predefined threshold ($th_2$), $p_r(x,y)$ remains as a part of segment. Ifthese conditions are not provided for any directions, $p_r(x,y)$ is eliminated from the segment.

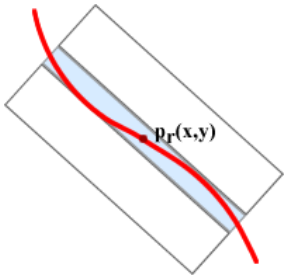
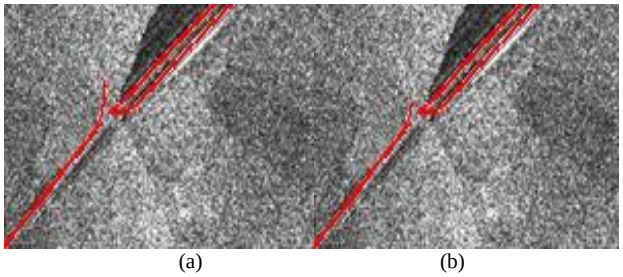



Fig. 6. Line Analysis Kernel.

**Initialization:** $th_1$=25, $th_2$=300, result=null
**for** i=1 **to** number of $p_r(x,y)$
**for** j=1 **to** number of kernel's directions
Calculate t1= total numbers of detected pixels as railroad which intersect with centre region of kernel
Calculate t2= total numbers of detected pixels as railroad which intersect with adjacent regions of kernel
**if** (t1> $th_1$ & t2< $th_2$) result=true **end if**
**end for**
**if** (result=true) remain $p_r(x,y)$ **else** eliminate $p_r(x,y)$ **end if**
**end for**

Fig. 7. Line Analysis Algorithm for Railroads.



(a) (b)

Fig. 8. Line Analysis Sample (A) Connected Railroad Segments (B) Image which is Obtained after Line Analysis Process.

This process is applied for all pixels which are parts of detected segments. In the experiments, $th_1$ and $th_2$ are used as 25 and 300 respectively. These thresholds are determined by experimental studies. Length of kernel, width of adjacent regions and width of centre region are 91, 11 and 2 pixels respectively. Width of centre region is determined according to maximum $w_1$ of detector model which is explained in section II.

Line analysis algorithm for railroads is shown in Fig. 7 and a line analysis sample is shown in Fig. 8. In Fig. 8(a), red regions denote connected railroad segments. Image which is obtained after line analysis process is shown in Fig. 8(b).

## V. Experimental Results

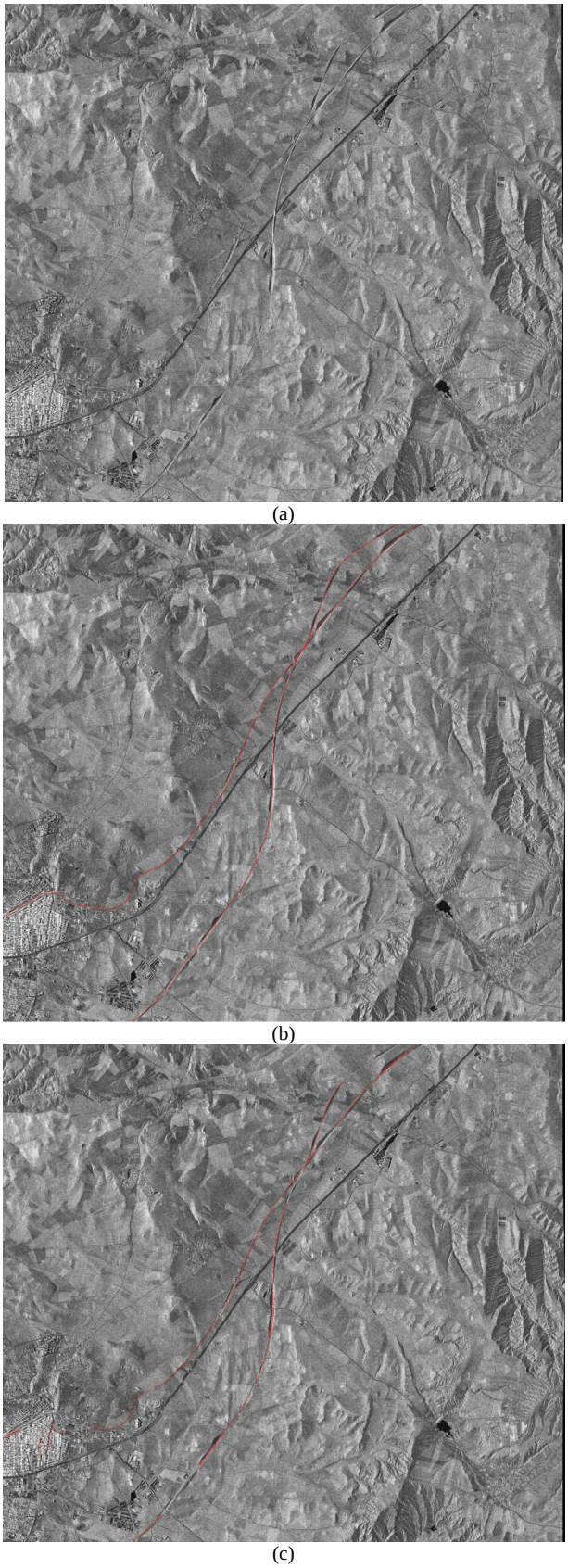
A software was developed for realizing and testing our method. We applied method to two different SAR images (spotlight mode, multi look ground range, HH polarization and up to 2m resolution). First image was acquired by TerraSAR-X over Polatlı (Ankara, Turkey). Second image was acquired by TerraSAR-X over Karaman (Turkey). These regions are selected since both of them include long railroads. Each of them covers a rural region of 10 km x 10 km. Images are resized, in other words, reduced in the ratio of 1/6 (totally 1/36) so first and second image become 2576 x 2299 pixels and 2553 x 2328 pixels respectively.

Sample results of images' railroad networks detection are shown in Fig. 9 and Fig. 10. In Fig. 9(b) and Fig. 10(b), red regions denote reference railroads. In Fig. 9(c) and Fig. 10(c), red regions denote detected railroads.

We compare real reference railroads with detected railroads pixel by pixel. Used formulas to calculate completeness and correctness values are defined in (4) and (5) respectively. These formulas are similar to the ones described in [30]. Reference pixels are determined by manually. A reference pixel is taken as matched reference pixel if there is a detected pixel in 3x3 pixels around and a detected pixel is taken as matched detected pixel if there is a reference pixel in 3x3 pixels around.

$$comp. = \frac{number\_of\_matched\_reference\_pixels \times 100}{number\_of\_reference\_pixels} \quad (4)$$

$$corr. = \frac{number\_of\_matched\_detected\_pixels \times 100}{number\_of\_detected\_pixels} \quad (5)$$

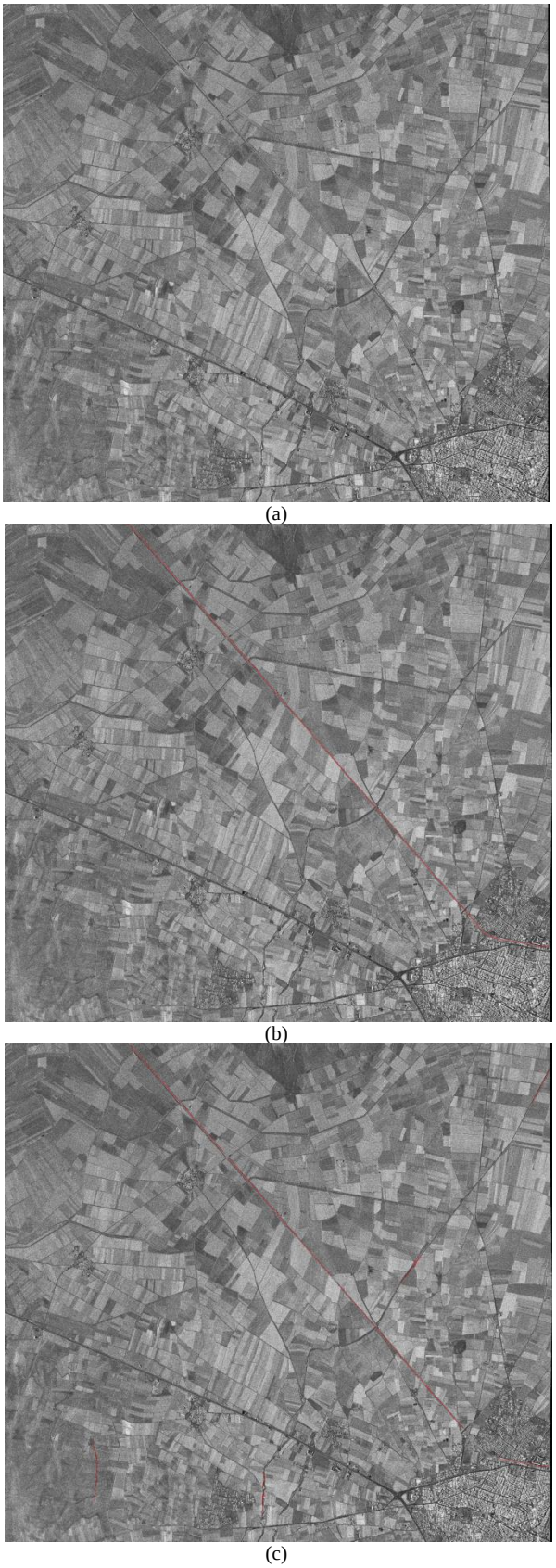


(a)

(b)

(c)

Fig. 9. First Image's Results (A) Original Image (B) Reference Railroads (C) Railroad Networks Detection ($th_c$=30 Completeness Value: 68.26 Correctness Value: 76.11).



(a)

(b)

(c)

Fig. 10. Second Image's Results (A) Original Image (B) Reference Railroads (C) Railroad Networks Detection ($th_c$=30 Completeness Value: 88.84 Correctness Value: 74.07).

TABLE III. EVALUATION OF RAILROAD NETWORKS DETECTION FOR THE FIRST IMAGE ($TH_C$=30)

| δ | After Connection | | After Line Analysis | | After Cleaning | |
|---|---|---|---|---|---|---|
| | Comp. | Corr. | Comp. | Corr. | Comp. | Corr. |
| 100 | 79.79 | 30.22 | 79.20 | 34.32 | 70.94 | 72.87 |
| 300 | 69.96 | 70.89 | 69.48 | 72.31 | 68.26 | 76.11 |
| 500 | 58.64 | 73.17 | 58.11 | 74.20 | 55.94 | 78.16 |

TABLE IV. EVALUATION OF RAILROAD NETWORKS DETECTION FOR THE SECOND IMAGE ($TH_C$=30)

| δ | After Connection | | After Line Analysis | | After Cleaning | |
|---|---|---|---|---|---|---|
| | Comp. | Corr. | Comp. | Corr. | Comp. | Corr. |
| 100 | 94.02 | 22.24 | 93.59 | 25.06 | 93.59 | 58.16 |
| 300 | 89.26 | 61.40 | 88.84 | 65.43 | 88.84 | 74.07 |
| 500 | 71.35 | 93.99 | 71.04 | 95.13 | 71.04 | 97.23 |

Evaluations of railroad networks detection are given in Tables III, IV, V and VI. After segment detection and connection processes, segments which have small sizes are deleted. The threshold relevant to deleted segments sizes is denoted as δ in tables. After connection process, line analysis and cleaning are applied. In cleaning process, segments whose sizes are equal or less than 200 pixels are deleted. Two different $th_c$ values (30 and 100) are used and results are evaluated separately. The results of $th_c$=30 are given in Tables III and IV and the results of $th_c$=100 are given in Tables V and VI.

Results of each processes for the first image are given in Tables III and V. Results of each processes for the second image are given in Tables IV and VI.

According to results, assessments occur as follows:

- When δ value increases, completeness decreases but correctness increases.
- After line analysis, completeness decreases but correctness increases.
- For the first image, after cleaning, completeness decreases but correctness increases. When δ is 100 pixel, these decrement and increment are more than other δ values.
- For the second image, after cleaning, completeness does not change but correctness increases. When δ is 100 pixel, this increment is more than other δ values.

TABLE V. EVALUATION OF RAILROAD NETWORKS DETECTION FOR THE FIRST IMAGE ($TH_C$=100)

| δ | After Connection | | After Line Analysis | | After Cleaning | |
|---|---|---|---|---|---|---|
| | Comp. | Corr. | Comp. | Corr. | Comp. | Corr. |
| 100 | 66.92 | 65.09 | 66.15 | 67.12 | 55.66 | 91.02 |
| 300 | 59.70 | 86.62 | 58.99 | 87.50 | 55.66 | 91.07 |
| 500 | 47.40 | 85.24 | 46.82 | 86.60 | 46.51 | 90.67 |

TABLE VI. EVALUATION OF RAILROAD NETWORKS DETECTION FOR THE SECOND IMAGE ($TH_C$=100)

| δ | After Connection | | After Line Analysis | | After Cleaning | |
|---|---|---|---|---|---|---|
| | Comp. | Corr. | Comp. | Corr. | Comp. | Corr. |
| 100 | 86.07 | 51.64 | 85.42 | 54.53 | 80.49 | 80.29 |
| 300 | 80.91 | 87.72 | 80.49 | 88.42 | 80.49 | 88.76 |
| 500 | 80.91 | 97.64 | 80.49 | 97.89 | 80.49 | 97.89 |

According to results, assessments occur as follows:

- When δ value increases, completeness decreases or does not change but correctness usually increases.
- After line analysis, completeness decreases but correctness increases.
- For the first image, after cleaning, completeness decreases but correctness increases. When δ is 100 pixel, these decrement and increment are more than other δ values.
- For the second image, after cleaning, completeness usually does not change but correctness increases or does not change. When δ is 100 pixel, this increment is more than other δ values.

When we compare the results obtained when $th_c$ value is equal to 100 with the results obtained when $th_c$ value is equal to 30, it is seen that the completeness value is generally lower, on the other hand, the correctness value is higher.

## VI. CONCLUSION

This study presents an efficient railroad networks detection method for SAR images. The method is tested by utilizing two SAR images. Completeness and correctness values are calculated for each one and obtained results are evaluated in detail. Proposed method can be modified to detect any continuous curvilinear structures.

A useful method will be developed in order to apply to reduce false detections so that correctness value will increases by not affecting completeness value. Furthermore, proposed method will be improved to work by using less constants.

## REFERENCES

[1] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek, and K. P. Papathanassiou, "A tutorial on synthetic aperture radar," IEEE Geosci. Remote Sensing Mag., vol. 1, no. 1, pp. 6–43, Mar. 2013.

[2] F. Tupin, H. Maitre, J. F. Mangin, J. M. Nicolas, and E. Pechersky, "Detection of linear features in SAR images: application to road network extraction," IEEE Trans. Geosci. Remote Sens., vol. 36, no. 2, pp. 434-453, Mar. 1998.

[3] J. Chanussot, G. Mauris, and P. Lambert, "Fuzzy fusion techniques for linear features detection in multitemporal SAR images," IEEE Trans. Geosci. Remote Sens., vol. 37, no. 3, pp. 1292-1305, May 1999.

[4] F. Dell'Acqua and P. Gamba, "Detection of urban structures in SAR images by robust fuzzy clustering algorithms: the example of street tracking," IEEE Trans. Geosci. Remote Sens., vol. 39, no. 10, pp. 2287-2297, Oct. 2001.

[5] B. Jeon, J. Jang, and K. Hong, "Road detection in spaceborne SAR images using a genetic algorithm," IEEE Trans. Geosci. Remote Sens., vol. 40, no. 1, pp. 22-29, Jan. 2002.

[6] P. Gamba, F. Dell'Acqua, and G. Lisini, "Improving urban road extraction in high-resolution images exploiting directional filtering,

perceptual grouping, and simple topological concepts," IEEE Geosci. Remote Sens. Lett., vol. 3, no. 3, pp. 387-391, Jul. 2006.

[7] D. Borghys, Y. Yvinec, C. Perneel, A. Pizurica, and W. Philips, "Supervised feature-based classification of multi-channel SAR images," Pattern Recognition Lett., vol. 27, no. 4, pp. 252-258, Mar. 2006.

[8] Y. Guo, Z. Bai, Y. Li, and Y. Liu., "Genetic algorithm and region growing based road detection in SAR images," in Proc. Third Int. Conference on Natural Computation, Haikou, 2007, pp. 330-334.

[9] H. Chaabouni-Chouayakh and M. Datcu, "Coarse-to-fine approach for urban area interpretation using TerraSAR-X data," IEEE Geosci. Remote Sens. Lett., vol. 7, no. 1, pp. 78-82, Jan. 2010.

[10] J. Cheng, W. Ding, X. Ku, and J. Sun, "Road extraction from high-resolution SAR images via automatic local detecting and human-guided global tracking," Int. Journal of Antennas and Propagation, Nov. 2012.

[11] C. He, Z. Liao, F. Yang, X. Deng, and M. Liao, "Road extraction from SAR imagery based on multiscale geometric analysis of detector responses," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 5, no. 5, pp. 1373-1382, Oct. 2012.

[12] M. Saati, J. Amini, and M. Maboudi, "A method for automatic road extraction of high resolution SAR imagery," Journal of the Indian Society of Remote Sensing, vol. 43, no. 4, pp. 697-707, Dec. 2015.

[13] M. Jiang, Z. Miao, P. Gamba, and B. Yong, "Application of multitemporal inSAR coveriance and information fusion to robust road extraction," IEEE Trans. Geosci. Remote Sens., vol. 55, no. 6, pp. 3611-3622, Jun. 2017.

[14] O. Hellwich and C. Streck, "Linear structures in SAR coherence data," in Proc. Int. Geosci. and Remote Sens. Symposium, Lincoln, USA, 1996, vol. 1, pp. 330-332.

[15] G. Franceschettia, A. Iodice, and M. Tesauroa, "From image processing to feature processing," Signal Processing, vol. 60, no. 1, pp. 51-63, Jul. 1997.

[16] L. Pigeon, B. Solaiman, K. P. B. Thomson, B. Moulin, and T. Toutin, "Human-experts rules modeling for linear planimetric features extraction in a remotely sensed images data fusion context," in Proc. IEEE Int. Geosci. and Remote Sens. Symposium, Hamburg, Germany, 1999, vol. 5, pp. 2369 - 2371.

[17] O. E. Okman, F. Nar, and C. Demirkesen, "Feature preserving sar despeckling and its parallel implementation with application to railway detection," in Proc. 9th European Conference on Synthetic Aperture Radar, Nuremberg, Germany, 2012, pp. 151-154.

[18] F. Liu, J. Shia, L. Jiao, H. Liu, S. Yang, J. Wu, H. Hao, and J. Yuan, "Hierarchical semantic model and scattering mechanism based PolSAR image classification," Pattern Recognition, vol. 59, pp. 325-342, Nov. 2016.

[19] R. Jin, W. Zhou, J. Yin, and J. Yang, "CFAR line detector for polarimetric SAR images using Wilks' test statistic," IEEE Geosci. Remote Sens. Lett., vol. 13, no. 5, pp. 711-715, May. 2016.

[20] R. Niu, X. Mei, L. Zhang, and P. Li, "Linear features extraction from remote sensing image based on wedgelet decomposition," in Proc. Fourth Int. Conference on Image and Graphics, Sichuan, China, 2007, pp. 508 - 512.

[21] F. Pacifici, M. Chini, and W. J. Emery, "A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification," Remote Sensing of Environment, vol. 113, no. 6, pp. 1276–1292, Jun. 2009.

[22] R. Beger, C. Gedrange, R. Hect, and M. Neubert, "Data fusion of extremely high resolution aerial imagery and LiDAR data for automated railroad centre line reconstruction," ISPRS Journal of Photogrammetry and Remote Sens., vol. 66, no. 6, pp. S40–S51, Dec. 2011.

[23] E. Karaman, U. Çinar, E. Gedik, Y. Yardımcı, and U. Halıcı, "Fourier based feature descriptors for railroad extraction from aerial images," in Proc. IEEE Int. Geosci. and Remote Sens. Symposium, Munich, Germany, 2012, pp. 6013 - 6015.

[24] Y. Hirogaki, T. Sohmura, H. Satoh, J. Takahashi, and K. Takata, "Complete 3-D reconstruction of dental cast shape using perceptual grouping," IEEE Trans. Medical Imaging, vol. 20, no. 10, pp. 1093-1101, Oct. 2001.

[25] Y. Qi, J. Guo, Y. Li, H. Zhang, T. Xiang, and Y. Song, "Sketching by perceptual grouping," in Proc. 20th IEEE Int. Conference on Image Processing, Melbourne, Australia, 2013, pp. 270-274.

[26] J. Wang, X. Yang, X. Qin, X. Ye, and Q. Qin, "An efficient approach for automatic rectangular building extraction from very high resolution optical satellite imagery," IEEE Geosci. Remote Sens. Lett., vol. 12, no. 3, pp. 487-491, Mar. 2015.

[27] Y. Qi, Y. Song, T. Xiang, H. Zhang, T. Hospedales, Y. Li, and J. Guo, "Making better use of edges via perceptual grouping," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015, pp. 1856-1865.

[28] Y. Xu, S. Tuttas, and U. Stilla, "Segmentation of 3D outdoor scenes using hierarchical clustering structure and perceptual grouping laws," in Proc. 9th IAPR Workshop on Pattern Recognition in Remote Sensing, Cancun, Mexico, 2016.

[29] Ş. Altay Açar and Ş. Bayır, "Effect of threshold values used for road segments detection in SAR images on road network generation," International Journal of Advanced Computer Science and Applications, vol. 8, no. 5, pp.287-291, May. 2017.

[30] C. Heipke, H. Mayer, C. Wiedemann, and O. Jamet, "Evaluation of automatic road extraction," Int. Archives of Photogrammetry and Remote Sens., vol. 32, pp. 47-56, 1997.

# Connectivity Restoration Techniques for Wireless Sensor and Actor Network (WSAN), A Review

Muhammad Kashif Saeed[*, #, 1], Mahmood ul Hassan[*, 2], Ansar Munir Shah[*, 3], Khalid Mahmood[*, #, 4]
[*]Department of Computer Science, IIC University of Technology, Phnom Penh, The Kingdom of Cambodia
[#]Department of Information Systems, King Khalid University, The Kingdom of Saudi Arabia

***Abstract*—Wireless Sensor and actor networks (WSANs) are the most promising research area in the field of wireless communication. It consists of large number of small independent sensor and powerful actor nodes equipped with communication and computation capabilities. Actors gather sensor's data and react collaboratively to attain application particular assignments. A powerful connected inter-actor network is required to coordinate its operations. Actor node may fail due to the battery depletion or any hardware failure and this failure may divide the network into disjoint segments. This problem can degrade the network performance but also reduce the efficiency and effectiveness of the network. To restore the network into its original state, the researchers have proposed many connectivity restoration techniques during last few years. This paper provides a brief review of the existing connectivity restoration techniques for WSANs with their advantages and limitations.**

***Keywords*—Wireless sensor networks; wireless sensor and actor networks; node failure; network partitioning; connectivity restoration; node movements**

## I. Introduction

Wireless Sensor networks (WSNs) have gained a substantial attention in recent years. The wide spectrum of WSNs applications has open many research areas and it is being considered the most substantial modern technologies of recent times. As a result, huge research work has been done in different areas of WSNs. Wireless Sensor and Actor Network (WSAN) is a new paradigm that has reflected a significant impact on recent wireless sensor technologies. WSAN consists of sensor and actor nodes. Sensor and actor nodes in WSAN can be static or mobile. Sensors have limited power and energy resources and are responsible for sensing the important data from an area of interest and transmit it to the actor / base station. Actors are more powerful and high performance nodes furnished with more energy and computation resources, and work in close collaboration with the sensors to perform the fruitful tasks. Figure 1 shows the communication design of WSAN, in which sensors sense data from the environment and send to actors. Processing of data can be performed by actors or may send to base station for further processing.

Sensor and actor nodes can be failed due to battery depletion, hardware failure, or any external attack. As a result, network gets partitioned into disjoint segments, resulting in performance degradation [1]. Therefore, connectivity restoration process is needed to be performed. Human intervention is very less in such applications and there is no observer, who observe the network and takes timely decision. Moreover, it is not an easy task to replace the faulty nodes with the new nodes, especially in a harsh environment like the battlefield or dense forest. So a comprehensive mechanism is required which replaces the failed node with any of its neighboring. Collaboration among the healthy actor nodes is required to restore the connectivity at the pre-failure level. Node's mobility can improve the overall performance of the network like connectivity, coverage and network lifetime [2]. These recovery processes cause messaging overhead as well. Normally, these algorithms deal with single actor node failure, and do not consider the efficiency of the resources and also lack of emphasis on recovery time.

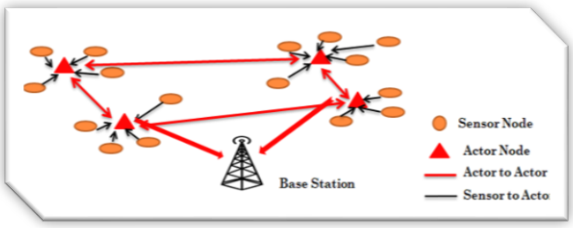



Fig. 1. WSAN Design

The main objective of the review paper is to provide a brief review of the connectivity restoration techniques for wireless sensor and actor network. Moreover, this review paper also provides a comparison of different connectivity restoration techniques for WSAN with their objectives, advantages and limitations.

Rest of the paper is organized as follows. Section II provides the overview on need of study. Section III contains the background of the study, whereas Section IV consists of the review of the existing techniques and finally Section V concludes the review paper.

## II. Need of Study

Wireless Sensor and Actor Network (WSAN) is implemented in different civilian and industrial applications. It has an exceptional architecture which differentiates it from the ordinary Wireless Sensor Network (WSN). Sensor Nodes have limited battery and power resources and are responsible for sensing and transmitting the sensed data to the actor nodes. Whereas, actors are more powerful and high performance nodes having high power and computational resources. Actor nodes have the ability to gather process and, send the collective and aggregated data to the base station. Actor nodes play a key role in WSAN. Therefore, special attention should be given to the actor nodes in order to increase the performance of the

network. Although, WSAN has a lot of benefits, but it is not free from obstacles and challenges. Communication of sensor to sensor, sensor to an actor, and actor to actor is very important. An efficient use and management of actor nodes can improve the overall performance of the network. The failure of an actor node causes disjoint segments in the network and sensed data could not reach to the base station. Therefore, connectivity restoration is the key problem to be addresses and solved.

## III. BACKGROUND

In Wireless Sensor and Actor Networks (WSANs), sensor and actor nodes work together to perform certain tasks. Actor network is incorporated with the sensor network to make the wireless sensor and actor networks [3]. WSANs can be affected due to the change of environment, any change in event detection, an actor mobility or failure of the actor due to the depletion of energy, attacks or any communication link issues. Failure of actors can divide the network into disjoint segments, and affect the whole network. Inter-actor nodes connectivity is significant [4]. Due to the powerful characteristics of actor nodes, it can be managed through the relocation of mobile nodes but the actor failure can damage the network more than the ordinary sensor nodes. It can also affect the loss of coordination and connectivity between the nodes and restrict the event handling. For example, an actor node may be destroyed due to the catastrophic damage or enemy attack which disconnects the network and needs to be addressed instantaneously. Deployment of some relay nodes in the area can be a solution to replace the faulty nodes but it is not feasible in risky areas like war zones. Therefore, the early recovery process can be initiated through involving existing actor and sensor nodes by self-restoration techniques. Nodes repositioning methods have been introduced by number of researchers for restoring partitioned networks.

A typical WSAN architecture can be categorized into semi-automated and an automated architecture. These architectures are based on data passing and decision making. In an automated wireless sensor actor network, sensor nodes sense events and then send facts to their associated actor nodes which act as the base station. In semi-automated WSANs, sensor nodes sense facts from the field, and transfer to the sink node where sink node process the data, and communicate with actors to perform the necessary tasks, if needed. So coordination is very important in WSANs.

Fault tolerance techniques work in distributed manners. Fault tolerant techniques enable a system to perform its operations properly after single or multiple node failure. Figure 2 shows the node types in fault management techniques. Fault tolerant techniques are divided into critical and non-critical which solve the problem using 1-hop, 2-hop or multi-hop nodes information. Fault tolerance detection mechanism can be classified as proactive and reactive or hybrid. In proactive method, fault and restoration processes are addressed in the network setup in which redundant and backup nodes are deployed to ensure fault tolerance [5]. In reactive techniques, it utilizes the in house resources of the network and performs the recovery process dynamically through repositioning of the nodes. It requires monitoring system to check the nodes status and recovery scope. Reactive techniques are divided into distributed or centralized and are discussed comprehensively in [6-8]. Moreover, some techniques require a single node whereas some require a block of nodes to be moved to restore the connectivity of the network. Detailed taxonomy of such techniques is shown in the Figure 3. Some of these techniques will be discussed later in the paper.

The impact of a sensor/ an actor node failure can be different, according to the node type and its importance in the network. Fault management detection algorithms and restoration procedures can be categorized into critical and non-critical nodes [9-11]. The failure of a critical node divides the network into disjoint segments. Most of the techniques describe the critical nodes by using 2-hop messages whereas some have the 1-hop message interchange to check the critical sensor/ actor nodes. A study was conducted using 1-hop to identify the critical actor nodes [12]. To identify the critical node, the proposed technique calculates the distance of actor nodes from their adjacent nodes. If the distance is lesser than the neighbor's communication range, an actor will be considered as a non-critical node, otherwise critical.

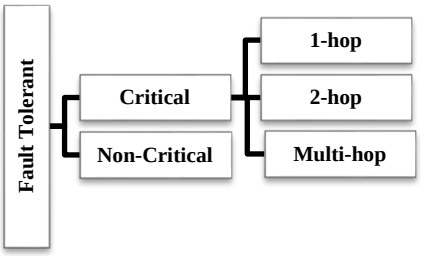



Fig. 2. Fault Management

Coordination is the important factor in communication like sensor to sensor, sensor to actor and actor to actor. Network efficiency can be improved by communication between actors to actor. In WSAN, a sensor and actor node can be static or mobile. During the network lifetime, nodes move, so topology management is equally important. Nodes failure can divide the network into small segments and create the coverage hole in the sensing area as well. Topology management techniques can be performed automatically for fault management.

## IV. REVIEW OF EXISTING TECHNIQUES

Actor node failure can divide the network of WSAN into segments. An actor may fail due to the fault in hardware, energy depletion, physical attacks or any communication link issues. Although, there are less chances of actor failure than sensor failure, but it can be controlled through relocation of some mobile nodes. Connectivity between the nodes and coordination will be lost in case of an actor failure, and leads to disjoint of the wireless sensor actor network. Fault tolerance is an ability of the network to do its work smoothly in response to node failure [13].

WSAN are mostly deployed in tough areas like the battlefield, dense forest or massive destruction areas and suppose to do work for the maximum period of time. Such networks are normally deployed in far areas from the main control Centre, so connectivity restoration in an efficient way is

a quite difficult process. Connectivity restoration process should be a distributed, self-healing and localized. Moreover, this process should be so fast that it reduces the impact of node failures, reduce overhead such that distance travelled, a number of messages while using limited energy supply. If a node travels too much to restore connectivity, then it will consume more energy and may also affect another network disconnection, especially if it is a cut-vertex. Moreover, it is difficult to find cut-vertex in large scale WSAN in a centralized and timely manner. So, connectivity restoration is a very challenging task in a distributed, localized, and an efficient manner. Reactive restoration techniques act passively and are initiated when a node failure occurs. No redundant resources are required in this case. Cooperative communication was introduced [14] which allowed a node to send a message away from its communication radius via its neighbor's. Two nodes can communicate with each other only if the received average Signal to Noise Ratio (SNR) is not less than the threshold. Strength of the signal reduces when we increase the transmission. Collaborative Single Node Failure Restoration algorithm (CSFR) [15] uses the cooperative communication approach to restore connectivity. It has low overhead but still has a long term process which consumes more energy and time. Therefore, it is not a suitable technique. In most of the reactive techniques, it reconnects a network by replacing a failure node with an appropriate backup node. It will be a recursive process that may relocate other nodes as well. It is mentioned earlier that only a cut-vertex node may break the network connectivity. A lot of techniques are available to detect whether the node is a cut vertex or not and then treat with its failure as well like NNN, DARA, PDARA, and PCR. In the improved version of PADRA, it forms a connecting dominating set (CDS). PADRA notifies in advance to a particular node about the partition occurrence in case of failure occurred.

Distributed Actor Recovery Algorithm (DARA) [16] finds a cut vertex by using two-hop neighborhood information. When a failure occurs, neighbors of the faulty node will select the most proper backup node. It will consider node's degree, distance and inform of its sibling nodes. Detailed process of identifying a cut vertex is not given in DARA technique in details. It is improved in PDARA in which it forms a connected dominating set (CDS). PDARA informs a particular node in advance whether a partition has occurred in case of failure. Nearest Non-Critical Neighbor (NNN) [17] decides whether it is cut-vertex or not and whether it preserves 2-hop neighbor's information of nodes. Distributed Partitioning Detection and Connectivity Restoration (DCR) algorithm [18] finds critical and non-critical actors in advance on the basis of information available locally and designates non-critical neighbor actors as backup actors. Once the failure of node occurs, the backup actor starts a recovery process which may include a coordinated method for relocation of multi actors. Recovery Algorithm for Multiple Nodes Failure (RAM) [18] handles two adjacent nodes failure simultaneously. It is a distributed hybrid technique which finds critical actors and assigns its backup nodes as well. Year-wise detailed comparison of different techniques is shown in Table 1. It covers techniques presented during the year 2007-2018 along with their objectives, node type, movement type, and node failure type. It shows that techniques are centralized or distributed and which technique has a capacity to restore connectivity of normal or cut-vertex node. This table also covers node movements like direct, cascade or block and also tells about the information that which technique has a capacity to restore connectivity of single, double or multi-node failure.

Recovery through inward Motion (RIM) [19] is a distributed technique which restores connectivity of critical and ordinary node. It maintains 1-hop information of each node; identifies nodes failure and starts recovery process by moving neighboring in cascade movement. All the one-hop neighboring nodes move towards the failed node till the distance is '' $R_C/2$''. It ignores the impact of coverage after restoring connectivity and does not differentiate between the positions of nodes. RIM is very simple and an efficient technique, but its performance reduces in the dense networks and messages overhead is be very high.

Node Recovery through Active Spare designation (NORAS) [20] is a recovery algorithm which can identify critical nodes and finds the backup nodes for such critical nodes before the failure occurs. After that, these backup nodes participate to restore connectivity. It also takes care of coverage and connectivity in an integrated manner. NORAS stores 2-hop information which causes the large communication overhead in the network. Moreover, the movement of nodes is increased as compared to other algorithms.

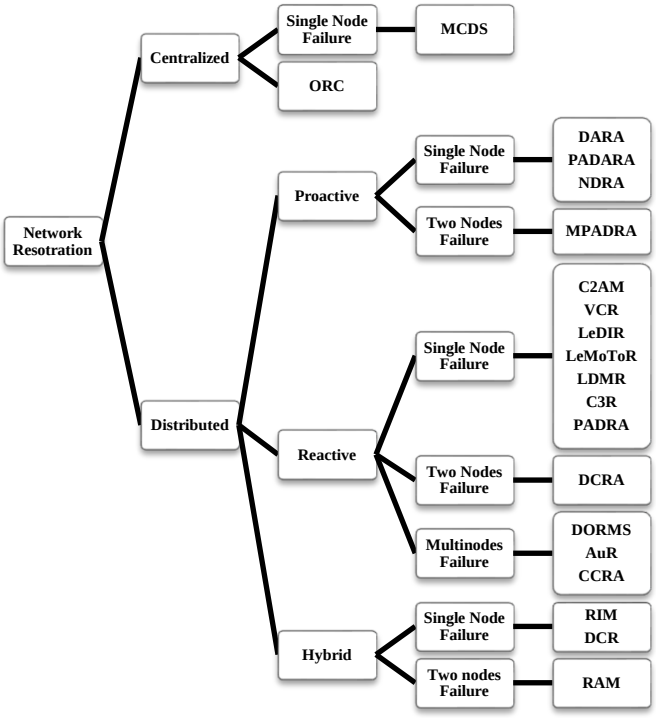



Fig. 3. Detailed taxonomy of techniques

$C^3R$-Coverage Conscious Connectivity Restoration [21] algorithm deals with the restoration of connectivity and coverage simultaneously. In this technique, failed node is replaced with the nearest node temporarily, and then it returns

to its position. A neighboring node comes forward to replace it. It deals only with a single node failure, and does not support simultaneous nodes failure. Unnecessary movements of the nodes are reduced by introducing Energy Centric Optimized Algorithm (ECR) which consumes a lot of energy. It performs better than RIM and NN because it localizes the failure recovery and introduces few changes in the network topology. Frequent back and forth movements of the nodes are happening to restore connectivity and coverage which is not energy efficient. It has adverse effects on the network. Moreover, some extra nodes are required in the network, which increases the cost also.

Least Movement Topology Repair Algorithm (LeMoToR) [23] is a localized and distributed technique which solves the network partitioning problem with less number of nodes movements and messaging overhead. It utilizes the path discovery activities at the time of connectivity restoration to identify the topology structure and takes suitable action accordingly. Faulty node is replaced with the neighboring node from the smallest disconnected block. This algorithm uses the recursive process to find the best route of recovery in which a lot of computation is required.

TABLE I. COMPARISON OF CONNECTIVITY RESTORATION TECHNIQUES

| Sr | Name of Technique | Year | Objective | Centralized/ Distributed | Movement Type | Node Type | Node Mobility | Node Failure Type |
|---|---|---|---|---|---|---|---|---|
| 1 | $C^2AP$ | 2007 | Connectivity | Distributed | Cascade | Any node | Mobile Sensor/ Actor | Single node |
| 2 | DARA | 2007 | Connectivity | Distributed | Cascade | Cut-Vertex | Mobile Sensor | Single node |
| 3 | NN | 2008 | Connectivity | Distributed | Direct move | Any node | Mobile Sensor | Single node |
| 4 | $C^2AM$ | 2009 | Connectivity | Distributed | Cascade | Any node | Mobile Sensor | Single node |
| 5 | RIM | 2010 | Connectivity | Distributed | Cascade | Anyone | Mobile Sensor | Single node |
| 6 | $C^3R$ | 2010 | Connectivity and Coverage | Distributed | To and fro movement | Any node | Mobile Sensor | Single node |
| 7 | NORAS | 2010 | Connectivity and Coverage | Distributed | Cascade | Cut-Vertex | Mobile Sensor | Single node |
| 8 | DORMS | 2010 | Connectivity with partial coverage | Distributed | Cascade | Any node | Mobile Sensor | Multi nodes |
| 9 | PADRA | 2010 | Connectivity | Distributed | Cascade | Cut-Vertex | Mobile Node/ Robot | Single node |
| 10 | MPADRA | 2010 | Connectivity | Distributed | Cascade | Cut-Vertex | Mobile Node/ Robot | Multi nodes |
| 11 | LeMoToR | 2011 | Connectivity | Distributed | Cascade | Cut-Vertex | Mobile Sensor/ Actor | Single node |
| 12 | DCRA | 2011 | Connectivity | Distributed | Block movement | Cut-Vertex | Mobile Sensor | Two nodes |
| 13 | DCR | 2012 | Connectivity | Distributed | Cascade | Cut Vertex | Mobile Sensor/ Actor | Single node |
| 14 | PCR | 2012 | Connectivity | Distributed | Cascaded or shifted | Cut Vertex | Mobile Sensor/ Actor | Single node |
| 15 | RAM | 2012 | Connectivity | Distributed | Cascaded or shifted | Cut Vertex | Mobile Sensor/ Actor | Two nodes |
| 16 | AuR | 2012 | Connectivity with partial coverage | Distributed | Cascade | Any node | Mobile Sensor | Multi nodes |
| 17 | LeDIR | 2013 | Connectivity | Distributed | Block movement | Cut-Vertex | Mobile Sensor/ Actor | Single Node |
| 18 | NNN | 2013 | Connectivity | Distributed | Cascade | Cut-Vertex | Mobile Sensor/ Actor | Single Node |
| 19 | DPCRA | 2014 | Connectivity | Distributed | Cascade | Cut-Vertex | Mobile Sensor/ Actor | Single Node |
| 20 | CC-IC | 2016 | Connectivity and Coverage | Distributed | Cascade | Any node | Mobile Sensor | - |
| 21 | CSFR-M | 2016 | Connectivity | Distributed | Cascade | Any node | Mobile Sensor | Single node |
| 22 | CCRA | 2016 | Connectivity | Distributed | Cascade | Cut-Vertex | Mobile Sensor | Multiple nodes |
| 23 | SFR-RNR | 2017 | Connectivity with partial coverage | Distributed | - | - | Mobile Sensor | - |
| 24 | HCR | 2017 | Connectivity | Distributed | Cascade | Cut-Vertex | Mobile Sensor/ Actor | Single node |
| 25 | EAR | 2017 | Connectivity | Distributed | Cascade | Cut-Vertex | Mobile Sensor/ Actor | Single node |
| 26 | DEENR[22] | 2018 | Connectivity with partial coverage | Distributed | Cascade | Any node | - | Single node |
| 27 | DCRMF | 2018 | Connectivity | Distributed | Cascade | Cut-Vertex | Mobile Sensor/ Actor | Single/ multi nodes |
| 28 | PRACAR | 2018 | Connectivity and Coverage | Distributed | Cascade | Any node | Mobile Sensor/ Actor | Single node |

There is no clear approach used to find the smallest block, either through depth first or greedy approach among the disconnected blocks. Moreover, communication overheads will be high while finding the smallest block.

Least Disruptive Topology Repair (LeDiR) [24] is another localized and distributed algorithm which can detect cut-vertex and execute recovery from node failure using path discovery and routing information. Neighboring nodes of the faulty node will recomputed their routing tables and develop enrolment decisions for the recovery process. Each node calculates the shortest path to other node and updates its information in the routing table. After the node failure, its one-hop neighbor's will check if the failed node is critical or not. Neighbor from the smallest block will move to replace the critical node. If more than one neighboring nodes are a part of the smallest block, then the nearest node from the faulty actor node will be selected to cope block movement. It consumes more energy because all the nodes of the block will participate in the recovery process. Moreover, smallest block calculation performed at the time of recovery which is another drawback of this algorithm.

Distributed Prioritized Connectivity Restoration Algorithm (DPCRA) [25] is used to restore connectivity and partitions by using a few numbers of nodes only. It identifies the negative effects of actors on partitions. The recovery process did locally while storing limited information in every node. Numerous backup nodes are used for the partition recovery in the network. This algorithm failed to address proper backup node selection criteria and as a result, there are more chances of failure of nodes. It can also affect the overall network performance and energy consumption as well.

Advanced-self-healing Connectivity Recovery Algorithm (ACRA) [26] defines the nature of the actor node that whether the failed node is cut-vertex or the connectivity of the nodes by depth first search technique. It will restore the connectivity of cut vertex node in which an actor node with more transmission power and high coverage area is selected to take part in the recovery process. Sensors and actor nodes are deployed randomly and form clusters. Every node has a system to detect failure of a normal node and also cut vertex actor nodes. When a cut vertex node fails, a neighboring cluster head (CHs) sends a recovery message to all the nearest nodes, towards the sink node, till it finds the next actor node or CH. A stable sensor CHs is selected as per GA based criteria among the neighboring nodes as a joining router for connecting partitioned network. This algorithm uses more energy because of the cluster heads. Senor nodes are involved in the recovery process on the basis of sensor resources, and there are more chances of nodes failure. Details of some important connectivity restoration techniques for wireless sensor and actor networks are presented in Table 2. Main objectives of some of vital techniques and their limitations summary are shown in Table 2. Limitations found in earlier algorithms have been addressed in the later techniques, whereas most of the techniques need researcher's attention to resolve these issues.

Hybrid Connectivity Restoration (HCR) algorithm [27] works proactively during selection process and reactively in motion phase. An actor selects the backup node through its 1-hop neighboring table and notifies the backup node to oversee this process. When a node fails, its backup node try to move to that position in order to restore connectivity. It is a localized process which is repeated until connectivity is restored. HCR select the backup node which has to travel short. Moreover, HCR tries to reduce the number of messages by forwarding node failure information to its backup node only. It is an effective scheme with low complexity. It deals single node failure at a time, handles sequential nodes failure only and doesn't handle the coverage issues.

Efficient actor recovery paradigm (EAR) [28] is a recovery technique which can differentiate between critical and non-critical nodes. It allocates an appropriate backup node from its neighbor which is chosen on the basis of its signal strength, and control in its surroundings. It is supported by three algorithms. Node Monitoring and Critical Node Detection (NMCND) algorithm which monitors nodes and tells about the nodes type, and also handles the packet forwarding process if the primary node fails. Network integration and Message Forwarding (NIMF) is introduced to send packet. Process Based Routing for Node Failure Avoiding Algorithm (PRNFA) was developed to handle the routing process in which redundant messages was reduced to avoided network congestion. The main goal of this algorithm is to improve the node recovery process while maintaining the Quality of Services (QoS).

Distributed autonomous connectivity restoration method based on the finite state machine (DCRMF) [29] is a technique which looks for critical nodes in the region and defines how to reposition the related nodes. It performs restoration autonomously. Critical nodes updating process is launched after every restoration process. It relocates the non-dominating nodes from the neighboring nodes, so the total moving distance is reduced. It can effectively reduce the movement overheads of the nodes in connectivity restoration process. A distributed localized connectivity restoration algorithm is introduced to handle the multiple nodes failure problem. A few nodes will be relocated with less moving distances.

Permanent Relocation Algorithm for Centralized Actor Recovery (PRACAR) [30] replaces the failed actor by one of its neighbor permanently. It is a self-route recovery algorithm which finds the optimum path to the sensors. Actor nodes are relocated permanently to a new location. This permanent placement of the redundant node at the position of the faulty node will stop the extra movements. This is an energy efficient technique which saves energy to do other key tasks of the network.

TABLE II. ACTOR FAILURE TECHNIQUES OBJECTIVES AND LIMITATIONS

| Sr | Name of Technique | Objectives | Limitations |
|---|---|---|---|
| 1 | DARA | Single node recovery with short total travel distance | • Do not handle actor failure.<br>• Do not provide a mechanism to detect cut-vertices nodes. |
| 2 | $C^2AM$ | Reduce total distance and minimize the message overhead | • Do not care coverage<br>• Not suitable for mission critical |
| 3 | PADRA/ ACR | • Localize the scope of recovery<br>• Reduce the message overhead and the total distance | • Nearest distance to the failed node which affect the overall network |
| 4 | RIM | • Minimize the total distance with fewer messages overhead. | • Don't considered coverage<br>• A lot of nodes moving<br>• Not suitable for multimode failure |
| 5 | DCR | • Minimize the scope of recovery, and total distance | • It manages the single failure at a time and doesn't consider energy. |
| 6 | PCR | • It uses a localized algorithm to recognize critical actors and designate backup nodes. | • It's focused on the sensor nodes and does not address the node failure. |
| 7 | NNN | • Total distance and cascade relocation overhead reduced | • It may involve the path increments among the nodes.<br>• Algorithm executes recursively which increase overhead. |
| 8 | LeDIR | • It minimizes the scope using path length validation | • It does not handle multiple nodes failure. |
| 9 | DPCRA | • Restore connectivity by using small number of nodes. | • Fails to manage backup node selection criteria, so there are more chances of nodes failure. |
| 10 | ACRA | • It is a clustered based algorithm. Actor node having high transmission power and coverage area is selected to participate in connectivity restoration. | • Actor uses higher transmission power, and not energy efficient. |
| 11 | DCRMF | • Decrease the movement over heads of the sensor-actor nodes.<br>• Nearest non-critical node replaced with the abnormal node | • Do not take care of coverage |
| 12 | LeDIR/ RNF | • It does not impose pre-failure overhead.<br>• It can recover from a single node failure at a time only. | • It doesn't deal with multiple nodes failure. |

## V. CONCLUSION

Maintaining internode connectivity in an efficient manner is very important in applications of wireless sensor and actor networks. Failure of actor nodes may disjoint the network and QoS of the network may degrade. There are many algorithms so far presented to restore connectivity of actor nodes, which were briefly discussed above. Some techniques use large number of nodes movement, more distance travel to restore connectivity, which consume more energy during the restoration process. A comprehensive review on state of the art connectivity restoration techniques is provided, in which different limitations and drawbacks in the schemes are discussed which may be addressed and may be helpful for the researches to develop new connectivity restoration techniques in the light of these guidelines.

## REFERENCES

[1] Mahmood, K., Khan, M. A., Shah, A. M., Ali, S., & Saeed, M. K. (2018). Intelligent On-Demand Connectivity Restoration for Wireless Sensor Networks. Wireless Communications and Mobile Computing, 2018.

[2] Younis, Mohamed, and Kemal Akkaya. "Strategies and techniques for node placement in wireless sensor networks: A survey." Ad Hoc Networks 6.4 (2008): 621-655.

[3] Akyildiz, Ian F., and Ismail H. Kasimoglu. "Wireless sensor and actor networks: research challenges." Ad hoc networks 2.4 (2004): 351-367.

[4] Kashi, Saeed Sedighian, and Mohsen Sharifi. "Connectivity weakness impacts on coordination in wireless sensor and actor networks." IEEE communications surveys & tutorials15.1 (2013): 145-166.

[5] Imran, Muhammad, et al. "Volunteer-instigated connectivity restoration algorithm for wireless sensor and actor networks." Wireless Communications, Networking and Information Security (WCNIS), 2010 IEEE International Conference on. IEEE, 2010.

[6] Younis, Mohamed, et al. "Topology management techniques for tolerating node failures in wireless sensor networks: A survey." Computer Networks 58 (2014): 254-283.

[7] Younis, Mohamed, and Kemal Akkaya. "Strategies and techniques for node placement in wireless sensor networks: A survey." Ad Hoc Networks 6.4 (2008): 621-655.

[8] Imran, Muhammad, et al. "Application-centric connectivity restoration algorithm for wireless sensor and actor networks." International Conference on Grid and Pervasive Computing. Springer, Berlin, Heidelberg, 2011.

[9] Abbasi, Ameer Ahmed, Mohamed F. Younis, and Uthman A. Baroudi. "Recovering from a node failure in wireless sensor-actor networks with minimal topology changes." IEEE Transactions on vehicular technology 62.1 (2013): 256-271.

[10] Morita, Kiyohiro, et al. "Sensor-actuator communication protocols in wireless networks." International Conference on Network-Based Information Systems. Springer, Berlin, Heidelberg, 2007.

[11] Abbasi, Ameer, Mohamed Younis, and Uthman Baroudi. "Restoring connectivity in wireless sensor-actor networks with minimal topology changes." Communications (ICC), 2010 IEEE International Conference on. IEEE, 2010.

[12] Imran, Muhammad, et al. "Localized motion-based connectivity restoration algorithms for wireless sensor and actor networks." Journal of Network and Computer Applications 35.2 (2012): 844-856.

[13] Sabri, Naseer, et al. "Towards smart wireless sensor actor networks: design factors and applications." Industrial Electronics and Applications (ISIEA), 2011 IEEE Symposium on. IEEE, 2011.

[14] Topology control in ad hoc wireless networks using cooperative communication. . *IEEE Transactions on Mobile Computing*, 5(6), 711-724.

[15] Wang, Huaiyuan, et al. "Adaptive connectivity restoration from node failure (s) in wireless sensor networks." Sensors 16.10 (2016): 1487.

[16] Abbasi, Ameer Ahmed, Kemal Akkaya, and Mohamed Younis. "A distributed connectivity restoration algorithm in wireless sensor and actor networks." Local computer networks, 2007. LCN 2007. 32nd IEEE conference on. IEEE, 2007.

[17] Haider, Noman, et al. "Performance analysis of reactive connectivity restoration algorithms for wireless sensor and actor networks." Communications (MICC), 2013 IEEE Malaysia International Conference on. IEEE, 2013.

[18] Imran, Muhammad, et al. "Localized motion-based connectivity restoration algorithms for wireless sensor and actor networks." Journal of Network and Computer Applications 35.2 (2012): 844-856.

[19] Younis, Mohamed, Sookyoung Lee, and Ameer A. Abbasi. "A localized algorithm for restoring internode connectivity in networks of moveable sensors." IEEE Transactions on Computers 59.12 (2010): 1669-1682.

[20] Vaidya, Ketaki, and Mohamed Younis. "Efficient failure recovery in Wireless Sensor Networks through active, spare designation." Distributed Computing in Sensor Systems Workshops (DCOSSW), 2010 6th IEEE International Conference on. IEEE, 2010.

[21] Tamboli, Neelofer, and Mohamed Younis. "Coverage-aware connectivity restoration in mobile sensor networks." Journal of network and computer applications 33.4 (2010): 363-374.

[22] ul Hassan, Mahmood, Khan, M. A., Ali, S., Mahmood, K., & Shah, A. M. (2018). Distributed Energy Efficient Node Relocation Algorithm (DEENR). International Journal of Advanced Computer and Applications, 9(3), 95-100.

[23] Abbasi, Ameer, Mohamed Younis, and Uthman Baroudi. "Restoring connectivity in wireless sensor-actor networks with minimal topology changes." Communications (ICC), 2010 IEEE International Conference on. IEEE, 2010.

[24] Abbasi, Ameer Ahmed, Mohamed F. Younis, and Uthman A. Baroudi. "Recovering from a node failure in wireless sensor-actor networks with minimal topology changes." IEEE Transactions on vehicular technology 62.1 (2013): 256-271.

[25] Ranga, Virender, Mayank Dave, and Anil Kumar Verma. "A hybrid timer based single node failure recovery approach for WSANs." Wireless personal communications 77.3 (2014): 2155-2182.

[26] Ranga, Virender, Mayank Dave, and Anil Kumar Verma. "A hybrid timer based single node failure recovery approach for WSANs." Wireless personal communications 77.3 (2014): 2155-2182.

[27] Yan, Ke, et al. "Hybrid connectivity restoration in wireless sensor and actor networks." EURASIP Journal on Wireless Communications and Networking 2017.1 (2017): 138

[28] Mahjoub, Reem K., and Khaled Elleithy. "Efficient Actor Recovery Paradigm for Wireless Sensor and Actor Networks." Sensors 17.4 (2017): 858.

[29] Zhang, Ying, Jun Wang, and Guan Hao. "An Autonomous Connectivity Restoration Algorithm Based on Finite State Machine for Wireless Sensor-Actor Networks." Sensors 18.1 (2018): 153.

[30] Khalid Mahmood, M. A. ‘’Permanent Relocation and Self-Route Recovery in Wireless Sensor and Actor Networks’’. International Journal of Advanced Computer Science and Applications (IJACSA), (2018) Vol. 9, No. 3.

# Design of Wearable Patch Antenna for Wireless Body Area Networks

Saqib Hussain, Saima Hafeez, Sajjad Ali Memon, Nasrullah Pirzada
Department of Telecommunication Engineering Mehran University of Engineering & Technology Jamshoro, Pakistan

***Abstract*—Wireless body area networks are being widely used due to the increase in the use of wireless networks and various electrical devices. A Wearable Patch antenna is used for enhancement of various applications for WBAN. In this paper, a low profile wearable microstrip patch antenna is designed and suggested for constant observation of human vital signs such as blood pressure, pulse rate and body temperature using wireless body area network (WBAN) technology. The operating frequency of the antenna is taken as 2.45 GHz which lies in industrial, scientific and medical (ISM) frequency band. Polyester textile fabric with a relative permittivity of 1.44 and thickness of 2.85 mm is used as a substrate material. The proposed antenna is designed to achieve better return loss, VSWR, gain and low value of specific absorption rate (SAR) as compare to other existing wearable antenna. The achieved antenna return loss at 2.45 GHz is about -10.52 dB and gain of 7.81 dB. The VSWR value achieved at 2.45 GHz is 1.84, which is good in terms of good impedance matching. Other antenna field parameters like 2D and 3D gain, radiation pattern, and SAR value have been calculated. High-Frequency Structure Simulator (HFSS) is used to design and simulate the proposed antenna.**

***Keywords*—High-Frequency structure simulator (HFSS); return loss; voltage standing wave ratio (VSWR); gain; specific absorption rate (SAR)***

## I. Introduction

Wireless Body Area Networks are being particularly used for the various real-time health monitoring applications. These networks include the use of wearable antennas for transmitting and receiving of the data for healthcare related systems. An antenna that is integrated into the clothing of the wearer is called a wearable antenna. A wearable antenna can be used in a variety of applications such as GPS navigation, military, monitoring of athletes fitness, telemedicine, satellite communication, digital watches, and RFID [1]–[3]. Advancements in the field of wearable electronics have progressed rapidly in recent times and as a result, intensive research activities are being held over body conformal antennas. These days, some vital signs of the human body such as heart rate, blood glucose, blood pressure, and electrocardiogram (ECG) need to be monitored regularly due to their severe implications on the human health. Therefore, multiple sensors can be placed on the human body to monitor those vital signs of the human body [4]–[6]. The body-worn sensors store information regarding various physiological parameters and transmit them to the wearable devices, which further transmit them to the nearest receiving node. On-body communication term is used when a wearable antenna communicates with a wearable medical device [7]. Whereas, off-body communication is said to be a communication between the on-body bio-medical transceiver device to external wireless transceiver device [8].

For the development of wearable antennas, 2.45 GHz of industrial, scientific and medical (ISM) frequency band is employed due to its global availability. The wearable antenna must be hidden and low profile for the convenience of the user. This entails a thinkable amalgamation of the antenna elements within daily life outfit. Microstrip patch antenna can be an ideal choice for the wearable antenna applications [9].

There are various benefits of microstrip patch antenna as it is lightweight, compact, flexible and able to resist mechanical strain without affecting antenna performance significantly [10]. A wearable antenna can be used to communicate between sensors and the human body. Due to its small and compact size, a wearable antenna is utilized in different applications such as for military, medical, healthcare, emergency services, and navigation. In a military application, it is used to establish communication links between different soldiers including sending images and videos, location tracking and army protection. Furthermore, medical applications are used to check and monitor health parameters of a patient and can communicate with each other or with the outside world [11].

Various types of wearable sensors are mounted on or implanted into the human body in order to access human vital signs information such as body temperature, blood pressure, and heartbeat. The medical information is then sent through sensors to the receiver at low frequency. After the reception of medical information, all the data is collected from the sensors mounted on the human body and then, sent to an external device. The doctor in the hospital or any remote location can look after the situation of the patient and suggest medicines immediately to improve the quality of healthcare [12]-[13].

The formation of this paper is as follows. Section II discusses the literature review and other works related to wearable microstrip patch antenna. The antenna designing steps are explained in Section III, IV, V, VI, and VII, respectively. Section VIII discusses the results and compares them with the other works. Finally, some conclusions are drawn in Section IX and future work suggestions are presented at the end.

## II. Literature Review

Microstrip patch antenna has many advantages as it is lighter in weight, low cost, low profile than the conventional

microwave antenna. The planar structure of the antenna provides ease of fabrication [14]. Presently, it has been observed that the microstrip patch antenna become an ideal choice for the wearable healthcare applications. However, human body tissues can affect the performance and efficiency of the antenna, therefore, the selection of the material used to design such kind of antenna plays a significant role [15]. Also, the antenna performance and the radiation pattern are greatly influenced by the absorbed energy. The Specific Absorption Rate (SAR) is used to measure the amount of power absorbed by human body tissues. According to Federal Communication Commission (FCC), the SAR value should be below 1.6 W/Kg averaged over 1 gm of tissue and in European Standard, its value should be 2 W/Kg averaged over 10 gm of tissue [16]-[17].

An e-textile patch antenna was designed by the authors [18] for the frequency of 5.8 GHz using jeans as a substrate material. The measured return loss ($S_{11}$), gain and SAR values are -21 dB, 3.05 dB, and 0.0111 W/Kg respectively. In [19], the authors have presented the flexible antenna design for the purpose of telemedicine applications. The authors used 2 substrate materials to design wearable antenna i.e. cotton and jeans. The jeans material has shown good results in terms of gain over cotton which is 5 dB compared to a gain of 3 dB for cotton. In [20], the authors have proposed a wearable patch antenna design using FR-4 as a substrate material covered by the jeans fabric as the outer layer. The measured antenna parameters such as return loss ($S_{11}$) of -15.28 dB and a gain of 5.209 dBi at a frequency of 2.4 GHz. Similarly, in [21], the authors have designed and fabricated the wearable antenna on a flexible substrate material known as denim gens. This antenna can be used to operate in various frequency bands such as L, S, C, and X with both horizontal and vertical polarization.

## III. Methodology

The proposed antenna is designed and simulated using the High-Frequency Structure Simulator (HFSS). The suggested antenna is a microstrip patch antenna, therefore its design in HFSS needs some geometrical and simulation parameters. The geometrical parameters of the proposed antenna are calculated using microstrip equations discussed in Section IV. However, the major simulation parameter of an antenna is a frequency that can be defined according to the application of the proposed antenna [22]-[23]. After the computation of geometrical parameters, the design of the suggested antenna can be modeled in the HFSS. The general methodology to model a patch antenna in HFSS is explained in the flow chart given below in Fig. 1.

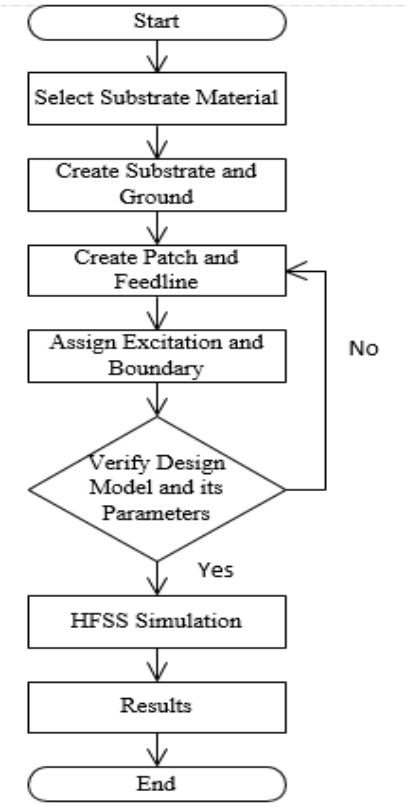



Fig. 1. The antenna simulation flowchart in HFSS

## IV. Design of Proposed Wearable Patch Antenna

Microstrip line inset feeding technique is utilized to design the proposed wearable patch antenna. The inset fed technique is used because it provides the planar structure to the antenna. The operating frequency of the antenna is taken as 2.45 GHz because it is unlicensed and can be used for a variety of applications. The patch antenna is fed by 50 Ω input impedance. This antenna is implemented on the polyester substrate material that has a low dielectric constant, which results in a reduction in the surface wave losses. The relative permittivity $\in_r$ of substrate material is 1.44, the thickness $h$ is 2.85 mm and the loss tangent $\tan\sigma$ is 0.01. The substrate dimension is $90\times90\ \text{mm}^2$. As microstrip inset fed technique is utilized to design the proposed antenna. Antenna geometrical parameters such as patch width $W_p$ and patch length $L_p$ have been computed using the following formulas mentioned below. [22].

$$W_p = \frac{c}{2 f_o}\left(\frac{2}{\in_r + 1}\right)^{\frac{1}{2}} \quad (1)$$

$$L_p = \frac{1}{2 f_o \sqrt{\in_{reff}\ \mu_o\ \in_o}} - 2\Delta L \quad (2)$$

Where $c$, $f_o$, $\in_r$, $\in_{reff}$, $\mu_o$ and $\Delta L$ is the speed of light, operating frequency, dielectric constant, effective dielectric constant, the permeability of the free space and extension length, respectively. Whereas, the extension length $\Delta L$ and the effective dielectric constant $\in_{reff}$ is computed by the subsequent equations [23].

$$\Delta L = 0.412h\left[\left(\frac{\in_{reff} + 0.3}{\in_{reff} - 0.258}\right)\left(\frac{\frac{W_p}{h} + 0.264}{\frac{W_p}{h} + 0.813}\right)\right] \quad (3)$$

$$\in_{reff} = \frac{\in_r + 1}{2} + \frac{\in_r - 1}{2}\left(1 + \frac{12h}{W_p}\right)^{\frac{-1}{2}} \quad (4)$$

After obtaining all the values of antenna geometrical parameters, the suggested wearable patch antenna is designed in antenna simulation software named as HFSS and polyester fabric is utilized as a substrate material whose loss tangent value is 1.44. The input impedance of the proposed antenna is characterized by 50 Ω.

## V. Substrate Material

The substrate material used to design wearable antenna is polyester. The main advantage of selecting the polyester material as a substrate is its flexibility. Furthermore, it is used in daily life fabric and readily available in the market. The material characteristics listed in TABLE I are as follows:

TABLE I. Substrate Material Design Values

| PARAMETERS | SYMBOL | VALUE |
|---|---|---|
| Dielectric constant | $\varepsilon_r$ | 1.44 |
| Loss tangent | tan σ | 0.01 |
| Thickness | h | 2.85 mm |

## VI. Dimensions of Patch Antenna

The dimensions of patch antenna play a pivotal role to make an effective antenna design in terms of efficient results. TABLE II illustrates the calculated parameters of the suggested patch antenna model.

TABLE II. Patch Antenna Design Values

| PARAMETERS | SYMBOL | VALUE |
|---|---|---|
| Operating Frequency | $f_o$ | 2.45 GHz |
| Patch Dimension Along x | $W_p$ | 55.43 mm |
| Patch Dimension Along y | $L_p$ | 47.9 mm |
| Substrate Thickness | $h$ | 2.85 mm |
| Substrate Dimension Along x | $W_s$ | 90 mm |
| Substrate Dimension Along y | $L_s$ | 90 mm |
| Inset Distance | $Y_o$ | 10 mm |
| Inset Gap | $G$ | 2 mm |
| Feed Width | $W_f$ | 3.3 mm |
| Feed Length | $L_f$ | 24 mm |
| Di electric constant of substrate | $\in_r$ | 1.44 |
| Input Impedance | $Z_o$ | 50 Ω |

## VII. Design of Human Phantom Model

A 3-layer human phantom model is created in HFSS for the calculation of specific absorption rate (SAR). The 3-layer human phantom model consists of 3 layers of human body tissues i.e. muscle, fat, and skin. The width of muscle, fat, and skin are 23 mm, 8 mm and 2 mm, respectively. TABLE III. illustrates the different values of human body tissues that are taken to create the human phantom model. In order to study SAR impact in the vicinity of the human body, the proposed wearable antenna is mounted on a 110×110 $mm^2$ body-phantom model. Fig. 2 illustrates the proposed antenna on human body phantom model.

TABLE III. Properties of Human Body Tissues

| Tissue | Permittivity ($\varepsilon_r$) | Conductivity (S/m) | Loss Tangent (tan σ) | Density ($Kg/m^3$) |
|---|---|---|---|---|
| Skin | 31.29 | 5.0138 | 0.2835 | 1100 |
| Fat | 5.28 | 0.1 | 0.19382 | 1100 |
| Muscle | 52.79 | 1.705 | 0.24191 | 1060 |

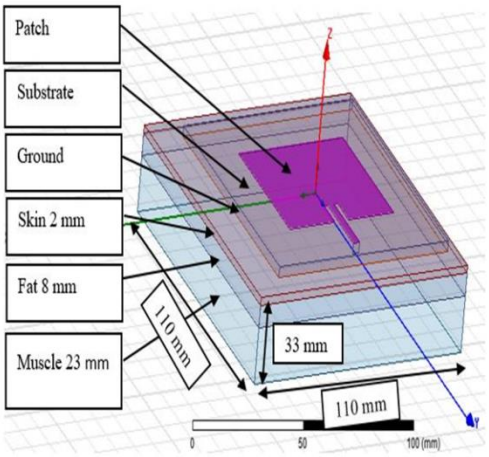



Fig. 2. Proposed Wearable Patch Antenna on Human Phantom Model

## VIII. Results and Discussion

The proposed wearable patch antenna is simulated in HFSS antenna simulation tool using simulation parameters tabulated in TABLE II. The antenna is characterized by 50 Ω input impedance. The scattering parameters ($S_{11}$) return loss of the simulated antenna is illustrated in Fig. 3. The minimum return loss curve value achieved at 2.45 GHz is -10.52 dB i.e. marked by $m_2$ and the achieved value is acceptable. Fig. 4 illustrates the 3D gain of the simulated antenna. At 2.45 GHz the peak gain of 7.81 dB has been obtained in the z-axis i.e. perpendicular to the antenna. Fig. 5 illustrates the 2-Dimensional gain of the antenna observed at 2.45 GHz. It can be seen that a peak gain of 7.81 dB i.e. marked by $m_1$ is achieved at 0°. This radiated power is analyzed in the far-field region of the antenna. The directivity of the proposed wearable patch antenna is illustrated in Fig. 6. The values obtained from the simulation are suitable for the health monitoring applications. As can be seen from Fig. 6 that the directivity is exactly perpendicular to the axis, which implies that the power is focused in only one lobe. Fig. 7 illustrates the radiation pattern of the antenna. The theta values are taken from 0° to 180°. It can be seen the major lobe of the antenna is radiating a majority of its power in the front direction, which concludes it has the high front to back ratio. In addition, it can be seen that the antenna has minimal minor lobes, which are very good for wearable applications.

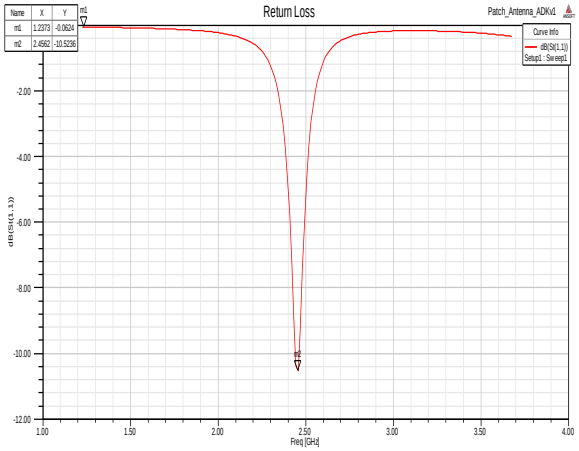



Fig. 3. Simulated Return Loss at 2.45 GHz

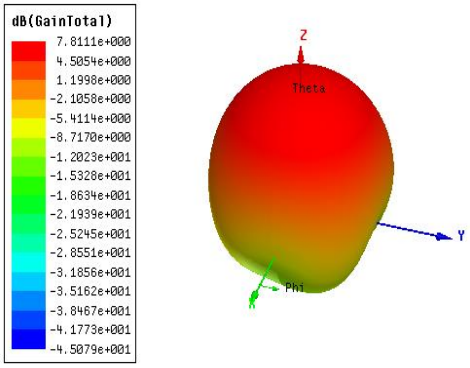



Fig. 4. Simulated 3D Gain at 2.45 GHz

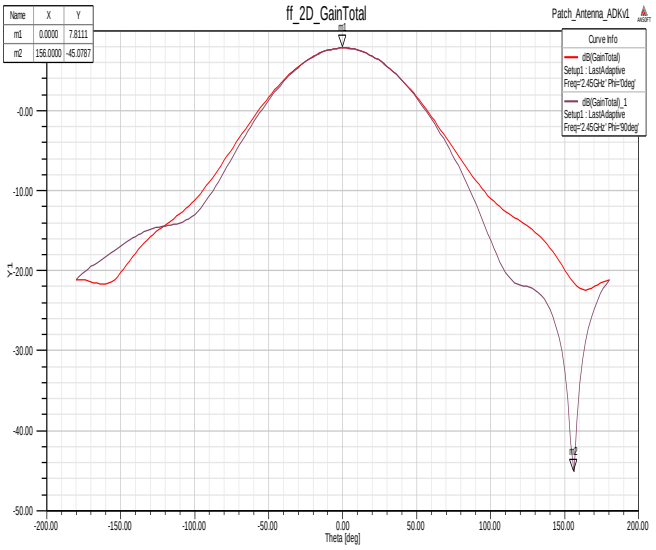



Fig. 5. Simulated 2D Gain at 2.45 GHz

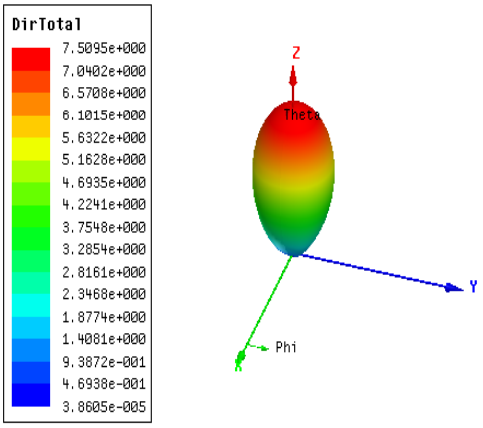



Fig. 6. Directivity of the Proposed Antenna

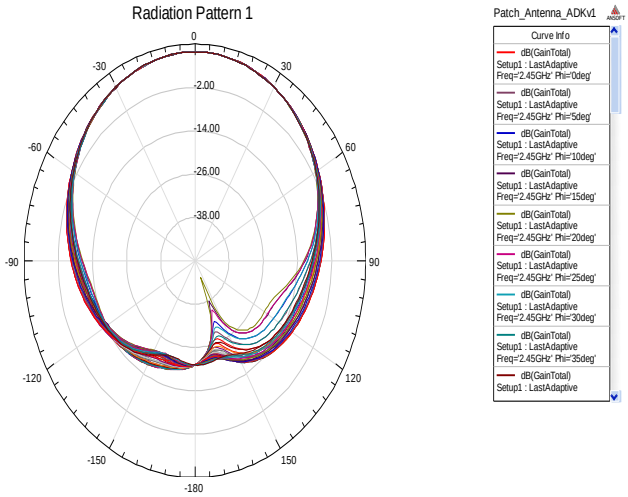



Fig. 7. Radiation Pattern of the Simulated Antenna

Fig. 8 illustrates the Voltage Standing Wave Ratio (VSWR) is an indication of the quality of the impedance match. In the proposed antenna, the VSWR value achieved at 2.45 GHz is 1.84 i.e. marked by $m_2$. Fig. 9 shows the measured value of SAR on the phantom model at 2.45 GHz. The achieved SAR value at 2.45 GHz is 0.0640 W/Kg averaged over 1 gm of tissue, which is under the limit of 1.6 W/Kg averaged over 1 gm of tissue.

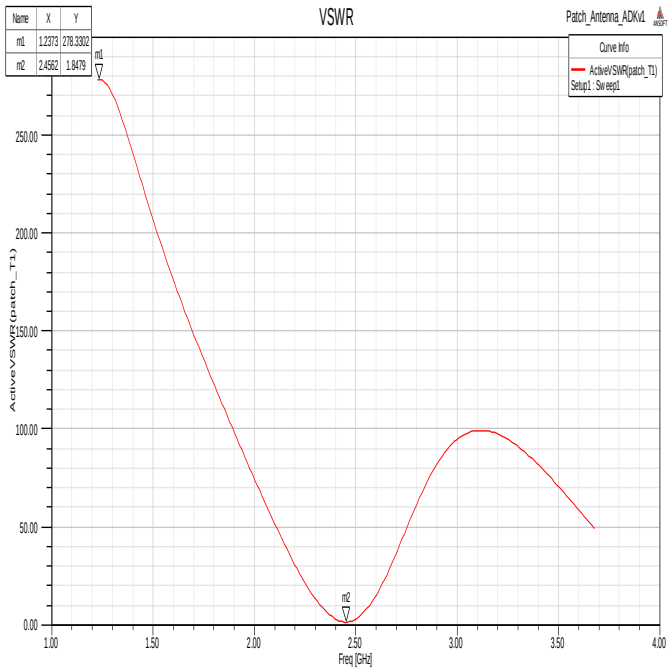



Fig. 8. VSWR of the Simulated Antenna

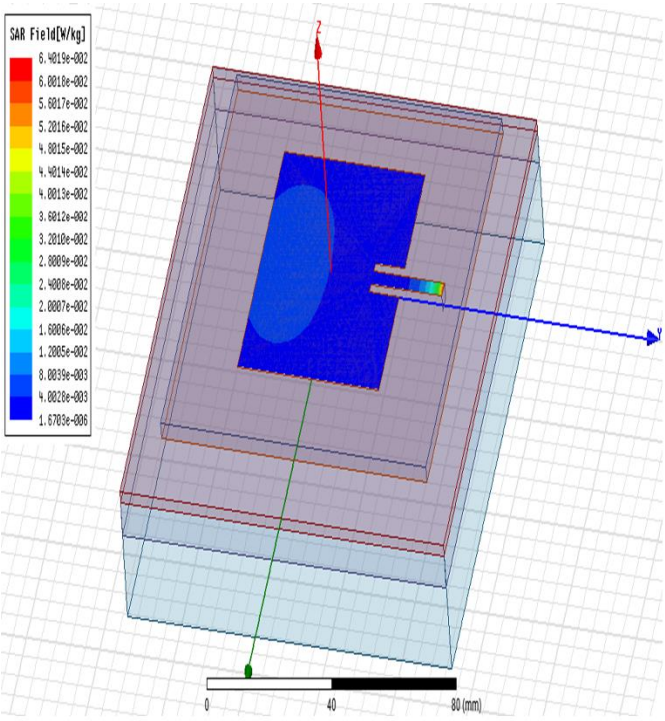



Fig. 9. SAR value of the Simulated Antenna at 2.45 GHz

TABLE IV. COMPARISON WITH OTHER RELATED WORKS

| Papers | Operating Frequency (GHz) | Size (mm) | Return Loss (dB) | Gain (dB) | VSWR | SAR (W/Kg) |
|---|---|---|---|---|---|---|
| Our Paper | 2.45 | 90 × 90 | -10.52 | 7.81 | 1.8 | 0.0640 |
| [18] | 5.8 | 40 × 40 | -21.52 | 3.05 | 1.1 | 0.0111 |
| [19] | 2.40 | 90 × 100 | -22.13 | 5.00 | 1.1 | - |
| [20] | 2.40 | 30 × 37 | -15.28 | 5.20 9 dBi | 1.4 | - |
| [21] | 1-10 | 120 × 120 | -9.54 | 6.49 | 2 | - |

TABLE IV shows the comparison of results of a wearable patch antenna with other related works. As can be seen from the table, the proposed antenna has an acceptable return loss, VSWR value and a high gain value compared to [18]-[21]. Specifically, the dimensions of the proposed antenna are slightly smaller than that of the [19] at the same operating frequency. Whereas, the return loss of the suggested antenna is better than that of the [21]. The antenna gain of the proposed antenna can be the certain factor value that is higher than all the other works mentioned in TABLE IV. Precisely, the gain of the suggested antenna is 4.76 dB, 2.81 dB, 2.5 dB and 1.32 dB higher than those of [18], [19], [20] and [21], respectively. Moreover, the specific absorption rate (SAR) value is quite low as compared to the FCC standard which is 1.6 W/Kg.

## IX. CONCLUSION

This paper proposes a wearable antenna that can be used for various applications such as monitoring of patients, navigation, and military applications. The rectangular patch antenna based on the inset fed patch technique has been designed on the textile material known as polyester. The substrate thickness is taken as 2.85 mm. The relative permittivity and loss tangent values are 1.44 and 0.01 respectively. The Inset fed technique is used because it provides planar structure and can be easily fed by 50Ω impedance. The overall antenna dimensions are 90×90 $mm^2$.

The measured return loss achieved at 2.45 GHz is -10.52 dB and a gain of 7.81 dB is obtained to ensure the efficient health monitoring. SAR is a very important parameter for wearable applications also SAR limit should have met the current standard set by IEEE which is 1.6 W/Kg averaged over 1 gm of tissue. The measured SAR value on the 3-layer human phantom model at 2.45 GHz is 0.0640 W/Kg averaged over 1 gm. of tissue. The radiation pattern of the simulated antenna was perpendicular to the axis of the wearer, which means that this antenna is feasible for wearable applications and cannot harm the human body tissues due to its high front to back ratio. Also, the wearable antenna is the best way to be applied for wireless body area network communication and it has vast applications to provide real-time health monitoring.

## X. FUTURE WORK

The future work can be done by improving the patch antenna bandwidth and efficiency. This can be done by increasing the height of the substrate material but it also increases the antenna dimensions. Also, the selection of textile

material plays a significant role to enhance the antenna efficiency in terms of gain and bandwidth. Moreover, the performance of an antenna quickly deteriorates under wet conditions in order to avoid this, waterproof materials can be used for future wearable communication designs.

## REFERENCES

[1] A. Khan, S. Bashir, and F. Ullah, “Electromagnetic Bandgap Wearable Dipole antenna with low Specific Absorption Rate,” in 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2018, pp. 1–4.

[2] H. R. Khaleel, H. M. Al-Rizzo, D. G. Rucker, and T. A. Elwi, “Wearable Yagi microstrip antenna for telemedicine applications,” 2010 IEEE Radio Wirel. Symp. RWW 2010 - Pap. Dig., pp. 280–283, 2010.

[3] A. Y. I. Ashyap et al., “Inverted E-Shaped Wearable Textile Antenna for Medical Applications,” IEEE Access, vol. 6, pp. 35214–35222, 2018.

[4] S. Bhattachaijee, S. Teja, S. R. B. Chaudhuri, and M. Mitra, "Wearable triangular patch antenna for ON/OFF body communication," 2017 IEEE Applied Electromagnetics Conference (AEMC), Aurangabad, 2017, pp. 1-2.

[5] S. Dumanli, "Challenges of wearable antenna design," 2016 46th European Microwave Conference (EuMC), London, 2016, pp. 1350-1352.

[6] R. Pei, J. Wang, M. Leach, Z. Wang, S. Lee and E. G. Lim, "Wearable antenna design for bio information," 2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Chiang Mai, 2016, pp. 1-4.

[7] N. H. M. Rais, P. J. Soh, F. Malek, S. Ahmad, N. B. M. Hashim, and P. S. Hall, "A review of the wearable antenna," 2009 Loughborough Antennas & Propagation Conference, Loughborough, 2009, pp. 225-228.

[8] K. Sajith, J. Gandhimohan, and T. Shanmuganantham, "A novel SRR loaded asymmetrical CPW fed ISM band wearable antenna for health monitoring applications," 2017 IEEE Applied Electromagnetics Conference (AEMC), Aurangabad, 2017, pp. 1-2.

[9] S. Sankaralingam and B. Gupta, “Development of Textile Antennas for Body Wearable Applications and Investigations on Their Performance Under Bent Conditions,” Prog. Electromagn. Res. B, vol. 22, no. November, pp. 53–71, 2010.

[10] A. Sabban, "New Wideband Printed Antennas for Medical Applications," in IEEE Transactions on Antennas and Propagation, vol. 61, no. 1, pp. 84-91, Jan. 2013.

[11] Salleh, Binti and Suriani Mohd Sam. “A Review of Antenna for On Body Wearable Communication and Its Application.”, Advanced Informatics School, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, Malaysia, (2016).

[12] K. Ito and Chia-Hsien Lin, "Dual-mode wearable antenna for medical applications," 2013 International Workshop on Antenna Technology (iWAT), Karlsruhe, 2013, pp. 372-375.

[13] X. Lin, S. Member, B. Seet, S. Member, and F. Joseph, “Flexible Fractal Electromagnetic Bandgap for Millimeter-Wave Wearable Antennas,” IEEE Antennas Wirel. Propag. Lett., vol. 17, no. 7, pp. 1281–1285, 2018.

[14] I. Singh and V. S. Tripathi, “Microstrip Patch Antenna and its Applications : a Survey,” International Journal of Computer Applications, vol. 2, no. 5, pp. 1595–1599, 2011.

[15] T. Patel and M. Sahoo, “Design Issues of Wearable Antennas,” International Journal of Science and Research (IJSR), vol. 5, no. 5, pp. 2015–2017, 2016.

[16] A. J. Agaliya, T. M. Neebha, and M. Nesasudha, "Efficient wearable antenna design by patch area extension for body area network applications," 2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, 2016, pp. 2130-2134.

[17] I. Gil and R. Fernandez-Garcia, “SAR impact evaluation on jeans wearable antennas,” 2017 11th Eur. Conf. Antennas Propagation, EUCAP 2017, pp. 2187–2190, 2017.

[18] M. I. Ahmed, M. F. Ahmed, and A. A. Shaalan, “Novel Electrotextile Patch Antenna on Jeans Substrate for Wearable Applications,” Progress in Electromagnetics Research C. 2018, Vol. 83, p255-265. 11p.

[19] G. Christina, A. Rajeswari, M. Lavanya, J. Keerthana, K. Ilamathi, and V. Manoranjitha, "Design and development of wearable antennas for telemedicine applications," 2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, 2016, pp. 2033-2037.

[20] Zaini, SR Mohd, and KN Abdul Rani. "Wearable Inset-Fed FR4 Microstrip Patch Antenna Design." IOP Conference Series: Materials Science and Engineering. Vol. 318. No. 1. IOP Publishing, 2018.

[21] S. K. Jain, N. Baviskar, N. Golait, and S. Jain, “Design of Wearable Antenna for Various Applications,” Antenna Test & Measurement Society (ATMS India-18), pp. 2–6, 2018.

[22] Faisal Ahmed Dahri, Riaz A. Soomro, Sajjad Ali Memon, Zeeshan Memon, and Majid Hussain Memon, “An Optimized Inset Feed Circular Cross Strip Antenna Design for C-Band Satellite Links” International Journal of Advanced Computer Science and Applications (IJACSA), 9(5), 2018.

[23] Asif Ali, Nasrullah Pirzada, Muhammad Moazzam Jawaid, and Sajjad Ali Memon, “Design and Simulation of a Rectangular E-Shaped Microstrip Patch Antenna for RFID based Intelligent Transportation” International Journal of Advanced Computer Science and Applications (IJACSA), 9(4), 2018.

# Product Feature Ranking and Popularity Model based on Sentiment Comments

Siti Rohaidah Ahmad[1], Nurhafizah Moziyana Mohd Yusop[4], Muslihah Wook[5], Arniyati Ahmad[6]
Department of Science Computer
Universiti Pertahanan Nasional Malaysia
Sungai Besi, Kuala Lumpur, Malaysia

Azuraliza Abu Bakar[2], Mohd Ridzwan Yaakub[3]
Centre for Artificial Intelligence Technology
Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia, Selangor, Malaysia

***Abstract*—This paper proposes the development of a model to determine feature popularity ranking for products in the market. Each feature that is reviewed by a customer has a relation to sentiment words present in the sentences within a customer review. Feature quantity of a product, derived from customer review dataset, cannot be used as a benchmark to determine customers' preferences since each feature is influenced by sentiment words that give it either a positive or negative meaning. A positive meaning shows that the feature is liked by user; and a negative meaning shows that it is disliked by user. This study finds that sentiment assessments by users play an important role in determining feature popularity ranking; and they affect the feature of a product. Thus, this study proposes the development of a model that takes into account the importance of sentiment assessments present in each sentence within a customer review of a product feature. A case study has been conducted in proving that the developed model is able to produce a list of product feature popularity ranking. Results of this experimental model is also put into simple comparative analysis with a few models from previous studies.**

***Keywords*—Product feature ranking; sentiment analysis; feature selection; sentiment word***

## I. INTRODUCTION

This study introduces feature ranking calculation and arrangement model; based on customer feedbacks on the advantages and disadvantages of a product feature. In this study, the satisfaction level of users on a product is identified through the use of positive or negative sentiment words, and the strength of customer reviews in the comment sections. Product feature is highlighted since users will usually look for important product information regarding the advantages and disadvantages of a product in its features. However, existing datasets only provide either positive or negative information about the product without a clear and detailed positive or negative feature list; and thus, lacking an explanatory aspect. It is conclusive to say that existing customer reviews could not really help customers decide whether they should buy the product or not.

For that matter, the model developed in this study examines customer review datasets by analyzing positive or negative sentiment words and sentiment strength values. The proposed model will be able to evaluate the forms of customer reviews on a product feature with either a negative or a positive form. Thus, the output of this model could ultimately help users effectively assess a product based on the analyzed features. The element of sentiment assessment in this model, which is based on customer reviews, will enable it to provide detailed feature ranking information for the perusal of customers and producers. For example: when the proposed model is applied on a Canon product (a camera), the model will classify and display a list of positive features such as *camera*, *photo*, *picture*; and a list of negative features such as *memory card*, *view finder*, *size* etc.

The analysis of feature ranking and popularity is very important to get a clear and detailed overview of the product features preferred by the consumers. However, the abundance of unanalyzed reviews make it difficult for users to assess a product. This problem also complicates the identification process of product features and real sentiment types for both consumers and manufacturers. Even when a product feature is mentioned repeatedly in a customer review, it does not really give the impression that the product feature is favoured by user. Conversely, if the feature rarely appears in customer review, then the product can be considered as unfavourable by users. Therefore, a summarized analysis on product features and sentiment words is very important to help product users and manufacturers make prompt decisions and save time [1].

Most of the previous studies on feature ranking in sentiment analysis are more focused on customers' assessments on product feature compared to the disadvantages and advantages of the product feature [2]. Findings from the analysis on previous literatures have shown that most researchers have studied feature ranking based on two aspects: feature importance in customer reviews; and positivity or negativity in sentiment categories.

Feature ranking technique [3] is based on feature importance and is determined by two factors: feature relevancy and feature frequency. *Hyperlink-induced topic search* (HITS) algorithm and *bipartite* graph are used to determine feature relevancy; whereas functional score value is used to determine feature frequency. However, this technique does not consider the sentiment category for each feature present in customer reviews. Additionally, Yu et al. [4] has developed a feature ranking technique based on the frequency of appearance for one particular feature in customer reviews. Features that are frequently reviewed in customer reviews are considered important features regardless of the feature sentiment category.

Eirinaki et al. [5] has developed *high adjective count* (HAC) algorithm to extract feature labelled as POS *noun*. This feature is frequently used in customer reviews to express opinions. HAC algorithm calculates the score for each feature; based on sentiment score calculation value, which is derived from *max opinion score* algorithm. Additionally, Jawadwala and Kolkur [1] have improvised on the work of Eirinaki et al. [5] by increasing the Senti-WordNet function to get an objective score value for each sentence. This function is used to identify sentence types: either subjective or objective. Other than that, feature ranking value is calculated based on opinion score value derived from Senti-WordNet. Similarly, the study by Ahmad and Doja [6] has utilized Senti-WordNet function to derive score values for each feature and calculate the overall orientation of the feature. This function will then be able to determine whether a feature is prone to be a more positive or negative sentiment. According to them, the feature tendency of the sentiment: either positive or negative, is determined by identifying the category of sentiment word that has a relation with the feature.

The remainder of this paper is organised as follows: Section 2 will discuss the methodology. Next, Section 3 will outline the proposed of model, and Section 4 will describe the experimental result. Section 5 will explain the discussion. Lastly, Section 6 will conclude this work.

## II. METHODOLOGY

The objective of this research is to propose a developmental model to calculate the list of feature ranking and population feature of a product. The components of each phase are depicted in Fig. 1. There are 6 phases in developing this model, which are:-

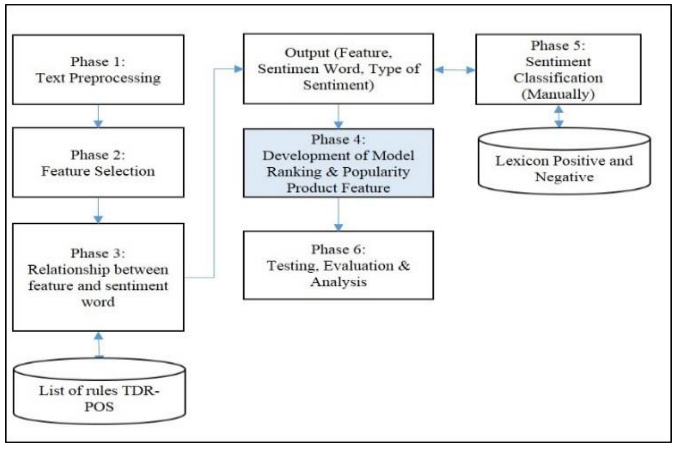



Fig. 1. Methodology of Developing Product Feature Ranking and Popularity Model

### A. Phase 1 – Text Preprocessing

Each data has to go through the data cleaning process which includes correcting words with spelling errors and correcting errors in word capitalization. After the data cleaning process has been completed, each data has to go through the part-of-speech tagging process to identify each of the word tags such as noun, verb, adverb, determiner, negation, etc. For each word, the noun tag in the sentences will be extracted and stored in a table. Refer to Fig. 2.

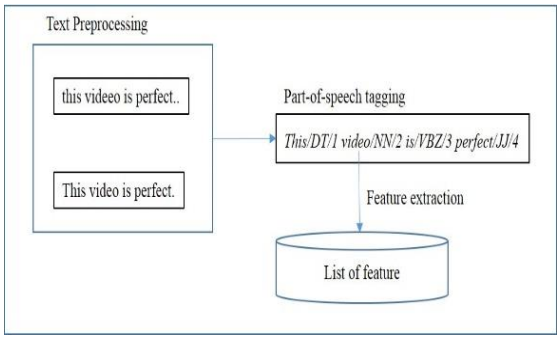



Fig. 2. Text Preprocessing

### B. Phase 2 – Feature Selection

Words stored in the list from phase 1 has to go through a feature selection process to choose the features that represent the actual dataset. In this study, ant colony optimization (ACO) algorithm is used as the feature selection technique because it has been proven to be effective in a previous study [7], [8].

### C. Phase 3 – Relationship between Feature And Sentiment Word

The process of identifying the relations between feature and sentiment word is important because it determines the real sentiment of a feature. This study uses a combination of typed dependency relations and part-of-speech tagging parameters to identify the relations between each word in the sentences from customer reviews. To identify this relation, an algorithm based on typed dependency relations and POS tagger has been developed. Observably, the combination between the concepts in POS tagger and typed dependency relations makes it easier to identify features and sentiment words that are related. Part-of-Speech Tagger (POS Tagger) is a software that performs the process of grammar labelling for each word in customer reviews. The labelling process would classify words into different types such as: noun, adverb, verb, adjective, determiner, conjunction, etc. On the other hand, typed dependency relations is the process of identifying the type of relations between one word and another in a sentence from a customer review. There are about 50 types of typed dependency relations in Stanford Parser. Among them are: NSUBJ, NMOD, ADVMOD, AMOD, CONJ, and others. Fig. 3. can be referred to for an explanation regarding the process of POS tagging and typed dependency relations.

Each feature and sentiment word relation that has been identified in sentences has to go through a checking process. This process identifies whether the sentiment word is grouped into positive or negative lexicon group. Finally, the feature and sentiment word relation is recognized whether it is positive or negative. This study uses a combination of typed dependency relations and part-of-speech tagging parameters [9] to identify the relations between each word in a sentence from the customer's reviews.

Example 1:

Sentence: This camera is perfect.
POS tagging: This/DT/1 camera/NN/2 is/VBZ/3 good/JJ/4
Typed dependencies relations:

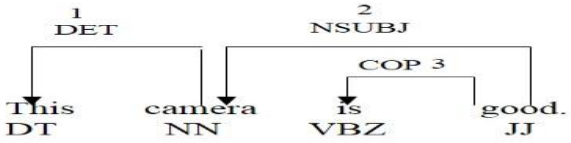



Fig. 3. Typed dependencies relations for the sentence: This camera is good.

### D. Phase 4 – Sentiment Classsification

Sentiment Classification in this study refers to the process of checking and ensuring that the feature and sentiment word pair which has been produced in section (d) is in its correct class: either in the positive or negative class. The output is manually checked with its dataset to ensure that the finding is consistent with the contextual information present in the dataset. The items checked in this phase are: feature, sentiment strength value, and sentiment class.

### E. Phase 5 – Development of the Product Feature Ranking and Popularity Model

This phase will be explained in The Proposed Product Feature Ranking and Popularity Model section.

### F. Phase 6 –Comparison and Analysis

This process of comparison in this study aims to test the effectiveness of the proposed feature popularity ranking model; whereby this model is tested against the model developed by Ahmad and Doja [6]. The process of analysis will later be thoroughly explained in the Base Model section below.

## III. The Propsed Product Feature Ranking and Popularity Model

This section explains the development of a calculation model for feature ranking as proposed in this study.

**Definition 4.1:** Product is defined based on its possession of its feature set; which is the product characteristics. For example, a Nikon product has a feature list of camera, battery, picture quality, flash, weight, etc. Below are the respective definitions of product and feature; presented in the form of equations (1) and (2):-

$$\text{Product} = P\{F\} \quad (1)$$

$$\text{Feature} = F = \{f_1, f_2, f_3, \ldots, f_i\} \quad (2)$$

**Definition 4.2:** Review is defined as the user review set; that is, $P = \{p_1, p_2, p_3, \ldots p_j\}$. Each user review contains feature *f*, sentiment word *ps*, and sentiment strength value *ks*, whereby $p_1 = \{(f_{i1}, ps_{i1}, ks_{i1}), \ldots (f_{im}, ps_{im}, ks_{im})\}$. Refer to Table 1 for sentiment strength values.

TABLE I. Sentiment Strength Values

| Positive value | Description | Negative value | Description |
|---|---|---|---|
| 3 | Strongest | -3 | Weakest |
| 2 | Medium strong | -2 | Medium weak |
| 1 | Strong | -1 | Weak |

### A. Feature Weightage

The frequency of feature occurrence in each dataset is counted. Each feature is categorized according to its type and number based on Table II below.

$$\text{Weightage} = ks_i * np_{fi}; \quad (3)$$

Where:

$ks_i$ = sentiment strength for user review ($p_i$), on feature $f_i$;

$npf_i$ = feature weightage, $f_i$, according to the feature frequency present in user review dataset;

### B. Total number of Sentiment Strength Weightage According to Fetaure Type

$$\text{Jks}_i = \sum_i^a Weightage f_i; \quad (4)$$

Where:

$Jks_i$ = total number of feature weightage based on sentiment strength which is according to feature type, $f_i$;

$a$ = total number of feature, $f_i$, according to feature type;

### C. Feature Ranking

$$\text{RF}_i = \text{Rank}(\text{Jks}_i); \quad (5)$$

Where:

$RF_i$ = all features in a dataset is ranked according to the frequency of feature occurring in user review, $p_i$, based on the total number of sentiment strength weightage, $Jks_i$. The order of feature ranking is arranged with the highest value ($Jks_i$) at the top and the lowest value ($Jks_i$) at the bottom.

The number of features for each user review dataset is quantified according to feature type. Table II exemplifies the process. Additionally, each feature is grouped according to its respective feature type. Based on this feature number, it is observable that the weightage values is determined based on the feature quantity. Table III exemplifies the process.

TABLE II. Example For Quantity of the Feature for Dataset Nikon

| Feature Type | Quantity |
|---|---|
| camera | 50 |
| size | 23 |
| picture | 15 |
| auto mode | 5 |

TABLE III. Weightage Value According to the Total Number of Feature Quantity

| Quantity of features | Weightage value ($npf_i$) |
|---|---|
| 75- above | 1 |
| 51-74 | 0.75 |
| 25-50 | 0.5 |
| 1-24 | 0.25 |

### D. Product Reputation Value

In this study, product reputation value is calculated according to the total number of positive and negative reviews. The reviews are separately counted before the total number of positive reviews is deducted from the total number of negative reviews. The final result is considered as the performance value for the product [10]. The algorithm in Fig. 4 is used to calculate and sort the feature ranking.

**Input:**
Product feature $f_j$;
Value of sentiment strength $ks_i$;
weight $npf_j$
**Output:**
Feature Ranking $f_j$

**0.** **START**
**1.** Get vectorA[$ks_i$, $npf_j$];
**2.** for (int bilVektorA=0; bilVektorA < bilrekodVektorA; bilVektorA++)
{
**3.** Get value of feature weight $fw_i$;
**4.** Get value of feature strength $fs_j$;
**5.** $fws_j=(fw_i * f_j)$;
**6.** Update vectorA[$fs_j$, $fw_j$, $fws_j$];
**7.** } End for
**8.** Get vectorB[$f_j$, $fws_j$];
**9.** for (int bilVektorB=0; bilVektorB < bilrekodVektorB; bilVektorB++)
{
**10.** $TOS_j$ =Aggregate $fws_j$ and group based on $f_j$;
**11.** }End for
**12.** ListofRanking =Rank($TOS_j$);
**13.** Return list of ranking;
**14.** **END**

Fig. 4. Algorithm for the calculation of feature ranking based on feature quantity, sentiment strength value and sentiment category

The purpose of weightage value is to gauge the importance of certain features; which is done based on the total number of feature quantity present in user review dataset. The bigger the total number of feature quantity, the higher the weightage value. The whole process of calculating feature ranking is based on feature quantity, sentiment strength value dan sentiment category.

## IV. Experimental Setup and Result

### A. Dataset

To test the effectiveness of this proposed model, a customer review dataset as compiled by [11] is used. The dataset comprises of reviews on five types of electronic products, as shown in Table IV. This dataset is written in English and taken from the Amazon website.

TABLE IV. Summary of Review Datasets

| Dataset | Number of manual product features | Number of review sentences |
|---|---|---|
| Nikon | 70 | 346 |
| Nokia | 100 | 546 |
| Apex | 104 | 739 |
| Canon | 100 | 597 |
| Creative | 170 | 1716 |

### B. Base Model

To evaluate the base model, the effectiveness of this ranking model is tested with feature popularity list; and the result is compared with the proposed model in [6]. The feature ranking calculation method as proposed by Ahmad and Doja [6] is based on the value derived from Senti-WordNet [12]. Each sentiment word with a relation to related feature is classified using Senti-WordNet. Additionally, each feature has a total weightage that is taken as the polarity value for each sentiment word; whereby the sentiments words are identified with the total number of sentences that contain the related features and sentiment words.

Total of weight (Wt) =

$$\sum_{n-1}^{d} Wt\ of\ positive\ features - Wt\ of\ negative\ features \quad (6)$$

whereby $d$ = Number of documents containing feature;

Results from the calculation model in (6) enable each feature to have its value identified with eitiher positive or negative. According to the researcher in [6], if the total weightage of the feature is identified to be of positive value, then the feature is likely viewed positively and preferred by users. Conversely, if the total weightage of the feature is identified to be of negative value, then the feature is likely viewed negatively by users. The feature ranking will be sorted in an ascending order whereby the highest positive value is placed at the top and the lowest negative value is placed at the bottom. The value is determined based on the total value of feature weightage.

### C. Result

This section lays out the results of a case study that was conducted to test out the effectiveness of the proposed calculation model on feature ranking and popularity. Table V displays the comparative analysis between three models: feature ranking based on quantity; the proposed feature ranking list model; and Ranking System model as proposed by Ahmad and Doja [6].

TABLE V. DIFFERENCES ON PRODUCT FEATURE RANKING BETWEEN FEATURE RANKING MODEL AND RANKING SYSTEM

| **Feature** | **Quantity** | **Feature** | **Feature Ranking Model** | **Feature** | **Ranking System** |
|---|---|---|---|---|---|
| camera | 43 | camera | 51.5 | camera | 20.38 |
| picture | 17 | picture | 7 | picture | 7.38 |
| picture quality | 10 | picture quality | 5.5 | picture quality | 5.50 |
| use | 10 | feature | 4.5 | use | 5.13 |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| sunset feature | 1 | indoor image | -0.5 | audio | -0.63 |
| transfer | 1 | system error | -0.5 | system error | -0.63 |
| txt file | 1 | transfer | -0.5 | lcd | -0.75 |
| view finder | 1 | lens cap | -1 | indoor picture | -0.75 |
| zoom image | 1 | 8mb | -1 | 8mb | -1.13 |
| **Product Reputation** | **202** | **Product Reputation** | **104.5** | **Product Reputation** | **75.50** |

The model using feature ranking based on quantity only displays the total number of feature that was reviewed by users; without any analysis on its value. This information would not be able very helpful to new users in evaluating the product. On the other hand, Table VI has displayed that the suggested model is able to determine the feature value: both positive or negative value; and thus, the popularity of feature strength, the feature level, and the feature frequency in customer review dataset can also be identified. A positive feature value will display high sentiment strength, high feature level and high frequency of appearance in customer review dataset. Conversely, a negative feature value will display low sentiment strength and feature level; and thus, will be ranked at a low position in the feature ranking list.

Table VI is a comparison on product reputation value between the proposed model and Ranking System. In overall, product reputation value for the proposed Product Feature Ranking model is found to be higher than Ranking System.

TABLE VI. DIFFERENCES ON PRODUCT REPUTATION VALUES BETWEEN PRODUCT FEATURE RANKING MODEL AND RANKING SYSTEM.

| **Dataset** | **Feature Ranking Model** | **Ranking System** |
|---|---|---|
| Nikon | +104 | +75.50 |
| Nokia | +160 | +103.75 |
| Apex | -98 | +33.48 |
| Canon | +157.25 | +92.58 |
| Creative | +253.25 | +145.65 |

Product reputation or item reputation is the total evaluation given by users; based on certain aggregation method. This study proposes a calculation model on product reputation value in which the sum of customer sentiment evaluation is calculated. In general, the reputation value for feature ranking and popularity model exhibits better overall value than Ranking System. The use of equations (3), (4) and (5) has shown a major change to the results of product reputation value.

Results from the proposed model are better than Ranking System.; including the dataset on Apex product. The huge difference in the Apex dataset is due to the high number of criticisms on *player* feature compared to other features. Even though the *player* feature receives the highest number of customer reviews among other features, this number does not mean that *player* feature is preferred by user. In fact, the *player* feature receives a lot of negative sentiment evaluations from user. Besides that, there is a marked difference of 89.9 on the reputation value of Creative dataset between the proposed model and the Ranking System. Feature list for Creative dataset such as *player, software, price, sound, battery, size, sound quality* has a high number of positive reviews from user. Thus, it affects the calculation of reputation values for Creative product. The situation here exemplifies that the proposed model yields better reputation calculation because the equations (3), (4) and (5) take into consideration the sentiment value provided by user.

## V. DISCUSSION

This study has clearly explained the proposed calculation model for feature ranking. The calculation function for feature ranking is developed based on three aspects: feature quantity, sentiment strength value, and sentiment category. In effectively analyzing the information found in user reviews, it is important to take into considerations the element of user ratings on the features of a product. In other words, even though the product feature might receive a high number of reviews from users, it does not mean that the product feature is preferred by users. Hence, a detailed analysis should be conducted in identifying the real overall sentiment type provided by users. It is very important to consider sentiment assessments given by users on product features since it actually reflects the product quality. Aside from that, the assessment also offers a real view of what is liked and disliked by users.

Understanding this situation, the proposed feature ranking model takes into consideration the sentiment assessments given by users. Results derived from this model could eventually assist in creating a list of product feature ranking that

represents actual user preferences. This information will be very useful and helpful to new users in deciding whether to buy a product or not. The advantage of this model is that: it uses existing data in user review datasets, which include: sentiment strength value for each feature, feature quantity, and sentiment category for each product feature. The information on sentiment strength value is derived from user assessments on their satisfaction level regarding a certain product. Additionally, in terms of generating feature ranking, this model is not dependent to the hierarchical concept or product ontology, and the Senti-WordNet calculation model in obtaining sentiment strength values.

## VI. Conclusion

This study proposes a new model to determine product feature ranking and popularity; based on the information in user feedback such as: sentiment strength assessment, feature quantity, and sentiment category. The developed model takes into consideration the effect of sentiment evaluation given by user. Based on this study, it is also conclusive to say that the quantity of features extracted from user reviews, whether the total number is big or small, cannot be used as an indicator that the product is liked or otherwise. The results derived from this calculation model can help users assess a product based on product characteristics or features. For producers, the resultant information from the model can be used to improve the quality of products produced.

## Acknowledgment

The authors gratefully acknowledge Universiti Pertahanan Nasional Malaysia for supporting this research project.

## References

[1] M. K. Jawadwala and S. Kolkur, “Feature Ranking in Sentiment Analysis,” International Journal of Computer Applications, vol. 94, no. 13, pp. 42–49, 2014.

[2] A. Abdel-Hafez, Y. Xu, and D. Tjondronegoro, “Product Reputation Model: An Opinion Mining Based Approach,” in 1st International Workshop on Sentiment Discovery from Affective Data (SDAD 2012), 2012, pp. 16–27.

[3] L. Zhang, B. Liu, S. H. Lim, and Eamonn O’Brien-Strain, “Extracting and Ranking Product Features in Opinion Documents,” in Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010), 2010, pp. 1462–1470.

[4] J. Yu, Z.-J. Zha, M. Wang, and T.-S. Chua, “Aspect Ranking : Identifying Important Product Aspects from Online Consumer Reviews,” Computational Linguistics, pp. 1496–1505, 2011.

[5] M. Eirinaki, S. Pisal, and J. Singh, “Feature-based opinion mining and ranking,” Journal of Computer and System Sciences, vol. 78, no. 4, pp. 1175–1184, 2012.

[6] T. Ahmad and M. N. Doja, “Ranking System for Opinion Mining of Features from Review Documents,” International Journal of Computer Science Issues, vol. 9, no. 4, pp. 440–447, 2012.

[7] S. R. Ahmad, A. Abu Bakar, and M. R. Yaakub, “Metaheuristic Algorithms for Feature Selection in Sentiment Analysis: A Review,” in Science and Information Conference (SAI), 2015, pp. 222–226.

[8] M. H. Aghdam, N. Ghasem-Aghaee, and M. E. Basiri, “Text Feature Selection using Ant Colony Optimization,” Journal Expert Systems with Applications: An International Journal, vol. 36, no. 3, pp. 6843–6853, 2009.

[9] S. R. Ahmad, A. Abu Bakar, and M. R. Yaakub, “Detecting relationship between features and sentiment words using hybrid of typed dependency relations layer and POS tagging (TDR Layer POS Tags) algorithm,” International Journal on Advanced Science, Engineering and Information Technology, vol. 6, no. 6, pp. 1120–1126, 2016.

[10] A. Jøsang, R. Ismail, and C. Boyd, “A survey of trust and reputation systems for online service provision,” Decision Support Systems, vol. 43, no. 2, pp. 618–644, 2007.

[11] M. Hu and B. Liu, “Mining Opinion Features in Customer Reviews,” in Proceeding AAAI’04 Proceedings of the 19th national conference on Artifical intelligence, 2004, vol. 21, no. 2, pp. 755–760.

[12] A. Esuli and F. Sebastiani, “SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining,” Proceedings of the 5th Conference on Language Resources and Evaluation, pp. 417–422, 2006.

# Interest Reduction and PIT Minimization in Content Centric Networks

Aadil Zia Khan
Lahore University of Management Sciences, Lahore, Pakistan

***Abstract*—Content Centric Networking aspires to a more efficient use of the Internet through in-path caching, multi-homing, and provisions for state maintenance and intelligent forwarding at the CCN routers. However, these benefits of CCN's communication model come at the cost of large Pending Interest Table (PIT) sizes and Interest traffic overhead. Reducing PIT size is essential since larger memory sizes have an associated cost of slower access speeds, which would become a bottleneck in high speed networks. Similarly, Interest traffic may lead to upload capacity getting filled up which would be inefficient as well as problematic in case of traffics having bidirectional data transfers such as video conferencing. Our contribution in this paper is threefold. Firstly, we reduce PIT size by eliminating the need for maintaining PIT entries at all routers. We include the return path in the packets and maintain PIT entries at the egress routers only. Further, we use Persistent Interests (PIs), where one Interest suffices for retrieving multiple data segments, in order to reduce PIT entries at the egress routers as well as to reduce Interest overhead. This is especially useful for live and interactive traffic types where packet sizes are small leading to a large number of pipelined Interests at any given time. Lastly, since using PIs affects CCN's original transport model, we address the affected aspects, namely congestion and flow control and multi path content retrieval. For our congestion scheme, we show that it achieves max-min fairness.**

***Keywords*—Content centric networks; congestion control; scalability***

## I. Introduction

Content Centric Networking aims for a more efficient use of the Internet. It follows a pull based communication model - the content is split into multiple segments and the subscriber sends out a separate Interest packet for each segment. Unlike the current Internet, routing is based on content names as opposed to end hosts' IP addresses. The approach followed by Content Centric Networks (CCNs) has many benefits – in-path caching, multi-homing, and provisions for state maintenance and intelligent forwarding at the CCN routers, to name a few. However, CCN does have a problem of a large memory footprint and overhead traffic that needs to be addressed to get the maximum benefit from CCN's pull based communication.

### *A. Memory Footprint of the PIT*

PIT as a core module of the CCN architecture allows many benefits in particular, such as multicasting, forward and reverse path symmetry, and filtering of unsolicited and duplicate packets. However, CCN derives these benefits at the cost of high memory requirements. To put this statement in context and see how much of a problem large memory sizes can be, let us first look at the different memory types studied for CCNs.

Off chip SRAMs can store up to 210Mb of data and have an access time of 4ns. SRAMs have been shown to not follow Moore's Law. RLDRAMs, on the other hand, can store up to 2Gb of data but have an access time of 15ns. DRAM has an access time of 55ns, but it can store up to 10GB [1]. Assuming each PIT entry size is 100 bytes; SRAMs can only contain around 262.5K entries, whereas RLDRAMs can have 2.5M entries. Can 100 bytes for a PIT entry be considered a realistic figure? In [2], for two different data sets, authors give an average URL length of 18 and 53 characters. Note that in CCN, the lengths may be longer if you also include AS names in the content name for ease of routing. Also, in the future, if different languages are to be incorporated in the content names, Unicode encoding would be required instead of ASCII so in terms of bytes the length would double.

Let's examine the above figures in the context of the PIT. The PIT would have a large memory footprint. İn [2] authors observe from trace of a 20Gbps access link that the corresponding PIT had 1.5M entries. This figure is expected to go up as bandwidths and the number of links increase. The number of PIT entries in a router is roughly on the order of the following:

$$(N_{port} \times BW \times RTT) / Size_{packet} \quad (1)$$

$N_{port}$ is the number of ports that the switch has, BW is the bandwidth, RTT is the total round trip propagation delay, and $Size_{packet}$ is the size of data packets.

The contribution of each flow to the PIT size may be large in case of pipelined interests for live and interactive traffic where packet sizes are small. For example, in VoCCN, each client sends out around 50 interests per second [3]. Unlike live traffic, for traffic types where chunk sizes can be large, such as file sharing, the number of Interests can be kept small if we keep large segment sizes and allow packet fragmentation [4]. However large packet sizes have performance issues in face of packet loss. So we believe that the number of pipelined interests could be large in non-live traffic too.

Fitting PIT in the SRAM would be difficult and slow memory access makes DRAM or RLDRAM infeasible for CCN routers. Note that multiple memory accesses would be required for a PIT entry lookup. Another benefit of reducing PIT and FIB sizes would be a reduction in overall costs and energy requirements since fast access memories are expensive ($27/MB for SRAM as compared to $0.27/MB for RLDRAM) and consume more power (0.12W/MB for SRAM as compared to 0.027W/MB for RLDRAM) [1].

### B. Overhead of the Interest Traffic

The one Interest packet per Data packet communication model adopted by CCN causes other overhead issues as well. Firstly, frequent insertions and deletions increase CPU load and lead to a decrease in efficiency, causing increase in download times [5]. Secondly, high upload bandwidth usage is an outcome if packet sizes are small and bitrates are high. For 9Mbps videos with packet size of 1500bytes, the number of Interests generated per second would be 750, requiring 0.6Mbps upload capacity assuming 100 byte Interest packet. For video conferencing, assuming 1Mbps streaming rate and average 500 bytes packets [6], there will be around 250 packets per second, requiring 0.2Mbps upload capacity. If we consider an IPTV solution, it is possible that multiple channels are retrieved simultaneously for smooth channel transitioning [7]. In this case the upload rate would increase by a factor equal to the number of channels simultaneously pre-fetched. Constraining upload capacity usage would be especially important in cases where upload is used for data transmission as is the case in video conferencing or in cases where links have asynchronous upload and download capacities. Also, lost Interests due to a congested upstream link would affect the stream quality even if the downstream link had sufficient capacity available.

We believe it is essential to revisit the original architecture with these issues in mind so as to ensure that the architecture does not become a bottleneck in the future. In this regard, we reduce PIT size, upload bandwidth usage, and insertion/deletion operations by using PIs where necessary, and completely eliminating PIT entries where possible. To this end, we also address congestion and flow control and multipath forwarding issues that arise due to a departure from CCN's original pull based approach to a push based one.

Keeping in mind increasing bandwidths and an ever increasing share of long flow streaming traffic, we believe the benefits of this work would remain even in the long term.

Rest of the paper is organized as follows. First of all, we examine the related work in section 2. Our approach for PIT size reduction by including the return path inside the packets is proposed in section 3. Our PI based communication model, together with max-min fair congestion control is described in section 4. In section 5, we give a summary of our proposed changes to the CCN. Evaluation results are given in section 6. We conclude and describe our future direction in section 7.

## II. Related Work

PIT size reduction is recognized as an important goal in CCN research. State of the art focuses on the data structure and hardware aspect of PIT. We believe that an approach which can reduce or eliminate the dependence on PIT without losing the benefits PIT brings is not only ideal but achievable as well.

Recent works try to reduce PIT sizes by using counting bloom filters [8], mapping bloom filters [9], and encoded named prefix tree [2]. Distributing PIT over multiple cards within a router has also been explored [8]. Using fingerprints, instead of full names, at the core routers, has been proposed in [10] to reduce the PIT sizes. Our approach can be used together with all of these approaches to further reduce the memory requirements.

Authors in [11] explore Interest aggregation by using Interest Sets. Unlike them, we use PIs to reduce PIT size as well as Interest traffic. The problem with Interest aggregation is that a single Interest loss would affect several data packets. Also, in case of traffics having a large number of pipelined Interests, for instance HiDef IPTV, we would have large Interest Sets which would introduce great complexity at the router. This complexity would be because a large Interest Set would lead to a traffic spike. Also if an Interest Set has several Interests, the router would have to check each corresponding data packet in its Content Store which would require several memory lookups. Multipath forwarding would also be an issue since an Interest Set can only be forwarded to one source. Unlike their work, we also address multipath forwarding and congestion and flow control issues arising due to doing away with the one interest per data packet model. Similarly, authors in [12] also use Interest aggregation, however at the router, these Interests are internally translated into multiple PIT entries thus the PIT size is not reduced and only the Interest traffic is reduced.

A different approach to reducing PIT sizes is taken in [13], where RTT awareness at the routers is used to reduce the residence time of PIT entries that have not been responded to.

In our congestion control scheme, the routers inform of the paths bottleneck fair share bandwidth. This approach has been studied in the context of Internet in works such as RCP[14] and XCP[15] where the returning data packet carries the bottleneck information. This bottleneck information is echoed back to the sender so that transmission rates may be adjusted accordingly. Compared to the approaches studied in the context of the current Internet, our approach would react quickly to changing network conditions for two reasons. Firstly, since CCN's breadcrumb approach eliminates forward and reverse path asymmetry, instead of echoing the rate information in Data packets, the Interest packets can be used to convey the bottleneck rate. Secondly, as we show in section IV, in CCN we can keep count of the exact number of flows as well as the flows bottlenecked on that router, thus enabling sharing of residual capacity amongst the bottlenecked flows only.

PIs have been proposed before [16]. Unlike our work, they do not address issues, such as congestion and flow control and multipath forwarding, arising due to a push based approach.

## III. PIT Reduction using Stateful Packets

PIT preforms three functions. Firstly, it eliminates forward and reverse path asymmetry. Secondly, it allows multicasting. When an Interest is received at a router but there is already a corresponding PIT entry, this implies a multicast scenario. The router will drop the Interest and update the PIT entry. When the corresponding data packet is received, copies will be made and forwarded on all the faces at which the Interests were received. Thirdly, PIT prevents unnecessary forwarding. When an unsolicited data packet is received which does not have a PIT entry, it would be dropped. Likewise if a duplicate Interest is received (having the same nonce) on a different face, it would be ignored and no further action would be taken. Our

approach allows these benefits to be achieved even if we eliminate the PIT entries at the non-egress routers. We now describe our approach followed by a discussion of how the above mentioned benefits can still be achieved.

### A. Interest Packet Transmission

We propose specifying the return path in the Interest packet's header using a PATH tag. When an Interest is received at an ingress router, instead of making an entry in the PIT, the ID of the node as well as the face at which the Interest was received is entered in the label header. The packet is then forwarded towards the egress router. Instead of making an entry in the PIT, each intermediate node appends only the ID of the face at which the Interest was received to the PATH tag.

When the Interest reaches the egress node, one option would be to append the face id to the PATH tag and forward the packet without making any entry in the PIT. While this would lead to complete elimination of the PIT, thus saving memory, it would also lead to the loss of benefits mentioned above that PIT brought. Thus we propose maintaining a PIT entry only at the egress routers. This PIT entry would also include the complete reverse path from the egress to the ingress routers within that AS. PATH tag would then be cleared before forwarding the packet into the next Autonomous System (AS).

In case multiple Interests, having the same ingress node ID in the PATH tag, are received at an egress router, the PIT entries would be combined to allow multicasting. To do this, the PATH tag saved in the PIT entry would be updated so that last entry corresponding to the ingress ID and the face includes all the relevant face IDs.

### B. Data Packet Transmission

When a Data packet is received at the egress router, the PIT would be consulted and the reverse path inserted in the packet's PATH tag. Based on this field, each downstream router would determine the face at which the packet needs to be forwarded. In this manner, the packet would either reach the egress router of the downstream AS or the content requester.

### C. Discussion

As shown in Fig.1, suppose an Interest packet is received at node R1 face 1. It would be forwarded to the egress node R4 as per our approach. When the Interest reaches the egress node, the PATH tag shows for each intermediate node the faces at which this Interest was received. The PIT entry would include this return path. Suppose another Interest is received at R1 face 2. This Interest would also be forwarded towards R4. However, from the PIT entry, R4 would infer that this is the same Interest but for a different face. So the entry would be updated. In the figure 1, <R1:1,2> implies that the returning Data packet has to be forwarded to node R1 faces 1 and 2.

The above approach ensures that the benefits of PIT are not lost as follows. When a Data packet is received at the egress router, the PIT would be consulted and incase it is unsolicited, it would be dropped. Otherwise, the packet would follow the reverse path specified in the PIT.

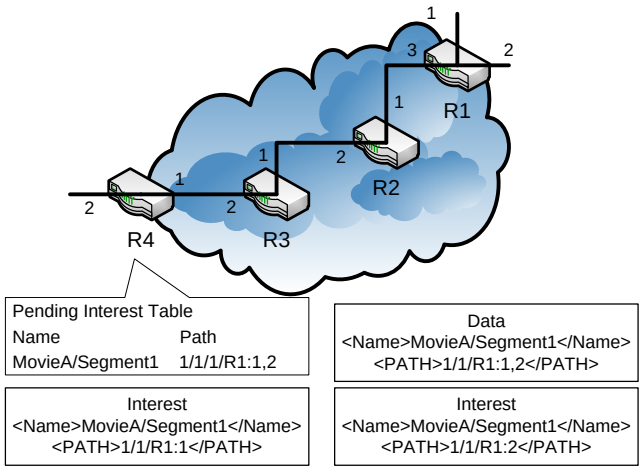



Fig. 1. PIT and Packets Structure.

For multicasting, note that if an Interest packet is received at the egress router and there is already a corresponding entry (or it is a duplicate Interest as determined by the nonce), the Interest would be dropped and the PIT entry updated. This would ensure multicasting at the gateway link. In case multiple Interests were received at an ingress router, to prevent duplicate data packets flowing from the egress to the same ingress node, the egress node would infer from its PIT entry if an Interest had been received at multiple faces of the same ingress router. This information would then be conveyed to the ingress router using the packet's PATH tag as already explained. This approach would reduce link stress [19] at the egress link and also on the path from the egress to the ingress router. However, link stress would not be eliminated in a scenario where two copies of the same data flow from the egress gateway to two different ingress gateways having some overlap in their paths. This is a tradeoff. We believe that minimizing memory requirements of CCN routers is important and link stress reduction at the gateway links is sufficient because link stress in the intra-AS paths would not be so problematic because load balancing approaches will be in place.

## IV. Interest Reduction using Persistent Interests

The approach discussed in the previous section reduces PIT entries at the non-egress nodes. The proposed approach would work well for static content, such as file sharing, where the data chunks already exist, can be large, and each chunk can be requested as required and from whatever source desired. Note that packet fragmentation [4] would allow Data packet sizes larger than the link MTU.

For live and interactive traffic, the Data packets have to be small leading to many pipelined Interests. VoCCN forwards 50 interests per second - the rate going higher if we also include video data. If we could replace these pipelined Interests with one Interest, we would save upload bandwidth as well as reduce the number of entries in the egress router's PIT. Also, decrease in the number of read write operations on the PIT would reduce processor load improving performance especially for video traffic. For this purpose, we adopt the concept of PIs to aggregate the Interests belonging to the same flow.

### A. Session Management

To initiate content retrieval, the end user would send out a PI registration packet, which would also include the rate at which the application seeks to receive data. The receiving router would make a corresponding entry in the PIT and forward the PI packet to the next upstream router towards the content store. Once the transmission has been completed, a tear down message would be transmitted by the end user, indicating session closure.

During the session, the user would periodically send PI keep-alive messages to keep the session alive as well as to inform the content source of the allowed transmission for congestion and flow control. In case a deadline for a PI packet is missed, or a deregistering message is received, the node would remove the corresponding PIT entry.

### B. Congestion and Flow Control

Similar to approaches like XCP and RCP, for congestion control, we propose an approach where each router maintains fair share bandwidth and the sender is informed of the bottleneck's fair share.

The content subscriber sends out a PI packet including the allowed bitrate specified in CURBOTRATE tag in the Interest packet. The receiving router would compare its own bottleneck fair share with value specified in the packet's CURBOTRATE tag, and update the latter if it is less before forwarding. Each router calculates the fair share bandwidth (FS) for each of its links using the following equation:

$$FS_t = (TotalCap/N) + (ResidualCap_{t-1}/N) \quad (2)$$

In the above equation, TotalCap refers to the link capacity, while $ResidualCap_{t-1}$ refers to the total link capacity minus the capacity used in the previous measurement cycle. N is the number of flows passing through that link. N is basically the number of PIT entries of type PI for that link.

When the Content Source receives the Interest, it would forward the data stream at CURBOTRATE.

We observe in our experiments that the convergence to min max fair share for each flow would be slow if we use the above equation for sharing residual capacity among flows. For this reason we propose an improved residual capacity sharing approach based on the observation that state maintenance inherent in the CCN's architecture [18] can enable us to share the residual capacity among bottlenecked flows only instead of sharing among all flows.

For this purpose, we introduce a PREVBOTRATE tag in the Interest packet which would include the path bottleneck for the previous cycle, and the router would also maintain a count of bottlenecked flows (BOTFLOWCOUNT) for each link. The PIT entry would include a flag variable specifying if that flow is bottlenecked on that router or not. Each time a Persistent Interest is sent, the sender would reflect the CURBOTRATE value so that it becomes the PREVBOTRATE value in the next PI keep alive packet. Now when a PI packet is received, the router would check if the value in the PREVBOTRATE tag is greater than or equal to the FS. If it is, it would imply that the flow is being bottlenecked on that link. PIT entry would be checked and if the bottlenecked flow flag is not set, the flag would be set and BOTFLOWCOUNT would be incremented. If the flag is already set, it implies that the flow has already been counted in the BOTFLOWCOUNT so no action would be taken. As opposed to this if a PI packet is received and PREVBOTRATE tag is less than the FS, it would mean that the flow is bottlenecked upstream. In this case if the PIT entry's flag is set, it would be set to false and the BOTFLOWCOUNT value would be decremented. If the flag is not set, no further action would be taken. The packet would then be forwarded upstream after modifying the CURBOTRATE tag if needed as per the approach described above. In this approach the equation for calculating the FS would be as follows.

$$FS_t = (TotalCap/N_{all}) + (ResidualCap_{t-1}/N_{bot}) \quad (3)$$

In the above equation, $N_{all}$ is the number of flows passing through that link while $N_{bot}$ is the number of flows bottlenecked on that link.

This approach requires PIT entries in all routers. We believe this is acceptable because firstly, multiple entries belonging to the same flow would be replaced by a single entry, and secondly this would allow multicasting on all nodes which would be especially important for reducing link stress in live and interactive traffic types. It should be noted that PI would only be used in live/interactive streaming types where the either the encoder can modify bitrate based on network feedback or the content has multiple copies encoded at different discrete rates and the corresponding rate can be sent based on the feedback. Thus for a multicast scenario, if multiple requests are received, based on different bottleneck rates, it is possible that different bitrates may be requested. Thus they have to be treated as different flows. For future work, we plan to address these issues and explore other issues arising when using this approach for multimedia streaming.

### C. Multipath Download

PI approach would change the way multipath retrieval is done in CCNs. We handle this by splitting the session into multiple streams if needed and using regular expressions in the PIT and PI packets. For instance if the content can be obtained from 4 different faces, the segment numbers requested at face would be as the expression below and corresponding entries would be made in the PIT.

$$segment_{id} \% 4 == face_{id} \quad (4)$$

For example, if the face id is 2 and the total number of faces is 4, the above expression would basically be requesting only those segment numbers which when divided by 4 would return a remainder of 2 and these segment would only be requested from the content source reachable through face 2. Other regular expressions may also be used.

## V. Changes to CCN

Our approach eliminates the need for maintaining PIT at the non gateway routers for traffic types similar to file transfer. In this regard we need to modify the way Interest and Data packets are forwarded and PIT entries are made. For intra-AS transmission purposes, we include a PATH field in the Interest and Data packets to include the path to be followed from the

egress to the ingress node. PIT entries also need to be modified so that each entry also includes the paths that need to be followed by the returning packets.

For live and interactive traffic, our approach replaces multiple PIT entries belonging to the same flow with a single entry. We use PIs so that content stream may be retrieved using a single Interest. Since this would affect how congestion control is handled in CCN, we also propose a router specified bottleneck rate based rate control. For this purpose, we require each router to include local information regarding each link's fair share bandwidth, number of flows on that link, number of bottlenecked flows, and residual capacity available from the previous measurement cycle. We also require Interest packets to include fields for return path's current bottleneck fair share rate as well as the bottleneck fair share in the previous transmission cycle.

Lastly, multipath retrieval would be different if we use PIs. For this purpose, we propose splitting the stream and using regular expressions to match streams with the requests in PIT entries and Interest packets.

## VI. Experimental Evaluation

To make a case and see how much PIT size reduction can be achieved by using PIs, we took dataset from CAIDA[1] related to four 10Gbps links which measured the packets per second as well as the number of flows per second flowing through the links during a one hour period once every month. The former corresponds to the number of PIT entries using CCN's original approach while flows per second roughly translate to the number of PIT entries using our PI based approach. Dividing them gives us the reduction factor. For the readings provided, we get a maximum 6.56 fold PIT reduction with an average of 2.4 fold assuming a 100ms RTT. If we assume a 200ms RTT, the number increases to a maximum of 13.13 fold PIT reduction with an average of 4.89 fold reduction.

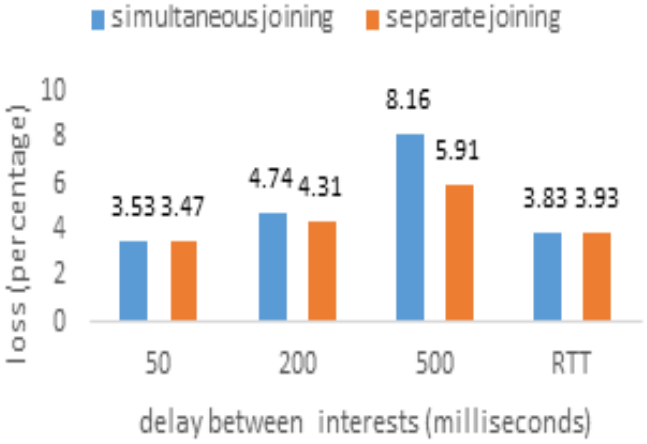



Fig. 2. PI Frequency vs Packet Loss.

For evaluation purposes, we implemented our PI approach in ndnSim[17]. First of all we studied the effects of frequency of PI keep alive messages on packet loss. We chose to measure packet loss because it would be representative of how quickly the sending rate would be updated in face of changing network conditions. We used a dumbbell topology consisting of four subscribers requesting content from four different sources sharing a bottleneck link having 1mbps capacity. All other links also had 1mbps capacity. Router queue limit was set to 10 packets. The RTTs for the four subscribers were 142ms, 62ms, 242ms, and 44ms respectively. The duration was 10 seconds and after 5 seconds, the bandwidth of node1's access link fell to 80kbps. For experimentation, we used two different scenarios. In the first scenario, each node joined at the same time, while in the second, the nodes joined at 0, 1, 2, and 3 seconds respectively. Results are shown in Fig. 2.

From the Fig. 2 we observe that the best results are obtained if the bottleneck probing PI packet is sent more than once per RTT. However, for our scenario, sending the probe once per RTT also gives comparable results. For less frequent probes, the sender would be slow to react, leading to a higher loss rate. Also, as shown in the figure, loss rate is lower for the scenario in which all the nodes join at different times. This is because in our scenario each node joined after a delay of one second after the previous nodes joining. This gave sufficient time for the network to converge to the fair share before the next node joins.

For the simultaneous joining scenario, we also studied the throughput each node got so as to measure fairness. For this purpose, we used a dumbell topology where each link had 10Mbps capacity and RTT was set to 200ms for each node. Three nodes subscribed to a continous stream whereas the fourth node followed an on-off transmission pattern where the on period varied from 500ms to 1.5s. The off periods also varied in a simillar manner. The duration was 10 seconds. Fig. 3 shows the average throughput each node got. The slight difference is because of packet loss.

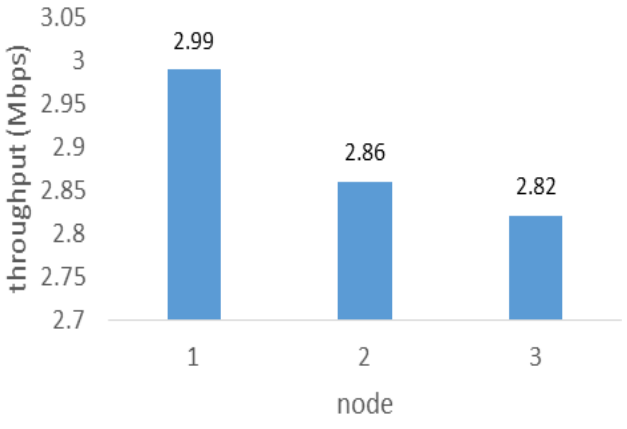



Fig. 3. Fairness.

[1] http://www.caida.org/data/passive/trace_stats/

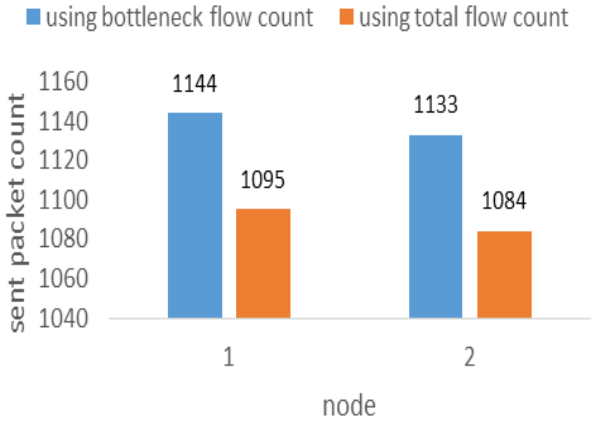



Fig. 4. Residual Capacity Sharing among Bottlenecked Flows.

Lastly, we explored the total number of packets transmitted in the scenario where the residual capacity was shared between all flows and compared it with the scenario where the residual capacity was shared only between the flows bottlenecked on that link. Sent packet count is used because it shows how quickly the sender converges to the fair share in case some flow's bottlneck shifts from the bottlenecked link to some upstream link. For this purpose we used the same dumbbell topology as above, however the the duration was 3 seconds. Also after one second, the bottleneck for node 3 and 4 shifts upstream and drops to 80kbps so that the original bottleneck link now has excess residual capacity which can be shared between flows belonging to nodes 1 and 2. As shown in the Fig. 4, more packets are transmitted if we distribute the residual capacity between the bottlenecked flows only.

## VII. Conclusion and Future Work

In this paper, we focused on PIT size reduction. For this purpose we proposed maintaining the return path inside the packets so as to eliminate the need for PIT in non egress routers. For traffic types where the number of pipelined Interests would be large, such as in multimedia streaming and video conferencing, we proposed using PIs where one Interests would be sufficient to retrieve the entire stream. We also proposed a congestion control approach to work with PIs. Our congestion control approach has been presented as a proof of concept in order to show the feasibility of the Persistent Interest approach.

For future work, we intend to focus on the congestion control aspect and experiment on larger topologies and in greater detail. We also plan to apply our fair share bottleneck probing approach for bit rate selection in rate adaptive video streams.

## References

[1] D. Perino, and M. Varvello, "A reality check for content centric networking," in ACM Sigcomm Workshop on Information-Centric Networking (ICN) 2011.

[2] H. Dai, B. Liu, Y. Chen, and Y. Wang, "On pending interest table in named data networking," In Proceedings of the eighth ACM/IEEE symposium on Architectures for networking and communications systems (ANCS) 2012.

[3] V. Jacobson, D. K. Smetters, N. H. Briggs, M. F. Plass, P. Stewart, J. D. Thornton, and R. L. Braynard, "Voccn: Voice over content centric networks," in Workshop on Re-architecting the Internet (ReArch) 2009.

[4] A. Narayanan, N. Nathan, D. Oran, and G. Tsudik, "Secure Fragmentation for Content-Centric Networks," CCNxCon 2013.

[5] M. Wahlisch, T. C. Schmidt, and M. Vahlenkamp, "Bulk of Interest: Performance Measurement of Content-Centric Routing," in Proc. of ACM SIGCOMM, Poster Session, 2012.

[6] Implementing QoS Solutions for H.323 Video Conferencing over IP. http://www.cisco.com/c/en/us/support/docs/quality-of-service-qos/qos-packet-marking/21662-video-qos.html

[7] D.A.G. Manzato, and N.L.S. da Fonseca, "A survey of channel switching schemes for IPTV," Communications Magazine, IEEE , vol.51, no.8, pp.120,127, August 2013.

[8] W. You, B. Mathieu, P. Truong, J.-F. Peltier, and G. Simon, "DiPIT: a distributed Bloom-Filter based PIT table for CCN nodes," In ICCCN, 2012.

[9] Z. Li, K. Liu, Y. Zhao and Y. Ma, "MaPIT: An Enhanced Pending Interest Table for NDN With Mapping Bloom Filter," in IEEE Communications Letters, vol. 18, no. 11, pp. 1915-1918, Nov. 2014.

[10] H. Yuan and P. Crowley, "Scalable Pending Interest Table design: From principles to practice," IEEE INFOCOM 2014 - IEEE Conference on Computer Communications, Toronto, ON, 2014, pp. 2049-2057.

[11] X. Jiang, and J. Bi, "Interest Set Mechanism to Improve the Transport of Named Data Networking," in Proceedings of SIGCOMM, 2013.

[12] D. Byun, B.-J. Lee, and M.-W. Jang, "Adaptive flow control via interest aggregation in CCN," in Proc. IEEE ICC 2013.

[13] Y. Tan, Q. Li, Y. Jiang and S. Xia, "RAPIT: RTT-Aware Pending Interest Table for Content Centric Networking," 2015 IEEE 34th International Performance Computing and Communications Conference (IPCCC), Nanjing, 2015, pp. 1-8.

[14] N. Dukkipati, M. Kobayashi, R. Zhang-Shen, and N. McKeown, "Processor sharing ows in the Internet," Thirteenth International Workshop on Quality of Service, 2005.

[15] D. Katabi, M. Handley, and C. Rohrs, "Congestion control for high bandwidth-delay product networks," Computer Communication Review, vol.32, no. 4, pp. 89-102, 2002.

[16] C. Tsilopoulos and G. Xylomenos, "Supporting Diverse Traffic Types in Information Centric Networks," In ACM SIGCOMM Workshop on Information Centric Networking (ICN 2011), 2011.

[17] A. Afanasyev, I. Moiseenko, and L. Zhang, "ndnSIM: NDN simulator for NS-3," NDN, Technical Report NDN-0005, 2012.

[18] A. Z. Khan, S. Baqai and F. R. Dogar, "QoS aware path selection in content centric networks," 2012 IEEE International Conference on Communications (ICC), Ottawa, ON, 2012, pp. 2645-2649.

[19] A. Z. Khan and S. Baqai, "A Case for Reducing Link Stress in a Multimedia Streaming Service Backbone," The IEEE symposium on Computers and Communications, Riccione, 2010, pp. 945-947.

# Image Retrieval System based on Color Global and Local Features Combined with GLCM for Texture Features

Jehad Q. Alnihoud
Department of Computer Science,
Al al-Bayt University

***Abstract*—In CBIR (content-based image retrieval) features are extracted based on color, texture, and shape. There are many factors affecting the accuracy (precision) of retrieval such as number of features, type of features (local or global), color model, and distance measure. In this paper, a two phases approach to retrieve similar images from data set based on color and texture is proposed. In the first phase, global color histogram is utilized with HSV (hue, saturation, and value) color model and an automatic cropping technique is proposed to accelerate the process of features extraction and enhances the accuracy of retrieval. Joint histogram and GLCM (gray-level co-occurrence matric) are deployed in phase two. In this phase, color features and texture features are combined to enhance the accuracy of retrieval. Finally, a new way of using K-means as clustering algorithm is proposed to classify and retrieve images. Two experiments are conducted using WANG database. WANG database consists of 10 different classes each with 100 images. Results of comparing the proposed approach with the most relevant approaches are promising.**

***Keywords*—CBIR; color histogram; GLCM; K-means; WANG database***

## I. INTRODUCTION

Visual cues rather than textual description of images is used in CBIR to extract features automatically from query image and image database and the most similar images are ranked and retrieved. Color and texture features are considered as the most frequently used in CBIR that because it is easy to extract features based on these visual contents. Moreover, shape [1], and regions [2] are broadly considered to index image based on contents. These low level features of an image are directly related to the contents of the image. In [3], they propose CBIR system based on color distribution. The proposed system relies on the probability of identical pixel colors as compared with its neighbors and the difference among pixels. Moreover, they deployed K-means Clustering algorithm group pixels. The approach overcome problems related to image displacement and rotation.

There are three major methods for color analysis and retrieval. The most common method is global-based or global color histogram. Approaches relies on this method such as [4, 5] are efficient but suffers from lack of information related to spatial distribution of colors. The second method relies on dividing images to equal blocks or regions such as [1, 6] and utilizes local color histogram for each region to extract color features. These approaches provides better information related to spatial distribution of colors. The third method is segmentation based, in which the division of the image is done using clustering algorithm such as [7, 3]. Complexity of features extraction is considered as the major drawback of such approaches.

Texture is used as an image discriminator utilizing visual properties of homogeneity rather than the existence of distinctive color [8]. There are many texture features such as contrast, correlation, energy, uniformity, density, coarseness, and homogeneity. As stated in [9], the most common texture descriptors are statistical model and transform-based. Statistical model extracts statistical features based on spatial dependence of texture's grey-level. The co-occurrence matrix used in [3, 10] is considered as an example of statistical model based. Discrete wavelet transform (DWT) is the most dominant example of transform based model. In [17], they propose CBIR system by integrating many features extracted using DWT, HSV color feature, cumulative color histogram, dominant color descriptor (DCD) and GLCM. Moreover, they propose using adjacency matrix of a bipartite graph to the sub-blocks of query image and image database. The proposed approach achieves better results as compared with the most relevant retrieval schemes. In [18], they propose CBIR system based on evolutionary algorithm. The proposed system discriminate images based on shape, color and texture features. The system is evaluated by precision-recall value of the retrieved results. In [19], they studied the effect of color spaces on CBIR using wavelet decomposition of each color channel. They recommended to use Lab color space and color features extracted using color moments. In [20], a two-leveled CBIR system is proposed. The proposed approach uses color features, edge detection, morphological operator (Dilate), and 2D-DWT (wavelet transform). WANG database is used to test the proposed CBIR system. Experimental results show superiority of the proposed approach as compared with other CBIR systems.

Most of the previously mentioned approaches used color and texture features. In this paper attempt is made to overcome some limitations related to spatial consideration of color distribution, disregard unrelated parts of image which may affect the accuracy of retrieval, and better utilizing of k means as clustering algorithm to enhance the retrieval accuracy.

In this paper a new and efficient CBIR system of two phases is proposed. In the first phase the emphases is given to global color features utilizing color histogram. Moreover, a new cropping technique is proposed to eliminate the least significant pixels within the image. This phase allows huge reduction of dissimilar images to be considered in the second phase. However, in the second phase images are sliced to three different regions and for each region the most significant features of color and texture based on joint histogram and gray level co-occurrence are calculated. Finally, a modified way of using K-means clustering is proposed to retrieve the final set of similar images as compared to the query image. The rest of the paper is organized as follows. Section 2 reviews the most important global and local color features to be extracted. Section 3 explains the texture features extraction using GLCM. Proposed approach is discussed in section 4. Section 5 presented similarity measures and experimental results. Finally, section 6 draws the conclusion.

## II. Color Feature Extraction

In CBIR number and type of features extracted from image play a significant role in increasing or decreasing the precision of image matching and retrieval. In this section the major extracted color features are discussed to enhance the understanding of the proposed system. As shown in [11], color is easy to implement and it considered as the most reliable visual content as compared with texture and shape. It is robust to background compilation and not affected by image size and orientation. Color histogram is the most prominent approach for color features extraction. Color histogram demonstrates the color distribution in image and it requires less computational cost as compared with other color approaches. Lack of spatial information is considered as the major drawback of color histogram approach.

As stated in [12], color histogram of image is represented using vector H= {H(0), H(1), …,H(j),…, H(N)}, where j denotes the color bin, N represents the number of bins, and H(j) denotes the total number of pixels at j[th] color bin. In this research, images are converted to HSV color space first, then color histogram is used to extract global features of cropped images. Furthermore, statistical-based features (color moments) are extracted for each image regions of the cropped images. These statistical features as follows:

- *Mean***:** the mean value for each color channel (H, S, and V) is calculated as follows:

$$M_i = \frac{1}{N}\sum_{j=1}^{N} P_{ij} \quad (1)$$

Where, *N* is the number of pixels in the image and $P_{ij}$ is the j[th] pixel value of the image at the $i^{th}$ color channel

- *Standard deviation:* the standard deviation value for each color channel (H, S, and V) is calculated as follows:

$$\boldsymbol{\sigma_i} = \sqrt{\left(\frac{1}{N}\sum_{j=1}^{N}\left(\boldsymbol{P_{ij}} - \boldsymbol{M_i}\right)^2\right)} \quad (2)$$

Where, *N* is the number of pixels in the image and $P_{ij}$ is the j[th] pixel value of the image at the $i^{th}$ color channel

- *Entropy:* the entropy value for each color channel (H, S, and V) is calculated as follows:

$$Entropy = -\sum_i P_i log_2 P_i \quad (3)$$

Where, $P_i$ is the probability of the difference between 2 adjacent pixels.

## III. Texture Features Extraction

As stated in [9], GLCM elements measures the relative of gray level occurrences between pairs of pixels within the image in different directions. The co-occurrence matrix is considered as useful way of extracting second order statistics. These statistics related to human view of texture.

In this research GLCM is utilized to extract texture statistical features (contrast, energy, correlation, and homogeneity). Moreover, GLCM is applied for images sub-regions in order to enhance accuracy of matching utilizing localization of features rather than relying on global texture features only. As an example, the GLCM is calculated to the following image in figure 1 using the following MATLAB code.

The MATLAB code is shown in figure 2.

```
%read image
I= imread('beach.jpg');
% convert rgb image to gray
R1 = rgb2gray(I);
% resize image
R2 = imresize(R1, [200 200]);
% create gray-level comatrix with default offset [0 1]
GL = graycomatrix(R2,'Offset',[0 1]);
%Extract properties of gray-level co-occurrence matrix
Ca = graycoprops
(GL,{'contrast','Correlation','Energy','homogeneity'})
```

Fig. 1. GLCM calculation (MALLAB code)

Gray level co-occurrence matrix of the sample image is [8 x 8] as follows:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1034 | 149 | 3 | 0 | 0 | 0 | 0 | 0 |
| 137 | 3606 | 700 | 17 | 0 | 1 | 0 | 0 |
| 4 | 659 | 13833 | 738 | 78 | 22 | 5 | 0 |
| 1 | 29 | 654 | 7024 | 684 | 75 | 35 | 0 |
| 0 | 10 | 52 | 687 | 4796 | 404 | 63 | 3 |
| 0 | 1 | 26 | 81 | 409 | 1892 | 185 | 1 |
| 0 | 0 | 1 | 12 | 59 | 211 | 1397 | 9 |
| 0 | 0 | 0 | 0 | 1 | 8 | 4 | 0 |

Fig. 2. Sample image

Table 1 shows statistical features of the sample image:

TABLE I. STATISTICAL FEATURES

| Contrast | Correlation | Energy | Homogeneity |
|---|---|---|---|
| 0.2168 | 0.9404 | 0.1809 | 0.9192 |

## IV. PROPOSED APPROACH

In this section a complete description of the proposed approach is presented.

### A. Phase I: Global Color Histogram

*a)* ∀ Image (I) ∈ Db (database) do:

*1)* Convert the RGB images to HSV. Figure 3 shows the original image converted to HSV color model and the different histograms related to hue, saturation, and value. As stated in [13], RGB image is converted to HSV color space using the following formulas:

$$H = cos^{-1}\frac{\frac{1}{2}[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-B)(G-B)}} \quad (4)$$

$$S = 1 - \frac{3[\min(R,G,B)]}{R+G+B} \quad (5)$$

$$V = \left[\frac{R+G+B}{3}\right] \quad (6)$$

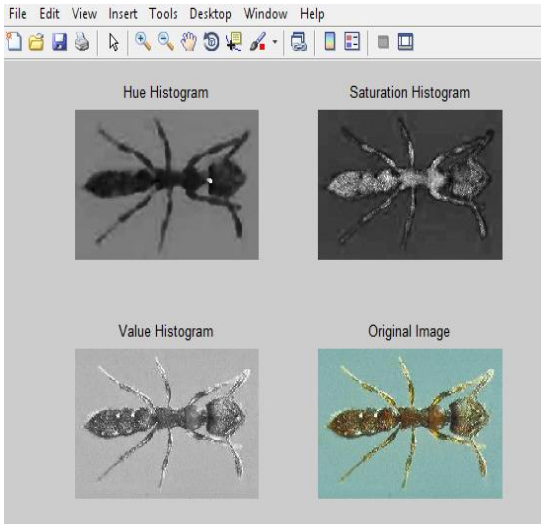



Fig. 3. RGB to HSV color conversion

*2)* Eliminate the least significant pixels of the original image (*I*) by applying the following steps to create rectangular mask with proper dimension. Figure 4 illustrates the cropping process.

- Identify cropping starting points (x , y) by applying the following formulas:

$$x = \frac{1}{8} * original_{width} \quad (7)$$

$$y = \frac{1}{8} * original_{height} \quad (8)$$

***Note:*** different scaling factors were tested and $(\frac{1}{8})$ gives the best results.

- Identify the new width and new height using the following formulas:

$$newWidth = original_{width} - x \quad (9) \quad (9)$$

$$newHeight = Oroginal_{height} - y \quad (10) \quad (10)$$

- Crop the original image using (x, y, newWidth, and newHeight).

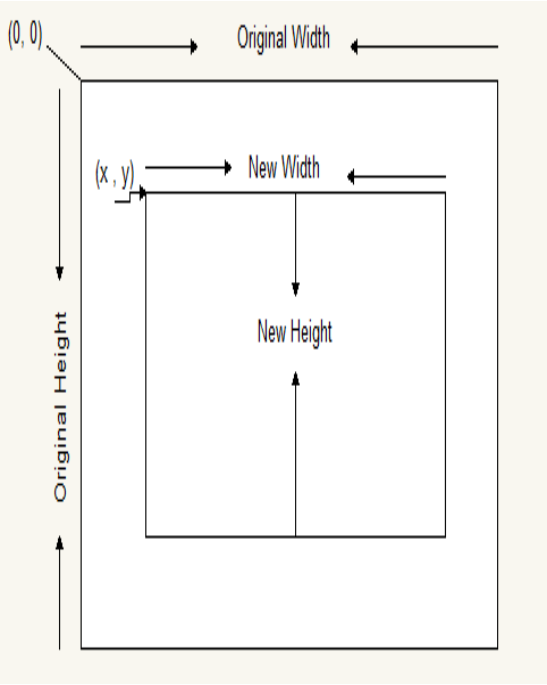



Fig. 4. Cropping process

*3)* Apply global color histogram to the cropped image in (2) using (H-8 bins, S-2 bins, and V-2 bins)

*b)* Apply steps (1-3) to query image.

*c)* Do apply bin to bin comparison between image query and image database then:

- Calculate the similarity based on Euclidean distance measure.
- Consider the most similar images- using proper threshold value as candidate images to be used in Phase-II.

### B. *Phase II: Joint Histogram and Gray Level Co-occurrence Matric (GLCM)*

*1)* Resize candidate images (CI) retrieved via Phase-I to be 200 x 200 pixels.

*2)* Starting from the Top-Left Corner, divide each image in (1) to 3-regions as follows:

R1= 50 x 50 Pixels.
R2= 100 x 100 Pixels.
R3= 200 x 200 Pixels.

*3)* $\forall R_i \in CI,\ i = 1, 2, 3$. Do extract:

- $[\mu_h, \mu_s, \mu_v]$, mean values for H, S, and V.
- $[\sigma_h, \sigma_s, \sigma_v]$, standard deviation values for H, S, and V.
- $[E_h, E_s, E_v]$, entropy values for H, S, and V.

This step generates 9-D feature vector for each region. Total number of features is 27-D for each image. R3 represents the whole image.

*4)* Texture based retrieval :

$\forall R_i \in CI$ Do:

- Apply GLCM (gray-level co-occurrence matrix) to the image.
- Normalize GLCM.
- Extract the following features:

*a)* Contrast using the following formula:

$$Cont = \sum_{i,j=0}^{N-1} P_{ij}(i-j)^2 \quad (11)$$

*b)* Correlation using the following formula:

$$Corr = \sum_{i,j=0}^{N-1} P_{ij} \frac{(i-\mu)(j-\mu)}{\sigma^2} \quad (12)$$

*c)* Energy using the following formula:

$$En = \sum_{i,j=0}^{N-1} (P_{ij})^2 \quad (13)$$

*d)* Homogeneity using the following formula:

$$Homog = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(i-j)^2} \quad (14)$$

Where, $P_{ij}$ is the probability of ij element of the normalized symmetrical GLCM, N is the number of gray-levels in the image, $\mu$ the GLCM mean, and $\sigma$ is the variance of intensity.

This step generates 12-D feature vector. At the end of this step, each image is represented by 39-D feature vector generated by step 3 and 4.

*5)* Do apply steps (1-4) to query image.

*6)* Retrieve similar images by deploying a new way of using K-means clustering as shown in figure 5.

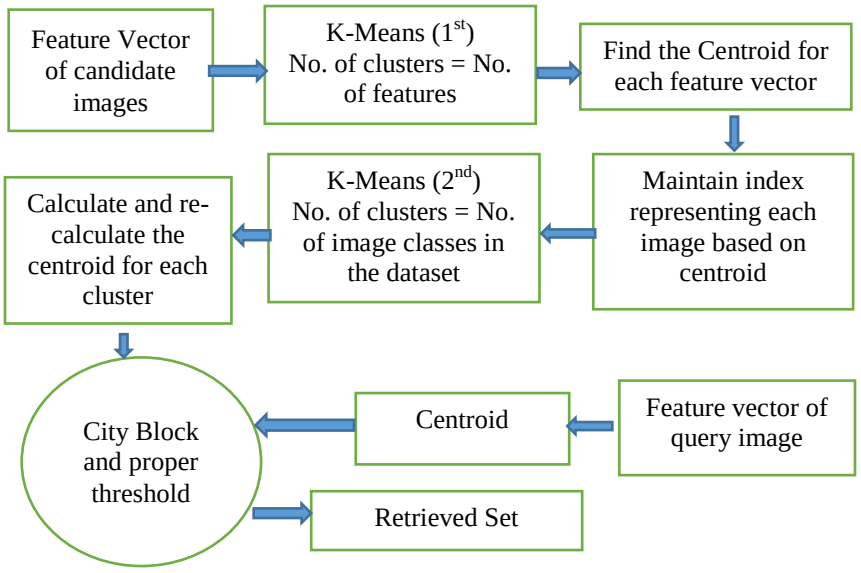



Fig. 5. Retrieving final set using K-means

As stated in [9], K-means as clustering algorithm is one of the unsupervised learning algorithms. It classifies data set into number of clusters in simple way. Users of K-means need to define k centroids based on number of chosen clusters. It is preferable to place these centroids away from each other. Each point in the data set is associated to the proper cluster by calculating the distance between this point and nearest centroid. The process continued by re-calculating k new centroid till no changes are done. To overcome some of problems related to user interference in original K-means

clustering, a two phases K-means clustering is proposed as previously shown in figure 5. In this research number of clusters is chosen based on number of features and image classes, then cluster centroid is calculated and re-calculated based on the feature vector of image data set. As for query, centroid is calculated based on its feature vector and the distance between the centroid of query image and different clusters is measured. Consequently, the image class is identified first and similar images are ranked based on distance from small to large.

## V. Similarity Measure and Experimental Results

### A. Similarity Measure

The similarity between two feature vectors can be measured using one of the many available distance functions such as Euclidian distance, City block distance, Minkowski, and Mahalanobis [14]. Using different distance functions may slightly increases or decreases precision of retrieval. In this research, Euclidean distance and City block are chosen because of ease of use and they achieve better results as compared with other distance measure. As shown in [15], Euclidean distance function effectiveness and efficiency is noticeable because of that it is considered as one of the most common similarity measures used in image retrieval. It calculates the distance between two feature vectors as follows:

$$D_{euclidean} = \sqrt{\sum_{i=1}^{n} (|I_i - D_i|)^2} \quad (15)$$

City block distance is calculated as:

$$D_{city} = \sum_{i=1}^{n} \left| I_i - D_i \right| \quad (16)$$

Where, *I* and *D* are two feature vectors of size n.

To measure the effectiveness of any CBIR system, the precision and recall are the most common parameters to be considered [16]. Precision is defined as follows:

$$Precision = \frac{No.of\ Releveant\ Retrieved\ Images}{No.of\ All\ Retrieved\ Images}$$

While, recall is defined as:

$$Recall = \frac{No.of\ Relevant\ Retrieved\ Images}{No.of\ All\ Relevant\ Images\ in\ the\ Catogery}$$

Precision measures accuracy of the retrieved set as compared to the query image, while the recall measures the ability of the system to retrieve as much as possible of similar images related to query image from the database. Precision and recall are affected by the threshold values because number of retrieved images may increases or decreases based on these values. If the accuracy or the precise matching is our target, then the threshold value (tolerance) is chosen to be small. As a result, the number of retrieved images will be less. Whereas, for the general purpose CBIR systems, the similarity rather than the precise matching is the target of such systems. Consequently, threshold values are chosen to be large. In the following section, precision and recall are measured based on different threshold values.

### B. Experimental Results

The proposed system is implemented using MATLAB R2010a and WANG database [3]. WANG database containing 1000 JPEG Corel images of 10 categories, each with 100 images. These categories as buses, elephants, flowers, and horses represent wide range of image classes. Categories such as dinosaurs, and flowers contain objects in contrast with the background, while food and Africa contain objects that mixed with background. As stated in [13], this database is widely used because of reasonable size and classes categorizations.

#### 1) Experiment One

In order to test the proposed approach many images randomly selected as query images. Figure 6 shows sample of test images belong to different categories available in WANG data set.

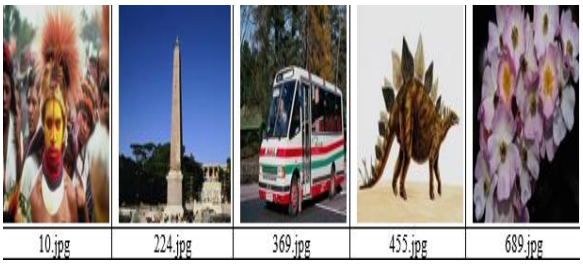



Fig. 6. Test images

Precision and recall are affected by the number of retrieved images which is increases or decreases based on the chosen threshold values. Figure 7 shows sample of retrieved set, image query which belongs to africa category is shown at the upper left corner. Precision of the retrieved set is 88% at recall level 21%.

Fig. 7. Sample of retrieved set (Africa)

Figure 8 shows sample of retrieved set, image query which belongs to dinosaur's category is shown at the upper left corner. Precision of the retrieved set is 92% at recall level 22%.


Fig. 8. Sample of retrieved set (Dinosaurs)

Figure 9 shows the experimental results of the proposed approach at different recall levels. It is reasonable that the precision is affected by recall level since as recall level goes higher, the retrieval of unrelated images are increased.

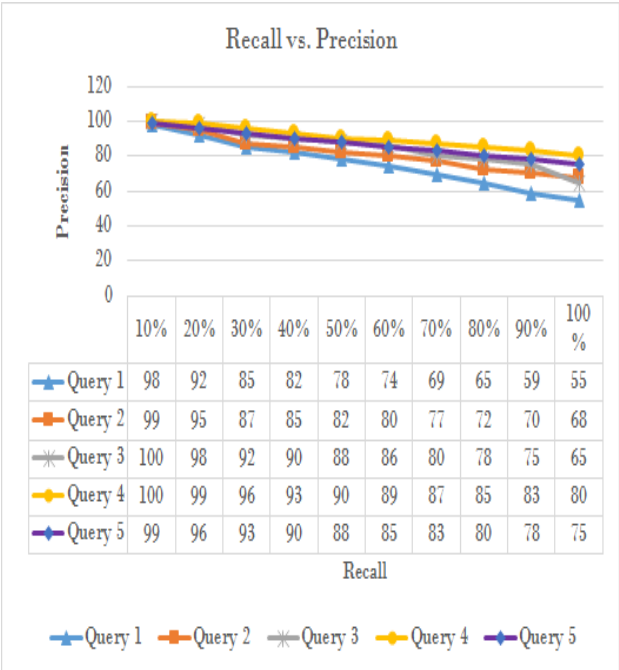



Fig. 9. Recall vs. Precision

The average precision is calculated for each test image and the overall average precision is measured to show the retrieval accuracy of the proposed system. The result is compared with two of the most relevant approaches as it can be seen in table 2.

TABLE I. 1ST COMPARISON WITH SOME OF THE RELATED CBIR METHODS

| | **1st Method [17] Yossef *et. al* [2012]** | **2nd Method [18] Jadhav and Pail [2012]** | **Proposed Method** |
|---|---|---|---|
| Average precision | 0.703 | 0.525 | 0.84 |

Table 2 shows that the proposed approach achieves higher average precision rate as compared with [17] and [18]. Phase-I as part of the proposed approach enhances the precision dramatically since filtration process based on color global features normally precise and eliminates unwanted candidate images to be considered in Phase-II.

*2) Experiment Two*

Once again WANG database is used in the second experiment. In Giveki *et. al.* [19] and Jehad [20], they tested all categories of image database available in WANG. In order to be consistent with the proposed approaches in [19] and [20], all categories of image database are tested with the same test images used in [20]. These categories are (Buses, Elephants, Flowers, Horses, Dinosaurs, Buildings, Food, Mountains, Beaches, and Africa) each with 100 images. Test images are shown in Figure 10.

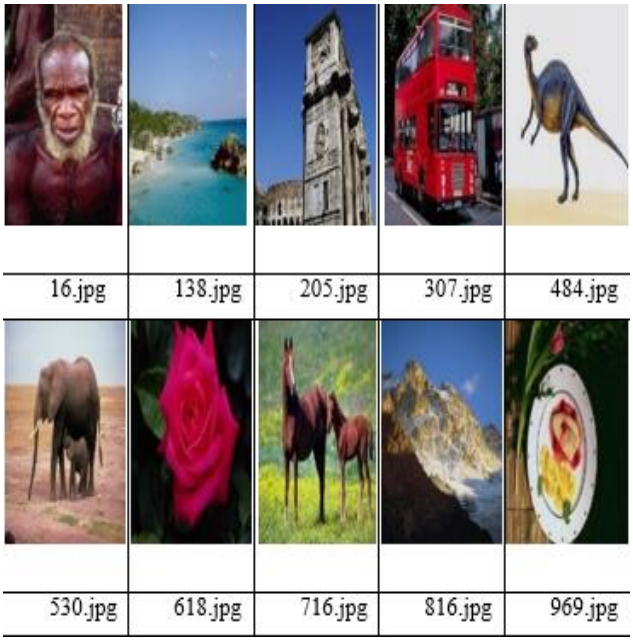



Fig. 10. Test images (2nd experiment)

Figure 11 shows sample of retrieved set, image query which belongs to horse's category is shown at the upper left corner. Precision of the retrieved set is 92% at recall level 22%.

Fig. 11. Sample retrieved set (2nd Experiment)

Figure 12 shows the precision of retrieval to the image query (138.jpg) which belongs to the Beach category. The precision was measured at different recall levels and the average precision is 80%.

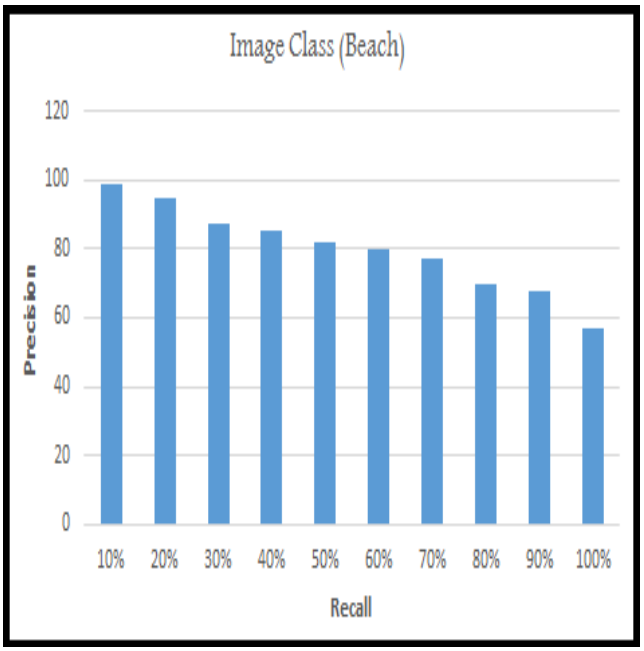



Fig. 12. Image class (Beaches)

Figure 13 shows the precision of retrieval to the image query (816.jpg) which belongs to the Mountains category. The precision was measured at different recall levels and the average precision is 75%.

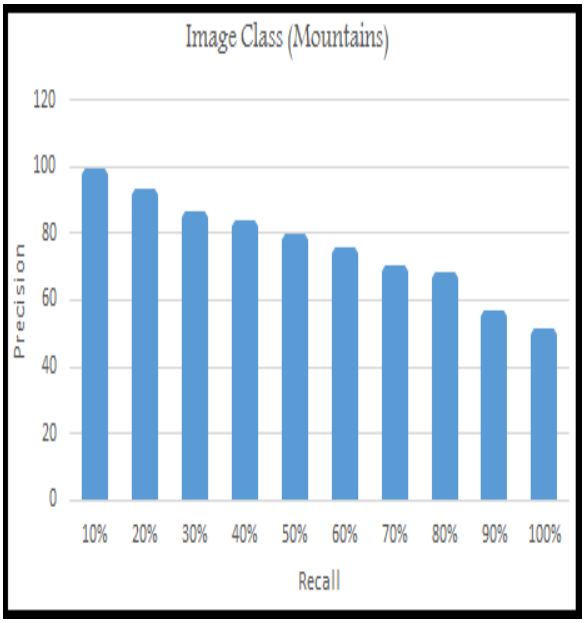



Fig. 13. Image class (Mountains)

Figure 14 shows the precision of retrieval to the image query (484.jpg) which belongs to the Dinosaurs category. The precision was measured at different recall levels and the average precision is 99%.

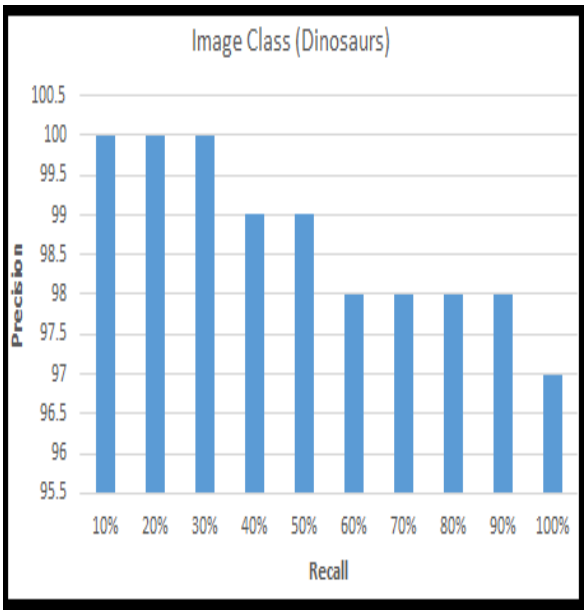



Fig. 14. Image class (Dinosaurs)

It is observed that, the proposed approach is succeed to achieve high precision rate with image categories in which objects are mixed with the image background such as Mountains, Beaches, and Africa and it achieves better precision rate with image categories in which objects are in contrast with the image background such as Buses, and Dinosaurs. Because of space limitations the results of testing are summarized in the following figure. Figure 15 presents the average precision of test images. Moreover, the overall average precision is shown.

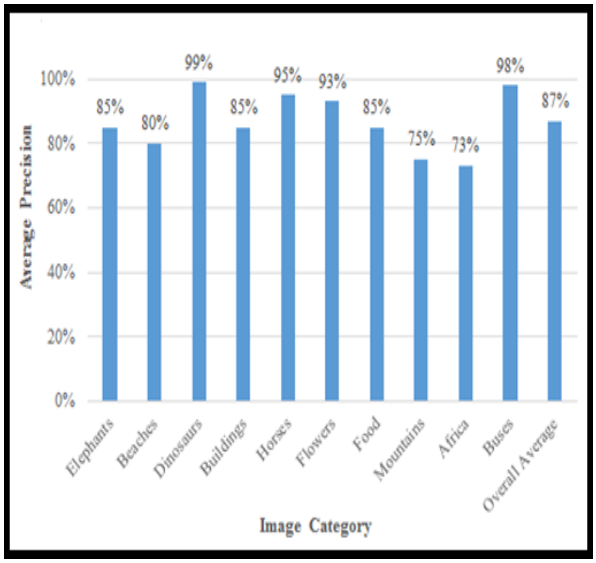



Fig. 15. Average precision of the proposed approach (2nd Experiment)

In [19], they propose CBIR system using wavelet transform. They tested different color models, the best results are obtained using HSV-wavelet with average precision 64%. In [20], a two levels CBIR approach is proposed utilizing color features and 2D-DWT (wavelet transform). The best results are obtained in the 2nd level with average precision 83%. Accordingly, the proposed approach in this paper achieves 87% overall average precision which is considered higher than the previous approaches.

## VI. Conclusion

In this paper a new CBIR system utilizing color and texture features is proposed. HSV color space is deployed and color global and local features are extracted in two phases. Initially, color histogram is used in the first phase with the cropped images. As a result the number of candidate images to be considered in phase two are reduced dramatically. Dividing images to regions or sub-images and extract features using joint histogram and gray-level co-occurrence matric (GLCM) is proposed in phase two. Lastly, K-means as clustering algorithm is utilized in an innovative manner to classify and retrieve images. WANG database is used to test the proposed approach and the results obtained is better as compared with the most relevant approaches. Exploring techniques related to shape features extraction and integrating them with the proposed system is left as a future work.

## References

[1] K. Iqbal, M. O. Odetayo, and A. James, "Content-based image retrieval approach for biometric security using colour, texture and shape features controlled by fuzzy heuristics," Journal of Computer and System Sciences, vol. 78, no. 4, pp. 1258-1277, 2012.

[2] M. Nascimento, V. Sridhar, and X. Li, "Effective and efficient region-based image retrieval," Journal of Visual Languages and Computing, vol. 14, no. 2, pp. 151-179, 2003.

[3] C. Horng Lin, R. Tai Chen, and Y. Kuan Chan, "A smart content-based image retrieval system based on color and texture feature," Image and Vision Computing, vol. 27, no. 6, pp. 658-665, 2009.

[4] C. Man Pun, and C. Fong Wong, "Fast and robust, color feature extraction for content- based image retrieval," International Journal of Advancements in Computing Technology, vol. 3, no. 6, 2011.

[5] M. J. Swain, and D. H. Ballard, "Color indexing," International Journal of Computer Vision, vol. 7, no. 1, pp. 11-32, 1991.

[6] M. Stricker, and M. Orengo, "Similarity of color images," In SPIE Storage and Retrieval for Image and Video Databases III, vol. 2420, pp. 381- 392. 1995.

[7] R. O. Stehling, M. A. Nascimento, and A. X. Falcao, "Cell histograms versus color histograms for image representation and retrieval," Knowledge and information Systems, vol. 5, no. 3, pp. 315-336, 2003.

[8] C. C. Gotlieb, and H. E. Kreyszig, "Texture description based on co–occurrence matrices," CVGIP, vol. 51, no. 1, pp. 70–86, 1990.

[9] S. Sridhar, and M. Gowri, "Color and Texture Based Image Retrieval", ARPN Journal of System and Software, vol. 2, no. 1, pp. 1-6, 2012.

[10] N. Jhanwar, G. Chaudhuri, G. Seetharaman, and B. Zavidovique, "Content based image retrieval using motif co-occurrence matrix," Image and Vision Computing, vol. 22, no. 14, p.1211-1220, 2004.

[11] S. M. Youssef, "Ictedct-cbir: Integrating curvelet transform with enhanced dominant colors extraction and texture analysis for efficient content-based image retrieval," Computers and Electrical Engineering, vol. 38, no. 5, pp. 1358-1376, 2012.

[12] R. Nag, M. Aatif, A. Kazi, and Z. Khan, "Color and Texture Based Image Retrieval," International Journal for Research & Development in Technology, vol. 7, no. 4, pp. 724-729, 2017.

[13] Y. Mistry, D. T. Ingole, M. D. Ingole, "Content based image retrieval using hybrid features and various distance metric", Journal of Electrical Systems and Information Technology. 2017, http//dx.doi.org/10.1016/j.jesit.2016.12.009.

[14] E. U. Scott, Computer vision and image processing. International Edition, Prentice-Hall, Inc, 1998.

[15] F. Malik, and B. Baharudin, "Analysis of distance metrics in content-based image retrieval using quantized histogram texture features in the dct domain", Journal of King Saud University Computer and Information Sciences, vol. 25, no. 2, pp. 207-218, 2013.

[16] H. Muller, W. Muller, S. Marchand-Maillet, T. Pun, and D. McG. Squire, "A framework for benchmarking in visual information retrieval," International Journal on Multimedia Tools and Applications, vol. 21, no. 1, pp. 55-73, 2003.

[17] S. M. Youssef, Mesbah S., Mohmoud Y. M., "An efficient content-based image retrieval system integrating wavelet-based image sub-block with dominant colors and texture analysis", In 8th International Conference on Information Science and Digital content Technology (ICIT), vol. 3, IEEE, pp. 518-523, 2012.

[18] S. M. Jadhav, V. Patil, "An efficient content-based image retrieval (CBIR) system based on evolutionary programming (EP)", IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), pp. 310-315, 2012.

[19] D. Giveki, A. Soltanshahi, F. Shiri, and H. Tarrah, "A new content based image retrieval model based on wavelet transform", Journal of Computer and Communications, vol. 3, pp. 66-73, 2015. http://dx.doi.org/10.4236/jcc.2015.33012.

[20] J. Q. Alnihoud, "Image Retrieval Model Based on Color and Shape Features Utilizing 2D Wavelet Transform", World of Computer Science and Information Technology Journal (WCSIT), vol. 7, no. 4, pp.20-25, 2017.

# Assessing Trends of Existing Research Contribution Towards Internet-of-Things

Bhagyashree Ambore (Assistant Professor)[1], Dr. Suresh L (Principal and Professor)[2]
Dept of computer science & Engg.
Cambridge Institute of Technology
Bangalore, India

***Abstract*—With the growing demands of system automation, technology integration, and non-human intervention technique, Internet-of-Things (IoT) has evolved as a boon and value-added services over pervasive computing. IoT comprises a highly complex system that integrates ubiquitous computing with low-powered data capturing devices via a gateway. Along with various forms of unimaginable advantages, IoT is also associated with a huge list of ongoing problems. The prime objective of the paper is to gauge the effectiveness of existing works of literature being carried out towards mitigating the issues of IoT. The paper illustrates the most frequently explored research topic and less regularly explored topic in IoT for providing a true picture of existing research trends. The paper also idealizes some of the research gaps that have been extracted after reviewing the existing literature.**

***keywords*—Bandwidth; cloud computing; energy; internet-of-things; security; sensor network**

## I. Introduction

The area of network and communication system is in the faster process of evolution and has witnessed tremendous advancements in its technologies in most recent years. IoTor commonly known as IoT is one of such technological advancement, which has been born by the amalgamation of network and various forms of devices to capture data [1]. Technically speaking, IoT is a sophisticated network of low-powered embedded devices with the large connectivity of the network. Fig.1 shows one of the typical schemas of IoT. The schema shows four different top-down layers. The bottom layer is the sensors, which is responsible for capturing the raw information from the environment, which is connected to the IoT devices in the next upper layer. The data is finally aggregated from IoT devices and forwarded to the wired or wireless terminal, which uses gateway services to process the data to the top layer. The top layer is the data center which offers various forms of cloud-based services over the internet to the users [2]. Hence, IoT primarily assists the sensors to capture the information and forward it to the user using cloud services [3]. However, IoT is not that easy as it seems like. The term *thing* in IoT is not necessarily sensors; it could be any form of a low-powered device. However, it should be understood that IoT is a very futuristic technology that enables fair communication between two different kinds of machines with the incorporation of the higher degree of system automation and thereby avoid intervention from human [4].

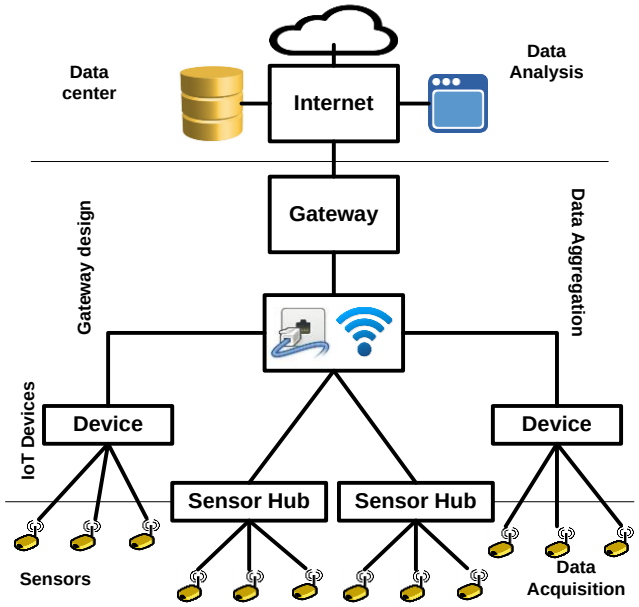



Fig. 1. IoT Typical Schema.

Usage of IoT will lead better task automation, better return on investment, however, it is also associated with issues, e.g., i) IoT doesn't provide any standardization for tagging, ii) IoT has the most challenging complexity in its system design ( a minor fault in the system will result in entire malfunction), iii) privacy and security standard could be easily compromised, iv) device compatibility is another severe issues in IoT, and v) lack of human intervention. Although IoT is a very novel technology, it is also one of the burning topics among the researchers. Hence, this paper intends to understand the existing research work towards IoT and targets to understand the effectiveness of the existing research contribution. The prime contribution of this paper is to identify the scale of the effectiveness of existing research work. Section II discusses the important information pertaining to IoT very briefly with respect to its essential characteristics and existing frameworks. The standard IoT architectures are discussed in Section III. Section IV discusses existing survey work in IoT followed by existing research trends in Section V. Section VI addresses the research gap identified from critical analysis of existing research trend in IoT. Finally, Section VII concludes the paper with the idea of future work.

## II. ESSENTIALS OF IOT

IoT is one of the upcoming trends of technological advancement, which has started created news right from now. IoT is the connectivity of various forms of electronic devices, which are terms as *the thing* with a unique identification. This electronic device senses certain physical attribute, process them using various software and transmits the information using heterogeneous or homogeneous networking protocols [5]. The prime purpose of IoT is not only to sense but also to perform certain controlling mechanism over the things remotely. IoT essentially creates a bridge between computing and the physical world. The application areas of IoT involves smart cities, home automation, smart manufacturing, health care, wearables, automotive, transportation, etc. [6]. Various technologies assist in enabling the operationality in IoT applications. RFID (Radio Frequency Identification) is one of the best enabling technology in IoT that connects various objects with internet [7]. Near-Field Communication is another enabling technology of IoT applications [8]. The existing system also uses cost-effective communication schemes, e.g., QR codes, optical tags, etc. [9]. Bluetooth and sensors are the most recent evolution of the existing enabling technologies of IoT. Along with novelty, cut-edge features, there are associated challenges with IoT.

### A. Characteristics of IoT

The application that uses IoT basically encounters various challenging scenarios (discussed in the next sub-section). Owing to the novelty in the technology, it is essential to understand the most critical characteristics of IoT, which are as discussed briefly as follows,

- *Usage of Intelligence:* The fundamental design principle of IoT is based on autonomous control and Ambient Intelligence. The future applications of IoT are expected to be highly self-organization with the interoperable virtual object, where circumstances, context, and environments play a significant role.
- *Seamless Connectivity:* The different embedded devices that are connected by IoT are required to possess undisturbed accessibility.
- *Potential Sensing Capability:* Sensing a particular physical world attribute is one of the essential features of IoT. The sensed data are considered as input for the majority of IoT applications.
- *Massive Number of things:* The upcoming IoT comprises of connectivity of more than millions of sensing and controlling devices.
- *Energy Efficiency*: Incorporations of the higher degree of energy conservation is another significant feature of IoT as the majority of the IoT applications operate on adverse environmental condition that calls for unattended operation.
- *Secure Network:* Owing to the inclusion of multiple forms of networking and data processing protocols, potential security features become one mandatory target in IoT.

### B. IoT Frameworks:

The framework involved in IoT assists in formulating the interaction between the devices (or *things*) and permit for better-sophisticated supportability of distributed computing. Some of the well-known structures of IoT are:

- *Jasper:* It is one of the frequently used frameworks that provide an operational platform of rending the communication system among the devices. It is used on automotive applications in IoT utilizing the cloud. The framework assists car manufacturing organization to surveil the defects, insignificant correctness features in automotive. It also checks for successful security incorporations in transportation. A reputed organization, e.g., GE aviation, Coca-Cola, Audi, etc. already use Jasper for offering better services in their products [9].
- *Arrayant:* It is a form of the framework that assists in connecting the services or products with the manufacturer using the internet. It is delivered along with the framework with respect to SaaS. It also consists of a managed cloud for assisting device connectivity, software toolkits for developing web applications, and software library to connect the device with service on the internet [10].
- *AggreGate:* It is a computational framework that is used for managing various forms of embedded devices with multiple forms of data. It is mainly used in the manufacturing organization. Along with controlling various devices, it also offers automation, network management, monitoring attendance, managing data center, managing fleet, controlling sensor network, management of the mobile device, and controlling physical access [11].
- *Xively:* It is another frequently used framework for IoT that has the potential to connect any devices for carrying out communication with other particular devices using the internet. It also offers a cloud-based service (e.g., PaaS) for IoT-based services, e.g., data services, security engine, directory services, etc. Xively can be used with open source libraries with hardware and various APIs [12].
- *Carriots:* It is a software framework with the uniqueness of application hosting and features of PaaS for IoT applications. The framework is known for its capability of collecting valuable data from the devices and then processes it to make it suitable for a specific IT infrastructure. It is characterized by custom alerts, device management, SDK applications, API management, data export, etc. [13].
- *Everything:* It is another typical IoT framework that can access data and control any for any physical devices. It can perform integration of tags, SDK, and controllers. The administration, as well as analytics, also characterizes it. This framework is used for real-time data management, managing various product connections, integration with multiple forms of

hardware devices connected with internet, Cloud PaaS, Analytics and Administration, and security (or access) control [14].

## III. IOT ARCHITECTURE

The architecture followed by an organization to undertake IoT solution for their business to run is termed as *Reference Architecture*. It is a customizable architecture that defines the essential characteristics of required performance, critical functional requirements, execution, and security incorporation with industry-based standards [15]. Fig.2 highlights existing reference architecture of IoT, which is the base of all the well-defined architectures. One of the essential blocks of this architecture is called as *Reference Model*, which is again build up by three components, i.e., business vision, IoT Reference Model, and IoT Reference Architecture. The usage of IoT Architectural Reference model is shown in Fig.3, where it can be seen that it acts as enabling IoT architectural schema in system design of IoT-based applications. The system design, however, takes the input from the use cases and requirements which depend on the business concerns. The system design is also enabled by significant engineering strategies of usage of multiple technologies to make it operational. All these processes finally lead to the generation of concrete architecture for a specific, concrete architecture of IoT. The business vision component comprises all the essential requirements of business that finally acts as the industry standard to control the architecture. The component of *the Reference Model* furnishes higher abstraction level for supporting a comprehensive understanding of IoT domain. The component of the reference architecture is considered as the building block of all the major architectures of IoT. Various existing architectures, concerns of business as well as solutions are considered as input for existing IoT Architectural Reference Model via *SOTA* (Software updates Over The Air) using cloud services. After performing extrapolation, all these preliminary requirements transform itself into a single and joint requirement for reference architecture of IoT. Hence, the most critical part of the IoT architecture is to perform a unified understanding of multiple domains of IoT in terms of modeling.

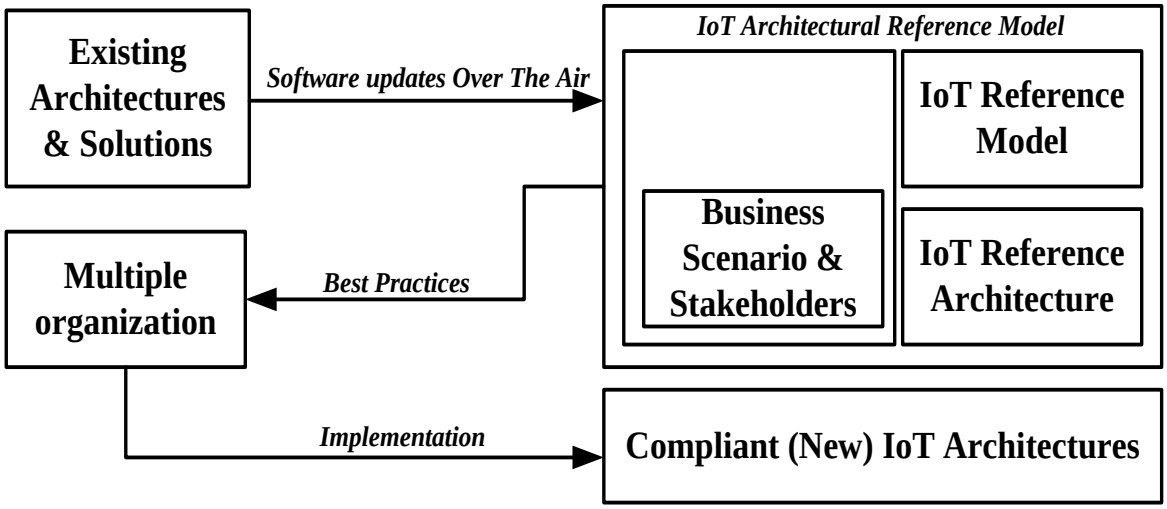



Fig. 2. Existing Reference Architecture of IoT [15].

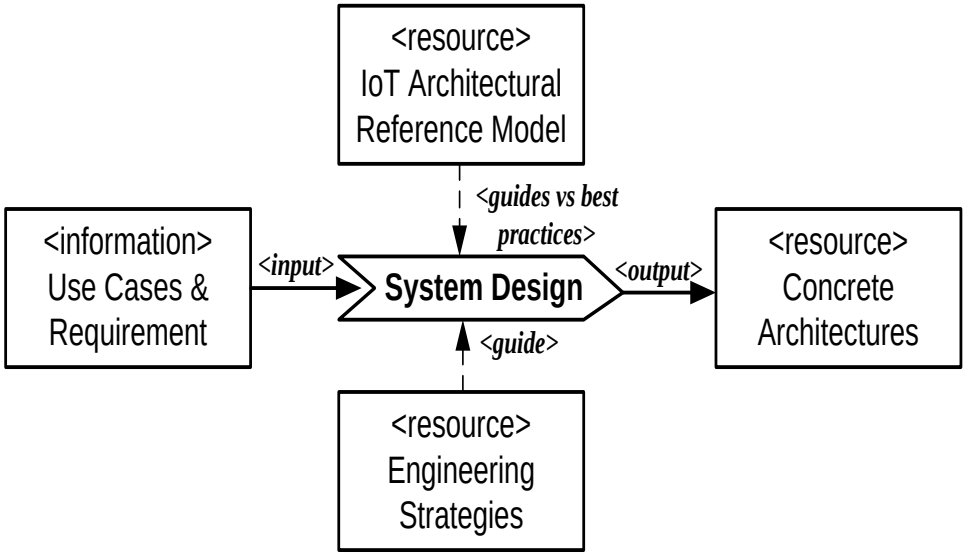



Fig. 3. Formation of Concrete Architecture [15].

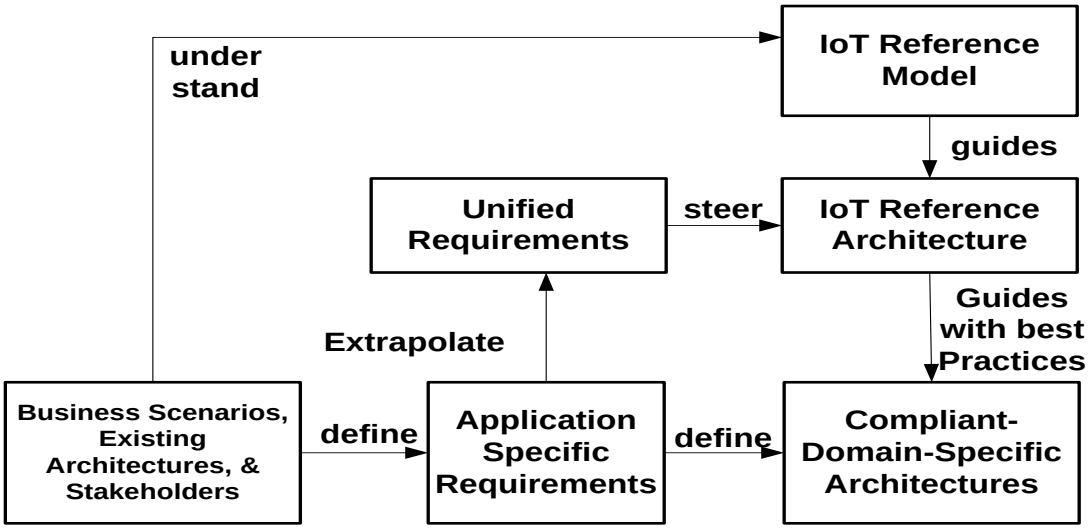



Fig. 4. High Level Design of IoT Reference Architecture [15].

The high-level design of IoT reference architecture is showcased in Fig.4. The backbones of this design level are two important controlling, i.e., i) *dynamic modeling* and ii) *functional modeling*. The first form of modeling results in Reference Model while the second form of modeling assists in Reference Architecture. One interesting feature of IoT architecture is also to provide use cases along with the generation of concrete architectures for a specific application. Another benefit of this architecture is its response system that provides the feedback of inconsistencies (if present within the architecture).

The design of the IoT reference model is carried out using the approach of the Spiral model[16]. The process of abstraction being carried out in higher level with respect to the domain is carried out using industry standard methodology, e.g., SysML [17], Model Driven approach [18], Aspect Oriented approach [19], and Pattern based approach [20]. Out of all these methodologies, Model-driven approach is widely used in the development of IoT architectures. The prime task of Model-driven approach is to transforms any platform independent model to platform-specific model. The aspect-oriented approach is another frequently used methodology in IoT architecture design which performs segregation of all critical supporting functions from the core logic of function in IoT architecture. The pattern-based approach is another frequently selected method in designing IoT architecture. It mainly reutilizes various recursive solutions to sort out generically evolved issues in IoT architecture functionalities. The following Table.1 gives the recently adopted methodologies for IoT architecture.

TABLE I. FREQUENTLY USED THE METHODOLOGY IN IOT ARCHITECTURE

| Methodologies | Responsibility |
|---|---|
| Model-Driven | Transformation for generic to specific architectures |
| Aspect-oriented | Functionality delineation |
| Pattern-based | Testing method efficiency |

Hence, deployment of IoT architectural model significant uses design patterns to make the implementation lot easier. Therefore the advantages of the use of existing IoT Reference architectural Model are as follows:

- *Cost Effective deployment:* IoT architecture provides a universal ground for multiple IoT system on one single IoT Reference Model.
- *The capability of Cognition:* IoT architectures can provide significant information about the generated architectural robustness.
- *Easier Deployment:* It can generate various specific concrete IoT architectures that make the decision of implementation quite easier.
- *Benchmarking:* The reference model of IoT Architecture can be used as benchmarks for the application specific IoT architectures.

## IV. EXISTING SURVEY WORKS IN IOT

This section discusses the existing survey work in the IoT domain to find the proliferation and advancement of research work in the same field. As IoT is one the most interesting research topic, it has attracted various researchers in the past to undergo investigation on the same. There are around 17 review papers associated with IoT issues and advancement; however, the present paper disucsses13 most relevant survey paper on IoT published most recently. However, judging the informative contents of the survey paper was the most difficult one. Hence, we choose to use a Likert scale of 1-5 (1-less information to 5 in more informative) to understand the following informative parameters,

- $P_1$ represents extents of theoretical discussion
- $P_2$ represents the extent of the implementation discussion
- $P_3$ comparative discussion of multiple studies
- $P_4$ identification of research gap

TABLE II. EXISTING SURVEY WORK ON IOT

| Year | Author | $P_1$ | $P_2$ | $P_3$ | $P_4$ |
|---|---|---|---|---|---|
| 2010 | Atzori et al. [21] | 5 | 1 | 0 | 0 |
| 2010 | Yang et al. [22] | 5 | 0 | 0 | 0 |
| 2013 | Aggarwal et al. [23] | 5 | 1 | 0 | 0 |
| 2013 | Said and Masud [24] | 3 | 1 | 0 | 0 |
| 2013 | Perera et al. [25] | 4 | 3 | 2 | 1 |
| 2014 | Whitmore et al. [26] | 2 | 4 | 0 | 0 |
| 2014 | Kumar & Patel [27] | 2 | 2 | 0 | 0 |
| 2014 | Pande & Padwalkar [28] | 2 | 0 | 0 | 0 |
| 2014 | Shah & Ambareen [29] | 2 | 0 | 0 | 0 |
| 2015 | Gazis et al. [30] | 2 | 0 | 0 | 0 |
| 2015 | Botta et al. [31] | 3 | 3 | 0 | 0 |
| 2015 | Fremantle & Scott [32] | 2 | 3 | 1 | 0 |
| 2015 | Karagiannis et al. [33] | 2 | 1 | 0 | 0 |
| 2016 | Bizanis & Fernando [34] | 2 | 3 | 1 | 0 |
| 2016 | Luong et al. [35] | 4 | 4 | 2 | 0 |
| 2017 | Akpakwu et al. [36] | 5 | 3 | 1 | 1 |
| 2017 | Al-Turjman[37] | 1 | 2 | 0 | 0 |
| 2017 | Ni et al. [38] | 4 | 4 | 1 | 0 |
| 2017 | Xu et al. [39] | 2 | 4 | 1 | 0 |
| 2017 | Verma et al. [40] | 4 | 4 | 1 | |
| 2017 | Tokognon et al. [41] | 2 | 4 | 1 | 0 |
| 2017 | Sezer et al. [42] | 4 | 3 | 1 | 0 |
| 2017 | Udoh and Kotonya [43] | 4 | 0 | 0 | 0 |
| 2018 | Alioto and Shahghasemi [44] | 4 | 3 | 1 | 0 |
| 2018 | Jeon et al. [45] | 2 | 4 | 1 | 0 |
| 2018 | McKee et al.[46] | 1 | 1 | 0 | 0 |
| 2018 | Saari et al. [47] | 1 | 1 | 0 | 0 |
| 2019 | Benkhelifa et al.[48] | 1 | 2 | 0 | 0 |

The objective of our search for the survey paper was essential to understand the best work done to date in the form of implementation and unsolved problems with respect to the research gap. Table 2 will give a complete highlight of our investigation for existing review papers on IoT.

Table 2 shows that review work done by Perera et al. [25] is the only work to date who have discussed the identification of the research gap. There are only two researchers Fremantle & Scott [32] and Perera et al. [25] found to address the comparative analysis of multiple priorly presented research work. Apart from work done till the year 2015, there is a diversified area on which survey was carried out with respect to IoT. The different categories are software-defined network (Bizanis & Fernando [34]), data aggregation (Luong et al. [35]), 5G (Akpakwu et al. [36]), localization (Al-Turjman [37]), security of fog computing (Ni et al. [38]), clustering techniques (Xu et al. [39]), analytics (Verma et al. [40]), etc. However, more work in focused on communication and its associated performance, but less towards security approaches. All the survey papers published to date has more emphasis on the theoretical aspects of IoT with a lesser context of discussing implementation work by other researchers from the viewpoint of solutions.

## V. EXISTING RESEARCH TRENDS

This section discusses the exiting research trends in IoT application. The prime reason for this attempt is to understand what is the scale of the effectiveness of the investigations that have been already made. To adhere to the standard of research manuscript, we investigated the availability of research papers from only reputed international Journals only.

TABLE III. RESEARCH ARCHIVES IN SPRINGER FOR 2010-18

| | |
|---|---|
| Chapter | 71,022 |
| Article | 24, 939 |
| Protocols | 110 |
| Reference work entry | 1134 |
| Book | 29 |
| Book Series | 1 |

TABLE IV. RESEARCH ARCHIVES IN SCIENCEDIRECT FOR 2010-18

| Year | No. of Journals |
|---|---|
| 2010 | 2566 |
| 2011 | 2914 |
| 2012 | 3113 |
| 2013 | 3439 |
| 2014 | 4051 |
| 2015 | 4736 |
| 2016 | 5032 |
| 2017 | 6167 |
| 2018 | 5353 |

Table 3-4 shows the amount of the existing research work being carried out in Springer and ScienceDirect. Reputed IEEE Xplore was found to exhibit the hit of 23,168 manuscripts with the keyword “Internet of Things” published between 2010 and 2018, where there is 18,434 conference paper, 3,889 journals, 551 early access articles, and 275 e-books. With similar keywords, there are 20, 928 journals in Science Direct, 24, 939 Journals on Springer, and 25, 250 journals in ACM Digital Library. The above values are approximated owing to the match with the keywords. The relevancy of the actual content is quite less when appropriately checked with the numbers and abstract of such massive numbers. Below sections and discussion in it will give the correct values of existing research trends in IoT.

### A. Frequently Investigated Problems

At present, there is a certain research area which has successfully received the attention of the research communities. This section discusses the problems that have been investigated most frequently in the area of IoT applications.

- *Cloud-based Integration:* Cloud is the prime backbone of the IoT-based applications. Lopez and Macias [49] have carried out a study of IoT framework considering cloud as the major component is the framework. A unique study carried out y Poorter et al. [50] has presented a technique which leverages SOA (Service-Oriented Architecture) of IoT, where cloud services play crucial roles. The study conducted by Mitton et al. [51] has also emphasized on the joint integration of wireless sensors as well as cloud using simulation-based study. Kim et al. [52] have developed another joint integration of cloud and mobile networks using game theory. The studies have also focused on resource allocation dynamically. The significance of virtualization in the cloud has been emphasized by

Abeele et al. [53] by integrating with the sensor network. Hence, there are various archives in reputed journals with more number of publications being focused on the integration of cloud and sensors mainly. However, there are various other computing devices which could also be used to be integrated with the cloud apart from sensors, which has not been focused on existing studies.

- *Data mining Services:* Emphasis on data mining and existing techniques of data mining approaches over IoT can be seen in the most recent review of Chen et al. [54]. This paper has reviewed more than 100 publications that have focused on implementing data mining techniques over cloud integrated with IoT domain. Khan et al. [55] have focused on analytics of large and heterogeneous data over the cloud to be used in IoT applications. Bin et al., [56] have also discussed various research trends on the usage of data mining on IoT applications. The authors have discussed various standard data mining models that are frequently used in IoT. Deployment of semantics in knowledge-extraction process was seen in the study of Bove et al. [57]. The authors have applied it to the RFID based network in IoT. Hachem et al. [58] have studied various schemes of ontologies in IoT. The study has essentially discussed three forms of ontologies used in IoT, i.e., global ontology, device ontology, physics domain ontology, and estimation ontology. Serrano et al. [59] have discussed prominent challenges in IoT with respect to interoperability on semantics. Cretu [60] have presented a semantic web-based application for smart cities of IoT. Hence, numerous studies have repeatedly been investigated in exploring better data mining approaches in IoT.

- *Middleware-based approach:* Integration of large network with numerous devices to capture data is the prime function of IoT, where middleware plays a crucial role in data processing, load balancing, and security management. Work done by Huo et al. [61] has discussed the significance of middleware-based application. Such forms of studies were mainly focused on interoperability and issues associated with the integration of the IoT devices. Lim and Park [62] have developed unique middleware services for performing the sharing of significant resources while integrating cloud and sensors in IoT-based applications. Hachem et al. [63] have introduced a novel middleware system that is motivated by service-oriented architecture in IoT. The middleware was used for mainly governing a large number of the mobile devices in IoT applications.

### B. Less Explored Problems

There are few problems in IoT that has been less emphasized in the research area of IoT. Following are some of the problems that have received quite a less attention in the research work.

- *Bandwidth Issues:* It should be noted that 99% of the applications over IoT runs over wireless connectivity. The system allows machine-to-machine communication using existing wireless standards, e.g., Bluetooth, LTE, WLAN, RFID, etc. IoT comprises numerous consumer devices that are connected via the internet. The existing commercial users who depend on wide area network will need to expand their channel capacity soon to fill the gap in bandwidth. At present, the usage of the mobile application, services, and the network is tremendously on the rise and is already creating havoc in traffic management. However, adding to data communication in IoT will further increase the channel capacity, which is quite practically difficult to increase or manage. Moreover, in reality, less than 1% of the available bandwidth existing in the network is being utilized in IoT application. Most recently, the need of larger amount of bandwidth in IoT-based applications is supported by the 3G/4G network. However, it cannot support the integration of the heterogeneous physical devices in IoT, e.g., sensors. The networking and telecommunication services at present don't bear enough capacity to carry the increasing load of traffic of IoT. Bandwidth is one of the significant factors that can potentially impact the performance of IoT applications. The preliminary impact of poor bandwidth will come over data center. Although data centers are there for massive storage, it is not ready for incoming or outgoing data transmission from IoT based applications. It is quite challenging to understand as if 1000 sensors producing data on every one second if IoT comes in commercial usage. Hence, such forms of IoT data transmission may result in high degradation enough to jam the entire services to one data center in 1 day itself. This example is cited only for the sensor network; hence it is almost nightmare to consider other forms of sensing device which captures and transmits data in every second. Hence, although there is massive research work done over conventional bandwidth management system in the normal network as well as cloud, it is essential to emphasize even for IoT also, which has not received considerable attention.

- *Energy Issues:* The IoT devices are majorly low-powered hardware with resource constraint. A closer look into the existing works of literature found that 85% of the existing research work has been focused on using wireless sensor network and rest 15% towards RFID in IoT domain. A battery with limited lifetime powers both the forms of devices (sensors and RFID reader/tags). In a wireless sensor network, it is said that the core design of the sensor is built based on the radio-energy model [64]. According to radio-energy model, it is believed that energy parameter is closely linked with communication in one sensor. This will mean that if the energy dissipates unwantedly than the communication will degrade too thereby reducing the network lifetime of sensors. The biggest problem in IoT pertaining to sensors is related to heterogeneous profiling of its devices. Energy consumption for sensors is quite different from that of the RFID-based device as well as mobile devices. Hence, although at present we have solution towards controlling energy drainage, it is the only applicable inhomogeneous network. No standard

energy aware technique can ensure an efficient controlling of power dissipation. Moreover, such devices are often free from human intervention, which will mean that if the devices are saturated with its battery drainage than there is no way that it can be physically or remotely recharges. Although there is an exception in this case too owing to energy harvesting technologies. However, even energy harvesting technology will require standard external storage and a robust algorithm to decide the need for charging dynamically, when needed. A simple node doesn't have that much memory to execute such complex and sophisticated algorithms. Hence, there is a more significant gap between memory, energy, and computational requirements in IoT devices. Although there is extensive literature approximately 71,117 published between 2010-2018, there are few research implementations towards conserving energy control over IoT-based applications.

- *Security Issues:* Security has always been a constant concern right from the beginning in IoT-based applications. At present, there are more than 10,000 research papers published in the last five years related to security protocols in a wireless network, but very few studies that have proven robust security techniques over IoT-based applications. To enable a better range of security, it is essential that IoT devices must have better access control mechanism, robust and scalable firewall system, effective intrusion detection/prevention system, and potential and fail-proof authentication of IoT devices with secured booting of devices. Researchers have addressed none of these in the last five years or before that. Owing to the inclusion of heterogeneous devices, it is almost impossible to develop a generic algorithm that can provide full-fledged security solutions to the entire ranges of IoT devices. Hence, developing a robust security protocol for 1000 (example) sensors of different types will be a huge expenditure and is quite infeasible owing to integration problems or data processing problems. Moreover, there is no assurity that the developed security protocol can resist the potential threats over the internet, a place where almost every day, thousands of malicious Trojans takes birth and reproduce in the network without even any single alarms. Moreover, there is a bigger trade-off between the security protocols and communication in existing IoT applications. There is a need for the cost-effective solution, which is quite a far from really looking into the existing trends of solutions. Usage of AES, SHA, DES and all form of cryptographic algorithms are already in use, which is already reported of various security threats. Moreover, the dependency of cloud-based services posses another reason for security breaches in IoT applications.

Hence, it is important to understand the extent of research work being carried out in the above three areas in IoT, which have received less attention.

*1) Studies on Bandwidth Issue*: Bandwidth plays an important role in the communication module of IoT. Table.5 highlights the existing studies in bandwidth issues in IoT. At present, there have been various studies that have focused on bandwidth issues on WLAN [65] along with an emphasis on bandwidth allocation schemes [66]. There are also studies focused on optimizing bandwidth on wireless sensor network [67]. However, studies on bandwidth management in IoT are quite a few to find. This section will discuss 11 research papers that are found to be associated with bandwidth management in IoT. Jin et al. [68] have discussed the emergence of various impediments that calls IoT to possess more work towards bandwidth management. Authors have also theoretically discussed 4 types of architecture, i.e., ubiquitous network, application layer overlay network, autonomous network, service-oriented network. Athreya et al. [69] have presented a technique that allows the devices connected in IoT to organize themselves. The authors have also presented an empirical formulation of self-adaptation with reprogrammable interfaces. The [69] framework analysis and control agent that is connected with radio agent and link agent are given mainly to perform self-organization of IoT devices. In the end, the authors have also discussed the various challenges associated with self-configuration of self-organization of IoT devices. Studies considering the wireless sensor network and its possible involvement in IoT are seen in the work carried out by Zhou et al. [70]. Although the work has focused on minimizing energy consumption among the sensors, the study was performed with a problem identification of bandwidth allocation. The authors have presented simple empirical modeling with outcome tested using energy. Deepalakshmi and Rajaram [71] have introduced a tree-based technique to reserve a good amount of bandwidth in the multistage network like IoT. Xu et al. [72] have addressed the bandwidth problems in an IoT-based multimedia streaming application using delay parameter over the sensor network. The outcome of the study was also compared with a round robin to find reduced computational complexity. Most recently, a research paper of Zachariah et al. [73] has discussed the practical problems of IoT, which is related to the gateway between software and hardware. This problem has a close connection with the IoT for not supporting devices with low-bandwidth. The authors have used the protocol of Bluetooth (IEEE 802.15.4) for profiling gateway. Saeed et al. [74] have presented a novel technique that supports the integration of multiple IoT devices with a focus on task scheduling. Thomas and Irvine [75] have carried out an investigation of bandwidth allocation consider LTE networks as well as the sensor network. Khan et al. [76] have presented a technique for reserving bandwidth over the cloud. The concept is very much close to IoT applications. Jun et al. [77] has developed a scheme for bandwidth allocation for IoT along with cloud using game theory. Yang et al. [78] have discussed the dependencies of the bandwidth factor with respect to IoT based applications. The recent work carried out by Xu et al. [79] has addressed the usage of orthogonal frequency division multiplexing for optimizing the use of bandwidth over noisy channels connected with IoT-device.

*2) Studies on Energy Issues*: Majority of the devices connected in IoT are operated in low power with batteries, whose lifetime is quite limited. In the existing system, there are various mechanisms that address the issues of energy consumption. There is couple of studies that correlates energy problems with wireless sensor network [80], [81], [82], [83], [84].

Studies concerning the control and management of energy in the viewpoint of IoT applications are less significant in existing works of literature. Table 6 discusses the existing studies on energy management in IoT. Karnouskos [85] have discussed the smart grid applications and discussed the prolonging challenges in IoT. A similar direction of the work is also carried out by Weiss [86]. Sun et al. [87] have presented a scheme that can govern the energy consumption owing to frequent dynamics of the duty cycle with respect to the sensor network. The authors have discussed a greedy technique to accomplish energy conservation. Machado et al. [88] have presented a unique communication protocol in IoT, where the quality of the established link decides the richness of the communication. The study outcome was found with increased packet delivery ratio with energy efficiency. Gorlatova et al. [89] have presented a technique that harvests energy from the kinetic sources on IoT devices. The authors have used real-time prototypes of sensors to perform energy management. Devasenapathy et al. [90] have investigated the possible influence of directionality of antenna as well as energy harvesting on IoT devices. The study introduced a technique to understand the amount of energy required for exploring neighbor nodes. Pabbuleti et al. [91] have essentially investigated some of the prominent security aspects and developed a framework to evaluate the amount of energy required to process it. Usage of WLAN is one of the cost-effective solutions for powering up the low powered devices to IoT. The study introduced by Kellogg et al. [92] is of similar direction. The authors have used real-time Wi-Fi routers to investigate the rate of communication. The performance parameters of the study were tested with the data rate, which is found to decrease with an increase of the distance between the IoT devices and WLAN router. Bin and See [93] have presented a design of control system using a middleware system. Kim et al. [94] have presented a unique framework of energy management for home automation applications in IoT. The study is found to use middle-based approach for conserving energy. The implementation of the study is made over real-time hardware for optimizing power requirements over the photovoltaic panel. Hence, there are various studies that have focused on energy management of the IoT enabled devices in terms of networking. Conserving the maximum amount of residual power is extremely important for IoT enabled applications pertaining to healthcare and industrial automation. Alsaryrah et al. [95] have presented an optimization-based solution towards addressing the energy problems in IoT. Mozaffari et al. [96] have also presented a technique of energy efficiency considering the case study of aerial vehicles. The recent literature by Roy et al.[97] have presented a discussion on sustainable IoT factors where a communication strategy has been presented to support IoT-based communication. Shafique et al. [98] have discussed the importance of energy harvesting in IoT devices using Rectenna-based approach. A complete prototype has been designed and fabricated for this purpose.

TABLE V. EXISTING STUDIES ON BANDWIDTH ISSUES IN IOT

| Authors | Techniques | Advantages | Limitation |
|---|---|---|---|
| Jin et al. [68] | Conceptual Discussion about network architectures | Theoretically sound discussion | No focus on implementation |
| Athreya et al. [69] | Framework for Self-Configuration IoT nodes | Empirical Modelling | No focus on implementation |
| Zhou et al. [70] | Empirical modeling of bandwidth allocation | Reduced bit-error-rate | The outcome doesn't discuss data delivery, No comparative analysis, and complex computational process due to the iterative method. |
| Deepalakshmi and Rajaram [71] | Tree-pruning for bandwidth management | Better delay performance | Applicability of this algorithm in the heterogeneous network, e.g., IoT is not discussed. |
| Xu et al. [72] | Resource allocation, delay-aware | Reduced computational complexity | The study is done considering homogeneous sensor network. |
| Zachariah et al. [73] | Bluetooth based gateway profiling | Technique supports IoT devices with low bandwidth | Numerical Outcomes and Analysis not discussed in the paper. |
| Saeed et al. [74] | Task scheduling with bandwidth management | Achieved higher RTT values | Numerical Outcomes and Analysis less focused |
| Thomas and Irvine [75] | Bandwidth allocation for LTE based sensor network | Better data dissemination by the simulation study | The outcome is measured with packet dropped only. |
| Khan et al. [76] | Pricing Method to reserve bandwidth | Simple scheduling technique | Applicability of this algorithm in the heterogeneous network, e.g., IoT is not discussed. |
| Jun et al. [77] | Game theory based resource allocation | Supports cellular network and cloud | Applicability of this algorithm in the heterogeneous network, e.g., IoT is not discussed. |
| Xu et al. [79] | orthogonal frequency division multiplexing | Enhance data rate | No benchmarking or extensive analysis |

TABLE VI. EXISTING STUDIES ON ENERGY ISSUES IN IOT

| Authors | Techniques | Advantages | Limitation |
|---|---|---|---|
| Karnouskos [85] | Conceptual discussion | Cost effective Smart Grid design in IoT | Numerical outcomes, benchmarking not discussion |
| Weiss [86] | Conceptual discussion | -N/A- | Numerical outcomes, benchmarking not discussion |
| Sun et al. [87] | Decision making for energy control | Reduced duty cycle, good energy conservation | Computational complexity is higher for the greedy approach |
| Machado et al. [88] | Energy consumption, quality of link-based routing | Reduces energy consumption | Less Applicability on the heterogeneous network |
| Gorlatova et al. [89] | Energy harvesting technique | Better energy consumption | Computational complexity is higher, no benchmarking |
| Devasenapathy et al. [90] | Neighbor discovery using the directional antenna | Energy efficient | Scalability issues not addressed |
| Pabbuleti et al. [91] | Energy minimization for security protocols | Computation overhead optimization | Scalability, benchmarking, complexity not discussed. |
| Kellogg et al. [92] | The energy requirement for connecting IoT devices with WiFi | Cost-effective solution to reuse WLAN in IoT | Reduction in Data rate, dependability on |
| Bin and See [93] | Middleware based energy management | Cost effective home automation | Numerical outcomes, benchmarking not discussion |
| Kim et al. [94] | Middleware based energy management | Better power optimization | Computational complexity is higher, no benchmarking |
| Alsaryrah et al. [95] | Optimization-based | Energy efficiency | Computational complexity is not carried out |
| Mozaffari et al. [96] | Energy efficiency | Better trajectory performance | Computational complexity is not carried out |
| Shafique et al. [97] | Rectenna | Maximum power transmission | No extensive analysis to proof device robustness |

*3) Studies on Security Issues*: Security has always played a critical role in any networking applications and services. For more than a decade there has been extensive research on security protocols, but owing to novelty in the IoT domain, there is an open research question about the success factor of existing security techniques. IoT-based application posses multiple forms of low-powered devices which have their capability of performing encryption and so is its supportability with its connecting network. The biggest challenge in this regards is how to provide a safe encryption mechanism on multiple devices in IoT.

At present, there is already a massive research work being carried out in enabling technologies of IoT, i.e., Wireless Sensor Network, RFID, etc. The recent review work on security issues and challenges involved in secure routing is discussed in [98] [99], while security challenges in RFID based applications are discussed in [100][101][102]. This paper discusses the available research papers that address the security issues in the IoT domain. Katagi and Moriai [103] have presented a discussion on cryptography for securing IoT applications. Khajuria and Andersen [104] have presented a typical encryption technique for securing IoT enabled wireless devices. Developed over FPGA, the authors have used AES (Advanced Encryption Standard) to incorporate security. Yang et al. [105] have adopted PKI (Public Key Infrastructure) and identity-based cryptography for supporting data processing in IoT applications. Saied et al. [106] have presented a security technique using a trust factor for securing communication over IoT domain. The outcome of the study was evaluated with respect to the level of trust in increasing time factor. Kim [107] has presented a unique ciphering scheme at a minimal cost of hardware. Markmann [108] has presented a technique using a smaller length of the digital signature for securing IoT based networks. The study has also used identity-based encryption and outcome is evaluated using energy consumption. Shafagh et al. [109] have adopted homomorphic encryption for promoting privacy on IoT applications. The technique was found to have better compliance of reduced memory usage. Discussion of various standard libraries of encryption is carried out by Kumar et al. [110]. The study has contributed to understanding effective libraries of cryptography to be implemented for securing IoT applications over the internet. Similar usage of homomorphic usage is discussed by Shafagh et al. [111]. Dinu et al. [112] have developed a security framework using block encryption process on real-time ARM processor to testify its effectiveness on IoT applications. Huang and Mu et al. [113] have developed a secure protocol to safeguard RFID-based communication in IoT. The focus of the study was to mitigate forged tag and reader attack, tracking attack, and desynchronization attack using a new distribution of secret key mechanism in cryptography. Hence, it can be seen that there are a good amount of studies that have focused on securing communication over an IoT-based application. The researches formed towards security issues in IoT are given Table.7.

TABLE VII. EXISTING STUDIES ON SECURITY ISSUES IN IOT

| Authors | Techniques | Advantages | Limitation |
|---|---|---|---|
| Katagi and Moriai [103] | Cryptographic-based technique | Applicable for low-powered IoT devices | No numerical analysis presented |
| Khajuria and Andersen [104] | Advanced Encryption Standard | Supports hardware acceleration | No numerical analysis presented |
| Yang et al. [105] | Public Key Infrastructure, Identity-based Cryptography | Lower algorithm complexity | No numerical analysis presented |
| Saied et al. [106] | Trust-based security | Resilient against selfish behavior in IoT | Algorithm complexity not discussed, no benchmarking |
| Kim [107] | Inverse-independent ciphering scheme | Lightweight security protocol | Not resilient against key compromise attacks in IoT |
| Markmann [108] | Digital Signal, identity-based encryption | Storage compliant, | Not enough validation for security keys |
| Shafagh et al. [109] | Homomorphic encryption | Lower memory usage | Higher processing time, less extensive analysis of outcome |
| Kumar et al. [110] | Study of encryption libraries | Good theoretical knowledge about tools | Doesn't have the reflection about its effectiveness on IoT. |
| Shafagh et al. [111] | Homomorphic encryption | Lower memory usage | Higher processing time, less extensive analysis of outcome |
| Dinu et al. [112] | block encryption | Lightweight ciphering process | Not resilient against physical attacks in IoT |
| Huang and Mu et al. [113] | Key distribution | Lightweight ciphering process | No numerical analysis presented |

## VI. RESEARCH GAP IDENTIFICATION

This section discusses the existing research gap towards IoT. The discussion made in this section is an actual outcome of the review of the literature discussed in prior sections.

- *Less Focus on Bandwidth:* Bandwidth or the channel capacity is one of the critical requirements to make an operational success of existing and upcoming IoT-based applications. The existing studies are more focused on various schemes, but practical implementations and applicability on real-time are still questionable. The literature has less focus on numerical analysis with a few comparative performance analysis, for which reason, existing studies can be just treated as better theoretical guidelines but is quite risky to implement followed by enhancing it. Almost all the studies are done in simulation-based, where there is quite less rationale or justification of the values of parameters with almost no validation of the outcomes. Another bigger problem is an adoption of performance parameters. A better schema of channel capacity will lead to the reduction of propagation delay and increase in throughput. This fact is not found in any outcomes of recent implementations on bandwidth management in IoT.

- *Availability of Energy Conservation Scheme:* The existing studies on energy conservation were found to have various implementations towards energy harvesting schemes mainly related to sensors. However, there was no discussion of any possible connection between energy and communication performance. The standards of wireless sensor networks use first/second order radio-energy model, which means the slightest improvement in energy conservation should also enhance the quality of data transmission and data delivery performance. This phenomenon should be included in the performance assessment of any research work focusing on energy efficiency in IoT. Majority of the prior papers have discussed the usage of sensors but without considering these performance parameters. Moreover, other research gaps explored in the studies pertaining to energy conservation schemes are less applicability on the heterogeneous network, higher computational complexity, no benchmarking, and no addressing of scalability issues.

- *Poor Security Standards:* As discussed, the security systems applied over IoT applications are not able to cater up to the potential vulnerability of the malicious codes that circulates on the internet. Some of the papers discussed that WLAN is one of the cost-effective technology assisting in communications in IoT, but it should be known that WLAN uses security protocols like WEP, WPA, TKIP, etc., which are quite obsolete and all are majorly reported of serious attacks. The existing security techniques used in wireless sensor networks are only developed for securing homogeneous connectivity and never heterogeneous connectivity. Hence, the applicability of existing security standard on multiple IoT devices is not resistive against potential threats in IoT applications and calls for serious investigations.

## VII. CONCLUSION

The paper has explored the research trends in IoT applications. The paper starts with briefing the essential characteristics of IoT and various research problems associated with it. From the research analysis it is found that IoT has attracted attention among the research communities, but at the same time, there are also some areas where it has received less focus, e.g., bandwidth, energy conservation, and poor security standards. There is a massive set of research work in all these issues in non-IoT-based applications; however, effective focus

on this with respect to IoT is quite less effective. The paper has explored the hidden problems associated with IoT after reviewing all the significant literature published most recently. With the analysis of the existing researches research problems were incorporated which can be considered for future research. The future work will be towards proposing a novel probabilistic design to schematically parameterize various significant issues in IoT especially emphasizing on channel capacity, energy, and security problems and evolve up design principles to mitigate the issues. In order to accomplish the above mentioned goal, following objectives are targeted viz. i) to apply a probabilistic and strategic decision-making model for signifying the tradeoff between channel capacity and energy efficiency in IoT, ii) to develop an energy-aware trust derivation scheme for securing wireless sensor networks for IoT application, and iii) to provide a method of risk strategy analysis to stimulate the nodes' cooperation thereby minimizing the overhead and maximizing the efficiency suitable for sensors in IoT.

The futuristic scope of the research study is presented as follows

- A system modeling of a novel energy-effective intruder detection and isolation scheme can be analytically designed using robust decision-making principle to address the research gap.
- A simple and yet sophisticated scheme can be formulated for dynamic bandwidth optimization scheme that could offer a higher degree of energy-efficiency.

## REFERENCES

[1] O. Vermesan, P. Friess, Internet of Things: Converging Technologies for Smart Environments and Integrated Ecosystems, River Publishers, 2013

[2] F. Behmann, K. Wu, Collaborative Internet of Things (C-IoT): For Future Smart Connected Life and Business, John Wiley & Sons, 2015

[3] M. Aazam, E-N Huh, M.S. Hilaire, "Cloud of things: Integration of IoT with Cloud Computing", *Springer Journal of Robots and Sensor Clouds*, 2015

[4] S. C. Mukhopadhyay, Internet of Things: Challenges and Opportunities, Springer Science & Business Media, 2014

[5] J. Holler, V. Tsiatsis, C. Mulligan, S. Avesand, From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence, Academic Press, 08-Apr-2014

[6] S. Evdokimov, B. Fabian, O. Gunther, RFID and the Internet of Things: Technology, Applications, and Security Challenges, Now Publishers Inc, 2011

[7] M. Hendry, Near Field Communications Technology and Applications, Cambridge University Press, 18-Dec-2014

[8] M. Henderson, G. Romeo, Teaching and Digital Technologies: Big Issues and Critical Questions, Cambridge University Press, 09-Oct-2015

[9] http://www.jasper.com/iot-service-platform/control-center. Accessed on 20th-Oct, 2015

[10] http://www.arrayent.com/platform/iot-platform-overview/. Accessed on 20th-Oct, 2015

[11] http://aggregate.tibbo.com/solutions/iot-platform.html. Accessed on 20th-Oct, 2015

[12] https://xively.com/. Accessed on 20th-Oct, 2015

[13] https://www.carriots.com. Accessed on 20th-Oct, 2015

[14] https://evrythng.com/. Accessed on 20th-Oct, 2015

[15] M. Bauer, N. Bui, F. Carrez, "Introduction to the Architectural Reference Model for the Internet of Things." Retrieved from http://www.iot-a.eu/public on 20th Oct 2015

[16] A. Bassi, M. Bauer, M. Fiedler, Enabling Things to Talk: Designing IoT solutions with the IoT Architectural Reference Model, Springer, 28-Oct-2013

[17] S. Friedenthal, A. Moore, R. Steiner, A Practical Guide to SysML: The Systems Modeling Language, Morgan Kaufmann, 23-Oct-2014

[18] J. Whittle, T. Clark, T. Kuhne, Model Driven Engineering Languages and Systems, Springer- 14th International Conference, 2011

[19] D. Mouheb, M. Debbabi, M. Pourzandi, Aspect-Oriented Security Hardening of UML Design Models, Springer, 22-Apr-2015

[20] S. Balandin, S. Andreev, Y. Koucheryavy, "Internet of Things, Smart Spaces, and Next Generation Networks and Systems", *Springer- 14th International Conference*, 01-Aug-2014

[21] L. Atzori, A. Iera G. Morabito, "The Internet of Things: A survey", *Elsevier-Computer Networks*, vol.54, pp.2787–2805, 2010

[22] D-L Yang, F Liu,Y-D Liang, "A Survey of the Internet of Things", Researchgate-The 2010 International Conference on E-Business Intelligence, 2010

[23] C. C. Aggarwal, N. Ashish, A. Sheth, "The internet of things: a survey

from the data-centric Perspective", Book Chapter in "Managing and Mining Sensor Data", *Springer*, 2013.

[24] O. Said, M. Masud, "Towards Internet of Things: Survey and Future Vision", *International Journal of Computer Networks*, Vol.5, Iss.1, 2013

[25] C. Perera, A. Zaslavsky, P. Christen, D. Georgakopoulos, "Context Aware Computing for The Internet of Things: A Survey", *IEEE Communications Surveys & Tutorial*, 2013

[26] A. Whitmore, A. Agarwal, L. D. Xu, "The Internet of Things—A survey of topics and trends", *Springer Journal*, 2014

[27] J. S. Kumar, D. R. Patel, "A Survey on Internet of Things: Security and Privacy Issues", *International Journal of Computer Applications*, Vol.90, No 11, March 2014

[28] P. Pande, A. R. Padwalkar, "Internet of Things –A Future of Internet: A Survey", International Journal of Advance Research in Computer Science and Management Studies, Vol.2, Iss.2, February 2014

[29] P. G. Shah, J. Ambareen, "A Survey of Security Challenges in Internet of Things (IoT) Integration with WSN", Australian Journal of Wireless Technologies, Mobility & Security, 2014

[30] V. Gazis, M. Gortz, M. Huber, "A Survey of Technologies for the Internet of Things", IEEE, 2015

[31] A. Botta, W. Donato, V. Persico, "Integration of cloud computing and Internet of Things: A survey", Future Generation Computer Systems, 2015

[32] P. Fremantle, P. Scott, "A security survey of middleware for the Internet of Things", Open Access of PeerJ Preprints, 2015

[33] V. Karagiannis, P. Chatzimisios, F. V-Gallego, "A Survey on Application Layer Protocols for the Internet of Things", Transaction on IoT and Cloud Computing, 2015

[34] N. Bizanis and F. A. Kuipers, "SDN and Virtualization Solutions for the Internet of Things: A Survey," in IEEE Access, vol. 4, pp. 5591-5606, 2016.

[35] N. C. Luong, D. T. Hoang, P. Wang, D. Niyato, D. I. Kim and Z. Han, "Data Collection and Wireless Communication in Internet of Things (IoT) Using Economic Analysis and Pricing Models: A Survey," in *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2546-2590, Fourthquarter 2016.

[36] G. A. Akpakwu, B. J. Silva, G. P. Hancke and A. M. Abu-Mahfouz, "A Survey on 5G Networks for the Internet of Things: Communication Technologies and Challenges," in *IEEE Access*, vol. 6, pp. 3619-3647, 2018.

[37] F. Al-Turjman, "Positioning in the Internet of Things Era: An overview," *2017 International Conference on Engineering and Technology (ICET)*, Antalya, 2017, pp. 1-5.

[38] J. Ni, K. Zhang, X. Lin and X. S. Shen, "Securing Fog Computing for Internet of Things Applications: Challenges and Solutions," in *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 601-628, Firstquarter 2018.

[39] L. Xu, R. Collier and G. M. P. O'Hare, "A Survey of Clustering Techniques in WSNs and Consideration of the Challenges of Applying

Such to 5G IoT Scenarios," in *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1229-1249, Oct. 2017.

[40] S. Verma, Y. Kawamoto, Z. M. Fadlullah, H. Nishiyama and N. Kato, "A Survey on Network Methodologies for Real-Time Analytics of Massive IoT Data and Open Research Issues," in *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1457-1477, thirdquarter 2017.

[41] C. Arcadius Tokognon, B. Gao, G. Y. Tian and Y. Yan, "Structural Health Monitoring Framework Based on Internet of Things: A Survey," in *IEEE Internet of Things Journal*, vol. 4, no. 3, pp. 619-635, June 2017.

[42] O. B. Sezer, E. Dogdu and A. M. Ozbayoglu, "Context-Aware Computing, Learning, and Big Data in Internet of Things: A Survey," in *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 1-27, Feb. 2018.

[43] I. S. Udoh and G. Kotonya, "Developing IoT applications: challenges and frameworks," in *IET Cyber-Physical Systems: Theory & Applications*, vol. 3, no. 2, pp. 65-72, 6 2018.

[44] M. Alioto and M. Shahghasemi, "The Internet of Things on Its Edge: Trends Toward Its Tipping Point," in *IEEE Consumer Electronics Magazine*, vol. 7, no. 1, pp. 77-87, Jan. 2018.

[45] K. E. Jeon, J. She, P. Soonsawad and P. C. Ng, "BLE Beacons for Internet of Things Applications: Survey, Challenges, and Opportunities," in *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 811-828, April 2018.

[46] D. W. McKee, S. J. Clement, J. Almutairi and J. Xu, "Survey of advances and challenges in intelligent autonomy for distributed cyber-physical systems," in *CAAI Transactions on Intelligence Technology*, vol. 3, no. 2, pp. 75-82, 6 2018.

[47] M. Saari, A. M. bin Baharudin, P. Sillberg, S. Hyrynsalmi and W. Yan, "LoRa — A survey of recent research trends," *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 2018, pp. 0872-0877.

[48] E. Benkhelifa, T. Welsh and W. Hamouda, "A Critical Review of Practices and Challenges in Intrusion Detection Systems for IoT: Towards Universal and Resilient Systems," in *IEEE Communications Surveys & Tutorials*.

[49] E. A. Lopez and J A. G Macías, "Mashing up the Internet of Things: a framework for smart environments", *Springer- EURASIP Journal on Wireless Communications and Networking*, vol.79, 2012

[50] E.D. Poorter, I. Moerman and P. Demeester, "Enabling direct connectivity between heterogeneous objects in the internet of things through a network-service-oriented architecture", *Springer- EURASIP Journal on Wireless Communications and Networking*, vol.61, 2011

[51] N. Mitton, S. Papavassiliou, A. Puliafito, and K. S Trivedi, "Combining Cloud and sensors in a smart city Environment", *Springer- EURASIP Journal on Wireless Communications and Networking*, vol.47, 2012

[52] S. Kim, "Nested game-based computation offloading scheme for Mobile Cloud IoT systems", *Springer- EURASIP Journal on Wireless Communications and Networking*, vol.229, 2015

[53] F. V. Abeele, J. Hoebeke, G. K. Teklemariam, "Sensor Function Virtualization to Support Distributed Intelligence in the Internet of Things", *Springer Science-Wireless Personal Communication*, vol.81, pp.1415–1436, 2015

[54] F. Chen, P. Deng, J. Wan, "Data Mining for the Internet of Things: Literature Review and Challenges", *Hindawi Publishing Corporation International Journal of Distributed Sensor Networks*, 2014

[55] Z. Khan, A. Anjum, K. Soomro and M. A. Tahir, "Towards cloud based big data analytics for smart future cities", *Springer-Journal of Cloud Computing, Advances, Systems, and Applications*, Vol.4, Iss.2, 2015

[56] S. Bin, L. Yuan, W. Xiaoyi, "Research on Data Mining Models for the Internet of Things", *IEEE*, 2010

[57] E. Bove, A. Cinquepalmi, D. De Filippis, "A semantic-based framework for RFID-assisted port supply chains", *Toward Emerging Technology for Harbour sYstems and Services*, 2014

[58] S. Hachem, T. Teixeira, V. Issarny, "Ontologies for the Internet of Things", *ACM/IFIP/USENIX* 12th International Middleware Conference, Dec 2011

[59] M. Serrano, P. Barnaghi, F. Carrez, "Internet of Things , *European Research Cluster On The Internet Of Things* ,2015

[60] L-G. Cretu, "Smart Cities Design using Event-driven Paradigm and Semantic Web", *Informatica Economică*, vol. 16, no. 4/2012

[61] C. Huo, T-C Chien, and P H. Chou, "Middleware for IoT-Cloud Integration across Application Domains", *IEEE design and test of computers*, August 2013

[62] Y. Lim and J. Park, "Sensor Resource Sharing Approaches in Sensor-Cloud Infrastructure", *Hindawi Publishing Corporation International Journal of Distributed Sensor Networks*, 2014

[63] S. Hachem. "Service-Oriented Middleware for the Large-Scale Mobile Internet of Things.Mobile Computing". *Universite de Versailles-Saint Quentin en Yvelines*, 2014.

[64] X. Yan, X. Liu, "Evaluating the energy consumption of the RFID tag collision resolution protocols", Springer, Vol. 52, Issue. 4, pp. 2561-2568, 2013

[65] N. Singh, A. K. Singla, "Bandwidth Management in IEEE 802.11 WLAN: A Survey", *International Journal of IT, Engineering and Applied Sciences Research*, Vol.1, No. 1, October 2012

[66] N.Snehalatha, S.A. Julia, P. Rodrigues, "Survey of Bandwidth Management Techniques", *International Journal of Science and Modern Engineering*, Vol.1, Iss.8, July 2013

[67] R. Manjuparkavi, A. Ramya, K. Kalaignanam, P. Sivakumar, "Bandwidth Optimization in Wireless Sensor Networks – A Survey", *Middle-East Journal of Scientific Research*, vol. 23, Iss.7, pp. 1334-1340, 2015

[68] J. Jin, J. Gubbi, T. Luo, and M. Palaniswami, "Network Architecture and QoS Issues in the Internet of Things for a Smart City", *IEEE-International Symposium on Communications and Information Technologies*, 2012

[69] A. P. Athreya, B. DeBruhl, and P. Tague, "Designing for Self-Configuration and Self-Adaptation in the Internet of Things", *IEEE-International Conference Conference on Collaborative Computing: Networking, Applications and Worksharing*, 2013

[70] Yang Zhou, Chuan Huang, Tao Jiang, Wireless Sensor Networks and the Internet of Things: Optimal Estimation With Nonuniform Quantization and Bandwidth Allocation, IEEE SENSORS JOURNAL, VOL. 13, NO. 10, OCTOBER 2013

[71] R. Deepalakshmi, and S. Rajaram, "Effective Heuristic Algorithm For Dynamic Routing And Bandwidth Management Under Quality Of Service Constraints In Multistage Interconnection Networks", *American Journal of Applied Sciences*, vol.11, Iss.3, pp.414-424, 2014

[72] J. Xu, Y. Andrepoulos, Y. Xiao, "Non-stationary Resource Allocation Policies for Delay-constrained Video Streaming: Application to Video over Internet-of-Things-enabled Networks", *IEEE Journal on Selected Areas in Communication*, 2014

[73] T. Zachariah, N. Klugman, B. Campbell, "The Internet of Things Has a Gateway Problem", *ACM*, 2015

[74] A. Saeed, M. Ammar, K. A. Harras, and E. Zegura, "Vision: The Case for Symbiosis in the Internet of Things", *ACM*, 2015

[75] D. Thomas, J. Irvine, "Connection and Resource allocation of IoT Sensors to cellular technology- LTE", *IEEEConference on PhD Research in microelectronics and Electronics*, 2015

[76] A. M. Khan, X. Vilaca, L. Rodrigues, and F. Freitag, "Towards Incentive-Compatible Pricing for Bandwidth Reservation in Community Network Clouds", *GECON*, 2015

[77] H. Jun, Y. Ying, Y. Huifang, "Context-aware resource allocation for device-to-device communications in cloud-centric internet of things, Journal of Chongaing University of Posts and Telecommunication, vol.27, no.4, 2015

[78] W. Yang *et al*., "Narrowband Wireless Access for Low-Power Massive Internet of Things: A Bandwidth Perspective," in *IEEE Wireless Communications*, vol. 24, no. 3, pp. 138-145, June 2017.

[79] T. Xu and I. Darwazeh, "Non-Orthogonal Narrowband Internet of Things: A Design for Saving Bandwidth and Doubling the Number of Connected Devices," in *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 2120-2129, June 2018.

[80] J-Y Chang and P-H Ju, “An efficient cluster-based power saving scheme for wireless sensor networks”, Springer- EURASIP Journal on Wireless Communications and Networking, vol.172, 2012

[81] S. K. Gharghan, R. Nordin, and M. Ismail, “A Survey on Energy Efficient Wireless Sensor Networks for Bicycle Performance Monitoring Application”, Hindawi Publishing Corporation Journal of Sensors, 2014

[82] X. Xu, L. Shu, M. Guizani, M. Liu, and J. Lu, “A Survey on Energy Harvesting and Integrated Data Sharing in Wireless Body Area Networks”, Hindawi Publishing Corporation International Journal of Distributed Sensor Networks, 2014

[83] G. Zhou, L. Huang, W. Li, and Z. Zhu, “Harvesting Ambient Environmental Energy for Wireless Sensor Networks: A Survey”, Hindawi Publishing Corporation Journal of Sensors, 2014

[84] N. Javaid, M. B. Rasheed, M. Imran, “An energy-efficient distributed clustering algorithm for heterogeneous WSNs”, Springer- EURASIP Journal on Wireless Communications and Networking, vol.151, 2105

[85] S. Karnouskos, “The cooperative Internet of Things enabled Smart Grid”, paper of SAP research, 2010

[86] M. Weiss, “Leveraging residential energy management through the Internet of Things”, ESF Exploratory Workshop on The Internet of Things for a Sustainable Future, Vielsalm, Belgium, May 2011

[87] Z. Sun, C. H. Liu, C. Bisdikian, “QoI-Aware Energy Management in Internet-of-Things Sensory Environments”, IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, 2012

[88] K. Machado, D. Rosario, E. Cerqueira, “A Routing Protocol Based on Energy and Link Quality for Internet of Things Applications”, Sensors vol.13, pp.1942-1964, 2013

[89] M. Gorlatova, J. Sarik, M. Cong, “Movers and Shakers: Kinetic Energy Harvesting for the Internet of Things”, Technical report of ArXiv, 2013

[90] S. Devasenapathy, R V. Prasad, V. S Rao, “Impact of antenna directionality and energy harvesting rate on Neighbor Discovery in EH-IoTs”, IEEE- Consumer Communications and Networking Conference, 2013

[91] K. Pabbuleti, D. Mane and P. Schaumont, “Energy Budget Analysis for Signature Protocols on a Self-powered Wireless Sensor Node”, Springer Journal, 2014

[92] B. Kellogg, A. Parks, S. Gollakota, “Wi-Fi Backscatter: Internet Connectivity for RF-Powered Devices”, ACM-SIGCOMM, 2014

[93] F. C. Bin, “Energy Awareness Architecture for IoT Home Sensor Network”, The 3rd National Graduate Conference (NatGrad2015), Universiti Tenaga Nasional, Putrajaya Campus, 8-9 April 2015.

[94] J. Kim, J. Byun, D. Jeong, “An IoT-Based Home Energy Management System over Dynamic Home Area Networks”, Hindawi Publishing Corporation International Journal of Distributed Sensor Networks, 2015

[95] O. Alsaryrah, I. Mashal and T. Y. Chung, "Bi-Objective Optimization for Energy Aware Internet of Things Service Composition," in *IEEE Access*, vol. 6, pp. 26809-26819, 2018.

[96] M. Mozaffari, W. Saad, M. Bennis and M. Debbah, "Mobile Unmanned Aerial Vehicles (UAVs) for Energy-Efficient Internet of Things Communications," in *IEEE Transactions on Wireless Communications*, vol. 16, no. 11, pp. 7574-7589, Nov. 2017.

[97] S. S. Roy, D. Puthal, S. Sharma, S. P. Mohanty and A. Y. Zomaya, "Building a Sustainable Internet of Things: Energy-Efficient Routing Using Low-Power Sensors Will Meet the Need," in *IEEE Consumer Electronics Magazine*, vol. 7, no. 2, pp. 42-49, March 2018.

[98] K. Shafique *et al.*, "Energy Harvesting Using a Low-Cost Rectenna for Internet of Things (IoT) Applications," in *IEEE Access*, vol. 6, pp. 30932-30941, 2018.

[99] A. Araujo, J. Blesa, E. Romero and D. Villanueva, “Security in cognitive wireless sensor networks. Challenges and open problems”, *Springer- EURASIP Journal on Wireless Communications and Networking*,vol. 48, 2012

[100] F.Mezrag, M.Benmohammed, B.Bouderah, “A Short Survey on Secure Routing Protocols in Hierarchical Cluster-Based Wireless Sensor Networks”, *International Research Journal of Engineering and Technology*, Vol.02, Iss.03, June-2015

[101] I. Erguler, E. Anarim, and G. Saldamli, “A SalientMissing Link in RFID Security Protocols”, *Hindawi Publishing Corporation EURASIP Journal on Wireless Communications and Networking*, 2011

[102] Md Monzur Morshed, A. Atkins and H. Yu, “Secure ubiquitous authentication protocols for RFID systems”, *Springer- EURASIP Journal on Wireless Communications and Networking*, vol.93, 2012

[103] M. Katagi and S. Moriai, “Lightweight Cryptography for the Internet of Things”, 2011. Retreived from https://www.iab.org/wp-content/IAB-uploads/2011/03/Kaftan.pdf

[104] S. Khajuria and B. Andersen, “Authenticated Encryption for Low-Power Reconfigurable Wireless Devices”, *Journal of Cyber Security and Mobility*, Vol. 1, 189–203., 2012

[105] L. Yang, P. Yu,W. Bailing, “The Internet of Things Security Architecture Based IBE Integration with the PKI/CA”, *SERC*, 2013

[106] Y. B. Saied, A. Olivereau, D. Zeghlache, “Trust management system design for the Internet of Things: A context-aware and multiservice Approach”, *Elsevier-Sciverse ScienceDirect*, vol.1, 2015

[107] H. Kim, K. Kim, “Toward an Inverse-free Lightweight Encryption Scheme for IoT”, CISC-W'14, 2014

[108] T. Markmann, “Securing Constrained Networks with ID-based Cryptography and Short Signatures”, Related Work Report for Amazon WebServices, 2014

[109] H. Shafagh, A. Hithnawi, S. Duquennoy, “Poster: Towards Encrypted Query Processing for the Internet of Things”, *ACM*, 2015

[110] U. Kumar, T. Borgohain, S. Sanyal, “Comparative Analysis of Cryptography Library in IoT”, *International Journal of Computer Applications, International Journal of Computer Applications*, 2015

[111] H. Shafagh, A. Hithnawi, A. Droscher, “Talos: Encrypted Query Processing for the Internet of Things”, ACM SenSys, 2015

[112] D. Dinu, Y. L. Corre, D. Khovratovich, “Triathlon of Lightweight Block Ciphers for the Internet of Things”, ePrints, 2015

[113] P. Huang and H. Mu, “A High-security RFID Grouping Proof Protocol, International Journal of Security and Its Applications”, Vol.9, No.1 pp.35-44, 2015

# Text Clustering using Ensemble Clustering Technique

Muhammad Mateen[1], Junhao Wen[2], Sun Song[4]
School of Big Data and Software Engineering,
Chongqing University,
Chongqing, 401331, China

Mehdi Hassan[3]
Department of Computer Science,
Air University,
Islamabad, 44000, Pakistan

***Abstract*—Clustering is being used in different fields of research, including data mining, taxonomy, document retrieval, image segmentation, pattern classification. Text clustering is a technique through which text/ documents are divided into a particular number of groups, so that text within each group is related in contents. In this paper, the idea of ensemble text clustering of majority voting is defined. For this purpose, different clustering methods such as fuzzy c-means, k-means, agglomerative, Gustafson Kessel and k-medoid are used. After performing the pre-processing of the documents, inverse document frequency (IDF) has been achieved by the provided dataset. The achieved IDF is considered as input to the clustering algorithms. Dunn Index and Davies Bouldin Index have been calculated which are applied to analyze the usefulness of the proposed ensemble clustering. In this work, a dataset "Textclus" which contains four different classes, history, education, politician and art as a text is applied. Additionally, another dataset "20newsgroups" is also applied for analysis. The clustering quality measures have also been calculated from the proposed ensemble clustering results. The attained results show that the proposed ensemble clustering outperforms the other state of the art clustering techniques.**

***Keywords—Agglomerative; document clustering; ensemble clustering; gustafson kessel; inverse documents frequency; text clustering***

## I. Introduction

Clustering is mostly used in the area of pattern recognition and information retrieval. Text clustering is a technique through which text/ documents are divided into a particular number of groups, so that text within each group is related in contents [1]. The goal of text clustering us to make a set containing relative data objects in a particular way like kind of text, a group of text, etc.

In the text clustering unsupervised technique of data is needed for clustering. Usually, text clustering techniques use characteristics like sequences, words, phrases from the documents to apply the clustering [2]. Text clustering is an interesting and advance research area because the availability of a huge amount of information in electronic forms [3]. A lot of several applications have been designed in literature which are applied for document clustering [4]. Several techniques have been used in text/document clustering are given as under:

- Frequent pattern-based clustering
- Constraint based clustering
- Partitioning
- Grid-based
- Model-based
- Hierarchical
- Density-based
- Fuzzy clustering

Currently, text clustering is widely used to solve the problem raised by the area of the database, the main purpose of this research is to scalable clustering of a heterogeneous and multi-dimensional type of data [5]. To cluster out the data, there are various clustering techniques. Hard and soft clustering methods are used for document clustering according to their attributes [6]. There is a close relationship among clustering methods and several fields including heterogeneous data analysis, web application, and DNA analysis in computational biology, white blood cells and red blood cells clustering [7]. Hisham Al-Mubaid et al. [8] proposed a new clustering technique based on a successful feature selection method and logic based learning method Lsquare. Yanjun Li et al. defined a feature selection technique and after that evaluated with K-means clustering algorithm for different standard datasets. Their results showed that TCFS with choice of internal representations (CHIR) has better accuracy of clustering in the form of purity and f-measure [9].

The rest of the paper is managed as; related work of study is discussed in section II, section III is based on proposed technique and in section IV, the experimental results with the help of graphical representation are discussed. In section V, finally the research outcome is concluded.

## II. Related Work

A lot of research work has been dedicated to improve and develop the text clustering algorithms. In [10] researchers proposed a novel clustering technique which is based on mode seeking without any parameter. In the mean algorithm, mode seeking technique is especially used for clustering the data; it is also helpful and makes the system efficient when the parameters are not defined in a proper way. In this work, the researcher proposed ensemble clustering contains the repetition of the recent mode seeking technique of kNN. Mode seeking algorithm is more convenient and faster than the regular mean shift, for the high level of dimensional data. Mode seeking algorithm is also famous about the robustness about parameter selection and high level of dimensional data. The proposed method is achieved with the help of the consensus algorithm. Consensus algorithm process is based on two main steps. (1)

Randomly initialize the subsets with different parameters to run the multiple clustering and this session is considered as kNN ensemble mode seeking (2) merge all the results to calculate the consensus of whole repeated clustering.

The complex structure of the dataset includes data dimensionality and distribution managed by clustering algorithms. In [11] authors proposed a new data clustering technique that is based on the k-nearest neighbor chain (KNNC) using heuristic rules. With the inspiration of the PageRank, the algorithm Researcher used random walk design to calculate the value of data points. After that, based on valuable data points, researcher designed a KNNC to arrange the K-nearest neighbors with respect to distance and defined two heuristic approaches to obtain the suitable number of groups having same characteristics of objects and also initial clusters. In the first rule of heuristic approach is the distance of K-nearest neighbor chain which demonstrates the level of separation of groups with convex curves and secondly, the distance of internal compactness of a group to the nearest neighbor of KNNC. The proposed clustering technique achieved better performance than famous clustering algorithms.

Text mining is widely used in the field of social networks, opinion mining considered emotion analysis. It plays an effective role in the evaluation of emotions and opinion in texts. Opinion methodology normally depends on an expression lexicon, which is a group of predefined keywords that define emotions. Opinion mining gather required emotional words defined in advance and also have some complexity to classify the sentences that involve a judgment without the use of any emotional keywords. In [12] the researcher proposes a novel emotion analysis technique, based on hidden Markov design with text data. In this research, there is a text-based representation of emotions via ensemble TextHMMs. It explains hidden variables with the help of semantic group information. For the reflection of diverse models, an ensemble technique is applied that is based on TextHMMs classifier.

Consensus clustering is increasing attention to the heterogeneous data analysis. The co-association technique is used to define the consensus clustering as a graph partitioning problem. In [13] proposed spectral ensemble clustering (SEC) to influence the benefit of co-association for information integration but works in an efficient way. Algorithmic complexity exponentially reduced with the combination of SEC and weighted K-means clustering. The proposed big data grouping technique is introduced to overcome the challenges based on incomplete partitions with the extension of SEC. Multi-view and ensemble clustering techniques demonstrate the advantage of SEC with other big data clustering techniques.

In recent years, [14] ensemble clustering has become an attractive method for robust clustering. In existing ensemble clustering techniques there is a limitation that all the clusterings are regarded as the same as dependability, which caused as vulnerable to the low-level base clusterings. To globally analyze and weight the clusterings, there are some tasks has been performed which are also used to neglect the diversity of groups inside the base clustering. There is a common issue about to analyze the reliability of clusterings and utilize the neighbor diversity in the ensemble technique to improve the consensus quality, particularly, in the scenario, when there is no chance to obtain the data features or precise prediction about data distribution.

To overcome this problem [15] proposes a new ensemble clustering technique on the basis of ensemble-driven clustering estimation and locally weighted strategy. Particularly, the doubtfulness of every cluster is predicted by considering the clustering labels in the internal ensemble through an entropic principle. A new approach named as ensemble-driven clustering is explained and locally weight co-association matrix has defined the conclusion of an ensemble of diverse groups.

In the last few years, fuzzy c-means cluster ensemble and random projection approaches have been designed for data clustering with high dimensionality. Random projection is widely used to reduce the dimensionality because of its efficiency and simplicity. A large amount of space is required to store the huge affinity matrix, and earn large computational time to cluster out the affinity matrix.

In [15], to reduce the dimensionality of huge data, the researcher designed a framework with the combination of fuzzy c-means and random projection based on cluster ensemble. The framework uses the collective agreement to cumulative fuzzy partitions. With the use of internal and external cluster indices, the fuzzy partitions are ranked with random projections. The most excellent partition in the levels of the queue is the core partition achieved by cluster ensemble.

In [16], tree-ensemble clustering technique is introduced for the analysis of static CRAFTER, datasets to handle the high dimensionality. CRAFTER is useful to tackle numerical and categorical features concurrently and analyze the size and dimensionality of datasets.

CRAFTER influences the features of a tree-ensemble to tackle high dimensionality and mixed attributes. The class probability estimation technique is used for the representation of data points of clustering. According to the limitations of the ensemble clustering, one of them is the independence of basic clusterings and ignorance between their relationships. On the other, there is a lack of information corporation between the local and global relationship of clusterings, especially, when renovating from one point to another of the similar matrix between basic clusterings. In [17] a new ensemble clustering technique is introduced as relative density path accumulation (MRDPA). In this technique, density nearest-neighbor and relative k-nearest neighbor are used to create basic clusterings. The clusterings represent multi-scale features for the input dataset of k into the RNKD. To investigate the global information in the creative k-nearest neighbor graph, and finally generated via consensus function.

The basic purpose of ensemble clustering is to combine the various essential parts into the consent. In the related work of different articles, it has been explained that to increase the numerous partitions caused a lower variance and better performance for ensemble clustering. In this scenario, for the given dataset, the best partition among the different partitions is still a challenging problem.

In [18], the researcher proposed a novel approach to solving this problem. The author introduced the infinite ensemble clustering (IEC) to drop out the noising data and represents the infinite partitions. In this technique, de-noising auto encoder is used to generate the prospective representation of infinite partitions. For resultant clustering, concatenation of deep features is applied to k-means.

In [19], researcher introduced a novel technique for ensemble (subspace) clustering with high dimensionality of text data. This technique implements the integration of two stages for feature representation of data including topics and words to make clusters. Ensemble clustering is effective to enhance the strength of clusters. This approach based on topic modeling to lead the two steps attributes the representation of data and make various ensemble components. With the use of both words and topics to cluster out the text data, significant clusters can be achieved with the weight of topics and words in every cluster.

## III. Proposed Techniques

In the field of information retrieval, text clustering is an important area of research to categorize and understand the unstructured textual data. In this research, the ensemble clustering technique is investigated. The ensemble clustering is based on k-means, agglomerative, fuzzy c-means, k-medoid, and Gustafson Kessel clustering and has obtained different clustering results separately of a specific data; observed that all results were different from each other's [20]. These processes are used for the quality and performance of clustering algorithms, and these stages are necessary to complete the clustering algorithm [21]. The proposed ensemble clustering technique consists of four various stages and depicted in Figure 1.

*1)* In data collection phase some processing operations are applied at the given data include crawling, indexing, filtering etc. which are helpful in document clustering. It also indexes the documents to store the data and access in an efficient way and screen out the unnecessary data for example stop words.

*2)* In pre-processing phase some specific operations have been performed which are used to make the text into a meaningful format like, vector-model, graphical model.

*3)* Text clustering phase divides a set of text into a specified number of clusters, as each dataset specified number of clusters having different features.

*4)* In post-processing phase, core applications are included through which documents can be clustered.

To obtain accurate results, the ensemble clustering technique based on majority voting scheme is proposed. To ensure that ensemble clustering results are useful for text clustering. Dunn Index (DI) and Davies Bouldin Index (DBI) are evaluation parameters indicate that ensemble clustering technique offers superior results as compared to individual technique.

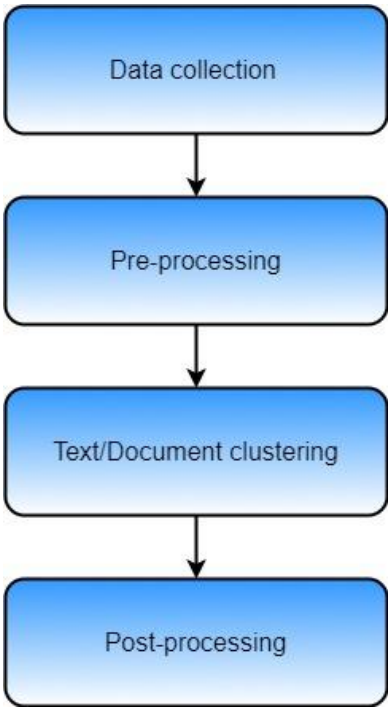



Fig. 1. Process of Clustering

The following clustering techniques have been used for text clustering. A brief description of each algorithm used in ensemble clustering is given as under:

### A. K-mean Algorithm

K-means clustering is a very popular and widely used technique which is successfully being used in image segmentation, computer vision and object qualification etc [22].

**Algorithm:** K-means

1. Cluster centroids $\mu 1, \mu 2 \ldots \mu k \in \mathbb{R}n$ arbitrarily.
2. Recur till union {

   For each $i$, set

$$c^{(i)} := \arg\ \min ||x^{(i)} - \mu_j\ ||^2.$$

For every j, set

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$$

}

### B. Fuzzy C-Means Algorithm

Fuzzy c-means (FCM) is a technique of soft clustering in which a portion of data belongs to more than one clusters [23].

Algorithm: Fuzzy c-means

$Input: X, c, m$

$Output: U, V$

$Initialize\ V$

$While: max_1\ £\ \{||v_{k,new} - v_{k,old}||^2\} > \hat{I}\ do$

$$m_{ij} = \left[ \sum_{k=1}^{c} \left( \frac{||x_j - v_i||}{||x_j - v_k||} \right)^{\frac{2}{m-1}} \right]^{-1} ,''i, j$$

$$v_i = \frac{\sum_{j=1}^{n} (m_{ij})^m x_j}{\sum_{j=1}^{n} (m_{ij})^m} ,''i$$

where U is the $(c \times n)$ partition matrix, $V = \{v_1, \ldots, v_c\}$ is the set of c cluster center in $R^d$, $m > 1$ is the fuzzification constant and $\|\cdot\|_A$ is an inner product A-induced norm. $\mu_{ij}$, $v_i$ are iterated until algorithm terminates.

### C. Hierarchical Agglomerative Clustering Algorithm

Hierarchical clustering algorithms are either bottom-up or top-down approaches. Bottom-up algorithms are treated as a pair of clusters which are merged until all clusters are merged into a single cluster which contains all documents. Hierarchical agglomerative clustering is more commonly used in information retrieval than top-down clustering [24, 25].

$$K = arg_{K'} min[RSS(K') + lK] \quad (1)$$

Where K′ denotes cut of the hierarchy that results in K′ clusters, RSS represents the residual sum of squares and λ is a consequence for each additional cluster. Another measure of distortion can be used instead of RSS.

### D. K-medoids Algorithm

K-medoid comes in an algorithm related to the k-means algorithm, except when fitting the centers *C1,…Ck*, restrict the attention to the points themselves [25].

The initial guess for centers *C1,… Ck* (e.g., randomly select *K* of the points *X1,…Xn*), then repeat:

*1)* Minimize over C: for each i=1,…n, find the cluster center ck closest to *Xi* and let *C(i)=K*

*2)* Minimize over *c1,…ck*: for each *k=1,… K* let *Ck=Xk*, the medoid of points in cluster k, i.e., the point Xi in cluster k that minimizes

$$\sum c(j) = k \| X_j - X_i \|_2^2$$

stop when within-cluster variation doesn't change

In words:

*1)* Cluster (label) each point based on the closest center

Replace each center by the medoid of points in its cluster

### E. Gustafson-Kessel Algorithm

The Gustafson-Kessel (GK) algorithm is a dominant clustering method with numerous applications in different domains including, classification, system identification, and image processing. The Gustafson-Kessel is a technique that is used to combine each cluster with both a matrix and a point, correspondingly shows the cluster center and its covariance. But the fuzzy c-means make the inherent assumption that clusters are spherical, the Gustafson- Kessel algorithm is not subject to this restriction and can categorize ellipsoidal clusters [26].

The Gustafson–Kessel algorithm is based on iterative optimization of an objective function of the c-means type:

$$J_m = \sum_{j=1}^{n} \sum_{i=1}^{k} \mu_{ij}^m d_{ij}^2$$

Where $J_m$ defines within group sum of squared errors.

### F. Ensemble Clustering

Ensemble clustering combines a set of clustering from the same data set and generates a final clustering. The goal of ensemble clustering is to improve the quality of individual data clustering. In this paper, majority voting based ensemble clustering is applied. For this purpose, five clustering techniques are employed which are already defined on specific data sets to obtain individual results. After implementation, the ensemble clustering results are compared with individual techniques. It has been observed from ensemble clustering results that it generates better clustering results.

## IV. Experimental Results And Discussion

In this paper, five various clustering algorithms and two datasets are used to obtain the ensemble clustering, but here k-means, agglomerative, fuzzy c-means, k-medoid and Gustafson Kessel clustering techniques. The ensemble clustering is obtained from these five clustering techniques. To access the clustering quality measure, Dunn Index and Davies Bouldin Index are computed for each technique. Results indicate that the ensemble clustering technique offers superior results as compared to all others state of the art technique. All experiments have been performed on Matlab 2015(b) on a PC with RAM 8GB and Windows 7.

### A. Datasets

In this research, two datasets are used, one of which is a private dataset named it "Textclus". The “Textclus” dataset has four classes, education, history, art, and politician. A total of sixty files have been collected by various websites. The proposed ensemble clustering technique has been evaluated at the obtained dataset “20newsgroups” freely available [27]. The obtained experimental results on the basis of both datasets are depicted in TABLE I.

TABLE I. Datasets

| Clustering Algorithms | Dataset: Textclus | | Dataset: 20newsgroups | |
|---|---|---|---|---|
| | ***DI*** | ***DBI*** | ***DI*** | ***DBI*** |
| ***K-Means*** | *0.3738* | *0.6505* | *0.0707* | *0.4845* |
| ***FCM*** | *0.0702* | *1.7287* | *0.0118* | *0.3229* |
| ***Agglomerative*** | *0.4386* | *0.8118* | *0.3538* | *0.3743* |
| ***K-Medoid*** | *0.0473* | *1.1102* | *0.0903* | *0.4845* |
| ***Gustafson Kessel*** | *0.0154* | *1.7038* | *0.0484* | *0.7409* |
| ***Ensemble Clustering*** | *0.7675* | *0.4454* | *0.7879* | *0.1748* |

### B. Clustering Quality Measures

It is mandatory for any clustering technique to evaluate the performance by using some standard quality measures. For this purpose, Dunn Index (DI) and Davies Bouldin Index (DBI) are used as clustering quality measures. These clustering quality measures have been computed from the resultant ensemble clustering results as well as the other algorithms results used for the ensemble. A high value of DI indicates better clustering whereas; the low value of DBI represents better clustering.

### C. Scenario-I Performance Comparison of Proposed Ensemble Clustering at Textclus Dataset

In this section, first of all, k-means is used which is a part of partitioning methods on the specific dataset; k-means is mostly used for hard clustering. Cluster labels are obtained from k-means and stored it for comparison with other results, after that agglomerative algorithm is used on text data which is a part of hierarchical methods, clustering labels are also obtained from the agglomerative algorithm and then stored it for later use. Next, fuzzy c-Means is applied which is also a part of partitioning methods but mostly used for soft clustering. Similarly, this algorithm to the given data set is applied and then obtained clustering labels for future experiments, after that, k-medoid algorithm is used on the same dataset and obtained results for later use. At the end, the fifth algorithm which is Gustafson Kessel algorithm technique is used and obtained results, stored it for later use. After obtained clustering labels, from all five clustering labels, the ensemble clustering technique is applied to obtain better results, the ensemble cluster labels are obtained by used of majority voting. First, cluster labels are obtained and then applied quality parameters on all clustering labels which are obtained from, k-means, agglomerative, fuzzy c-means, k-medoid, Gustafson Kessel and ensemble clustering technique. In this research, Dunn Index and Devis Bouldin Index are used as quality parameters.

In figure 2, the graph represents the clustering quality of dataset "Textclus". A high value of DI indicates better clustering whereas; the low value of DBI represents better clustering. So Ensemble clustering results are compatible with Dunn Index and Davies Bouldin Index.

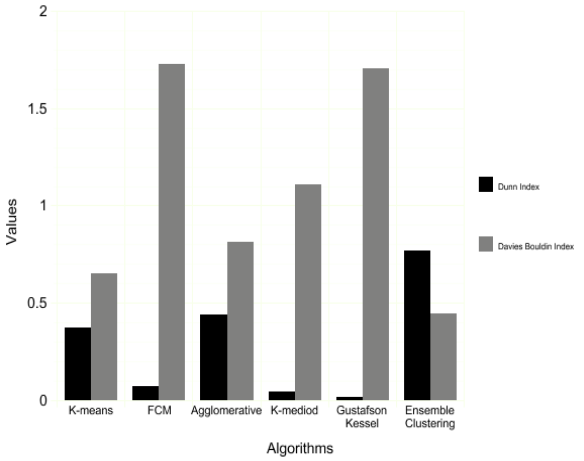



Fig. 2. Clustering quality of Dataset “Textclus”

### D. Scenario-II Performance Comparison of Proposed Ensemble Clustering at 20newsgroups Dataset

For the experiment, another dataset "20newsgroups" is also used. Clustering results are obtained using "20newsgroups" and observed that ensemble clustering technique is better than all above techniques which are used before. The high value of DI indicates better clustering whereas; the low value of DBI represents better clustering. So it can be observed that ensemble clustering results are according to the standards of Dunn Index and Devis Bouldin Index which are defined below. After application of these parameters, different results are obtained from different clustering labels.

In this paper, a new clustering technique is introduced which is based on major voting. Various clustering techniques are applied including, agglomerative, fuzzy c–means, k-medoid, Gustafson Kessel and k-means. Inverse document frequency is calculated from given dataset, and used as an input for clustering algorithms, experimental results are obtained using above clustering algorithms then observed that all were different from each other; at the end, ensemble clustering technique is proposed for better results on the basis of majority voting. After that, Dunn Index and Davies Bouldin Index algorithms are applied for quality performance. The results can be observed by bar graph with both data sets.

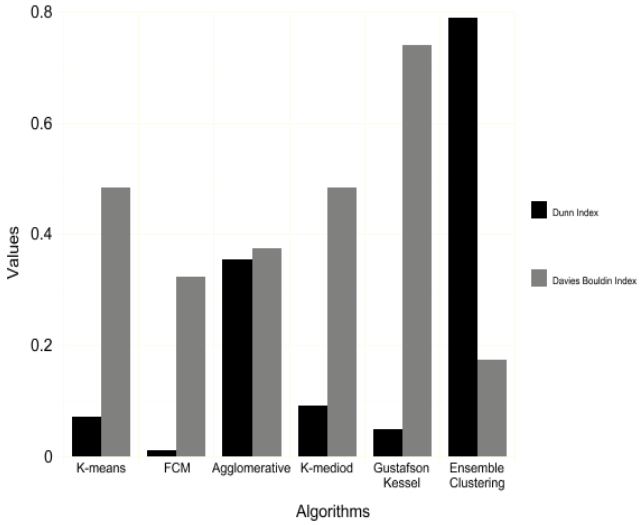



Fig. 3. Clustering quality of Dataset “20newsgroups”

In figure 3, the graph represents the clustering quality of dataset "Textclus". A high value of DI indicates better clustering whereas; the low value of DBI represents better clustering. So ensemble clustering results are compatible with Dunn Index and Davies Bouldin Index.

## V. Conclusion and Future Work

In this paper, clustering is discussed that is a group of similar objects. Five clustering methods are applied on datasets, first on "Textclus" then on "20newsgroups" and obtained individual results. After that an ensemble clustering technique is proposed based on major voting, to enhance the performance of text clustering. Cluster quality parameters are applied named as Dunn Index and Devis Bouldin Index. A

high value of DI indicates better clustering whereas; the low value of DBI represents better clustering. During experiments on specified datasets, results of five clustering techniques represented in the above graphs, could not fulfill the requirements of DI and DBI, but ensemble clustering using majority voting technique proved fruitful for text clustering with better clustering results. Therefore, ensemble clustering is found better than five clustering techniques named as k-means, fuzzy c-means, agglomerative, k-medoid, and Gustafson Kessel. For future work, the ensemble clustering technique can be applied on text streams/web data clustering to separate contents and find the extremism content in it.

## ACKNOWLEDGMENT

This research was supported by the Basic and Advanced Research Projects in Chongqing, China under Grant No.61672117.

## REFERENCES

[1] Hau, C.C., Handbook of pattern recognition and computer vision. 2015: World Scientific.

[2] Cheng, G., J. Han, and X. Lu, Remote sensing image scene classification: benchmark and state of the art. Proceedings of the IEEE, 2017. 105(10): p. 1865-1883.

[3] Mather, P. and B. Tso, Classification methods for remotely sensed data. 2016: CRC press.

[4] Chen, F., et al., Data mining for the internet of things: literature review and challenges. International Journal of Distributed Sensor Networks, 2015. 11(8): p. 431047.

[5] Zuech, R., T.M. Khoshgoftaar, and R. Wald, Intrusion detection and big heterogeneous data: a survey. Journal of Big Data, 2015. 2(1): p. 3.

[6] Witten, I.H., et al., Data Mining: Practical machine learning tools and techniques. 2016: Morgan Kaufmann.

[7] Perez-Guaita, D., et al., Multispectral Atomic Force Microscopy-Infrared Nano-Imaging of Malaria Infected Red Blood Cells. Analytical chemistry, 2018. 90(5): p. 3140-3148.

[8] Nguyen, D.B., New methods for classification, prediction, and feature weighting in bioinformatics. 2015: University of Houston-Clear Lake.

[9] Rajinikanth, T. and G.S. Reddy, A SOFT SIMILARITY MEASURE FOR K-MEANS BASED HIGH DIMENSIONAL DOCUMENT CLUSTERING. IADIS International Journal on Computer Science & Information Systems, 2017. 12(1).

[10] Myhre, J.N., et al., Robust clustering using a kNN mode seeking ensemble. Pattern Recognition, 2018. 76: p. 491-505.

[11] Lu, J., Q. Zhu, and Q. Wu, A novel data clustering algorithm using heuristic rules based on k-nearest neighbors chain. Engineering Applications of Artificial Intelligence, 2018. 72: p. 213-227.

[12] Kang, M., J. Ahn, and K. Lee, Opinion mining using ensemble text hidden Markov models for text classification. Expert Systems with Applications, 2018. 94: p. 218-227.

[13] Liu, H., et al., Spectral ensemble clustering via weighted k-means: Theoretical and practical evidence. IEEE Transactions on Knowledge and Data Engineering, 2017. 29(5): p. 1129-1143.

[14] Huang, D., C.-D. Wang, and J.-H. Lai, Locally weighted ensemble clustering. IEEE transactions on cybernetics, 2018. 48(5): p. 1460-1473.

[15] Rathore, P., et al., Ensemble fuzzy clustering using cumulative aggregation on random projections. IEEE Transactions on Fuzzy Systems, 2018. 26(3): p. 1510-1524.

[16] Lin, S., B. Azarnoush, and G. Runger, CRAFTER: a Tree-ensemble Clustering Algorithm for Static Datasets with Mixed Attributes and High Dimensionality. IEEE Transactions on Knowledge and Data Engineering, 2018.

[17] Li, E., et al. Ensemble Clustering Using Maximum Relative Density Path. in Big Data and Smart Computing (BigComp), 2018 IEEE International Conference on. 2018: IEEE.

[18] Liu, H., et al. Infinite ensemble for image clustering. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: ACM.

[19] Zhao, H., et al., Ensemble subspace clustering of text data using two-level features. International Journal of Machine Learning and Cybernetics, 2017. 8(6): p. 1751-1766.

[20] Wazarkar, S. and B.N. Keshavamurthy, A survey on image data analysis through clustering techniques for real world applications. Journal of Visual Communication and Image Representation, 2018. 55: p. 596-626.

[21] Mohebi, A., et al., Iterative big data clustering algorithms: a review. Software: Practice and Experience, 2016. 46(1): p. 107-129.

[22] Learned-Miller, E., et al., Labeled faces in the wild: A survey, in Advances in face detection and facial image analysis. 2016, Springer. p. 189-248.

[23] Sengupta, S., S. Basak, and R.A. Peters. Data Clustering using a Hybrid of Fuzzy C-Means and Quantum-behaved Particle Swarm Optimization. in Computing and Communication Workshop and Conference (CCWC), 2018 IEEE 8th Annual. 2018: IEEE.

[24] Thulasiram, K., S. Ramakrishna, and M. Jayakameswaraiah, To Assess the Performance of EAHC Algorithm Using Sensor Discrimination Dataset for the Improvement of Data Mining System. 2018.

[25] Chen, M., Fuzzy reasoning based evolutionary algorithms applied to data mining. 2015, North Dakota State University.

[26] Masoudi, P., Application of hybrid uncertainty-clustering approach in pre-processing well-logs. 2017, Rennes 1.

[27] "Dheeru, D.a.K.T., Efi",, "{UCI} Machine Learning Repository",. "2017",

# Profile-Based Semantic Method using Heuristics for Web Search Personalization

Hikmat A. M. Abdeljaber
Department of Computer Science
College of Computer Engineering and Sciences
Prince Sattam bin Abdulaziz University
AlKharj, Saudi Arabia

***Abstract*—User profiles play a critical role in personalizing user search. It assists search systems in retrieving relevant information that is searched on the web considering the user needs. Researchers presented a vast number of profile-based approaches that aims to improve the effectiveness of information retrieval. However, these approaches are syntactic-based which fail to achieve the user satisfaction. By the means that the search results do not meet user preferences, due to the fact that the search is keyword-based rather than semantic-based. Exploiting user profiles with the application of semantic web technology into personalization might produce a step forward in future retrieval systems. By adopting profiling approach and using ontology base characteristics, a semantic-based method using heuristics and KNN algorithm is proposed. It engages searching ontology base domains *horizontally* and *vertically* to discover and extract the closest *concept* to the meaning of the query keyword. The extracted *concept* is used to expand the user query to personalize the search result and present the customized information for individuals.**

***Keywords*—Semantic search method; user profile; heuristics; web search personalization; information retrieval**

## I. INTRODUCTION

Information Retrieval (IR) is a highly active research field; it depends significantly on the web as the main source of information. It involves assisting users to find information from vast amount of information resources on the Web. However, finding relevant information that satisfies users' query is a vital problem [1, 2]. Search engines perform "one size fits the all", in which the submitted query of keywords returns the same result to users of different interests. Many query keywords may have the same syntax but different semantics (homonyms). These keywords could be located at different horizontal and vertical domains of ontology base of semantic web. They could have different meanings under different concepts within the same domain.

Therefore, the main research focus of this paper is how to find the domain that reflects the closest meaning of the query keyword from various domains of ontology base. Then how to find the *concept* that reflects the closest meaning of that query, from different *concepts* within the same domain. For instance, the query keyword 'cell' may be found across various *horizontal* domains such as computer, biology and politics. In addition, the same query 'cell' may be found under various *vertical concepts* within the computer domain including processors, excel and company name. Accordingly, finding which particular *horizontal* domain and particular *vertical concept* within that specific domain for query keyword 'cell' at the same search session. Once the *concept* and domain of the query keyword are found, then it can take an advantage of both in the query expansion to retrieve better search results. This paper presented a novel approach to address these two issues.

Therefore, it is important to optimize the means of personalizing the web search and locating relevant documents tailored for individual users. Web search personalization approaches and techniques have been reviewed in [3-7]. Generally, there are two types of approaches to personalizing the search results, first is by the user query modification and secondly is by the search results re-ranking. On the other hand, there are many techniques that are based on the web contents, web link structure, browsing history, user profiles and user queries. These techniques have been widely used to implement personalized web search models [3, 6]. These models include, but not limited to, the use of hybrid of fuzzy set and ant colony optimization as described in [8]. Firefly algorithm is used to create and choose the cluster based optimal ranked clicked URLs to recommend a set of terms that expand the query search [9]. Ant colony optimization with a genetic algorithm were used by [10] to rank web pages. [11] has used a hybrid of genetic algorithm and back propagation neural network to classify user queries to clusters for web page recommendations. However, the proposed model is implemented using a hybrid of artificial intelligence (AI) heuristics and K-nearest neighbour (KNN) algorithm to extract semantic *concepts* from the ontology base using the user profile to expand the query.

User profile is the major method used for personalizing web search as presented in [12-14]. For capturing user's current information need, [15] have represented user's activities in the form of time-sensitive profile. It integrates both the current and the recurrent interactions with the search engines such as submitted queries, reformulated queries, and clicked results within a session search. User interactions are taken into account under the assumption that recent performed ones are more related to the current needs than to the foregoing ones. Authors in [16] have proposed a Funnel Mesh-5 algorithm. It constructs a search string by taking into account the context of information need and the user intention. This information is identified by the user profile then it is used to generate a personalized disambiguated search string for query expansion. In [17], the user profile is created based on the user

search behaviour using the web search logs and the eye tracking. It measures the user behaviour during the query session. Their system keeps on updating the user profile to build and enhance the user profile to suggest more relevant web pages to the user. However, my proposed approach maps the user profile content onto the ontology base of the semantic web technology to extract the closest *concepts* to the meaning of query keywords. Researchers have broadly studied the personalized techniques to represent users' information needs in the user models. These techniques are used not only for monolingual IR systems [5] but also for multilingual IR systems for re-ranking of web search results [18], browsing and searching the behaviour of polyglots [19], and personalizing the query expansion [20].

The idea of the semantic web aims at making the semantics of the web content machine understandable [21]. Semantic web standards and technologies can be used to enable the semantic search [22]. Semantic search, as an application of semantic web in the field of IR, has shown a significant potential in the function of improving the performance of retrieval. Compared with the traditional search that focuses on the frequency of word appearance, the semantic search attempts to understand the meanings hidden in the retrieved documents and users' queries. It works by adding the semantic tags into texts to structuralize and conceptualize the objects within documents [23]. Therefore, many semantic-based methods have been proposed in personalizing the web search. [24] used ontologies to improve the reliability of personalization through exploiting the formal semantics of query-based relevance processing, user preference representation, preference update, and result ranking. In [25], a semantic mapper is used to map the user query terms with personalized ontology, that were created from user web log file to identify semantic relation between user's queries.

However, this research combines the use of both the user profile and the characteristics of ontology base to develop an effective semantic search method. It purposes to deliver better and close results compared to the conventional method currently used in IR. Thus, the main objective of this paper is to implement the semantic-based search method. By acquiring the user intent implicitly by exploiting ontology base and the user profile to personalize the web search and enhance the contextual IR of web documents.

Semantic user profile is created in [26] for capturing scholar's interests, tasks, and competences in different research topics across different projects and publications in the scientific domain. The semantic user profile is modelled through an automated text mining pipeline approach of NLP for using it in semantic publishing applications as personalized web applications. [27] have developed a model to capture user context by generating the query context and the user context. Also, it introduced a forgetting factor to merge the independent user context in the user session for maintaining the evolution of user preferences. In contrast, my proposed model has applied AI heuristics techniques and KNN algorithm that implements a search method pluggable in the semantic web search applications. Therefore, this paper contributes mainly in the area of the semantic-based search methods in the semantic web technology. The development of the new semantic-based search method, as a web search service for the semantic web technology, characterized by its ability to interact with an ontology base, reason, infer, and consequently find and extract the closest meaning of the users' query keywords. It assists users to provide them with the results that are close to accurate, when they work with searching the search engines for specific information.

Query expansion is the process of adding more terms to an original query to attempt to refine the information search and improve the retrieval effectiveness [28-30]. The proposed model used the query expansion to improve the results by including *concepts*, extracted semantically from the ontology base, which lead to retrieving more relevant documents.

The remainder of this paper is structured as follows: Section II describes the structure of the system prototype. Section III illustrates the proposed semantic method for web search personalization. Section IV shows the experimental results and evaluation that were performed to validate the proposed approach. The paper is concluded and future work is presented in Section V.

## II. System Prototype

Information retrieval systems (IRSs) are text-based prototypes that use traditional methods to retrieve information. IRSs perform limited personalization for individual users and consequently provide irrelevant documents in terms of search precision. Combining the profile-based sematic search approaches and the AI techniques together to retrieve information is challenging. Therefore, the user profiles should use the semantic web technologies in IR process to improve the search results. This vision needs prototype to incorporate the semantic web tools and the user profile with the search system. Fig. 1 illustrates a high-level picture of the system prototype.

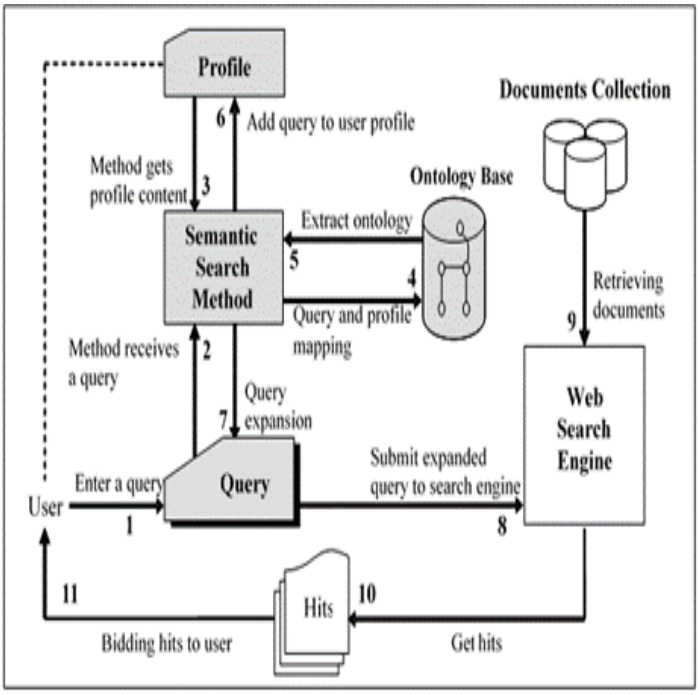



Fig. 1. Architecture of the system prototype.

The architecture of the system prototype consists of user query, semantic search method, user profile, and ontology base. The system prototype is described as follows: First, the user types a query. Second, the semantic search method receives the query. Third, the method obtains the user profile content. Fourth, the method maps the query and the user profile content onto the ontology base. Fifth, the method extracts the closest *concept* to the meaning of the query from the ontology base. Sixth, the method adds the query keywords to the user profile. Seventh, the method expands the query. Query expansion is performed to disambiguate the query by adding the extracted *concept* to the initial query automatically. Since this added information is originally acquired from the ontology base, it reflects the user needs. Eighth, the expanded query is submitted to the search engine. Ninth, the search engine retrieves the documents from the documents collection. Tenth, the search engine gets the hits and finally these hits are provided to the user. However, the proposed system prototype can be built and plugged into the typical IRS of web documents, such as search engines or metasearch engines, without effecting its standard operations.

## III. A Semantic Method for Web Search Personalization

The adopted approach needs a combination of five basic related components as follows: the user query, the user profile to maintain user query keywords, the semantic-based search method to extract from the ontology base the closest *concepts* to the meaning of query keywords, an ontology base to provide these *concepts* as contextual information for query expansion, and the web search engine to search the expanded query. The keywords of user profile and query are mapped onto the ontology base as shown in Fig. 2. The top level nodes of Fig. 2 are root domains of ontology base. The user profile keywords are marked with gray color in the solid circles and the query keyword is marked with a gray color in the dashed circle.

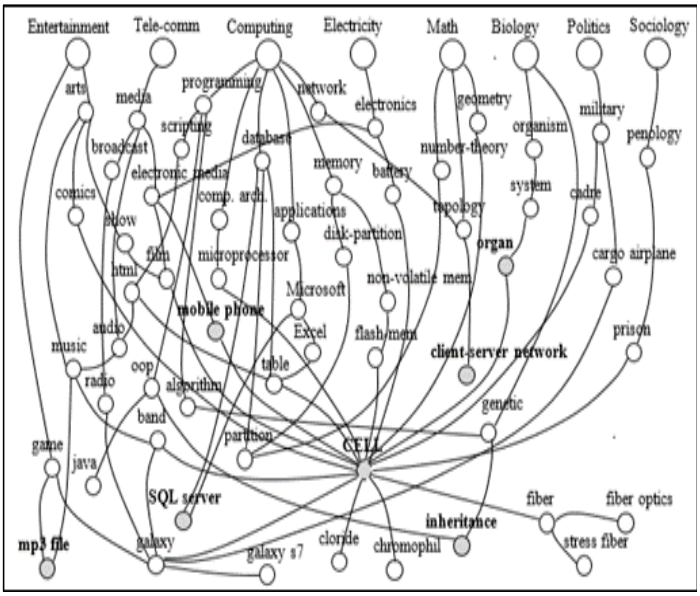



Fig. 2. An ontology base marked with user profile and query terms.

The proposed approach aims to find the closest *horizontal* root domain of the ontology base to the meaning of the given query and then search that domain *vertically* to determine the closest *concept* to the meaning of that query. To find the closest *horizontal* root domain, the information of long-term browsing history and short-term current browsing are utilized. The browsing history concerns the content of user profile whereas the current browsing concerns the user query itself. The heuristics technique is employed to make use of the user profile and query keywords to find the closest *horizontal* root domain. Heuristics provide many algorithms designed to traverse graph paths in order to discover the minimum cost path from a starting node to the goal. This work uses A* heuristic algorithm to find the closest *horizontal* root domain of the ontology base to the meaning of the query keyword. However, application of A* algorithm requires calculating two values for each node $n$ of the ontology base. The first value is known as $g(n)$ which is the distance value from the user profile or query keywords to the node $n$ and the second value is known as $h(n)$ which is the distance value from the root domains to the node $n$. Note that the node $n$ can be an ontology base node (concept), user profile keyword, or query keyword. Furthermore, keywords indicate the user profile and query keywords unless stated otherwise.

The distance from the keywords to the node $n$, $g\ (n)$, is calculated as follows:

$$g(n) = MIN_{\forall\ child\ i}\ \{\ distance_i\ \} + 1 \tag{1}$$

Where $\{\ distance_i\ \}_{\forall\ \text{child}\ i}$ is the set of distances from the keywords to all children of node $n$. Since the calculation of $g(n)$ starts from the user profile and query keywords, the initial value of $g(n)$ for each user profile keyword and query keyword is 0. It is important to note that the nodes which have paths with smaller values of $g(n)$ are closer to root domains.

The distance from the root domains to the node $n$, $h(n)$, is calculated as follows:

$$h(n) = MIN_{\forall\ parent\ j}\ \{\ distance_j\ \} + 1 \tag{2}$$

Where $\{\ distance_j\ \}_{\forall\ parent\ j}$ is the set of distances from the root domains to all parents of node $n$. Note that the root domains which have path to the query keyword are only involved in this step and the rest domains are excluded. Since the calculation of $h(n)$ starts from the root domains, the initial value of $h(n)$ for each root domain is 0. It is also important to note that the nodes which have paths with smaller values of $h(n)$ are closer to root domains.

A* algorithm is defined on the basis of (1) and (2) as follows:

$$f(n) = g(n) + h(n) \tag{3}$$

Where $f(n)$ is the evaluation function value of node $n$ that maintains its distance from the keywords to the root domains, $g(n)$ is the distance from the keywords to node $n$; and $h(n)$ is the distance from the root domains to node $n$. The value of $f(n)$ is computed based on the calculations of $g(n)$ and $h(n)$. There could be many paths from one keyword to one root domain in the ontology base. Each path may have different evaluation function value. However, this approach emphasizes on the *nearest* (smallest) of these values. Equation (4) is used to compute the *nearest evaluation function value* from each keyword to each root domain and the results are represented in a tabular form.

$$Nf(K_i)_{D_j} = \underset{\forall\ ancestor\ a}{MIN}\{f((K_i)_a \rightarrow D_j\}, \quad \forall i \forall j \tag{4}$$

Where $K_i$ is a keyword s.t $K_i \in \{K_1, K_2, \dots, K_l\}$ and $D_j$ is a root domain s.t $D_j \in \{D_1, D_2, \dots, D_m\}$. $MIN\ \{f((K_i)_a \rightarrow D_j)\}$ is the smallest value of all evaluation function values that are ancestors of $K_i$ ($(K_i)_a, \forall\ ancestor\ a$) and in the path from $K_i$ to $D_j$. This smallest value is assigned to $Nf(K_i)_{D_j}$ as the nearest heuristic evaluation function value from $K_i$ to $D_j$. Equation (4) repetitively computes the nearest values from all keywords corresponding to all root domains.

Frequencies of the user profile keywords are another source of heuristic information that help to find the closest *horizontal* root domain of the ontology base for a given query. The keyword with higher frequency has more weight than the one with lower frequency. To involve this information in the process, each $Nf(K_i)_{D_j}$ resulted from (4) is multiplied by the frequency of its corresponding user profile keyword ($K_i$) as given in (5).

$$(Nf(K_i)_{D_j})_{freq} = Nf(K_i)_{D_j} \times (K_i)_{freq} \tag{5}$$

Where $(Nf(K_i)_{D_j})_{freq}$ is the nearest value from $K_i$ to $D_j$ taking into consideration the frequency value of $K_i$, $Nf(K_i)_{D_j}$ is the nearest value from $K_i$ to $D_j$, and $(K_i)_{freq}$ is the frequency of user profile keyword $K_i$. The nearest values that are generated by (5) are then represented in a tabular form as depicted in Table I.

The frequency values given in Table I are assumed values for the purpose of clarification. The root domains $D_7$ and $D_8$ are included in Table I but they are excluded from Table II because they have no nearest evaluation function values (i.e. they have no paths to the user profile keywords). In addition, the query keyword 'CELL' which is included in Table I is also excluded from Table II because it is assumed that it is being inquired for the first time and it is not included in the user profile and hence it's $(Nf('CELL')_{D_j})_{freq}$ is 0, for all $D_j$, since its frequency value is 0.

TABLE I. THE NEAREST HEURISTIC EVALUATION FUNCTION VALUES OF KEYWORDS CORRESPONDING TO ONTOLOGY BASE ROOT DOMAINS.

| keywords / root domains | $K_1$ (mp3 file) freq. 2 | $K_2$ (mobile phone) freq. 1 | $K_3$ (SQL server) freq. 2 | $K_4$ (CELL) freq. 0 | $K_5$ (inheritance) freq. 1 | $K_6$ (client- server) freq. 3 | $K_7$ (organ) freq. 1 |
|---|---|---|---|---|---|---|---|
| | $(Nf(K_1)_{D_j})_{freq}$ | $(Nf(K_2)_{D_j})_{freq}$ | $(Nf(K_3)_{D_j})_{freq}$ | $(Nf(K_4)_{D_j})_{freq}$ | $(Nf(K_5)_{D_j})_{freq}$ | $(Nf(K_6)_{D_j})_{freq}$ | $(Nf(K_7)_{D_j})_{freq}$ |
| $D_1$(Entertainment) | 4 | | | – | | | |
| $D_2$(Tele-Comm.) | 6 | 3 | | – | | | |
| $D_3$(Computing) | 6 | | 4 | – | 3 | 9 | |
| $D_4$(Electricity) | | 3 | | – | | | |
| $D_5$(Mathematics) | | | | – | | 6 | |
| $D_6$(Biology) | | | | – | 2 | | 3 |
| $D_7$(Politics) | – | – | – | – | – | – | – |
| $D_8$(Sociology) | – | – | – | – | – | – | – |

Since the smallest values are the closest in meaning between keywords and root domains, they are better than the highest values. Therefore, a *normalization* process as given in (6) is needed to convert the smaller values to become more valued than higher values.

$$NNf(K_i)_{D_j} = \underset{\forall\ domains\ d}{MAX}(Nf(K_i)_d) - (Nf(K_i)_{D_j})_{freq} + \underset{\forall\ domains\ d}{MIN}(Nf(K_i)_d) \tag{6}$$

Where $NNf(K_i)_{D_j}$ is the *normalized nearest value* of the keyword $K_i$ for the domain $D_j$. The $MAX(Nf(K_i)_d)$ and $MIN(Nf(K_i)_d)$ are the maximum and minimum nearest values of the keyword $K_i$ associated with all the root domains, respectively. The $(Nf(K_i)_{D_j})_{freq}$ is the current nearest value of the keyword $K_i$ for the domain $D_j$. The normalized values are then maintained in Table II.

TABLE II. THE NORMALIZED NEAREST HEURISTIC EVALUATION FUNCTION VALUES OF KEYWORDS CORRESPONDING TO ONTOLOGY BASE ROOT DOMAINS.

| keywords / root domains | $NNf(K_1)_{D_j}$ | $NNf(K_2)_{D_j}$ | $NNf(K_3)_{D_j}$ | $NNf(K_5)_{D_j}$ | $NNf(K_6)_{D_j}$ | $NNf(K_7)_{D_j}$ |
|---|---|---|---|---|---|---|
| $D_1$(Entertainment) | 6 | | | | | |
| $D_2$(Tele-Comm.) | 4 | 3 | | | | |
| $D_3$(Computing) | 4 | | 4 | 2 | 6 | |
| $D_4$(Electricity) | | 3 | | | | |
| $D_5$(Mathematics) | | | | | 9 | |
| $D_6$(Biology) | | | | 3 | | 3 |

Each $NNf(K_i)_{D_j}$ in Table II may have different values for different root domains. Equation (7) is used to calculate the *degree of closeness* of *each* keyword $K_i$ for *each* connected root domain $D_j$ with respect to all root domains.

$$(NNf(K_i)_{D_j})_{close} = \frac{NNf(K_i)_{D_j}}{\sum_{j=1}^{m} NNf(K_i)_{D_j}}\ ,\ \forall i \forall j \qquad (7)$$

Where $(NNf(K_i)_{D_j})_{close}$ is the degree of closeness of each $K_i$ for each connected $D_j$, $NNf(K_i)_{D_j}$ is the normalized nearest value from $K_i$ to $D_j$, and $\sum_{j=1}^{m} NNf\ (K_i)_{D_j}$ is the summation of all normalized nearest values from $K_i$ to all its associated root domains $D_j$ , s.t $i = \{1, 2, \dots, l\}$ and $j = \{1, 2, \dots, m\}$ . The results of processing (7) are maintained in Table III.

TABLE III. THE DEGREE OF CLOSENESS VALUES OF KEYWORDS CORRESPONDING TO ONTOLOGY BASE ROOT DOMAINSNS.

| keywords / root domains | $(NNf(K_1)_{D_j})_{close}$ | $(NNf(K_2)_{D_j})_{close}$ | $(NNf(K_3)_{D_j})_{close}$ | $(NNf(K_5)_{D_j})_{close}$ | $(NNf(K_6)_{D_j})_{close}$ | $(NNf(K_7)_{D_j})_{close}$ |
|---|---|---|---|---|---|---|
| $D_1$(Entertainment) | 6/14 | | | | | |
| $D_2$(Tele-Comm.) | 4/14 | 3/6 | | | | |
| $D_3$(Computing) | 4/14 | | 4/4 | 2/5 | 6/15 | |
| $D_4$(Electricity) | | 3/6 | | | | |
| $D_5$(Mathematics) | | | | | 9/15 | |
| $D_6$(Biology) | | | | 3/5 | | 3/3 |

While (7) calculates the degree of closeness of each keyword $K_i$ for each connected root domain $D_j$, (8) given below, calculates the *degree of participation* of *each* keyword $K_i$ for *all* connected root domains with respect to *all* user profile keywords. This process is performed by dividing the total count number of non-zero normalized nearest values of a keyword $K_i$ for all root domains by the total count number of non-zero normalized nearest values of all keywords for all root domains. The result values are then maintained in Table IV.

$$\left(NNf(K_i)\right)_{part} = \frac{\left|NNf(K_i)_{D_j} \neq 0\right|_{\forall j}}{\left|NNf(K_i)_{D_j} \neq 0\right|_{\forall i \forall j}} \qquad (8)$$

$$i = \{1, \dots, l\}, \qquad j = \{1, \dots, m\}$$

Where $\left(NNf(K_i)\right)_{part}$ is the degree of participation of a keyword $K_i$ for all connected root domains, $\left|NNf(K_i)_{D_j} \neq 0\right|_{\forall j}$ counts the non-zero values of a keyword $K_i$ for all root domains, and $\left|NNf(K_i)_{D_j} \neq 0\right|_{\forall i \forall j}$ counts the non-zero values of all keywords for all root domains.

TABLE IV. THE DEGREE OF PARTICIPATION VALUES OF KEYWORDS IN ONTOLOGY BASE ROOT DOMAINS.

| keywords / root domains | $\left(NNf(K_1)\right)_{part}$ | $\left(NNf(K_2)\right)_{part}$ | $\left(NNf(K_3)\right)_{part}$ | $\left(NNf(K_5)\right)_{part}$ | $\left(NNf(K_6)\right)_{part}$ | $\left(NNf(K_7)\right)_{part}$ |
|---|---|---|---|---|---|---|
| $D_1$(Entertainment) | 1/11 | | | | | |
| $D_2$(Tele-Comm.) | 1/11 | 1/11 | | | | |
| $D_3$(Computing) | 1/11 | | 1/11 | 1/11 | 1/11 | |
| $D_4$(Electricity) | | 1/11 | | | | |
| $D_5$(Mathematics) | | | | | 1/11 | |
| $D_6$(Biology) | | | | 1/11 | | 1/11 |
| **For all root domains** | **3/11** | **2/11** | **1/11** | **2/11** | **2/11** | **1/11** |

Finally, the *confidence* values that represent the closest meaning of the user profile keywords corresponding to the root domains are computed by multiplying (7) by (8) as shown in (9) and the results are maintained in Table V.

$$(NNf(K_i)_{D_j})_{conf} = (NNf(K_i)_{D_j})_{close} \times (NNf(K_i))_{part} \qquad (9)$$

Where $(NNf(K_i)_{D_j})_{conf}$ is the confidence value of the keyword $K_i$ corresponding to the root domain $D_j$ . The confidence values of all keywords for every root domain are summed up and the root domain that has higher value is the closest to the meaning of the query keyword. Table V clearly shows that 'Computing' is the closest root domain to the query keyword 'CELL' since its confidence value (0.3118) is the highest.

TABLE V. CONFIDENCE VALUES OF THE KEYWORDS CORRESPONDING TO THE ONTOLOGY BASE ROOT DOMAINS.

| keywords / root domains | $\left(NNf(K_1)_{D_j}\right)_{conf}$ | $\left(NNf(K_2)_{D_j}\right)_{conf}$ | $\left(NNf(K_3)_{D_j}\right)_{conf}$ | $\left(NNf(K_5)_{D_j}\right)_{conf}$ | $\left(NNf(K_6)_{D_j}\right)_{conf}$ | $\left(NNf(K_7)_{D_j}\right)_{conf}$ | **sum of confidence values for each root domain** |
|---|---|---|---|---|---|---|---|
| $D_1$(Entertainment) | 0.116 | | | | | | **0.116** |
| $D_2$(Tele-Comm.) | 0.077 | 0.09 | | | | | **0.167** |
| $D_3$(Computing) | 0.077 | | 0.09 | 0.07 | 0.07 | | **0.311** |
| $D_4$(Electricity) | | 0.09 | | | | | **0.09** |
| $D_5$(Mathematics) | | | | | 0.10 | | **0.108** |
| $D_6$(Biology) | | | | 0.10 | | 0.09 | **0.198** |

Once the closest root domain (e.g. 'Computing') is determined, the process turns to search that domain *vertically* to identify the closest *concept* to the meaning of the query keyword (e.g. 'CELL'). For this purpose, KNN algorithm is employed. Fig. 2 shows some *concepts* that could indicate the meaning of the query keyword 'CELL' under 'Computing' domain including 'band', 'microprocessor', 'table', and 'flash-memory'. Calculating similarity between the query 'CELL' and each of these *concepts* might be beneficial to find the closest *concept* to the meaning of that query. The similarity for K=2 is calculated by using Euclidean distance as follows:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (10)$$

Where $d(p,q)$ is the distance between the *concept* p and the query q.

The KNN is an effective classifier method [31] used to find out the distance (closeness) between the query 'CELL' and its direct upper *concepts*, where K is determined by the heuristics or features of these *concepts*. Two features are identified to calculate the distance: first, the number of root domains connected to the *concept* and second, the number of user profile keywords connected to the *concept*. Therefore, K here equals to 2. Table VI shows the direct upper *concepts* of the query 'CELL' that are connected to the 'Computing' domain (column 1) and their identified features (column 2 and column 3). In addition, Table VI shows the ranks (column 5) of the *concepts* based on their distances (column 4) from the query.

TABLE VI. THE FEATURES OF CONCEPTS AND THEIR RANKS BASED ON DISTANCES OF CONCEPTS FROM THE QUERY

| The concept | # Root domains connected to the concept | # User profile keywords connected to the concept | Distance | Rank |
|---|---|---|---|---|
| Band | 2 | 1 | $\sqrt{5}$=2.236 | 2 |
| microprocessor | 1 | 0 | $\sqrt{9}$=3 | 3 |
| Table | 1 | 2 | $\sqrt{1}$=1 | 1 |
| flash-memory | 1 | 0 | $\sqrt{9}$=3 | 3 |

The Euclidean distance of KNN emphasize that the smallest distance value is the closest *concept* to the meaning of the query. Table VI, clearly shows that the *concept* 'table' is the closest *concept* to the meaning of the query 'CELL'. Extending the query 'CELL' by adding the *concept* 'table' to the query would personalize the user search and therefore, improve the search effectiveness.

## IV. EXPERIMENTAL RESULTS AND EVALUATION

A semantic-based web search method is proposed. This method adopts profiles and uses an ontology base to personalize users search and improve the accuracy of their search results. Essentially, a conceptual hierarchy is used in the experiments as ontology base for providing a shared understanding of the searchable domains. It is significant in discovering the closest meaning of the query keywords. Query keywords entered by the users should be selected from the experimental concept hierarchy terms (i.e. concepts) since the proposed method based on the ontology base. Moreover, the method employs Google search engine in the evaluation process to show the effectiveness of sematic-based search approach over keyword-based search approach. Such evaluation needs users to enter queries and to judge the relevancy of returned hits. Furthermore, the method implicitly extracts the closest *concept* to the meaning of the query from the ontology base and appends it to the initial query, to form the expanded query. The expanded query forms the final query, which is entered in the Google search engine to retrieve the desired documents.

The experimental trials had been conducted to evaluate the effectiveness of the proposed method in retrieving the relevant information. The evaluation investigates the degree to which the stated objectives are achieved. Recall and precision are two evaluation measures identified for IR systems [32]. Where recall measures the ability of the system to present all the relevant items in the collection, and precision measures the ability of the system to present only those items from the collection that are relevant. The documents collection of Google search engine is used intensively in the experiment tests. Users enter their queries into Google's search text box and consequently, Google searches its collection and returns a list of hits to users. Using Google collection restricts the evaluation process to use cut-off/precision measure rather than recall/precision measure, because it cannot calculate the normal recall points since the number of relevant documents in Google collection is unknown. However, precision is defined formally as follows:

$$precision = \frac{Number\ of\ relevant\ items\ retrieved}{Total\ number\ of\ items\ retrieved} \times 100 \quad (11)$$

Cut-off points are made in the experiments for the first 100 documents of the search engine hits. Precision values are calculated at cut-off (the first) 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 documents. The user should count the number of documents that are relevant to his needs and divides this number by 10 to obtain the precision value of the first group (10 cut-off points). The last step is repeated for all other groups as well. The user should count the number of relevant documents and divides this number by 20 to obtain the precision value of the second group (20 cut-off points), and so on. Thereby, the cut-off points and precision are for a single query. However, to evaluate the retrieval method accurately, we run it for several distinct queries; and an average is used for the cut-off and precision values. It is important to note that the average here means the precision value of all distinct queries at the corresponding cut-off point. In order to find the precision value at all cut-off points, we take the rate of all of its averages.

This paper compares the effectiveness of the semantic-based search approach with the keyword-based search approach. In particular, it compares the effectiveness of the proposed semantic-based search approach with a well-known and standard text-based IRS such as Google search engine. The comparison is achieved by searching the web twice. The first search uses Google search engine without employing the proposed method in the search process and the overall average of results is computed. The second search uses Google search engine with employing the proposed method in the search process and the overall average of results is computed. The difference between these two rates shows how happened to be an improvement of one over the other.

The averages of precision values of all entered queries at cut-off (the first group of documents) are calculated for 10 groups of sizes ranging from 10 to 100 documents with increments of 10. This calculation is done for both the proposed profiling semantic-based search method and the text-based search method (Google) as shown in Table VII.

TABLE VII. THE PRECISION VALUES AT CUT-OFF POINTS FOR THE PROPOSED METHOD AND GOOGLE AND THEIR AVERAGES.

| Cut-off | The proposed method (Semantic-based search) | Google (Text-based search) |
|---|---|---|
| 10 | 0.87 | 0.47 |
| 20 | 0.84 | 0.45 |
| 30 | 0.82 | 0.42 |
| 40 | 0.81 | 0.41 |
| 50 | 0.77 | 0.40 |
| 60 | 0.75 | 0.38 |
| 70 | 0.71 | 0.36 |
| 80 | 0.69 | 0.35 |
| 90 | 0.66 | 0.35 |
| 100 | 0.63 | 0.33 |
| **Average** | **0.755** | **0.392** |

The average of precision values of all entered queries for each of the 10 groups employing the proposed method is calculated. Then, an identical process is repeated for Google without employing the proposed method. As shown in Table VII, the first column (cut-off) denotes the number of documents taken as cut-off points, the second column shows the average precision values at these cut-off points using Google employing the proposed method, and the third column (Google) shows the average precision values at these cut-off points using Google without employing the proposed method.

The average here means the precision value at the corresponding cut-off point. In order to find the precision value of the proposed semantic-based search method using the user profile at all cut-off points, we take the rate of all of its averages. The same process is repeated with Google. The rate (average) of the proposed method and the rate (average) of Google are calculated and presented in the last row of Table VII. The difference between these two rates shows the improvement of one over the other. Table VII shows that the profile-based semantic search method improves the search results about 36% over the text-based search method (Google).

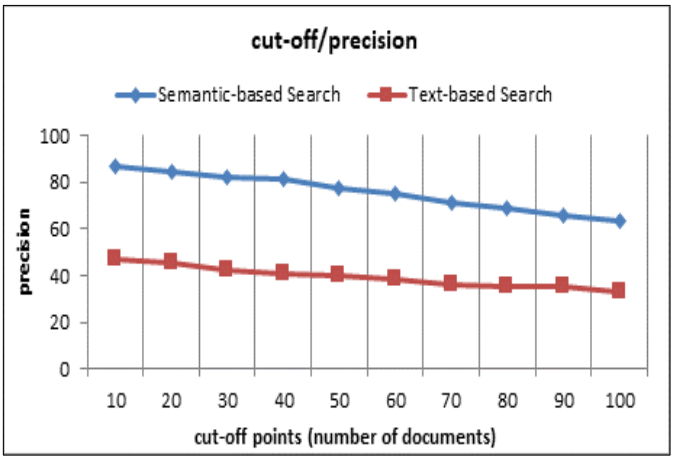



Fig. 3. Precision values at cut-off points for the proposed method and Google.

Fig. 3 is drawn based on Table VII. For each entered query, a cut-off/precision curve is drawn. These drawn curves are averaged to produce the final cut-off/precision shown in Fig. 3. The figure illustrates that the semantic-based IR is better than text-based IR. The figure shows 36% improvement on search results when employing the profiles and the heuristics in the retrieval process, especially when using ontologies for interpreting query terms, as recorded by the experiments. This aspect is promising to shift the web search engines from the text-based to the semantic-based information retrieval systems.

## V. CONCLUSION AND FUTURE WORK

Issuing a query to the web search system for retrieving relevant pages according to user preferences provides better result if the search method discover and extracts the *concept* from the ontology base that reflect the closest meaning of the query keyword. The proposed approach is significantly beneficial, especially when the query keyword found across several different *horizontal* root domains and under various different *concepts* of the same *vertical* root domain. The extracted *concept* is used to expand the query for personalizing the user's search. Incorporating the user profile and ontology base of the semantic web into the search process was the base of the proposed semantic-based search method. The heuristics and the KNN algorithm are applied to discover useful information in interpreting the query keywords. A profile-based personalized semantic search method shows a considerable improvement than text-based search in terms of search effectiveness, as recorded by the experiments. Despite the effectiveness and accuracy improvement of this approach, it has two limitations. First, the proposed search method is developed for handling only one-keyword size queries. Secondly, the effectiveness of the proposed approach is evaluated based on a comparison made with Google as a text-based search method. To get more accurate result, a comparison must be made with another semantic-based search method adopting different approach. For future work, identifying additional implicit features or heuristics and engaging knowledge management to discover knowledge from the heuristics information and representing them in user profile, may further improve the search results.

## ACKNOWLEDGMENT

This work was sponsored and supported financially by a research grant from the deanship of scientific research at Prince Sattam bin Abdulaziz University "REF 2014/01/2006".

## REFERENCES

[1] S. Chawla, "Search Engines: Information Retrieval on the Web," Everyman's Science, Indian Science Congress Association, Kolkata, India, vol. LI, no. 4, http://sciencecongress.nic.in/pdf/e-book/oct-nov-16.pdf (2016a). Accessed 2 Sep. 2017.

[2] H. A. Abdeljaber, "A profile-based semantic search strategy using an ontology base," Ph.D. Thesis, Universiti Kebangsaan Malaysia (UKM), Malaysia, 2009.

[3] J. Jayanthi and S. Rathi, "Personalized web search methods – a complete review," Journal of Theoretical and Applied Information Technology, vol. 62 no. 3, 2014.

[4] N. Borse, S. Patil, N. Agrawal, and R. Pachlor, "Survey on A Personalized Ontology Model for Web Information Gathering," International Journal of Emerging Technology and Advanced Engineering (IJETAE), vol. 3, no. 12, 2013.

[5] M. R. Ghorab, D. Zhou, A. O'connor, V. Wade, "Personalised information retrieval: Survey and classification," User Modeling and User-Adapted Interaction, vol. 23, no. 4, pp. 381-443, 2013. doi:http://dx.doi.org/10.1007/s11257-012-9124-1.

[6] B. Thomas, J. P. Jose, "A survey on web search results personalization," Compusoft, vol 4, no. 3, pp. 1582, 2015.

[7] V. Salonen and H. Karjaluoto, "Web personalization: The state of the art and future avenues for research and practice," Telematics and Informatics, vol. 33, no. 4, pp. 1088-1104, 2016. doi:10.1016/j.tele.2016.03.004.

[8] S. Chawla, "Use of hybrid of fuzzy set and ACO for effective personalized web search," International Journal of Advanced Research in Computer Science, vol. 8, no. 5, 2017a.

[9] V. Shakya and A. Sonker, "An optimal ranking approach for cluster based of clicked URLs using firefly algorithm for efficient personalized web search," International Journal of Advanced Research in Computer Science, vol. 8, no. 5, 2017.

[10] S. Chawla, "Web page ranking using ant colony optimisation and genetic algorithm for effective information retrieval," International Journal of Swarm Intelligence, vol. 3, no. 1, pp. 58, 2017b. doi:10.1504/IJSI.2017.082397.

[11] S. Chawla, "Application of genetic algorithm and back propagation neural network for effective personalize web search-based on clustered query sessions," International Journal of Applied Evolutionary Computation (IJAEC), vol. 7, no. 1, pp. 33-49, 2016b. doi:10.4018/IJAEC.2016010103.

[12] B. Mianowska and N. T. Nguyen, "Tuning user profiles based on analyzing dynamic preference in document retrieval systems," Multimedia Tools and Applications, vol. 65, no. 1, pp. 93-118, (2013;2012;). doi:10.1007/s11042-012-1145-6.

[13] A. Nanda, R. Omanwar, and B. Deshpande, "Implicitly learning a user interest profile for personalization of web search using collaborative filtering," Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 2, IEEE Computer Society; pp 54-62, 2014. doi:10.1109/WI-IAT.2014.80.

[14] S. Karpagam and P. N. Nesarajan, "Personalized web search protection using greedy algorithm," International Journal of Research in Computing Science, Engineering & Technology (IJRCSTE), vol. 1, no. 2, pp. 1-10, 2016.

[15] A. Kacem, M. Boughanem, and R. Faiz, "Emphasizing temporal-based user profile modeling in the context of session search," In: SAC, ACM, pp. 925–930, 2017.

[16] U. Gajendragadkar and S. Joshi, "Context sensitive search string composition algorithm using user intention to handle ambiguous keywords," International Journal of Electrical and Computer Engineering, vol. 7, no. 1, pp. 432-450, 2017.

[17] J. B. Baviskar, "Personalised web search using user behaviour," International Journal of Advanced Research in Computer Science, vol. 6, no. 1, pp. 185-187, 2015.

[18] M. R. Ghorab, D. Zhou, S. Lawless, and V. Wade, "Multilingual user modeling for personalized re-ranking of multilingual web search results, in Herder," E., Yacef, K., Chen, L. and Weibelzahl, S. (Eds), CEUR Workshop Proceeding of UMAP 2012, Montreal, Springer, Berlin, pp. 1-4, 2012.

[19] B. Steichen, M. R. Ghorab, A. O'Connor, S. Lawless, and V. Wade, "Towards Personalized Multilingual Information Access - Exploring the Browsing and Search Behavior of Multilingual Users," Proceedings from Conference on User Modelling, Adaptation and Personalization (UMAP 2014), Aalborg, Denmark, pp. 435–446, 2014. doi:10.1007/978--3--319-08786--3_39.

[20] D. Zhou, S. Lawless, X. Wu, W. Zhao, and J. Liu, "A study of user profile representation for personalized cross-language information retrieval," Aslib Journal of Information Management, vol. 68, no. 4, pp. 448-477, 2016. doi:10.1108/AJIM-06-2015-0091.

[21] R. D. Virgilio, F. Guerra, and Y. Velegrakis, Semantic search over the web, (1. Aufl.;2012;1; ed.). DE: Springer-Verlag (2012;2014;).

[22] A. Göker and J. Davies, Semantic search, Chichester, UK: John Wiley & Sons, Ltd. pp. 179-213. ch9, 2009. doi:10.1002/9780470033647.

[23] H. Dong, F. K. Hussain, and E. Chang, "A survey in semantic search technologies," 2nd IEEE International Conference on Digital Ecosystems and Technologies, pp. 403-408, 2008. doi:10.1109/DEST.2008.4635202.

[24] P. Castells, M. Fernández, D. Vallet, P. Mylonas, and Y. Avrithis, "Self-tuning personalized information retrieval in an ontology-based framework," vol. 3762, pp. 977-986, Berlin, Heidelberg: Springer Berlin Heidelberg, 2005. doi:10.1007/11575863_119.

[25] K. Makvana, P. Jay, P. Shah, and A. Thakkar, "An approach to identify semantic relations between user's queries in text retrieval," Proceedings of the Second International Conference on information and communication technology for competitive strategies, ACM, pp 1-6, 2016. doi:10.1145/2905055.2905271.

[26] B. Sateli, F. Löffler, B. König-Ries, and R. Witte, "Semantic user profiles: learning scholars' competences by analyzing their publications," International Workshop on Semantic, Analytics, Visualization. Enhancing Scholarly Data. Lecture Notes in Computer Science, Springer, Cham, vol. 9792, pp. 113-130, 2016.

[27] Z. Xu, H. Chen, and J. Yu, "Generating personalized web search using semantic context," The Scientific World Journal, vol. 2015, pp. 1-10, 2015. doi:10.1155/2015/462782.

[28] K. M. Fouad, A. R. Khalifa, N. M. Nagdy, and H. M. Harb, "Web-based semantic and personalized information retrieval," International Journal of Computer Science Issues (IJCSI), vol. 9, no. 3, pp. 266-276, 2012.

[29] K. Makvana, P. Shah, and P. Shah, "A novel approach to personalize web search through user profiling and query reformulation," International Conference on Data Mining and Intelligent Computing (ICDMIC), IEEE, pp. 1-10, 2014. doi: 10.1109/ICDMIC.2014.6954221.

[30] P. Jay, P. Shah, K. Makvana, and P. Shah, "An approach to identify user interest by reranking personalize web," Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, ACM, pp. 1-8, 2016. doi:10.1145/2905055.2905270.

[31] J. Gou et al, "A generalized mean distance-based k-nearest neighbor classifier," Expert Systems with Applications, vol. 115, pp. 356-372, 2019.

[32] R. Baeza-Yates and B. Ribeiro-Neto, Modern information retrieval, volume 463, ACM press, New York, 1999.

# Deep Learning Algorithm for Cyberbullying Detection

Monirah Abdullah Al-Ajlan
College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia

Mourad Ykhlef
College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia

***Abstract*—Cyberbullying is a crime where one person becomes the target of harassment and hate. Many cyberbullying detection approaches have been introduced, however, they were largely based on textual and user features. Most of the research found in the literature aimed at improving detection through introducing new features. However, as the number of features increases, the feature extraction and selection phases have become harder. On the other hand, no study has examined the meaning of words and semantics in cyberbullying. In order to bridge this gap, we propose a novel algorithms *CNN-CB* that eliminate the need for feature engineering and produce better prediction than traditional cyberbullying detection approaches. The proposed algorithm adapts the concept of word embedding where similar words have similar embedding. Therefore, bullying tweets will have similar representations and this will advance the detection. *CNN-CB* is based on convolutional neural network (CNN) and incorporates semantics through the use of word embedding. Experiments showed that *CNN-CB* algorithm outperform traditional content-based cyberbullying detection with an accuracy of 95%.**

***Keywords*—Cyberbullying; convolutional neural network; CNN; detection; deep learning**

## I. Introduction

With the proliferation of the internet and its anonymity nature, many ethical issues have emerged. Cyberbullying is among the most widely acknowledged problems by individuals and communities. It is defined as any violent, intentional action conducted by individuals or groups, using online channels repeatedly against a victim who does not have the potential to react [1]. Even though bullying has always been a critical issue and received much attention; the internet along with social media has only made the issue more critical and wide spread. This is because they open doors for predators and give them a 24/7 access to victims from all ages and backgrounds while keeping their identities anonymous [2]. For all the danger imposed by cyberbullying on victims and communities, this field of study is maturing, with a wealth of research and findings evolving every day. The vast range of existing cyberbullying studies are spanning fields like psychology, linguistics and computer science.

Psychologists recognized cyberbullying as being a phenomenon closely related to the well being of individuals. A study found in [3] where a total of 7000 students were examined, concluded that bullying contributes to higher levels of loneliness and lower levels of social well-being . Many psychologists were asked in [4] about the appropriate actions that need to be taken in response to the growing number of cyberbullying incidents and they were in favour of the automatic monitoring of cyberbullying.

Automatic monitoring of cyberbullying has gained considerable interest in the computer science field. The aim has been to develop efficient mechanisms that mitigate cyberbullying incidents. Most of the literature considered it to be a binary classification task, where text is classified as bullying or non bullying [5]. This is achieved through extracting features from text and feeding them to a classification algorithm. Many studies have addressed cyberbullying detection from different perspective, however, all falls under four features categories: content-based, user-based, emotion-based and social-network based features.

Even though the state of art in cyberbullying detection is rapidly evolving, there are many problems that has arisen. A fundamental issue still present is that most research attempt to improve the detection process by suggesting new features. However, this approach might generate huge number of features that require careful feature extraction and selection phases which lead to computational overhead. Moreover, features are not always easy to be extracted. In fact, features can be easily fabricated [6]. Another drawback is that they fail to adapt to the changing nature of language. Offensive words that are considered features in most detection approaches are not static and change over time. As a result, detection approaches must not rely on static features rather on more automated mechanisms. Despite the success of current approaches, a core problem has not been addressed. The semantic of words, their meaning and relations have been overlooked.

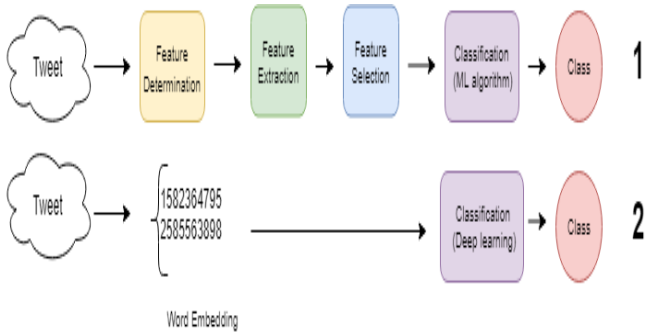



Fig. 1. Classical Machine Learning (1) vs CNN-CB (2).

In this article, we propose a convolutional neural network cyberbullying detection (*CNN-CB*) algorithm, which remedy the current unsolved problems. The primary goal is to develop an efficient detection approach capable of dealing with semantics and meaning and produces accurate result while keeping computational time and cost to a minimum. *CNN-CB* is based on deep learning which was on of MIT 10 Breakthrough Technologies Review in the year 2017 and 2013[7]. It is built upon the concept of convolutional neural network (CNN) which showed great success when applied to many classification tasks [8] [9] [10]. The most remarkable contribution is that *CNN-CB* is a cyberbullying detection algorithm that has shorten the classical detection workflow; it makes detections without any features. It transforms text into word embeddings and feeds them to a CNN. Previously, detection always started with feature extraction followed by, feature selection. Interestingly, *CNN-CB* has excluded these two steps and yet produced better result. Fig1 illustrates the traditional versus *CNN-CB* workflow.

This paper is organized as follows. Section II states the related work in cyberbullying. Then, section III describes *CNN-CB* in details. Section IV reports the experiments along with their results. Section V discusses the reported results. Finally, section VI concludes and summarizes this paper.

## II. Related Work

Cyberbullying detection has a rapidly growing literature, even though researches addressing bullying are traced back to early 2010. The rich literature in this field can be divided into three categories: content-based, user-based and network based detection. Each category will be covered briefly, and comprehensive summary is presented in table I.

### A. Content-Based Detection

Among the first to tackle bullying in social media is [11], where a framework was built to incorporate Twitter streaming API for collecting tweets and then classifying them according to the content. Their work combined the essence of sentiment analysis and bullying detection. As a first phase, tweets are classified as being positive or negative and then they are further classified as positive containing bullying content, positive without bullying content, negative containing bullying content, and negative without bullying content. For the sake of classification, Naïve Bayes was implemented and resulted in a relatively high accuracy (70%). Another later research found in [12], incorporated statistical measures namely (TFIDF) and (LDA) along with topic models in order to extract relevance in documents. However, they did not rely on statistical measures only but extracted content features like: bad words and pronouns. Other researchers in [13], continued to pursue cyberbullying detection from content-based perspective however, they introduced new features like: emotions icon and dictionary of hieroglyphs. Their approach was tested using many learning algorithms: Naïve Bayes, SVM and J 48. And the best result was recorded with SVM achieving an accuracy of 81%. Another research [14], presented a prototype system to be used by organization members to monitor social network sites and detect bullying incidents. The approach followed relied on recording bullying words and storing them in a database and then incorporate Twitter API to capture tweets and compare their content to the bullying material recorded earlier. Beside the promising innovative idea in their work, this prototype system has not been implemented yet.

### B. User-Based Detection

Many researchers believed that user information like age and number of tweets could indicate potentiality to harm others. In [15], researchers incorporated user information like number of tweets, number of followers and number of followings into the detection process. Their total features -user based and others- resulted in good predictions with an accuracy of 85%. Similarly, in [14] they added user age as feature along with a history of a user as a feature. They assume that if a user bullied in the past it is more likely for him to engage in bullying again. They investigated the effect of adding user features and concluded that it advances the recall with 5%. User-based features were also adopted in [16], where they added user gender and age to the feature set. The assumption was that different gender use different language and the people from different ages have different writing styles. Moreover, a new user feature was incorporated which was the user location.

### C. Network based Detection

An interesting perspective to cyberbullying detection studies the social structure of users. This starts by drawing network structure and deriving features from the graph. In [17], they focused on deriving features from social network graph. Features included: number of nodes indicating how large is the community and number of edges indicating how well connected is the community. Another research that addressed network based features is found in [12]. They used (Gephi) a graphical interface to visualise a user's connectivity based on the bullying posts. Then, they investigated the participants' role in the bullying, whether they are victims or predators.

## III. Proposed Algorithm

CNN-CB is an algorithm that advances current work in cyberbullying detection by adapting principles of deep learning instead of classical machine learning. CNN-CB architecture consists of four layers: embedding, convolutional, max pooling and dense which will be described in the following sub sections. The architecture and scope of every layer is shown in fig. 2. Its remarkable aspect is that it eliminates three classification phases previously employed by other detection algorithms, feature determination, extraction and selection. This is achieved through generating word embeddings (numerical vectors) for each word in a tweet and feeding them directly to a convolutional neural network. Detailed steps are explained in the following sub sections, and its pseudo code is represented in table II.

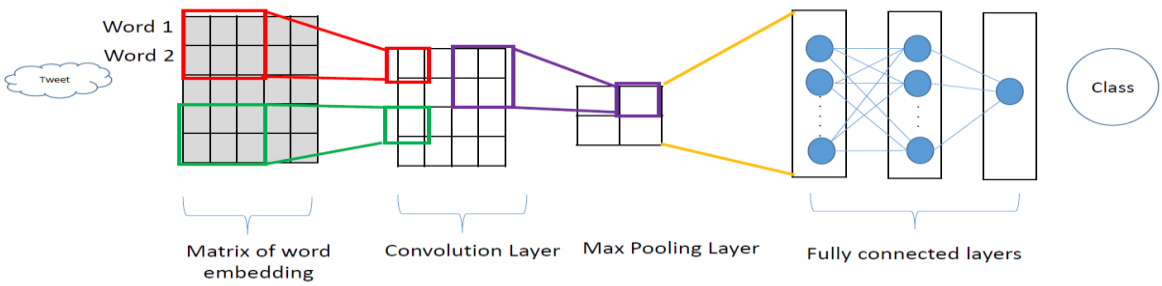



Fig. 2. CNN-CB Architecture.

TABLE I. CYBERBULLYING RESEARCH SUMMARY

| Paper | Data Mining Task | Algorithm Used | Features |
|---|---|---|---|
| [15] | Classification | SVM, J48, Naïve Bayes | - Text<br>- Profile<br>- User graph |
| [18] | Classification | SVM. | - Text<br>- Bag of words<br>- Author |
| [12] | Classification | Lib SVM | - LDA /TFIDF<br>- List of bad words<br>- pronouns |
| [19] | Classification | J48, Naive Bayes, SMO, Bagging and Dagging. | - Social network<br>- Text<br>- Part of speech |
| [20] | Classification | SVM | - Text<br>- Sentiment |
| [21] | Classification | C 4.5 | - Text<br>- List of bad words |
| [14] | Classification | SVM | - User<br>- Content<br>- Cyberbullying based |
| [22] | Classification | SVM, Naive Bayes | -Content<br>-Profile |
| [16] | Classification<br>Clustering | Fuzzy SVM | -User<br>-Location<br>-Text<br>-Media |
| [23] | Classification | Fuzzy logic | -Text<br>-Word statistics |
| [24] | Classification | SVM, J 48 | -Text<br>-TFIDF |

### A. Word Embedding

Word embeddings are a class of techniques used to generate numerical representation of textual material. A striking feature of word embedding is that they generate similar representations for semantically similar words. This remarkable feature enables a machine to actually understand what text means rather than dealing with it as strings of random numbers. In order to illustrate this great potential, fig 3 shows the similar words to word 'smart' along with their similarity score using word embedding provided by *Glove*. *Glove* [25] is one method of word embedding provided by google. This works by collecting millions of words and training a neural network to learn the similarity or differences in meaning.

In the proposed *CNN-CB*, embedding layer provided by *Keras* [26] was adopted rather than pre trained embedding like *Glove*. What distinguish this specific choice of embedding (*keras*) is that it is task specific. In other words, it takes all text (cleaned tweets in this case) and generates a vector space of vocabulary. Thus, it is easier -both in time and resource-to compute. The use of word embedding made *CNN-CB's* more advanced compared to traditional detection approaches since they incorporate semantics not just features extracted from raw text [27]. *Keras* embedding layer requires three parameters to be set prior to the construction of the vector space:

- Input dimension: specifies the total number of words in the vocabulary (whole corpus). This number is derived from the following. Let T be all tweets in the corpus.

$T = \{t_1, t_2, t_3 .. t_n\}$, n=number of tweets

*Input dimension* = length (Tokenized (T))

- Output dimension: specifies the size of the output vector from this layer.
- Input length: the length of each vector (maximum number of words per tweet). Twitter fixed maximum tweet length was not set, since this might change over time. Input length is calculated by using the following functions.

*Input length* = max (length for *t* in *T*)

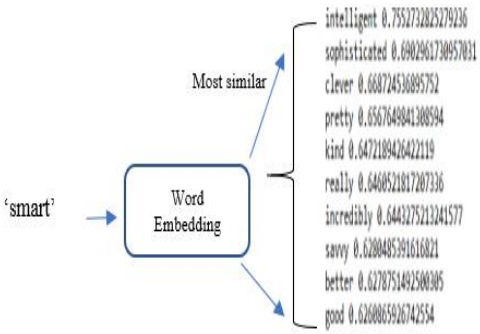



Fig. 3. Similar words to 'smart'.

### B. Convolutional Layer

The second layer after the embedding layer (in case of text) is the convolutional layer. It is the heart of a convolutional neural network. Its task is to convolve around the input vector to detect features, therefore, it compresses the original input vector while preserving valuable features. This is achieved by creating a set of matrices called *filters* of random numbers called *weights.* Each filter is then independently convolved around the original input vector creating many feature maps through elementwise multiplication with the part of the input it is currently on [28]. In order to calculate the resulting feature map, Let *V* be the input vector of words, and *F* be the filter of size *h*w*, then the elementwise multiplication is calculated according to the following equation.

$$(V*F)_{xy}=\sum_{i=1}^{h}\sum_{j=1}^{w} V_{ij} \cdot F_{x+i-1,\ y+j-1} \quad (1)$$

### C. Max Pooling Layer

What distinguishes CNN and gives it robustness and ability to deal with complex data like image and large corpus, is that it compresses the input to smaller matrices. This remarkable ability is achieved by both convolutional and max pooling layers; thus, they are used after one another. Max pooling matrix simply slides across the output of a convolutional layer and finds the maximum value of the selected area. In this way, only meaningful and clear features are preserved.

### D. Dense Layer

All layers described so far were concerned with shaping data (tweets in our case) and compressing them in a meaningful way. So far, no classification has been done. This is exactly the job of dense layers. As in neural network, dense layers are set of fully connected layers [14]. In other words, each neuron is connected to all other neurons in the following layer. The number of dense layers varies, however, the last one must have 2 neurons corresponding to the number of classes in this case.

TABLE II. CNN BASED DETECTION ALGORITHM (CNN-CB)

| *CNN-CB* |
|---|
| Input: List $T=t_1,t_2,t_3,\ldots t_n$ (n= number of tweets)<br>Number f (f=number of filters)<br>Number k (k=size of kernels)<br>Number p (p=pool size)<br>Number count (count=number of classes)<br>Number count (n=number of neurons) |
| Output: Number C=1 or 0 (0=no bullying,1=bullying) |
| **Begin** |
| Tokenize all tweets *Token* |
| Calculate total vocab, *vocab*=length (*Token*) |
| Calculate max tweet length, *len*=max (length for *t* in *T*) |
| Split tweets to testing *Test*, and Training *Train* |
| Encode training tweets |
| Encode testing tweets |
| Create *Embedding* layer, Embedding(vocab,len) |
| Create *Convolution* layer, Convolution(f,k) |
| Create *Max Pooling* layer, Pooling(p) |
| Create *Flatten* layer, Flatten() |
| Create *Dense* layer, Dense(n) |
| Create *Dense* layer, C=Dense(count) |
| **End** |

## IV. EVALUATION

Evaluation of the proposed algorithm aims to experimentally investigate crucial facts. First, that *CNN-CB* gives better results than traditional cyberbullying detection. Second, to evaluate other metrices like loss and recall. In order to have subjective evaluation, content-based detection was implemented for comparison with SVM algorithm. SVM was considered because a survey in [5], revealed that it is the mostly used in this domain. All experiments were run using *Windows* PC with 12 GB of *RAM*. All algorithms were programmed in *Python* [30] using *Spyder* environment [31]. *CNN-CB* was implemented using *Keras* [26] [29].

### A. Content-Based Detection (Cont)

There are many detection methods, however, a survey found in [5] stated that content-based methods are the most common with a total of 41 papers. Also, it has reported that SVM was the most common learning algorithm. The features included were: 1) the presence of bad words (bad words were retrieved from *noswearing.com* [32]); 2)the tweet's length;; 3)the presence of question marks since they indicate profane words; 4) the presence of exclamation marks since they indicate anger; 5) the presence of capital letters since they indicate anger.

### B. Dataset

The data set used in experiments were fetched from Twitter using Twitter streaming API [33]. A total of 39,000 tweets were retrieved from twitter public timeline. However, after annotating tweets, we found that there was an imbalanced class problem (very few bullying tweets). This has been solved by querying Twitter API with bad words from [32] so that it was more likely to return bullying tweets. After that, data were inspected and cleaned, removing duplicates and tweets with only pictures or URLs. A summary of the data collected for training and testing is presented in table III .

For data annotation, Figure 8 [34] human intelligence website was used. A job was posted, and sufficient instructions were given, and for quality purposes a test of 25 questions were required for a contributor to be accepted. Eventually, from those who succeeded the test with a percentage of 95%, two contributors were selected.

### C. Evaluation Metrics

Since cyberbullying detection is a classification task, the obvious choice of metric will be classification accuracy. However, this is an imbalanced class problem; so if we consider accuracy only as a metric then we might get an accuracy of 80% just be labelling all testing tweets with the majority class. This issue has been solved by considering two other metrics: recall and precision. All metrics are listed in the following equations.

$$accuracy=\frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$recall=\frac{TP}{TP+FN} \quad (3)$$

$$precision=\frac{TP}{TP+FP} \quad (4)$$

### D. Result

In this section, comprehensive comparison between three cyberbullying detection approaches was conducted. The aim here is to prove that the proposed algorithm CNN-CB advances the current state of cyberbullying detection by providing better predictions (higher accuracy) although it eliminates the need for feature engineering. The series of experiments starts by testing CNN-CB with different values of filters, kernels, pooling and neurons to prove that changing values changes the quality of prediction. This experiment is reported in table IV. Moreover, further experiments are conducted to test the CNN-CB model. Fig.4 and fig. 5 shows the model accuracy and loss during every epoch respectively. The loss used in here is the mean squared error The third experiment shown in table V, is conducted with the traditional approach of cyberbullying detection, specifically content-based detection cont, and provides a summarized overview about its performance. Remarkably, the advancement of CNN-CB over traditional approach is clearly reported in fig.5, fig.6 and fig.7.

TABLE III. DATASET DISTRIBUTION

| | Training | Testing |
|---|---|---|
| Bullying | 9,000 tweets | 2700 tweets |
| No-Bullying | 21,000 tweets | 6300 tweets |

TABLE IV. CNN-CB BPERFORMANCE BY VARYING VALUES OF FILTERS, KERNELS,POOLING AND NEURONS

| No. of filters | No. of kernels | Size of Pooling | No. of neurons | Accuracy (%) | Precision (%) | Recall (%) |
|---|---|---|---|---|---|---|
| 4 | 5 | 1 | 4 | 66 | 99 | 70 |
| 4 | 5 | 1 | 16 | 95 | 77 | 82 |
| **4** | **10** | **2** | **32** | **95** | **93** | **73** |
| 4 | 10 | 2 | 43 | 95 | 86 | 77 |
| 4 | 10 | 2 | 54 | 94 | 80 | 80 |
| 10 | 20 | 2 | 4 | 65 | 99 | 70 |
| 10 | 20 | 4 | 16 | 93 | 81 | 79 |
| 10 | 20 | 4 | 32 | 93 | 83 | 79 |
| 10 | 25 | 4 | 54 | 94 | 81 | 80 |
| 10 | 25 | 4 | 64 | 93 | 79 | 81 |

TABLE V. CONTENT BASED CYBERBULLYING DETECTION PERFORMANCE

| Algorithm | Accuracy (%) | Precision (%) | Recall (%) |
|---|---|---|---|
| *SVM* | *81.32* | *73* | *70* |

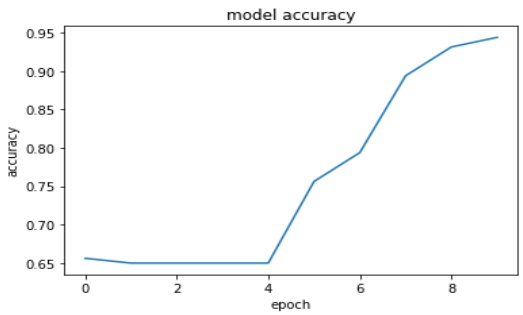



Fig. 4. Model Accuracy.

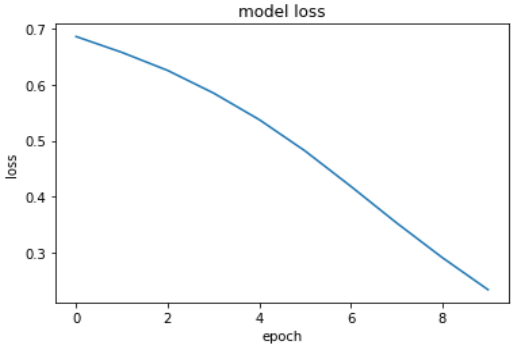



Fig. 5. Model Loss.

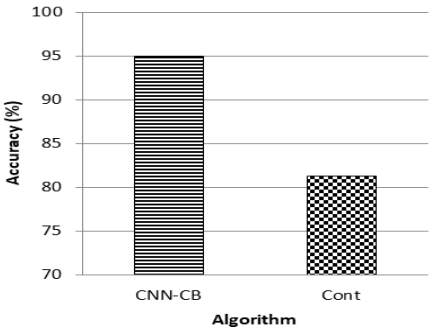



Fig. 6. CNN-CB vs Cont-SVM Accuracy.

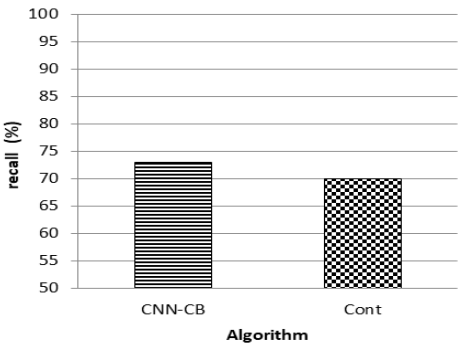



Fig. 7. CNN-CB vs Cont-SVM Recall.

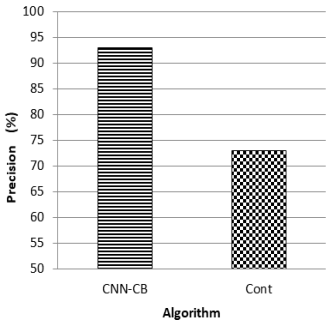



Fig. 8. CNN-CB vs Cont-SVM Precision

## V. Discussion

Cyberbullying detection has been addressed in the literature with classical machine learning approaches, mainly content-based ones. However, the conducted experiments showed that *cont-SVM* gave an accuracy of 81%, like results reported by others in the literature.

The performance of *CNN-CB* during epochs, was always raising. This is because learning is evolving with every epoch. The model started with an accuracy of 65% but rose to 95% after 10 epochs. Model loss which represents a measurement of miss classifications also proved that increasing the number of epoch improve quality of predictions.

When *CNN-CB* is compared to traditional cyberbullying approach *cont-SVM, CNN-CB,* reported better results in the three metrics accuracy, precision and recall. This is true for all variations of parameters proving that feature engineering elimination did not degrade the performance but in fact there was a noticeable improvement of about 12% accuracy. Among the three studied metrics, accuracy shows the most noticeable difference. On the other hand, recall has slightly differed between the two algorithms with them being in the 70s.

It also has been evident from table IV that changing the CNN structure has a strong impact on the resulting accuracy. Some variations produced an accuracy of 66% whereas, some produced 95%.

## VI. Conclusion

Technology revolution advanced the quality of life, however, it gave predators a solid ground to conduct their harmful crimes. Internet crimes have become very dangerous since victims are targeted all the time and there are no chances for escape. Cyberbullying is one of the most critical internet crimes and research proved its critical consequences on victims. From suicide to lowering victims' self-esteem, cyberbullying control has been the focus of many psychological and technical research.

In this article, the issue of cyberbullying detection on Twitter has been tackled. The aim was to advance the current state of cyberbullying detection by shedding light on critical problems that have not been solved yet. To the best of our knowledge, there has been no research that considered eliminating features from the detection process and automating the process with a CNN. The proposed algorithm makes cyberbullying detection a fully automated process with no human expertise or involvement while guaranteeing better result. Comprehensive experiments proved that deep learning outperformed classical machine learning approaches in cyberbullying problem.

As for future work, we would like to adapt the proposed algorithms for Arabic content. Arabic language has different structure and rules so comprehensive Arabic natural language processing should be incorporated.

## References

[1] R. Shetgiri, "Bullying and Victimization Among Children," Adv. Pediatr., vol. 60, no. 1, pp. 33–51, Jul. 2013.

[2] R. Donegan, "Bullying and cyberbullying: History, statistics, law, prevention and analysis," Elon J. Undergrad. Res. Commun., vol. 3, no. 1, pp. 33–42, 2012.

[3] D. Halpern, M. Piña, and J. Vásquez, "Loneliness, personal and social well-being: towards a conceptualization of the effects of cyberbullying/Soledad, bienestar social e individual: hacia una conceptualización de los efectos del cyberbullying," Cult. y Educ., vol. 29, no. 4, pp. 703–727, 2017.

[4] K. Van Royen, K. Poels, W. Daelemans, and H. Vandebosch, "Telematics and Informatics Automatic monitoring of cyberbullying on social networking sites : From technological feasibility to desirability," Telemat. INFORMATICS, 2014.

[5] S. Salawu, Y. He, and J. Lumsden, "Approaches to Automated Detection of Cyberbullying : A Survey," vol. 3045, no. c, pp. 1–20, 2017.

[6] T. Wu, S. Liu, J. Zhang, and Y. Xiang, "Twitter spam detection based on deep learning," Proc. Australas. Comput. Sci. Week Multiconference - ACSW '17, pp. 1–8, 2017.

[7] MIT Technology Review, "These are the 10 breakthrough technologies you need to know about right now," 2018. [Online]. Available: https://www.technologyreview.com/lists/technologies/2017/. [Accessed: 02-Mar-2018].

[8] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," arXiv Prepr. arXiv1412.1058, 2014.

[9] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in Advances in neural information processing systems, 2015, pp. 649–657.

[10] A. J. McMinn, Y. Moshfeghi, and J. M. Jose, "Building a large-scale corpus for evaluating event detection on twitter," in Proceedings of the 22nd ACM international conference on Information & Knowledge Management, 2013, pp. 409–418.

[11] H. Sanchez and S. Kumar, "Twitter bullying detection," ser. NSDI, vol. 12, p. 15, 2011.

[12] V. Nahar, X. Li, and C. Pang, "An effective approach for cyberbullying detection," Commun. Inf. Sci. Manag. Eng., vol. 3, no. 5, p. 238, 2013.

[13] C. Paper, "Methods for detection of cyberbullying : A survey," no. October, 2016.

[14] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. De Jong, "Improving cyberbullying detection with user context," pp. 2–5.

[15] E. A. Abozinadah, A. V Mbaziira, and J. H. J. Jr, "Detection of Abusive Accounts with Arabic Tweets," vol. 1, no. 2, 2015.

[16] V. Nahar, S. Al-Maskari, X. Li, and C. Pang, "Semi-supervised learning for cyberbullying detection in social networks," in Australasian Database Conference, 2014, pp. 160–171.

[17] P. K. Atrey, "Cyber Bullying Detection Using Social and Textual Analysis," pp. 3–6, 2014.

[18] S. O. Sood, E. F. Churchill, and J. Antin, "Automatic Identification of Personal Insults on Social News Sites," vol. 63, no. 4, pp. 270–285, 2012.

[19] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in Proceedings of the 3rd International Workshop on Socially-Aware Multimedia, 2014, pp. 3–6.

[20] R. Sugandhi, "Automatic Monitoring and Prevention of Cyberbullying Automatic Monitoring and Prevention of Cyberbullying," no. June, 2016.

[21] K. Reynolds, A. Kontostathis, and L. Edwards, "Using Machine Learning to Detect Cyberbullying."

[22] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Detection of Cyberbullying Incidents on the Instagram Social Network," 2014.

[23] B. S. Nandhini and J. I. Sheeba, "Online Social Network Bullying Detection Using Intelligence Techniques," vol. 45, pp. 485–492, 2015.

[24] K. Dinakar, B. Jones, C. Havasi, and H. Lieberman, "Common Sense Reasoning for Detection , Prevention , and Mitigation of Cyberbullying," vol. 2, no. 3, 2012.

[25] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on

empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[26] F. Chollet, "Keras," GitHub, 2015. [Online]. Available: https://github.com/fchollet/keras. [Accessed: 26-Aug-2018].

[27] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," J. Mach. Learn. Res., vol. 3, no. Feb, pp. 1137–1155, 2003.

[28] N. Majumder, A. Gelbukh, I. P. Nacional, and E. Cambria, "Deep learning-based document modeling for personality detection from text," IEEE Intell. Syst., vol. 32, no. 2, pp. 74–79, 2017.

[29] A. Géron, Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc.," 2017.

[30] "Python." [Online]. Available: http://www.python.org. [Accessed: 28-Aug-2018].

[31] "The Scientific Python Development Environment," 2018. [Online]. Available: https://www.spyder-ide.org/. [Accessed: 28-Aug-2018].

[32] "Swear Word List." [Online]. Available: https://www.noswearing.com/. [Accessed: 28-Aug-2018].

[33] "Python Twitter," 2016. [Online]. Available: https://python-twitter.readthedocs.io. [Accessed: 28-Aug-2018].

[34] Figure Eight Inc, "Figure Eight," 2018. [Online]. Available: https://www.figure-eight.com/. [Accessed: 28-Aug-2018].

# NADA: New Arabic Dataset for Text Classification

Nada Alalyani
Information Technology department
College of Computer and Information Sciences
King Saud University
Riyadh, KSA

Souad Larabi Marie-Sainte
Computer Science department
College of Computer and Information Sciences
Prince Sultan University,
Riyadh, KSA

***Abstract*—In the recent years, Arabic Natural Language Processing, including Text summarization, Text simplification, Text Categorization and other Natural Language-related disciplines, are attracting more researchers. Appropriate resources for Arabic Text Categorization are becoming a big necessity for the development of this research. The few existing corpora are not ready for use, they require preprocessing and filtering operations. In addition, most of them are not organized based on standard classification methods which makes unbalanced classes and thus reduced the classification accuracy. This paper proposes a New Arabic Dataset (NADA) for Text Categorization purpose. This corpus is composed of two existing corpora OSAC and DAA. The new corpus is preprocessed and filtered using the recent state of the art methods. It is also organized based on Dewey decimal classification scheme and Synthetic Minority Over-Sampling Technique. The experiment results show that NADA is an efficient dataset ready for use in Arabic Text Categorization.**

***Keywords*—Data collection; arabic natural language processing; arabic text categorization; dewey decimal classification; synthetic minority over-sampling**

## I. INTRODUCTION

Data collection consists of gathering information to assess the outcomes and validate the research study. The accuracy of data collection is crucial to keep the truth of research. Data collection is required in all research areas and studies such as mathematics, physics, humanity, business, computer science and many more.

Arabic Text Categorization is one application of Natural Language Processing in Computer Science that needs a huge amount of text documents to perform classification. Accessing to freely available corpus is a desirable aim. Unfortunately, these corpora are not easily found or not designed for Arabic Text Categorization such as Al-Dostor newspapers [1]. In other words, the existing corpora ([2], [3] and [4]) need modification before the usage. For example, increasing the number of classes, performing preprocessing techniques and providing the corpus with specific formats to facilitate the integration of the data. In fact, most of the existing Arabic corpora don't follow any technique necessary to organize the class hierarchy. This hierarchy helps illustrate the needed classes and keep corpus balanced to accomplish an accurate result. Moreover, some of the existing Arabic corpora are not dedicated for classification because either there are no defined classes such as 1.5 billion words Arabic Corpus [5], or the existing classes are not well defined ([6], [7], and [8]). Furthermore, most of the available corpora are published as raw data, which requires applying linguistic pre-processing operations such as cleaning, tokenization, normalization and stemming before use.

Consequently, the researchers in this field face a fundamental problem in comparing the results of their proposed methods with those of the state of the art techniques. This makes the validation step more difficult and time-consuming. So, it is extremely needed to propose a new Arabic corpus that overcomes the above limitations.

In this paper, we present NADA, a New Arabic Dataset built from two existing Arabic corpora and complemented with extra classes and documents. To cover the entire classes from different domains, the standard classification schemes (Dewey Decimal Classification scheme (DDC) [9]) is used to provide a logical hierarchy of classes needed in document classification. In addition, to reach a high classification accuracy, Synthetic Minority Over-Sampling Technique (SMOKE) [10] is applied to make the classes balanced. NADA is composed of 10 categories belonging to different domains, including Social science (e.g. economies, and law), Religious science (e.g. Islamic religion), Applied science (e.g. health), Pure science (e.g. Technology), Literature science, and Arts science (e.g. Sport). After the data was assembled and organized, the pre-processing methods and filtering are applied to make the data ready in ANLP and particularly ATC field.

This paper is organized as follows. Section 2 introduces the Arabic Language. Section 3 presents the Dewey Decimal Classification scheme. Section 4 surveys the existing Arabic corpus. Section 5 shows the formation of NADA corpus. Section 6 displays the experiment results and finally section 7 concludes this works.

## II. ARABIC LANGUAGE

Arabic is a complex language. It has diverse characteristics that make it different from the other languages. The Arabic word contains the diacritics placed above or below the letters rather than short vowels. However, these diacritics have been left in contemporary writing and expected to be filled in by the readers from their knowledge of the Arabic language [11]. Furthermore, in Arabic, many letters have a similar structure and are differentiated only by the existence and the number of dots. For example, the letters (b-ب, n-ن, t-ت) have the same structure but with different dot location and number. Moreover, the different shapes of Arabic letters depend on the placement of the letters in the word. Four shapes are found for 22 letters in Arabic, which are (word-initial, word-medial, and word-final). In Arabic, nouns and adjectives involve genders [12].

Another obvious complex characteristic of Arabic language is the richness of vocabulary. For example, the word "darkness" has 52 synonymous, "short" has 164, and 50 synonymous for the "cloud" [12].

## III. Dewey Decimal Classification

In order to arrange resources on the shelves and facilitate the retrieving process, the Dewey decimal classification scheme (DDC) can be used. The most usage of this scheme is in the libraries. DDC is a hierarchical number system that organizes all resources into ten main categories [9]. Each main category is then divided into ten sub-categories and so on. In this study, this scheme is used to help build NADA.

## IV. Related Works

The first step in text classification studies is data collection. The collected data must be suitable for the classification purpose. Data collection is required in each language performing text classification or other NLP applications. Many corpora can be found in English language (for example Newsgroup English benchmark [13], ACL Anthology Reference polish Corpus (ACLARC) [14], Reuters 21578 English corpus [15], and Reuters Corpus Volume 1 (RCV1) [16]) as long as in the other languages such that Chinese Souhu News corpus [17], Thai dataset [18].

In Arabic language, the state of the art studies presented a number of Arabic Corpora such that Al- Nahar [1], Al-Jazeera[2], Al-Hayat [3] and Al- Dostor newspapers [1], Hadith corpus [4], Akhbar-Alkhaleej corpus [2], Arabic NEWSWIRE [3], Quranic Arabic Corpus [4], corpus Watan-2004 [6], Khaleej-2004 [19], KACST Arabic corpus [20], BBC Corpus [7], CCN Corpus [8], Open Source Arabic Corpora (OSAC) [21] and Arabic corpus [4] that is composed of Watan-2004 and Khaleej-2004 corpora. Table 1 summarizes the existing corpora dedicated to ATC researches. Even though there are freely available Arabic corpora used in Arabic processing projects, most of them are either not suitable for text classification, or they might be appropriate for classification but still the data needs more filtering, processing and format conversion steps, which can negatively affect the classification accuracy.

On the other hand, few commercial corpora [5], are available but with extremely excessive cost. So, the need for developing free new corpora is critical in Arabic Text Categorization.

## V. NADA Dataset Setup

NADA corpus is collected from two existing corpora, which are Diab Dataset DAA corpus and OSAC corpus. DAA dataset has nine categories each of which contains 400 documents. Each category has its own directory that includes all files belonging to this category. These files have already been preprocessed and filtered [22]. The documents in each class of DAA corpus are considered in NADA corpus.

On the other side, OSAC dataset [21] has six classes each containing [500, 3000] raw documents. Each category has its own directory that includes all files belonging to this category.

The OSAC dataset is a raw data that requires preprocessing. For this, each text file is pre-processed as follows: 1) the digits, numbers, hyphens, punctuation marks and all non-Arabic characters are removed. 2) Some letters are normalized to unify the writing forms. 3) Arabic stop words like pronouns, articles, and prepositions are removed. 4) The light stemming is applied to the dataset to remove the entire affix and suffix from the word. However, Chen stemmer or Khoja algorithm for extracting the roots are not employed, because usually it is not valuable for Arabic text classification tasks, due to the conflation of various words to the same root form [12].

Furthermore, to reduce the dimensionality of the dataset, the recent new proposed Firefly based feature selection [23] is used. Firefly Algorithm is a well-known Artificial Intelligent technique applied to select the relevant words from a given document. This technique is applied to each document to reduce its size. The processed and filtered documents are considered in NADA dataset.

In this study, DAA and OSAC datasets are partitioned into two parts to building the training and testing data for the classification purpose. By this step, NADA corpus is constructed and becomes available for usage. This construction is based on DDC scheme to make its classes well organized. Figure 1 displays the hierarchy of NADA corpus; only the green classes and subclasses are considered in NADA. Furthermore, SMOTE technique is used to balance the classes and then increase the classification performance [10]. The data collection is summarized in Table 2 and 3. Table 2 shows the categories and the number of documents of OSAC and DAA datasets and Table 3 displays the content of the new corpus.

This corpus is available (https://www.researchgate.net/publication/326060650_NADA_A_New_Arabic_Dataset. DOI: 10.13140/RG.2.2.13606.01603) and can be found in various formats since the platforms and programming languages require different formats. It consists of three types of files including Attribute-Relation File Format (ARFF) file, classified text files and Sampled data file as follows:

- ARFF file: it is an ASCII file that involves a group of instances with a set of attributes. These instances are the text scripts that are involved in the text files. Each instance represents one text file. This file format is necessary to analyze and process the corpus using WEKA tool [5].
- Text files: each file involves Arabic script in a specific category. These text files are classified into 7 categories as shown in Table 3.
- Sampled file: to avoid imbalanced impact on classification results of the collected dataset, SMOTE [10] is used to balance the dataset classes. The impact of SMOTE is shown in Figure 2.

1 https://catalog.elra.info/en-us/repository/browse/ELRA-W0027/

2 alt.qcri.org/resources/aljazeeraSpeechCorpus/.

3 http://catalog.elra.info/product_info.php?products_id=632

4 https://sourceforge.net/projects/arabiccorpus/

5 https://catalog.ldc.upenn.edu/LDC2001T55

TABLE I. ARABIC TEXT CATEGORIZATION CORPORA

| Corpora | Sources | No. Of Text | No. Of Classes | Classes |
|---|---|---|---|---|
| KACST corpus [23] | Saudi Press Agency (SPA) | 1,526 | 6 | Social News, Cultural News, , General News, Economic News, Sports News, Political News |
| | Saudi News Papers (SNP) | 4,842 | 7 | Economic News, Cultural News, Social News, IT News Political, Sports News, News, General News |
| | WEB Sites | 2,170 | 7 | IT, NEWS, Economics, Religion, Medical, Cultural, Scientific |
| | Writers | 821 | 10 | Ten writers |
| | Discussion Forums | 4,107 | 7 | NEWS, IT, Religion, Economics, Medical, Cultural, Scientific |
| | Islamic Topics | 2,243 | 5 | Tafseer, Feqah, Aqeedah, Hadeeth, Linguistics |
| | Arabic Poems | 1,949 | 6 | Retha'a , Hekmah, Gazal, Hega'a, Madeh, Wasf |
| BBC Corpus | BBC Arabic website | 4,763 | 7 | Science & Technology, , World News, Middle East News Business & Economy, Sports, International Press, Art & Culture |
| CNN Corpus | CNN Arabic website | 5,070 | 6 | Business, Entertainments, Middle East News, World News Science & Technology, Sports |
| OSAC | Bbcarabic.com | 3012 | 1 | Economics |
| | Cnnarabic.com | | | |
| | Aljazeera.net | | | |
| | Banquecentrale.gov | | | |
| | Khaleej.com | | | |
| | Watan-2004 corpus | | | |
| | Hukam.net | 3233 | 1 | History |
| | Moqatel.com | | | |
| | Altareekh.com | | | |
| | Islamichistory.net | | | |
| | Saaid.net | 3608 | 1 | Education and family |
| | Naseh.net | | | |
| | CCA Corpus | 3171 | 1 | Religious and Fatwas |
| | EASC corpus | | | |
| | Moqatel.com | | | |
| | Islamic-fatwa.com | | | |
| | Saaid.net | | | |
| | Dr-ashraf.com | 2296 | 1 | Health |
| | CCA corpus | | | |
| | EASC corpus | | | |
| | W corpus | | | |
| | Kids.jo | | | |
| | Bbcarabic.com | 2419 | 1 | Sport |
| | Cnnarabic.com | | | |
| | Khaleej.com | | | |
| | Al-hayat.com | | | |
| | Arabastronomy.com | 557 | 1 | Astronomy |
| | Alkawn.ne | | | |
| | Bawabatalfalak.com | | | |
| | Nabulsi.com | | | |
| | Alkoon.alnomrosi.net | | | |
| | Lawoflibya.com | 944 | 1 | Low |
| | Qnoun.com | | | |
| | CCA corpus | 726 | 1 | Stories |
| | Kids.jo | | | |
| | Saaid.net | | | |
| | CCA corpus | | | |
| | Aklaat.com | 2373 | 1 | Cooking Recipes |
| | Fatafeat.com | | | |

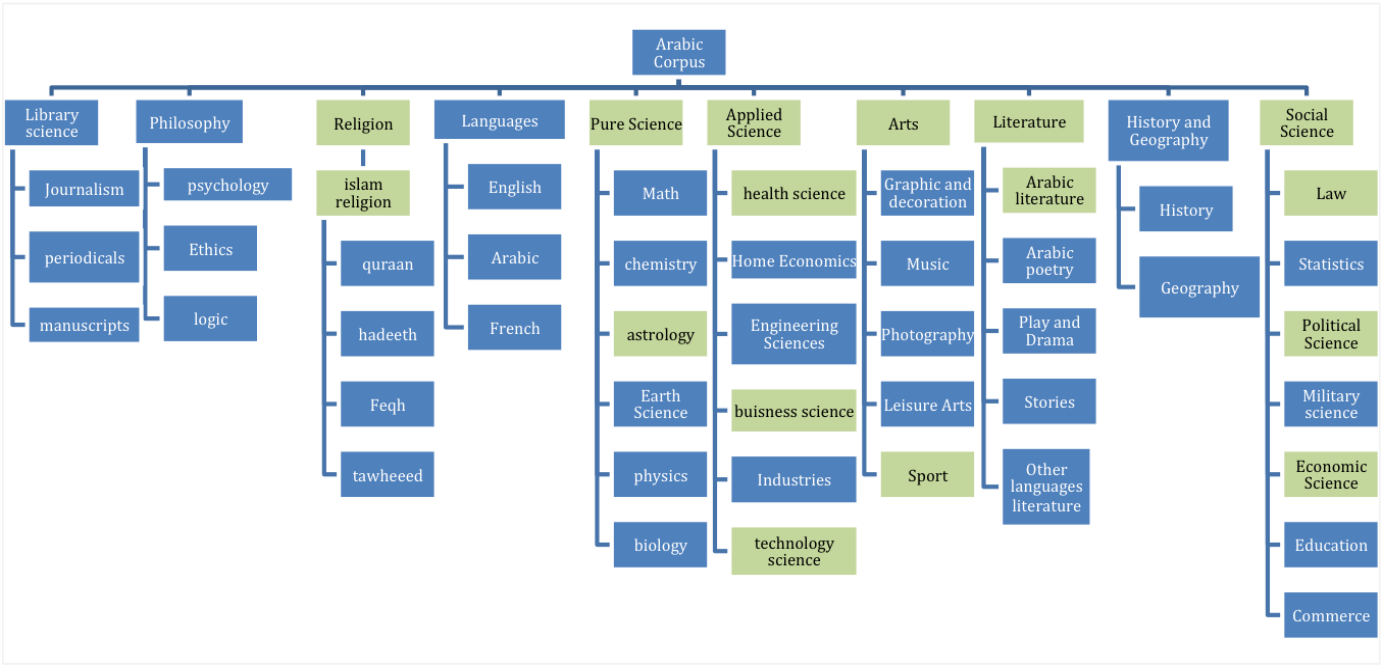



Fig. 1. NADA Corpus based on DDC Hierarchy

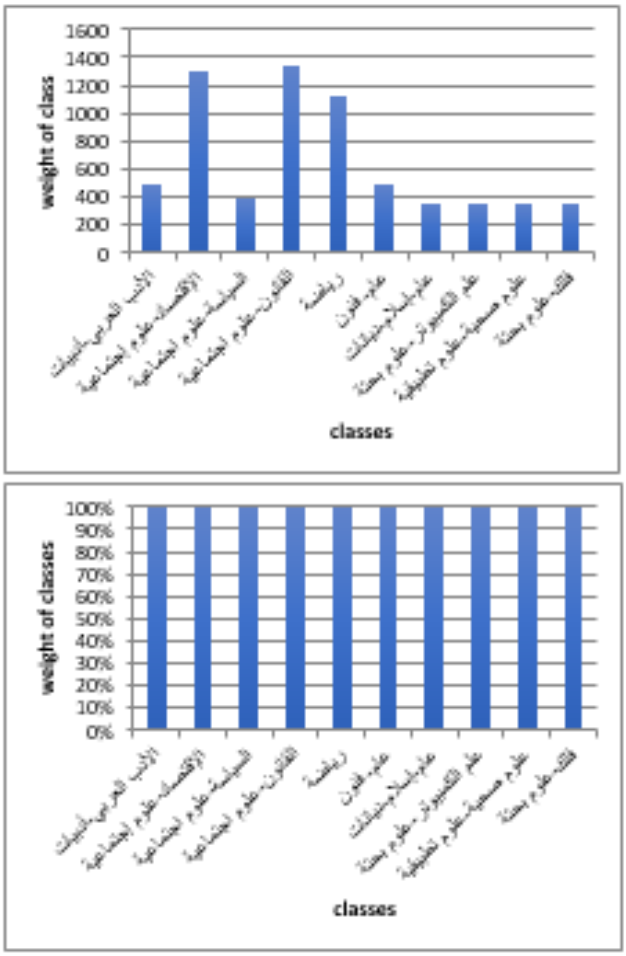



Fig. 2. The results before (above) and after (down) applying SMOTE sampling. Unbalanced data (above) and balanced data (down)

TABLE II. OSAC AND DAA ARABIC DATASETS

| New Arabic Dataset based classification experimental Results | | | | |
|---|---|---|---|---|
| OSAC | **Class** | **Precision** | **Recall** | **F-Measure** |
| | علوم إجتماعية-الإقتصاد | 0.965 | 0.985 | 0.984 |
| | علوم اجتماعية-القانون | 0.970 | 0.975 | 0.984 |
| | رياضة | 0.966 | 0.971 | 0.965 |
| | ديانات-إسلام-عام | 0.958 | 0.959 | 0.943 |
| | علوم تطبيقية-علوم صحية | 0.999 | 0.999 | 0.997 |
| | علوم بحتة-فلك | 0.996 | 0.996 | 0.996 |
| | **Weighted Avg.** | **0.982** | **0.982** | **0.982** |
| | **Correctly Classified Instances** | 98.1758 % | | |
| | **Incorrectly Classified Instances** | 1.8242 % | | |
| | **Consuming Time** | 86.95 seconds | | |
| DAA | **Class** | **Precision** | **Recall** | **F-Measure** |
| | أدبيات-الأدب العربي | 0.770 | 0.760 | 0.765 |
| | علوم إجتماعية-الإقتصاد | 0.675 | 0.856 | 0.755 |
| | علوم اجتماعية-السياسة | 0.485 | 0.436 | 0.459 |
| | علوم اجتماعية-القانون | 0.783 | 0.720 | 0.750 |
| | رياضة | 0.970 | 0.953 | 0.961 |
| | فنون -عام | 0.893 | 0.917 | 0.905 |
| | ديانات-إسلام-عام | 0.861 | 0.812 | 0.836 |
| | علومبحتة-علمالكمبيوتر | 0.863 | 0.805 | 0.833 |
| | علوم تطبيقية-علومصحية | 0.789 | 0.723 | 0.755 |
| | **Weighted Avg.** | **0.813** | **0.809** | **0.809** |
| | Correctly Classified Instances | 80.9087 % | | |
| | Incorrectly Classified Instances | 19.0913 % | | |
| | **Consuming Time** | 96.62 seconds | | |
| New corpus | **Class** | **Recall** | **Precision** | **F-Measure** |
| | أدبيات-الأدب العربي | 0.920 | 0.927 | 0.926 |
| | علوم إجتماعية-الإقتصاد | 0.908 | 0.884 | 0.871 |
| | علوم اجتماعية-السياسة | 0.948 | 0.950 | 0.944 |
| | علوم اجتماعية-القانون | 0.887 | 0.896 | 0.884 |
| | رياضة | 0.967 | 0.964 | 0.959 |
| | فنون -عام | 0.977 | 0.973 | 0.970 |
| | ديانات-إسلام-عام | 0.918 | 0.933 | 0.925 |
| | علومبحتة-علمالكمبيوتر | 0.912 | 0.925 | 0.917 |
| | علوم تطبيقية-علومصحية | 0.969 | 0.964 | 0.960 |
| | علوم بحتة-فلك | 0.967 | 0.973 | 0.925 |
| | **Weighted Avg.** | **0.939** | **0.939** | **0.932** |
| | **Correctly Classified Instances** | 93.8792% | | |
| | **Incorrectly Classified Instances** | 6.1208 % | | |
| | **Consuming Time** | 1467.62 seconds | | |

## VI. EXPERIMENTAL RESULTS

After CSV file is generated, it is converted into a sparse ARFF file format using TextDirectoryToArrf converter and StringToWordVector converter in WEKA (version 3-7-13). To measure the performance of classifying NADA, recall, precision and F1 measures are calculated and averaged using SVM classifier.

To apply the experiment, the training and testing data are required. So, the entire dataset is gathered in one ARFF file. Then, the data is divided into two partitions using percentage method, where the first partition is training data, with 60% of the dataset and the second partition is testing data with 40% of the dataset.

According to the result in Table 6, the classification accuracy of NADA is 93.8792% even though the classification accuracy of OSAC is 98.1758 % in Table 4. The result beyond the degradation of NADA's classification accuracy is due to the low accuracy of DAA where it is 80.9087 %, in Table 5 This can be explained by the fact that DAA is not well preprocessed and/or filtered which negatively affected the classification result.

For the running time, Tables 4, 5 and 6 show the time taken in classifying each dataset. The time required to classify Nada is 1467.62 seconds which is about 24 min and 28 seconds. This time is higher than the time needed for classifying OSAC and DAA datasets. This is because the number of instances in NADA dataset, which is 13066 instances is higher than that of OSAC and DAA datasets which are 3710 and 3600 instances respectively.

To conclude, NADA is well-organized dataset ready for use in ATC purpose and can be considered as a benchmark in this field of research and study.

TABLE III. NADA CORPUS COLLECTION

| NADA: New Arabic Dataset | | | | | |
|---|---|---|---|---|---|
| | **Class** | **Training Data** | **Testing Data** | **Total before SMOTE** | **Total after SMOTE** |
| NADA | Arabic Literature | 240 | 160 | 400 | 1142 |
| | Social science - economy | 822 | 485 | 1307 | 1307 |
| | Social science - politics | 240 | 160 | 400 | 1300 |
| | Social science - law | 986 | 658 | 1644 | 1644 |
| | Sport | 850 | 566 | 1416 | 1416 |
| | Art-General | 240 | 160 | 400 | 1300 |
| | General Religions - Islam | 309 | 206 | 515 | 1287 |
| | Applied science – computer science | 240 | 160 | 400 | 1300 |
| | Applied and health sciences | 257 | 171 | 428 | 1070 |
| | Pure Astronomy Science | 240 | 160 | 400 | 1300 |
| | **Total** | **4424** | **2886** | **7310** | **13066** |

TABLE IV. OSAC ACCURACY

| **OSAC Accuracy** | **Class** | **Precision** | **Recall** | **F-Measure** |
|---|---|---|---|---|
| | علوم إجتماعية-الإقتصاد Social science - economy | 0.965 | 0.985 | 0.984 |
| | علوم اجتماعية-القانون Social science - law | 0.970 | 0.975 | 0.984 |
| | رياضة Sport | 0.966 | 0.971 | 0.965 |
| | ديانات-إسلام-عام General Religions - Islam | 0.958 | 0.959 | 0.943 |
| | علوم تطبيقية-علوم صحية Applied and health sciences | 0.999 | 0.999 | 0.997 |
| | علوم بحتة-فلك Pure Astronomy Science | 0.996 | 0.996 | 0.996 |
| | **Weighted Avg.** | **0.982** | **0.982** | **0.982** |
| | | **Precision** | **Recall** | **F-Measure** |
| | **Correctly Classified Instances** | 98.1758 % | | |
| | **Incorrectly Classified Instances** | 1.8242 % | | |
| | **Running Time** | 86.95 seconds | | |

TABLE V. DAA ACCURACY

| **DAA Accuracy** | **Class** | **Precision** | **Recall** | **F-Measure** |
|---|---|---|---|---|
| | أدبيات-الأدب العربي Arabic Literature | 0.770 | 0.760 | 0.765 |
| | علوم إجتماعية-الإقتصاد Social science - economy | 0.675 | 0.856 | 0.755 |
| | علوم اجتماعية-السياسة Social science - politics | 0.485 | 0.436 | 0.459 |
| | علوم اجتماعية-القانون Social science - law | 0.783 | 0.720 | 0.750 |
| | رياضة Sport | 0.970 | 0.953 | 0.961 |
| | فنون -عام General Art | 0.893 | 0.917 | 0.905 |
| | ديانات-إسلام-عام General religions - Islam | 0.861 | 0.812 | 0.836 |
| | علومبحتة-علمالكمبيوتر Applied science – computer science | 0.863 | 0.805 | 0.833 |
| | علوم تطبيقية-علومصحية Applied and health sciences | 0.789 | 0.723 | 0.755 |
| | **Weighted Avg.** | **0.813** | **0.809** | **0.809** |
| | **Correctly Classified Instances** | 80.9087 % | | |
| | **Incorrectly Classified Instances** | 19.0913 % | | |
| | **Running Time** | 96.62 seconds | | |

TABLE VI. NADA ACCURACY

| **DAA Accuracy** | **Class** | **Precision** | **Recall** | **F-Measure** |
|---|---|---|---|---|
| | أدبيات-الأدب العربي Arabic literature | 0.920 | 0.927 | 0.926 |
| | علوم إجتماعية-الإقتصاد Social science- economy | 0.908 | 0.884 | 0.871 |
| | علوم اجتماعية-السياسة Social science - politics | 0.948 | 0.950 | 0.944 |
| | علوم اجتماعية-القانون Social science - law | 0.887 | 0.896 | 0.884 |
| | رياضة Sport | 0.967 | 0.964 | 0.959 |
| | فنون -عام Art-General | 0.977 | 0.973 | 0.970 |
| | ديانات-إسلام-عام General religions - Islam | 0.918 | 0.933 | 0.925 |
| | علومبحتة-علمالكمبيوتر Applied and computer sciences | 0.912 | 0.925 | 0.917 |
| | علوم تطبيقية-علومصحية Applied and health sciences | 0.969 | 0.964 | 0.960 |
| | علوم بحتة-فلك Pure Astronomy Science | 0.967 | 0.973 | 0.925 |
| | **Weighted Avg.** | **0.939** | **0.939** | **0.932** |
| | **Correctly Classified Instances** | 93.8792% | | |
| | **Incorrectly Classified Instances** | 6.1208 % | | |
| | **Running Time** | 1467.62 seconds | | |

## VII. Conclusion

This research study is performed to meet the extreme need of Arabic corpora and to overcome the difficulties faced by ANLP researchers especially in ATC field to find an appropriate corpus.

NADA is a New Arabic Dataset built from two existing Arabic corpora including OSAC and DAA datasets. This corpus followed a standard classification scheme (DDC) to provide logical hierarchy presentation of classes. NADA corpus is composed of 10 categories, which achieved 5 classes from the first level of DDC and some classes from the second level. To increase the classification performance, SMOTE technique is applied to balance the whole classes. This dataset passed through preprocessing and filtering steps to reduce researchers' efforts in rebuilding Arabic corpus. NADA is tested and validated using SVM classifier and three evaluation measures. The experiment results show that NADA is an efficient dataset for ATC purpose. This corpus can be extended by adding new classes and documents to increase its usage especially in Big Data and Deep Learning.

## Acknowledgments

This work was supported by the Machine Learning Research

Group; Prince Sultan University Riyadh; Saudi Arabia [RG-CCIS-2017-06-02].

## References

[1] K. Hamouda, (2013) "New Techniques for Arabic Document Classification," Heriot Watt University,.

[2] Abbas, M., &Smaili, K. (2005). Comparison of topic identification methods for Arabic language. Paper presented at the Proceedings of International Conference on Recent Advances in Natural Language Processing, RANLP.

[3] Khasawneh, R.T.; Wahsheh, H.A.; Al Kabi, M.N.; Aismadi, I.M., (2013)."Sentiment analysis of arabic social media content: a comparative study," in Internet Technology and Secured Transactions (ICITST), 2013 8th International Conference for , vol., no., pp.101-106, 9-12 Dec.

[4] Hijjawi, M.; Bandar, Z.; Crockett, K., (2013)."User's utterance classification using machine learning for Arabic Conversational Agents," in Computer Science and Information Technology (CSIT), 2013 5th International Conference on , vol., no., pp.223-232, 27-28 March

[5] Abu, Ibrahim. (2016)."1.5 Billion Words Arabic Corpus." [1402.1128] Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition, arxiv.org/abs/1611.04033.

[6] H. K. Chantar and D. W. (2011). Corne, "Feature subset selection for Arabic document categorization using BPSO-KNN," in Nature and Biologically Inspired Computing (NaBIC), 2011 Third World Congress on, pp. 546–551.

[7] R. Belkebir and A. Guessoum, (2013). "A hybrid BSO-Chi2-SVM approach to Arabic text categorization," in Computer Systems and Applications (AICCSA), 2013 ACS International Conference on, pp. 1–7.

[8] Grefenstette, G., &Tapanainen, P., (1994). 'What is a word, what is a sentence?: problems of Tokenisation, Rank Xerox Research Centre , pp. 79-87.

[9] J. Watthananon, (2014)."The relationship of text categorization using Dewey Decimal Classification techniques," in ICT and Knowledge Engineering (ICT and Knowledge Engineering), 12th International Conference on, pp. 72– 77.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, Vol 16, https://doi.org/10.1613/jair.953.

[11] Florian, R., Ittycheriah, A., Jing, H., & Zhang, T. (2003). " Named entity recognition through classifier combination", In Proceedings of the seventh conference on Natural language learning at HLT-NAACL, Volume 4, pp. 168-17.

[12] Larkey, L. S., Ballesteros, L., & Connell, M. E. (2002)." Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis", In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 275-282.

[13] M. J. Meena, K. R. Chandran, and J. M. Brinda, (2010). "Integrating Swarm Intelligence and Statistical Data for Feature Selection in Text Categorization," pheromones, vol. 1, p. 15.

[14] J. Nualart-Vilaplana, M. Pérez-Montoro, and M. Whitelaw, (2014). "How we draw texts: A review of approaches to text visualization and exploration," El Profesional de la Informacion, vol. 23, no. 3, pp. 221–235, May.

[15] Y. Xu, (2012), "A Comparative Study on Feature Selection in Unbalance Text Classification,", pp. 44–47.

[16] L. Zhang, Y. Li, C. Sun, and W. Nadee, (2013). "Rough Set Based Approach to Text Classification,", pp. 245–252.

[17] S. Puri, (2012)."A Fuzzy Similarity Based Concept Mining Model for Text Classification," arXiv preprint arXiv:1204.2061.

[18] Shalini Puri, Sona Kaushik, (2012) "An enhanced fuzzy similarity based concept mining model for text classification using feature clustering", Engineering and Systems (SCES) 2012 Students Conference on, pp. 1-6.

[19] M. Abbas, K. Smaili (2005) Comparison of Topic Identification Methods for Arabic Language, RANLP05 : Recent Advances in Natural Language Processing ,pp. 14-17, 21-23 september 2005, Borovets, Bulgary.

[20] A. Althubaity, A. Almuhareb, S. Alharbi, A. Al-Rajeh, and M. Khorsheed, "KACST Arabic text classification project: Overview and preliminary results," 2008.

[21] Saad, Motaz & Ashour, Wesam. (2010). OSAC: Open Source Arabic Corpora. 10.13140/2.1.4664.9288.

[22] Abuaiadah, D., El Sana, J., & Abusalah, W. (2014). On the impact of dataset characteristics on arabic document classification. International Journal of Computer Applications, 101(7).

[23] Larabi Marie-Sainte, S., Alalyani, N. (2018). Firefly Algorithm based Feature Selection for Arabic Text Classification. Journal of King Saud University - Computer and Information Sciences (JKSUCIS), 10.1016/j.jksuci.2018.06.004.

# An IoT based Warehouse Intrusion Detection (E-Perimeter) and Grain Tracking Model for Food Reserve Agency

Sipiwe Chihana
Department of Computer Science
Mulungushi University
Kabwe, Zambia

Jackson Phiri
Department of Computer Science
University of Zambia
Lusaka, Zambia

Douglas Kunda
Department of Computer Science
Mulungushi University
Kabwe, Zambia

***Abstract*—Zambia's agricultural sector through Food Reserve Agency (FRA) while still underdeveloped faces many challenges that range from marketing, spoilage, infestations, and theft at site, spillage and storage among others. The methods used by FRA in their business processes are largely manual as there are no systems in place. In order to help curb these problems, this paper proposed and developed novel methods that can be used to sense real-time warehouse intrusion and grain tracking within the FRA circulation. The IoT based prototype model made use of the APC220 transceiver, GSM, GPRS, RFID, PIR and cloud storage. To curb theft of grain at storage points, the system used motion sensing through the use of PIR sensors, wireless radio communication module and the GSM/GPRS technologies such that when anyone comes in the range of PIR sensor, then the sensor will send a logic signal to the microcontroller. Lastly, the RFID combined with GSM and Arduino microcontroller responsible for grain tracking. From the results obtained in the experiment conducted it is believed that once this technology is adopted, theft will be reduced and grain management in the FRA satellite Depots dotted around the country will improve.**

***Keywords*—Internet of things; motion sensing; RFID; cloud storage; GSM/GPRS***

## I. Introduction

Food is a fundamental human need and as such food security is cardinal in any country [1]. The Food and Agriculture Organization (FAO) defines food security as a condition whereby all people, at all times, have physical and economic access to sufficient, safe, and nutritious food to meet their dietary needs and food preferences for an active and healthy life [2]. In Zambia government through the FRA ensures countrywide food sustainability and offers market access for rural based small holder farmers by maintaining a sustainable countrywide strategic food reserve [3]. Additionally FRA acts as a macro-economic stabilizer for food grown in the country, such as maize which is the nation's staple crop.

The government of Zambia through FRA despite ensuring national food security, almost every year huge quantities of food are lost due to many factors such as spoilage, infestations, theft and spillage during transporting [4] in [1]. It has also been noted that FRA faces challenges in efficient management of inventory [1], this is according to the auditor general's report that was carried out between 2006 and 2009 to review the performance of FRA [5]. This was the first ever official audit report on food reserve agency grain management, according to the report FRA experienced maize shortages of 115, 516 × 50kg bags valued at ZMK4, 274, 092, 000 (old currency) at various depots. These losses were attributed to theft and grain spoilage at the warehouse depots [5].Most of the challenges reported were as a result of poor monitoring and a poor inventory management approach which is largely manual and paper based [1]. The report further stated that FRA faces challenges to effectively monitor and inspect its storage facilities and organized crime was another factor leading to shortages usually at the time of dispatch. Inorder to validate the report by the auditor general's findings as far as grain mismanagement is concerned, independent baseline studies by [1] and [6]were conducted between 2015 and 2017 respectively, this was in a sponsored joint research study by University of Zambia in partnership with Tshwane University of Technology in South Africa [6]. The baseline studies report that FRA as reported in the auditor general's findings still had challenges of manual report generation, lack of connectivity to remote warehouses [6], failure to track stock on demand [6], theft [1] and spoilage of stock due to lack of environmental monitoring [6] [1]. [7] Notes that efforts to improve upon food security subsequently reducing hunger especially in the world's poorest countries should give priority to the issue of crop losses.

However it is hardly possible to give priority to the issues of crop losses and ensure food security in the absence of proficient delivery, control and tracking mechanisms across the supply chain. As a result of the challenges identified at FRA, there is therefore need for better management of the grain through automation of the processes. [8] Augments that as the world is shifting towards new technologies and applications it is a necessary goal to adopt new technologies in agriculture as well. Therefore there is need to have a precise, and appropriate technique of tracking of the grain bags and managing the warehouse or storage facilities. This shall be done in order to reduce theft of the grain and ensure management and monitoring of the grain in a credible and efficient manner.

Our major objectives therefore in this paper are to design a warehouse Intrusion detection (E-Perimeter) and grain tracking Model based on IoT, cloud storage and mobile communication for securing grain. To also develop methods based on the

model for accessing grain status information in real-time and lastly to Develop a prototype application based on the web to provide real-time access to warehouse intrusion alerts and current warehouse grain quantities.

## II. Literature Review

This section explores the use of internet of things (IoT) and their application. The literature review creates a significant base for understanding the concept of IoT and their applications in intrusion detection and grain tracking. Lastly related works of previous research are presented.

### A. Internet of Things

With advancement in technology, many sectors have seen growth and improvement in service delivery. Today, many areas such as health, transport and logistics, retailing among others have adopted the use of internet of things (IoT). Agricultural sector has not been left out on this use of IoT. [9] Adds that special tools and strategies are handy for improvement of farming. According to the UN Food and Agriculture Organization [9], in order to feed the developing populace of the earth, the world will want to produce 70% extra food in 2050 than it did in 2006 [9]. To meet this demand, farmers and agricultural agencies are shifting towards the Internet of Things for analytics and greater manufacturing capabilities. Internet of Things (IoT) can play a great role in growing productivity, acquiring large global market and monitoring latest traits of crops [10] Defines IoT as the system of physical objects or things hooked up with hardware, software, sensors, and system connectivity which empowers these objects to gather and alternate information. IoT makes use of different kinds of protocols to work with exclusive objects. In a related study [11]adds that IoT is a network of ordinary objects which are embedded with technologies that helps to communicate and engage inside themselves and exterior environment, this in-turn affords Intellect to the objects thereby making people's lives convenient. IoT also provides networking to connect people, things, applications, and data through the Internet to enable remote control, management, and interactive integrated services [12].Fig 1 below shows a summary of the IoT definition.


Fig. 1. IoT Definition/Concept [13] [15] [16].

*1) Applications of IoT*: IoT technology has many applications and can be put to many uses, below is table 1 showing the IoT uses among others:

TABLE I. IoT smart applications and its services [12]

| No. | Application of Internet of things | |
|---|---|---|
| | ***Service Domain*** | ***Services*** |
| 1 | Smart Home | Entertainment, Internet Access |
| 2 | Smart Office | Secure File Exchange, Internet Access, VPN, B2B |
| 3 | Smart Retail | Customer Privacy, Business Transactions, Business Security, Business Security, B2B, Sales & Logistics Management |
| 4 | Smart City | City Management, Resource Management, Police Network, Fire Department Network, Transportation Management, Disaster Management |
| 5 | Smart Agriculture | Area Monitoring, Condition Sensing, Fire Alarm, Trespassing |
| 6 | Smart Energy & Fuel | Pipeline Monitoring, Tank Monitoring, Power Line Monitoring, Trespassing & Damage Management |
| 7 | Smart Transportation | Road Condition Monitoring, Traffic Status Monitoring, Traffic Light Control, Navigation Support, Smart Car support, Traffic Information Support, Intelligent Transport System (ITS) |

*2) IoT Architecture:* Architecture of IoT according to [13] in [11] is broadly classified into 4 layers. Figure 2 shows the architecture as classified by [13].

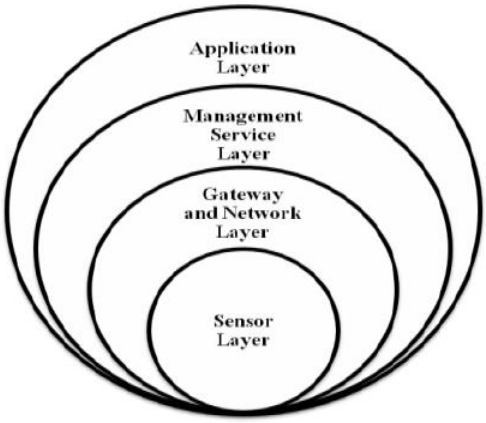



Fig. 2. Layered Architecture of IOT [11] [12] [14].

*a) Sensor Layer*

According to [15] and [12], this is lowest layer of IOT Architecture, which consists of sensor networks, embedded systems, RFID tags and readers or other smooth sensors which are one-of-a-kind varieties of sensors deployed in the field. Each of these sensors has identification and information storage (e.g. RFID tags), information collection (e.g. sensor networks).

*b) Gateway and Network Layer*

This layer is accountable for transferring the records gathered through sensors to the subsequent layer, this is according to [11]. The writer notes that this layer ought to support scalable and flexible requirements usual protocol for transferring facts from heterogeneous devices (Different sorts of sensor nodes). This Layer further ought to have high performance and sturdy network. Lastly it has to also assist more than one organization to dialog independently.

*c) Management Service Layer*

This layer acts as an interface between the Gateway - Network layer and the application layer in bidirectional mode [11]. It is responsible for gadget administration and information management and accountable for capturing massive amount of the raw data and extracting applicable information from the stored records as well from the real time data. Security and privacy of the facts should be ensured [11].

*d) Application Layer*

This is the top most layer of IoT which affords a user interface to access a range of purposes to one-of-a-kind users [11]. The functions can be used in more than a few sectors like transportation, health care, agriculture, supply chain, government, retail etc. [11] [14]

*3) IoT Elements*

*a) Sensing*

[16] Notes that the first step in IOT workflow is gathering data at a "point of activity." This can be information captured by an appliance, a wearable device, a wall set up control or any number of frequently found devices. [16] Further adds that the sensing can be biometric, biological, environmental, visual or audible (or all the above). The special component in the context of IOT is that the machine doing the sensing is not one that commonly gathered statistics in this way. Sensing technological know-how precise to this reason is required [16].

*b) Communication*

[16] Augments that IOT gadgets require ability for transmitting the information sensed at the device level to a Cloud-based provider for subsequent processing. This is where the great value inherent in IOT is created. This however requires either Wi-Fi (wireless LAN based totally communications) or WAN (wide area network… i.e. cellular) communications, [16].adds.

*c) Cloud based Capture*

[16] Observes that Gathered data is transmitted to a cloud based service where the information coming in from the IOT device is aggregated with other cloud based data to provide useful information for the end user. The facts being consolidated can be facts from different web sources as well as from others subscribing with comparable IOT gadgets [16].

*d) Delivery of Information*

The last step is delivery of useful information to the end user [16], this may be a customer, a commercial or an industrial user. It may also be another device in the M2M workflow. The aim in a customer use case is to provide the information in as simple and transparent a method as possible, [16] notes.

## III. Related Works

[17] Notes that India is one of the greatest agricultural land in the world with about 179.9 million hectares under cultivation. Still in India, meal grains are saved at warehouses using the usual manual method which leads to problems such as theft, rain, flood, variant in temperature and humidity, attacks of rodents, insects etc. the authors proposed and integrated smart sensing devices with Internet of Things (IOT) and Wireless Sensor Networks to preserve the quality and quantity of the stored products over time. According to [17] the device has capability of being controlled and monitored from remote location and delivering real time notification without human intervention. The proposed system made use of Raspberry Pi 3 Model B+, Sensors (PIR, Temperature and Fire), Ultrasonic Ranging Device, Web Camera, Buzzer, Software Requirements, Java, Win forms using c #. Net and Android studio. From this author we note that the focus was on monitoring grain quantity and quality by making use of temperature sensors for room temperature in which the grain is stored, PIR for motion sensing as well as fire sensors, however it is not clear on how the sensor data was being stored.

The study by [18] proposed an amusing and low-cost administration platform to recognize real-time monitoring and tracing for prepackaged meals for the supply chain based on Internet of Things (IoT) technologies and eventually make sure a nonthreatening and protected meals consumption environment. Following service-oriented architecture, a flexible layered architecture of tracking and tracing platform for prepackaged meals was to be developed. Besides lowering the implementation price whilst realizing fine-grained monitoring and tracing, a built-in solution of the usage of each the QR code and radio-frequency identification (RFID) tag was proposed. Furthermore, Extensible Markup Language (XML) was adopted to facilitate the information sharing amongst functions and stakeholders. In this study we observe the adoption of RFID and QR code for tracking and tracing of prepacked food in their supply chain, however the study does not show how the implementation was achieved as it only proposed the adoption of the said technologies.

[19] Proposed and developed a GSM Based Home Security System Using PIR Sensor, the system according to [19]was very simple. In their system PIR sensor was used for sensing and then microcontroller was used for controlling. GSM module was used for SMS and calling purpose. When anybody comes in range of PIR sensor, the sensor sends a signal to the microcontroller and takes control and perform a given job. [19] Adds that Home security system using PIR sensor can be been implemented for security concerns. The system would then make a call to the user when any intruder is detected inside a room. According to the author the results presented showed that the system was cheap, implementation is easy, Low power consumption and automated operation. However the gap noticed in this system was a lack of storage provision of the SMS alerts, storing alerts is important as it is easy to make reference to past intrusions and help make better decisions on how best security measures could be improved upon.

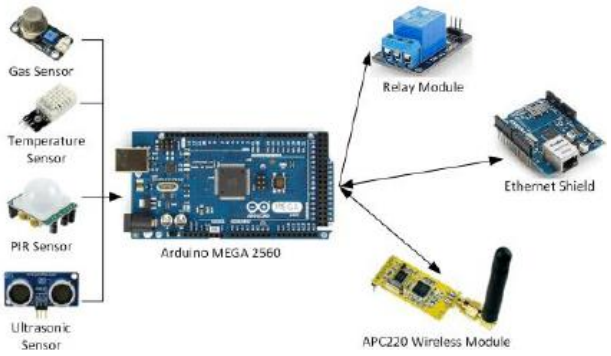



Fig. 3. Actual Hardware Implementation [20].

TABLE II. EXPERIMENTAL RESULTS OF APC220 WIRELESS COMMUNICATION MODULE [20]

| Scenario | Trials | | | |
|---|---|---|---|---|
| | *No of trials* | *Success trials* | *Failed trials* | *%Error (failed total trials)* |
| Line of sight | 10 | 10 | 0 | 0.00 |
| Across Building | 10 | 10 | 0 | 0.00 |

[20] In Malaysia proposed a smart home system using internet of things and four types of sensors including PIR, temperature, ultrasonic, and smoke gas sensor for automatic environmental control and intrusion detection. In their proposed paper, experiments on various sensors were conducted. Next, the communication channel using wireless and Ethernet modules were also discussed. Additionally, the authors said solar charger enhanced the availability of their prototype system. Results showed the effectiveness of proposed smart home system in the prototype and real life experiments [20]. Figure 3 shows the actual hardware implementation they had proposed.

[20] Also conducted an experiment to evaluate the wireless communication module on line of sight and across the building. Table II below shows results that were presented

The overall performance of the system was verified with very low percentage error [20], several experiments were conducted to evaluate the suitability of the selected sensors, as well as the data communication modules using Ethernet W5100 and APC220 Wireless communication. All the sensors and data communication modules were connected to Arduino as the IoT platform. The focus of study by [20] was muchly on experimenting and testing the effectiveness of selected sensors, wireless communication modules and Ethernet cables.

[21] Conducted an experiment called alarm based human motion detection, an embedded system which they used to provide security. [21]Adds that this system was better as opposed to having a manual security system, [21] notes that the use of alarm based detection system for detecting human motions to provide security reduces man power and is very affordable. When a person moving around the circuit, PIR sensor detected the change in the IR levels of surroundings then the system sends a signal to the microcontroller [21]. Features of their proposed system was that it is very cheap and affordable, It is efficient and portable that means it can placed at any place easily, very easy to use and Maintenance cost is low. It was observed that study by [21] made use of motion sensing to provide an alarm in case of intrusion at home as a means of security. Further in this system there is no provision of SMS alerts, phone call and storage of the alerts in case the home owner is not at home and an intrusion occurs.

[22]Aimed at making a smart surveillance system using Raspberry Pi along with PIR sensor and Raspberry Pi-Camera. PIR sensor was used to detect the motion whenever someone comes within its range. As soon as PIR Sensor detects the motion, Pi-Camera activates and captures an image. This image is then stored in the system and finds for a human face in the captured image using Python. [22] Further notes that by using motion detection, it saves the monitoring time and cost and has it has gained a lot of interests over the past few years. The system also uses GSM/GPRS technology, so that they can use long heave communication for the monitoring section and allow for the system to be pragmatic in real time.

[23] Proposed Internet of Things and Wireless Sensor Network that lead to agricultural transformation. They also designed, tested and explored an “Internet of Things“ based device which is capable of evaluating the sensed information and then conveying it to the intended user. The device is controlled and monitored from remote location and it can be applied in farming fields, grain stores and cold stores for security purpose [23]. In this device revealed sensors and electronic devices are integrated using Python scripts. And in this system IP based CCTV security cameras entail network connectivity for monitoring from far-flung locality.

[24] The paper proposes a novel security framework in light of Open source cloud server "things speak .com" and a minimal cost esp8266 Wi-Fi module. The task incorporates a PIR module which always monitors the Home or Work space .When the PIR module recognizes a trespasser it sends a flag to the microcontroller and the controller is associated with an Esp8266 Wi-Fi module and furthermore to an alert framework. The System transmits an alarm flag to think speak which gives an alarm motion on the clients' cell phone. The framework utilizes a second esp8266 module which is modified to go about as a web server and enables the client to actuate or deactivate the security framework by means of any gadget with the internet. The framework additionally utilizes a thumb print reader which controls the opening and the end of a locker entryway. In this way the framework utilizes Wi-Fi module and microcontroller to control the security framework from the client's cell phone.

[25] Acknowledges that the Internet of Things (IOT) is the driving force for a great deal of innovation and economic activities. The proposed project expands on the concept of IoT to agriculture. [25]Notes that farm equipment’s are costly and their theft is a very common problem faced by farmers. The purpose of the project was to make a device which would help them to monitor their farm equipment’s and prevent theft. Implementation consists of sensing parameters like motion of the object and location of the device and alerting them in case of any variation in their position or motion through mobile phone. A wireless network was setup between farm devices

and a central server which will further communicate with Android Application which would then receive alarm of theft and location of the device using global positioning system (GPS) of theft device.

[26] Proposed a prototype to reduce the theft of farmer input such as fertilizer under Farmer Input Support Program (FISP) in Zambia. The prototype used barcodes placed on bags to identify each bag and record the data in a database housed in the cloud. Upon distributing the bags to farmers, the database updates are made to reflect the changes in inventory levels. The input inventory management system was set to control the theft of inputs that stood at 56% in the warehouses. From the baseline study conducted by [26] the results show theft occurrence at the satellite depot, holding depot and during transportation. The main contribution in this paper was the use of cloud storage, the system did not however provide for warehouse storage mechanism despite the authors findings which showed high percentage of theft.

[27] Designed a wireless remote sensor network based on android things for monitoring environment conditions in storage warehouses for the Food Reserve Agency in Zambia. The system used sensor nodes to monitor temperature, humidity, water and motion. The temperature, humidity and water were used to monitor the state of the storage environment while the motion sensor was used to detect the presence of intruders. The data from the sensor nodes was sent to an aggregator node via ZigBee and later transmitted to the server in the cloud using a GSM network. The cloud storage service used was third party which may be inappropriate for storing data for a government agency. However the prototypes did demonstrate the use of ZigBee in environmental monitoring. In their system there is also no provision of the usage of the Web application for a more visualized sensor values.

[28] Proposed an Application of Internet of Things (IoT) in Supply chain management of Agricultural products. Here they designed a website for the farmer and manager. Using the Radio Frequency identification (RFID) technology, agricultural products are tracked at every stage of supply chain, The RFID tag is attached to the product container and scanned against RFID reader and ESP8266 at every stage of SCM. Once the tag is scanned that tag id is uploaded to the database. The product information is uploaded by the farmer through the manager. The main objectives were: Automatic identification of agricultural products, Global tracking and tracing of agricultural products, providing the complete information of the whole life cycle of the product to the farmer inorder to achieve transparency in the system. There exist a number of literature based on the implementation of IoT [29], [30], [31], [32], [33], [34], [35], [36], [37].

From the proposed and implemented experiments/ projects reviewed from literature, it has been observed that there is substantial potential to be tapped from the use of internet of things in agriculture and other sectors. We also note that there exist a number of research gaps in the sense that many authors merely proposed there prototypes and explained there benefits without putting these prototypes to full use. In line with our objectives and related literature reviewed we note that most experiments or projects did not make use of cloud storage, IoT Sensors have the ability to generate large volumes of data at a high rate and that poses a huge challenge because more storage space is required to keep the data, therefore cloud storage comes in to address the challenge. [38]Points out that having a centralized store of data puts real-time data in the hands of decision makers.

The literature reviewed was also the basis of the researcher's theoretical model for intrusion detection in the grain storage warehouse and grain tracking. For intrusion detection (e-perimeter) literature suggests Apc220 a wireless radio communication module as it offers no line of sight problems, GSM for SMS alerts and for data transfer to the cloud storage GPRS, Passive infrared(PIR) for motion detection, and open source microcontroller Arduino boards as our IoT platform as the hardware is less expensive. Apc220 transceiver offers no problem with line of sight and across buildings, has low power consumption, security and control capabilities, signals travel in the transmitter and receiver even when there is an obstruction which makes it suitable for security systems. For grain tracking literature suggests radio frequency identification (RFID) as it offers automatic identification of agricultural products, Global tracking and tracing of agricultural products. Also providing the complete information of the whole life cycle of the product so as to achieve transparency in the system, GPRS for transferring data to the cloud.

## IV. Proposed Model Prototype

The proposed model seeks to address the gaps in grain theft at FRA warehouse storage using motion sensing and grain tracking using RFID. The mapped current FRA business processes are presented in the study by [1].

The researchers proposed model is based on the current FRAs business processes but with a focus mainly on warehouse storage and grain tracking and tracing. The proposed IoT model is divided into two, intrusion detection which is achieved by putting in place an E-perimeter to monitor unauthorized entry in the restricted zone were the grain shall be stocked so that the grain is secured, the IoT technologies adopted for this module are Wireless data communication module APC220, GSM to alert management by sending an SMS once motion detection happens with the help of the motion sensors, GPRS via an HTTP GET REQUEST for data transfer to the cloud database and PIR (passive infrared) because these are mainly for intrusion detection, it is also capable of detecting any trespassing that occurs in the range of the motion sensor. The second module responsible for grain tracking, once the grain bags are tagged and ready for dis-patch, this module shall also be used for tallying so as to compare the number of dispatched bags against the received from one depot to another within the FRA circulation, this shall be achieved by the use of an IoT device called Radio Frequency Identification (RFID) to uniquely identify each bag, additionally we shall use GSM, GPRS modules that are embedded in the SIM808 communication module. Figure 4 below shows the proposed model business process for FRA.

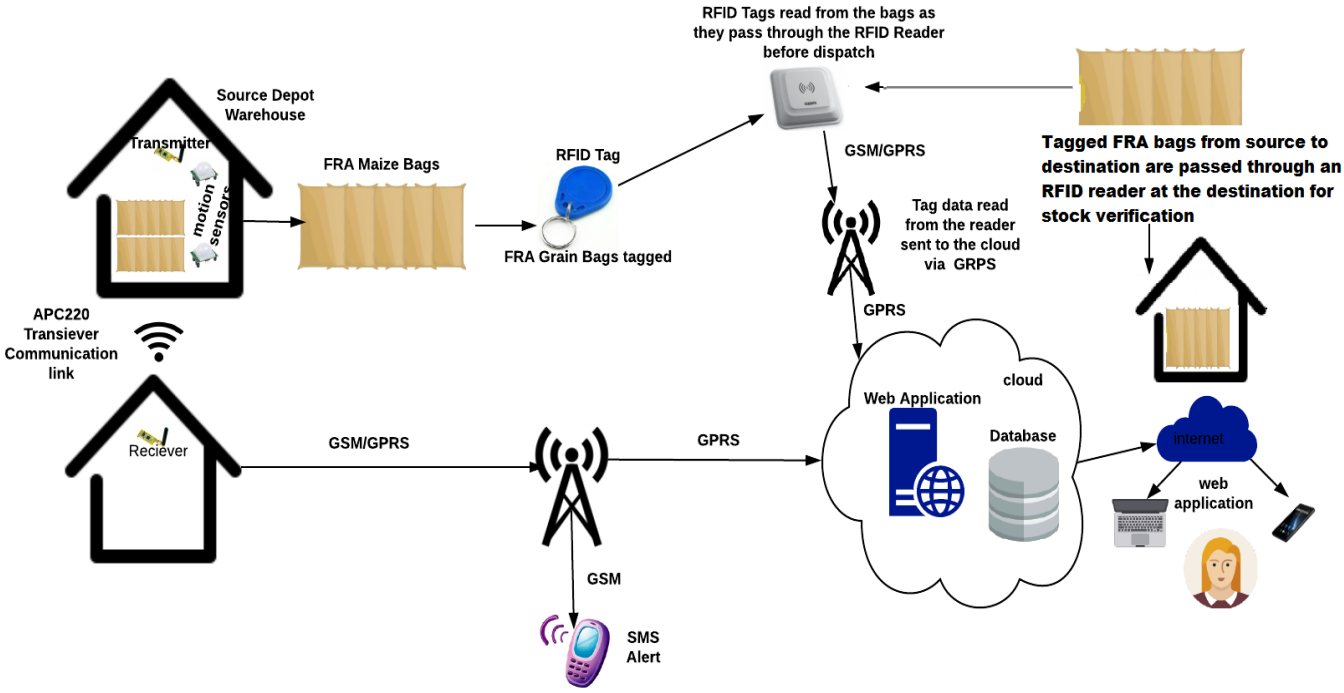



Fig. 4. Proposed IoT based Business Process Model for Warehouse Intrusion Detection and Grain Tracking.

## V. Materials and Methods

The main purpose of this research is to help food reserve agency solve the current theft problem by developing a grain tracking and e-perimeter system prototype that is cost effective. So the researcher made use of the components that have effective operation and usage. In this section, hardware parts, i.e. hardware components used for the project are outlined.

*1) Materials*: Material/ components for intrusion detection include:

- GSM/GPRS Module (SIM808)
- Arduino Uno (Micro controller)
- Passive Infrared (PIR) Sensors
- Apc220 Wireless Communication Modules

Material/ components for grain tracking include:

RFID Reader module

RFID tags

GSM/GPRS Module (SIM808)

Arduino Uno (Micro controller)

Software requirements

- PHP for the web application
- C++ for the microcontroller programming
- CSS Bootstrap for styling
- Java script
- MySQL remote database
- Apache remote Webserver
- Arduino IDE

*2) Methods:* The method adopted for testing and developing our prototype model was experiment based that enabled us to break down our work into modules namely the warehouse intrusion detection (e-perimeter) and the grain tracking module.

### *A. Warehouse Intrusion Detection (E-Perimeter)*

In this experiment a linear array of 2 Passive Infrared Sensors (PIR) are connected to a microcontroller (Arduino) at the transmitting end. The PIR constantly senses from an infrared inference given off by a human or animal. Upon breakage of the infrared flux, a signal is generated and sent to the micro controller using the digital connection bus. The embedded program parses the signal and transmits it to the receiving end as an integer “1” to signify intrusion or “0” to signify a void wirelessly using the included Omni-directional wireless radio receiver-transmitter. The receiving end is fitted with an Omni-directional receiver-transmitter, it receives the given flag; i.e. an integer signifying intrusion or a void. The embedded program then sends an SMS alert to a predefined cell phone number using the connected GMS module over the GSM network, simultaneously upon sending of an SMS alert via GSM network, GPRS using the HTTP GET request is used, the application on the remote server is invoked to parse the alert data and insert it into the cloud database for storage. A summarized working of this module is presented in a flow chart in figure 5 below.

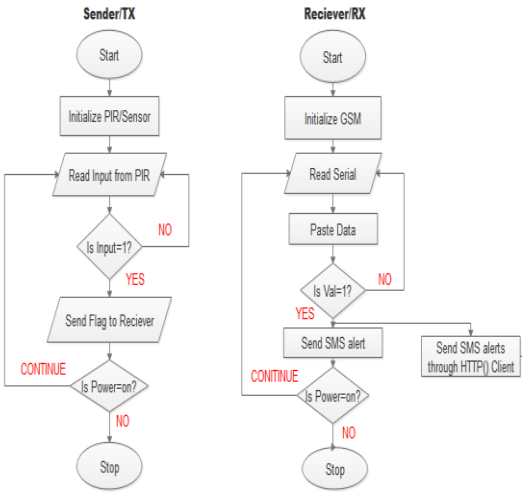



Fig. 5. Flow Chart Working of the Intrusion Detection Module.

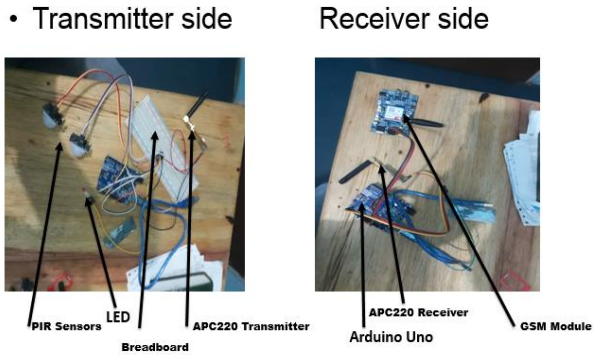



Fig. 6. Hardware Setup of the Intrusion Detection System.

### B. RFID Grain Tracking Module

The communication paradigm in the RFID setup is similar to that of the intrusion detection module mentioned above.

An RFID reader module is connected to a microcontroller (Arduino), bags of produce are tagged with RFID tags then each bag is scanned upon delivery as well as dispatch point. At the scanning point is the buzzer that gives a signal in form a buzz once the tag has been read. On delivery the quantity of the produce at a given warehouse is incremented in real-time in the cloud database. The opposite happens when a bag is scanned at dispatch. The RFID scanning device is embedded with a warehouse or depot id, this warehouse id is sent along with the tag id through a parameterized URL to the remote server via GPRS. Then the server side application on the application layer of IoT parses and stores the received parameters in the remote database. Further when an RFID tag is read, its id information is stored remotely by the application, the designated store id then is used to increment quantities in the related store. The RFID setup tracks the amount of produce at a specific warehouse to reflect current quantities. This information is overlaid on an interactive satellite map alongside other warehouse information. The application makes use of the Arduino microcontroller, GSM and GPRS module as well as RFID. The hardware setup of this module is depicted in figure 7 below, and the working algorithm depicted in a flow chart in figure 8.

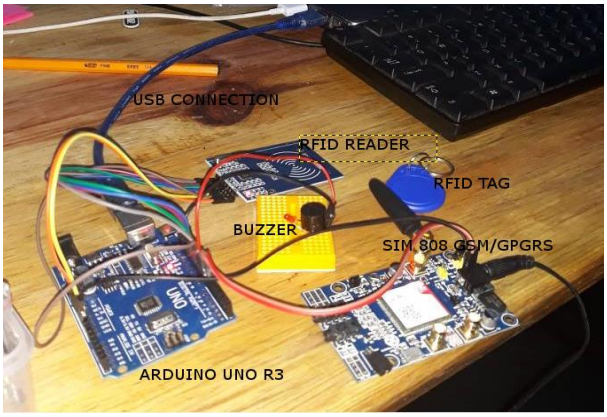



Fig. 7. Hardware Setup of the RFID Module.

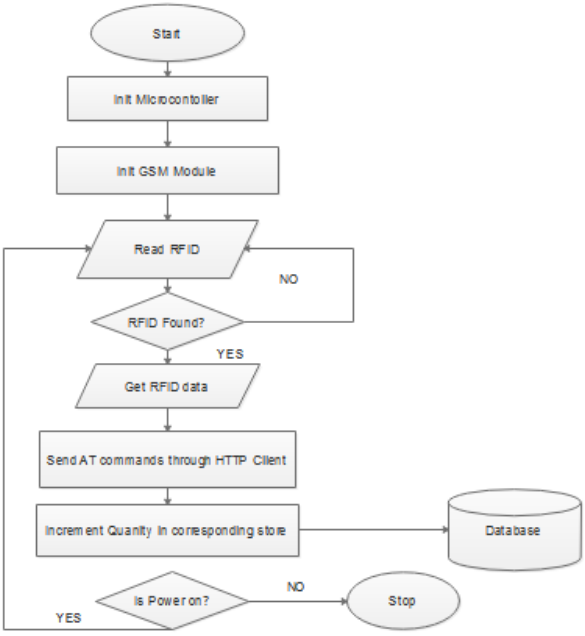



Fig. 8. Flow chart of RFID Grain Tracking.

### C. Result

#### a) E-Perimeter Results

The e-perimeter prototype hardware in figure 6 was put to a test. Below is the figure 9 showing how the intrusion can be detected.

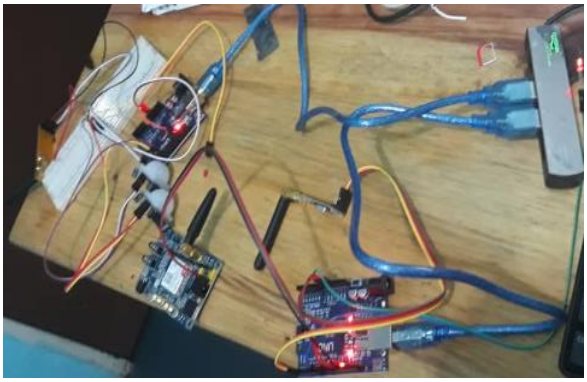


Fig. 9. Motion Intrusion Testing.

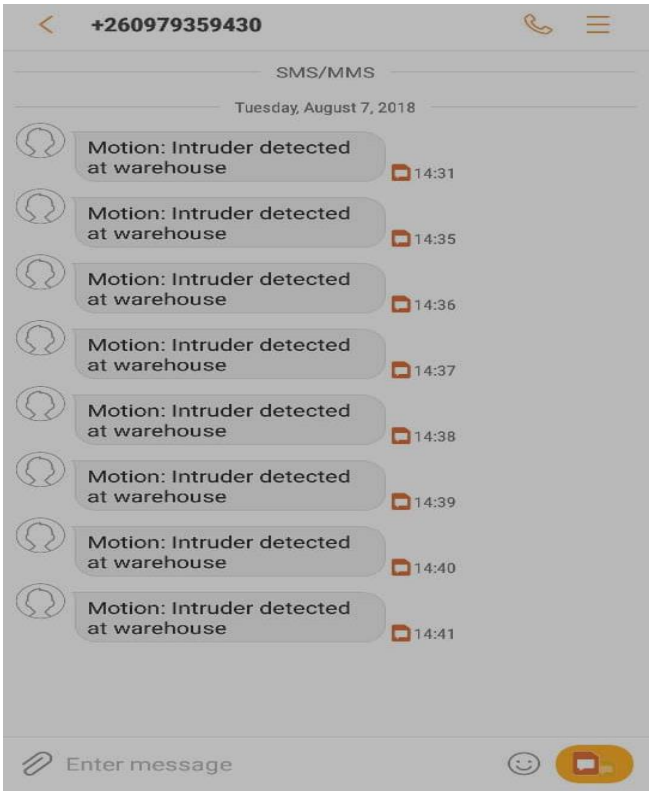



Fig. 10. SMS Alert to the Pre-defined Number.

To detect motion the intruder has to be in the range of 7 meters and any range below 7 meters because that is how far the PIR motion sensing can detect. However this setup forms a wireless sensor network, that is we have a sensor node, microcontrollers and transceivers, this implies that the communication between two apc220 transceivers should be wireless, so both the transmitter and the receiver have to be in the same network to communicate wirelessly. In the experiment of figure 10 motion sensors are set up at the sensor node, what was observed was an LED flashing light when an intruder comes in the range of the sensor node. As the motion got detected the transmitter on the sensor node transmitted the motion value to the receiver side in this network that received the sensor value and checked to confirm the received value and through the GSM module once the received value is confirmed to be that of motion an SMS alert through the GSM is sent to the designated number to alert them on the intrusion. The figure below shows the received alert message.

For monitoring the alerts on our web application we developed a web application which is hosted in the remote cloud server. The web application was developed using PHP, HTML, Bootstrap and JavaScript. The web application interface for alerts allows the administrators to view alerts of the warehouse intrusion, the warehouse name is captured so that it easy to know which warehouse had an intrusion, the intrusion time is also captured and the alert message. Figure 11 shows the alerts application view, the view gets updated as and when new intrusions are recorded.

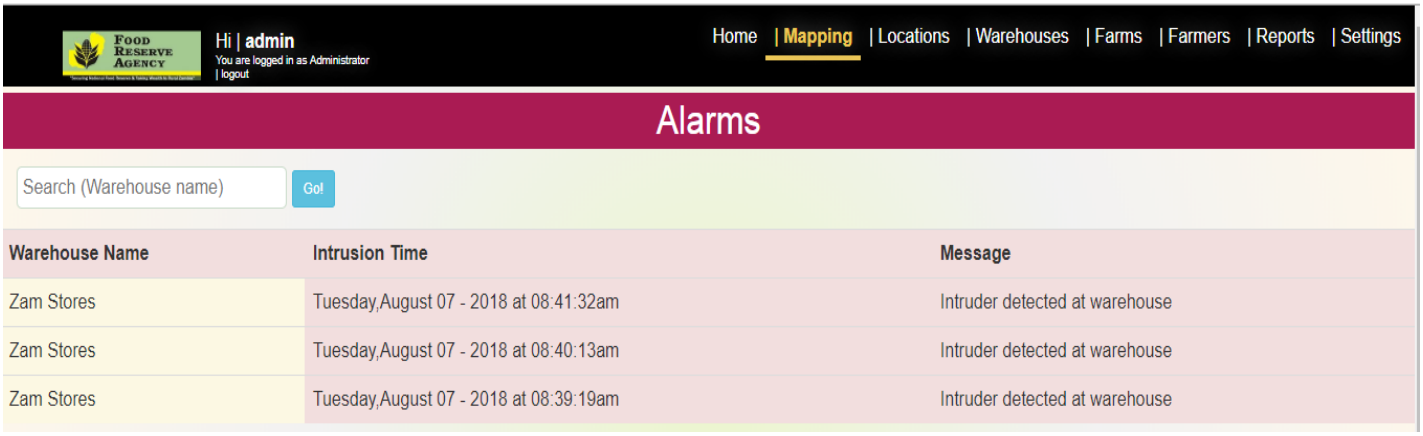



Fig. 11. Cloud Application of the Alarms.

*b) Grain Tracking using RFID Results*

The RFID tags are scanned against the RFID reader as can be seen in fig 12. After a successful scan the tag data is sent to the Arduino, the Arduino through sim808 GSM/GPRS module sends the scanned data to the cloud. Once the tag data is received and stored in the database, it is later overlaid on the google map embedded on the PHP webpage as spatial data together with other warehouse information. Quantity on the info window is representing the scanned RFID tags that are representing grain bags, the rest of the information showing on the info window of the warehouse is added by the admin upon creation of the warehouse. Fig 13 below shows the warehouse spatial data before scanning of the tags and Fig 14 after scanning.

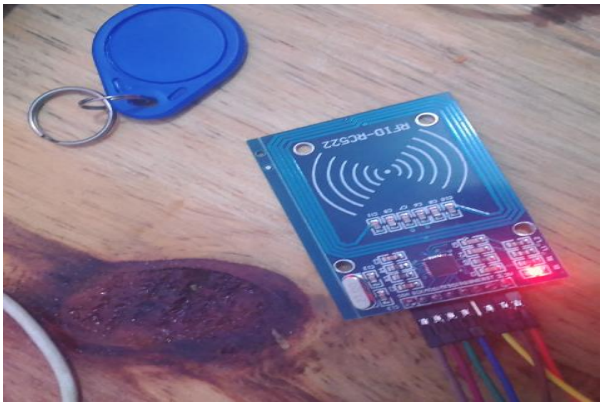


Fig. 12. Tag scanning Against the RFID Reader.

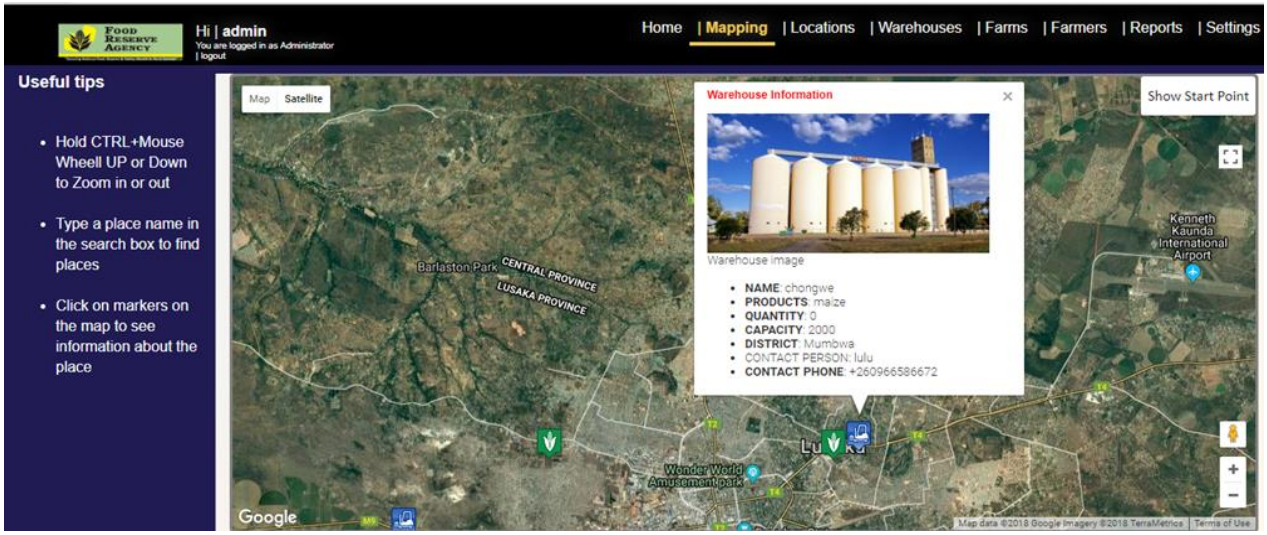



Fig. 13. Before Scanning.


Fig. 14. Quantity Overlaid on an Interactive Satellite Map Alongside Other Ware House Information.

### D. Findings

The experiment conducted was our proof of concept to ascertain the working of the prototype model that was proposed. From the experiment conducted it has been noted that the model proposed is not just a layout of the processes but the practicality behind it is possible as we have tested our modules and the methods we utilized proved beyond doubt that they are a better option for FRA grain management. The results achieved through the use of IoT is evidence enough that this is the route to take in order to reduce on theft at grain storage points(depots), further with the employment of RFID in FRA circulation security would be achieved as the grain moves from one depot to another verification by way of scanning and serving to the cloud as well as displaying the scanned quantities on the map will improve upon accountability as it would be easy to track the grain from source to destination in terms of its quantity. However from the intrusion detection (e-perimeter) module we the author observed that due to the blocking (synchronous) nature of the GSM module in use, the number of services that can be deployed on a single module would mean sacrificing the "real-time" paradigm which is paramount to monitoring in the e-perimeter application. This means when the perimeter is breached the GSM module must then begin to transmit alert messages by means of SMS and sending an HTTP REQUEST to the cloud application, this further implies changing the operation modes whenever communication occurs which in turn creates a blind-spot in the monitoring application as the GSM module ignores any other intrusions occurring while it is processing. This however can be overcome by implementing an array of GSM Modules as a future development in order to promote asynchronous operation of the alerting service, that is a single GSM module will be dedicated to performing a single function in the processing and transmission or alerts through appropriate channels hence blocking is alleviated since no module will have to wait for any in order to transmit an alert over a specific operation mode.

## VI. Discussion and Conclusions

From the proof of concept experiment carried out on the proposed IoT prototype based warehouse intrusion detection (e-perimeter) and grain tracking model, it has been observed that once these methods are adopted by the food reserve agency, food security will greatly improve as the sole aim of these methods are to reduce the chances of theft at FRA, and

the various ways in which these methods will bring about improvement in food security are that national resources will be preserved i.e. the money invested in buying of grain will be reduced as security mechanisms would be in place. Further the proposed methods will bring about efficiency in handling and managing grain and depots dotted around the country. These methods will also make it easy for FRA in terms of prosecuting would be offenders as and when theft is involved as there will be evidence readily available in form of the alert reports and other related data stored in the cloud database. Food reserve agency being the back born of Zambia's economy in terms of food security needs to improve on food security and management to meet the growing food needs due to population increase.

Based on the objectives of this study the authors arrived at the following conclusions that Internet of things based technologies RFID, WSN and cloud storage utilization reduces occurrences of theft and administrative errors, and aids in significantly improving efficiency and accountability. It was further drawn from the experiment and observation analysis on the model that the agency had an advantage of the modern proposed technology in this research.

## ACKNOWLEDGMENT

The authors would like to thank God almighty for the grace of good health, his mercies and grace upon us. Most importantly the grace he accorded us in our research endeavors.

## REFERENCES

[1] L. M. Cynthia and P. Jackson, "A Wireless Sensor Network Based Grain Inventory Management System for Zambia's Food Reserve Agency," IJIRSET, vol. 5, no. 3, pp. 3519-3526, March 2016.

[2] FAO, "Report of the World Food Summit," FAO, Rome, 1996.

[3] M. Nicole, S. T and J. Robert, ZAMBIAN SMALLHOLDER BEHAVIORAL RESPONSES TO FOOD RESERVE AGENCY ACTIVITIES, Lusaka, 2011.

[4] FAO, Global food losses and waste: Extent, Causes and Prevention, 2011.

[5] O. C. Anna, "FRA Audit report," Lusaka, 2009.

[6] C. Mulima and J. Phiri, "A Remote Sensor Network using Android Things and Cloud Computing for the Food Reserve Agency in Zambia," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 8, no. 11, pp. 411-418, 2017.

[7] W.-S. Charlie, J. T and M. W, "THE IMPACT OF REDUCING FOOD LOSS IN GLOBAL COLD CHAIN," University of Norttingham, Norttingham, 2015.

[8] Dr.N.Suma, S. R. Samson, S.Saranya, G.Shanmugapriya and R.Subhashri, "IOT Based Smart Agriculture Monitoring System," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 5, no. 2, p. 177 – 181, February 2017.

[9] V. N. Malavade and P. K. Akulwar, "Role of IoT in Agriculture," in National Conference On "Changing Technology and Rural Development" CTRD, 2016.

[10] E. C. Laxmirajam and D. R. S. Kawitkar, "A Survey: IoT based Vehicle Tracking System," International Journal of Advanced Research in Computer and Communication Engineering, vol. 6, no. 5, pp. 791-795, 2017.

[11] Soumyalatha and S. G. Hegde, "Study of IoT: Understanding IoT Architecture,Applications, Issues and Challenges," in 1st International Conference on Innovations in Computing & Networking (ICICN16), Bangalore, 2016.

[12] E Ankit, K. Jekishan and S. Chaudhary, "Internet of Things: Architecture and Research Challenges," CSI Communications, pp. 21-25, June 2016.

[13] Z. Ying-conga and Y. Jingb, "A Study on the Fire IOT Development Strategy," International journal of safety and security engineering, vol. 4, no. 2, p. 135–142, 2014.

[14] A. Khalid, "Internet of Thing Architecture and Research Agenda," International Journal of Computer Science and Mobile Computing, vol. 5, no. 3, p. 351 – 356, March 2016.

[15] D. Bandyopadhyay and J. Sen, "Internet of Things - Applications and Challenges in Technology and Standardization," Wireless Pers-commum, 2011.

[16] S. Choudhari, T. Rasal, S. Suryawanshi, M. Mane and P. S. Yedge, "Survey Paper on Internet of Things: IoT," International Journal of Engineering Science and Computing, vol. 7, no. 4, pp. 10564-10567, 2017.

[17] P. Kavya, K. Pallavi, M. Shwetha, K. Swetha and M. B. S, "Use of Smart Sensor & IoT to Monitor the Preservation of Food Grains at Warehouse," International Journal for Research Trends and Innovation, vol. 2, no. 6, pp. 449-454, 2017.

[18] Z. Li, G. Liu, L. Liu and X. Lai, "IoT-based tracking and tracing platform for prepackaged food supply chain," Industrial Management & Data Systems, vol. 117, no. 9, pp. 1906-1916, 2017.

[19] K. A. Rajani U S, "GSM Based Home Security System Using PIR Sensor," IJECT, vol. 8, no. 2, April-June 2017.

[20] S. ,. Teddy, R. Intan, Y. Husna, K. Mira and M. Hasmah, "Performance Evaluation of Smart Home System using Internet of Things," International Journal of Electrical and Computer Engineering (IJECE), vol. 8, no. 1, pp. 400-411, February 2018.

[21] K. Ajay, R. Prince, K. Justice and A. Mr, "MOTION DETECTION USING PIR SENSOR," International Journal of Scientific Development and Research (IJSDR, vol. 1, no. 5, pp. 103-106, 2016.

[22] A.Nivedha, V.Soundariya, J. C and K. Shaik, "Internet of Things - Smart Surveillance System using PIR Sensor with Raspberry Pi," International Journal on Future Revolution in Computer Science & Communication Engineering, vol. 4, no. 4, p. 22 – 24, April 2018.

[23] B. N. Tanmay and k. p. Pushpendra, "Development of IoT based Smart Security and Monitoring Devices for Agriculture," in 6th International Conference- Cloud System and Big Data Engineering (Confluence), 2016.

[24] Safa.H, P. ,. Sakthi, G. P. Vikkashini, S. Vishnupriya and Boobalan.T, "IOT based Theft Premption and Security System," International Journal of Innovative Research in Science,Engineering and Technology, vol. 5, no. 3, pp. 4312-4317, March 2016.

[25] P. S. Kushal, s. j. Mahendra, k. Santosh and D. A. Jain, "IoT based Theft Detection and Alarming System for Farm Equipment," Sinhgad Institute of Management & Computer Application(SIMCA), pp. 1-5, 2016.

[26] J. Chomba and J. Phiri, "Using the Cloud Architecture to Automate the Farmer Input Support Programme (Fisp) Inventory System," International Journal of Sciences: Basic and Applied Research (IJSBAR), vol. 28, no. 3, pp. 231-244, 2016.

[27] M. Chibuye and J. Phiri, "A Remote Sensor Network using Android Things and Cloud Computing for the Food Reserve Agency in Zambia," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 8, no. 11, pp. 411-418, 2017.

[28] S. Shambulingappa H and D. Pavankumar, "Application of IOT in Supply Chain Management of Agricultural products," International Journal of Engineering Development and Research, vol. 5, no. 3, pp. 863-867, 2017.

[29] P. A. Shinde and Prof.Mr.Y.B.Mane, "Advanced Vehicle Monitoring and Tracking System based on Rasberry Pi," in IEEE Sponsored 9th International Conference on Intelligent Systems and Control (ISCO), 2015.

[30] M. P. R. Joshi, M. V. V. Patil, M. P. S. Koli and M. B. S. Tade, "DEVICE TRACKING USING EMBEDDED GPS AND ZIGBEE TECHNOLOGY," International Journal For Technological Research In Engineering, vol. 4, no. 8, pp. 1175-1180, 2017.

[31] B. Kodavati, V.K.Raju, S. Rao, A.V.Prabu and T.Appa Rao, "GSM AND GPS BASED VEHICLE LOCATION AND TRACKING," International Journal of Engineering Research and Applications, vol. 1, no. 3, pp. 616-625.

[32] E. T and K. P, "The Smart Transportation using IoT and Intelligent Transport System in GPS Localization," International Journal of Innovative Research in Computer and Communication Engineering, vol. 4, no. 6, pp. 11197-11202, 2016.

[33] K. Priya, P. Manju, V. Mythra and S. Umamaheswari, "IoT Based Vehicle Tracking and Accident Detection System," International Journal of Innovative Research in Computer and Communication Engineering, vol. 5, no. 3, pp. 4424-4430, 2017.

[34] P. G. H, S. J. A, S. S. S., P. V. R. Kaushik and P. P. S.B., "IOT based Vehicle Tracking & Vehicular Emergency System- A Case Study and Review," International Journal of Advanced Research in Electrical,Electronics and Instrumentation Engineering, vol. 6, no. 10, pp. 8001-8012, 2017.

[35] T. S. Gunawan, I. R. H. Yaldi, M. Kartiwi and H. Mansor, "Performance Evaluation of Smart Home System using Internet of Things," International Journal of Electrical and Computer Engineering (IJECE), vol. 8, no. 1, pp. 400-411, 2018.

[36] Dhanalakshmi1 and a. s. leni, "instance vehicle monitoring and tracking with internet of things using arduino," international journal on smart sensing and intelligent systems special issue, pp. 123-135, 2017.

[37] "IoT Based Smart Security and Home Automation System," in International Conference on Computing, Communication and Automation (ICCCA2016, 2016.

[38] S. Alinani, P. Jackson, N. Mayumbo and K. M. M, "E-governance systems: A case study of the development of a small-scale Farmer database," Zambia Information Technology (ICT)Journal, vol. 2, no. 1, pp. 7-15, 2018.

# Communication System Design of Remote Areas using Openbts

Winarno Sugeng
Departement of Informatics Engineering
Institut Teknologi Nasional (Itenas)
Bandung – INDONESIA

Theta Dinnarwaty Putri
Departement of Informatics Engineering
Institut Teknologi Nasional (Itenas)
Bandung – INDONESIA

***Abstract*—OpenBTS is a software-based GSM BTS, which allows GSM cell phone users to make phone calls or send SMS (short messages), without using a commercial service provider network. OpenBTS is known as the first open source implementation of the GSM industry standard protocol. The OpenBTS network is a network that is easy to implement and also inexpensive in maintenance and installation. Communication using a cell phone is not only needed in urban areas but remote areas currently require. But the problem is that not all remote areas get services from commercial cellular operators. By implementing mobile phone communication using OpenBTS, remote communication is very likely to be implemented. In this research communication design was delivered using GSM mobile phones using OpenBTS with telephone and SMS services.**

***Keywords*—OpenBTS; GSM; communication; remote areas**

## I. Introduction

Communication is a process in which a person or several people, groups, organizations, and communities create, and use the information to connect with the environment and others [1]. Even so, they can exchange information.

Voice Over Internet Protocol (VoIP) is the easiest way to make phone calls over the internet by sending packets through a packet switched networks [7]. This technology does not use the Public Telephone Network (PSTN), but calls are made through internet protocol data networks. VoIP has enormous benefits of increasing savings, high-quality voice and video streaming and some other value-added services. The current VoIP software applications that have been used are Skype, Google talks and Windows Live Messenger, and other similar applications.

OpenBTS (Open Source Base Transceiver Station) is an open source GSM-based GSM software, which allows GSM phones to communicate without a network of commercial cellular operators [4]. OpenBTS is known as the first open source implementation of the standard GSM industry protocol. For transmitting power of 100mW, the price is around 25 million rupiah per piece, far below the usual Cellular BTS which is usually in the order of a million to several billion rupiahs.

One of the criteria in remote areas contained in article 2 of Permendikbud Number 34 of 2012 are not available and/or very limited in public facilities, educational facilities, health facilities, electricity facilities, information and communication facilities, and clean water facilities.

To enable remote communication, in this case, communication in areas using a mobile phone, this condition includes communication at the same time. OpenBTS which is a GSM-paketbased software which is open source and allows communication free of cost from commercial cellular operators. Communication using OpenBTS can be built independently.

## II. System Architecture

### A. OpenBTS Architecture

Basically, VoIP-based communication utilizing the internet network [7] is sufficient, but in the field, the communication infrastructure that used to exist is GSM-based, the communication device that is widely used is GSM-based. For this reason, further research is needed to bridge this problem. OpenBTS is the solution for this, where OpenBTS is able to unite internet-based communication and GSM. So that communication technology still uses VoIP but the advanced network uses GSM.

OpenBTS is a BTS using a software-based GSM (Global System for Mobile Communications) network, which allows GSM mobile phone users to make phone calls or short messages in this case SMS (Short Message Service), without using a network of telephone operators commercial mobile. open source software used to revolutionize mobile networks using traditional telco, traditional, complex protocols and hardware systems with flexible internet protocols and network architecture [8]. In this study used OpenBTS V5.0, which features GPRS Support, A3-A8-A5 / 1 authentication and encryption, handover support, and more structured files.

The BTS function for sending and receiving GSM signals is replaced by USRP (Universal Software Radio Peripheral) which is controlled by transceiver applications. All activities on layer 1 (one) on the BTS side will be processed by USRP. Furthermore, the upper layer functions of BSC and MSC will be replaced by OpenBTS software. Then the VLR and HLR functions as registers will be replaced by asterisk software. All software on this network communicates using SIP (session initiation protocol). USRP is a high-speed Digital Signal Processing (DSP) based radio-based radio platform for the practical development and deployment of flexible RF systems from DC to 6 GHz. USRP is hardware and currently has several versions. The most recent version uses Gigabit Ethernet so it can be stored on top of the tower easily in order to cover a wider area [2]. Figure 1 is the default OpenBTS architecture

[8]. Figure 2 is an OpenBTS system of Range Network type RAD1 production, in this case, used in research. The price of the RAD1 OpenBTS module is quite expensive for the size of the study, currently, around the US $ 3,000 and must be purchased directly from the United States (https://rangenetworks.com/get-started/), entered in the USRP1 category. But when compared to the price of commercial BTS is not based on open-source software that is widely used for the same specification OpenBTS price is very cheap, because the price of commercial BTS can reach US$ 300,000. Commercial BTS is about 100 (one hundred) times more expensive than OpenBTS.

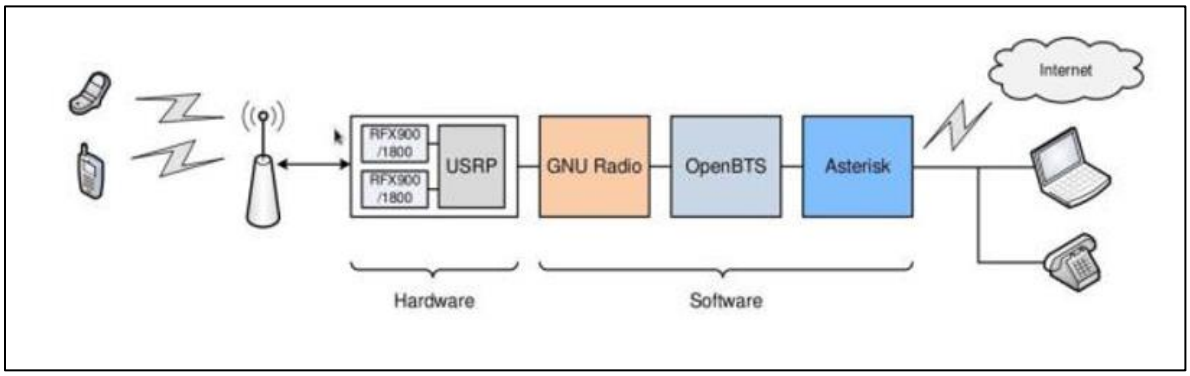



Fig. 1. OpenBTS standard architecture [2].

Fig. 2. OpenBTS module type USRP1 from Range Network

### B. UHD

UHD (USRP Hardware Driver) is a device driver provided by Research Ettus for use by USRP products. This product supports Linux, MacOS, and Windows platforms. Some frameworks include GNU Radio, LabVIEW, Matlab and Simulink using UHD [6]. Functions provided by UHD can also be accessed directly with the UHD API, which provides native support for C ++. UHD also provides the portability of the entire family of USRP products. Applications developed for a particular USRP model will support other USRP models if proper consideration is given to sample rates and other parameters. PyBOMBS (Phyton Build Overlay Managed Bundle System) is a management system for resolving dependencies or a collection of multiple applications for radio software including UHD installations.

### C. Asterisk

Asterisk is a free and open source framework for building communications applications that can transform ordinary computers into rich communications server or central phone features [5]. The Asterisk function used in OpenBTS is PBX Switching Core, which functions to handle incoming calls in the direction of Asterisk, calls can come from various interfaces and Asterisk can be used to create and deploy a wide range of telephony applications and services, including IP PBXs, VoIP gateways, calls center ACDs and IVR systems. Asterisk is released under the GNU General Public License (GPL) and is available for free download.

### D. Smqueue

OpenBTS requires to store and forward facilities such as the SMS Center of a conventional GSM network to send SMS. This function is represented by a smqueue application [3]. Smqueue uses the RFC 3428 standard as a method for sending Instant messages between users on the OpenBTS network. This standard is an extension of the SIP, therefore the procedure for connecting also follows the protocol.

## III. Design Process

The design process begins by installing USRP (Universal Peripheral Radio Software) which is equipped with USRP drivers using GnuRadio to introduce the software operating system used in Linux Ubuntu. After GnuRadio is installed and knows USRP, OpenBTS's open process is the main software that utilizes USRP. The main key of OpenBTS is an IP-based telephone center (Softswitch), in this study Asterisk was then a means of registering a GSM SIM card. The process is carried out with a GSM communication system like the services of commercial cellular operators in general.

The system used is a collection of several software and devices that communicate with each other. OpenBTS and all software that is on the list of needs for computers.

The entire system components designed and their interconnections are shown in the block diagram in Figure 3.

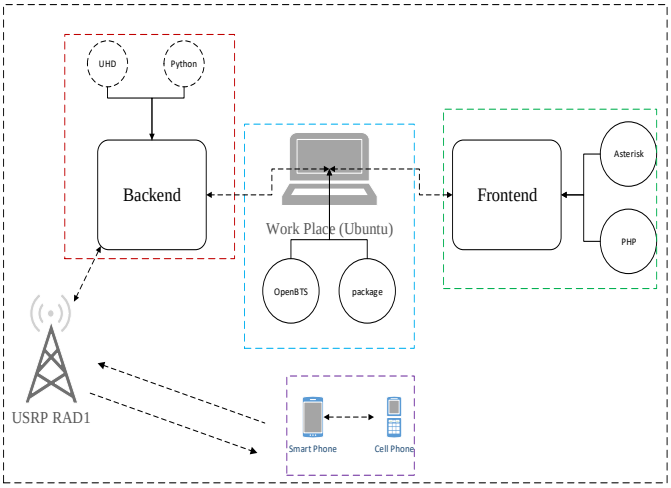



Fig. 3. System Block Diagram

From Figure 3 it can be seen that there are 3 main systems namely backend, Works Place, and frontend. The backend itself is a link between USRP RAD1 and workplace, using the UHD-hosted package. The workplace itself is a place where the operator/admin organizes the network through the OpenBTS interface and other supporting packages. While frontend is a system that regulates the relationship between users (Smartphones, cell phones) and other users.

The hardware component used is the RangeNetwork USRP type RAD1 (USRP1), while the software used is the Linux Operating System Ubuntu 14.04 32bit LTS, GNURadio, OpenBTS 5.0 (asterisk, sipauthserve, smqueue, openbts).

Before the installation process, some initial configuration is required. The first major configuration is USRP Handling, Ubuntu uses udev to handle hotplug devices, and in default does not provide non-root access to USRP. For this reason, several scripts are done to make USRP accessible to non-root, in this case, to handle USRP connect via USB both live and hotplug. If the USRP handling process is successful, the USRP will be recognized by the operating system.

The installation process starts with the installation of development packages so that the installation supports UHD, namely: libboost-all-dev libusb-1.0-0-dev python-cheetah doxygen python-docutils autoconf libtool libosip2-dev libortp-dev g ++ sqlite3 libsqlite3-dev erlang libreadline6 -dev libncurses5-dev libuhd-dev libuhd003 libboost-dev ntp. The installation process continues with the GNURadio, OpenBTS installation. After the installation process, the basic configuration process on SQLite is done to create the OpenBTS database. Before being activated, the system reboots so that all installations work. The final stage is the activation of the following applications:

*1)* Asterisk, as a telephone operator machine
*2)* Sipauthserve, as a registration machine for the SIP protocol
*3)* Smqueue, as an SMS engine
*4)* Openbts, as a BTS engine

## IV. Experiments And Results

### A. Telephone Card Registration

The use of the system starts with new users who have not registered in the OpenBTS system to register, by configuring the BTS option on the mobile phone in the BTS name option selected as the name of the built BTS, in this case, the name configured is OpenBTS, as shown in Figure 4.

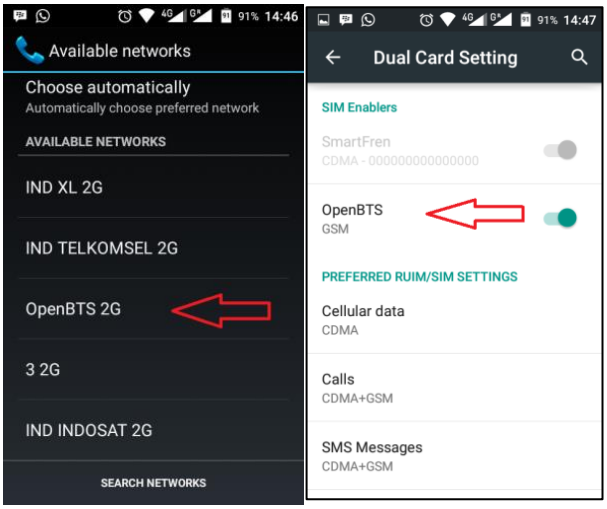



Fig. 4. Registration Process

The operator's technical and administrative process before registering is registering the IMSI and giving a telephone number, in this case mapping the selected telephone number to IMSI into the OpenBTS database.The process is as follows:

IMSI 1st mobile 510113157642463, this will be numbered 0201063001, the order is:

```
sudo asterisk -rx "database put IMSI IMSI510113157642463 0201063001"
sudo asterisk -rx "database put PHONENUMBER 0201063001 IMSI510113157642463"
```

IMSI 2nd mobile 510101421302212, this will be numbered 0201063002, the commands are:

```
sudo asterisk -rx "database put IMSI IMSI510101421302212 0201063002"
sudo asterisk -rx "database put PHONENUMBER 0201063002 IMSI510101421302212"
```

The test results after number mapping can be seen in the asterisk database, via the asterisk -rx command "database show" will read as follows:

```
/IMSI/IMSI510101421302212          : 0201063002
/IMSI/IMSI510113157642463          : 0201063001
/PHONENUMBER/0201063001            : IMSI510113157642463
/PHONENUMBER/0201063002            : IMSI510101421302212
```

After the registration process the operator then the telephone operator through the registration process, in this case, activates the IMEI sim card used which has been registered in the database server and the user will get a new alias or phone number that can be used in the OpenBTS network. The activation process by sending an SMS to number 101 with the contents of the telephone number message that has been mapped, if the number activation process is successful, the system will send an SMS sign of successful registration activation as shown in Figure 5.

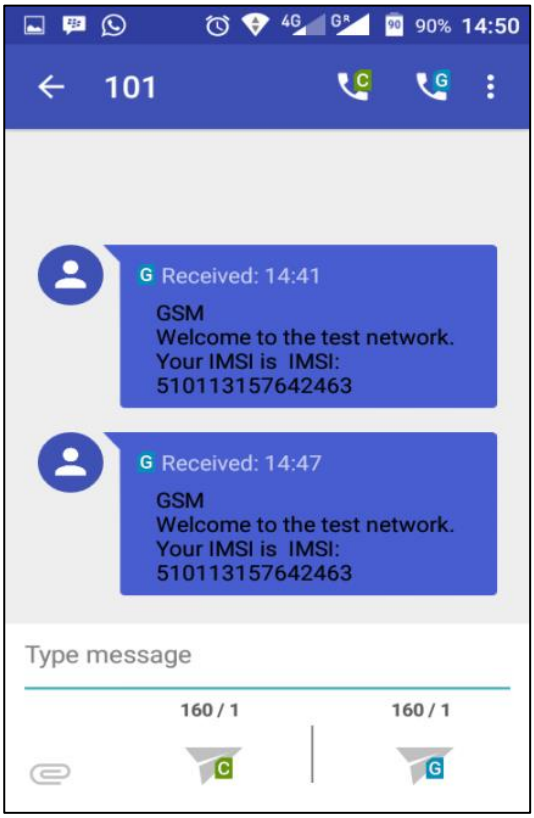



Fig. 5. Successful Registration

## B. Telephone Service

Telephone services starting with the system will check the IMSI data and the destination number to be registered or not. If it is registered, it will be connected to the destination number and the phone record data will be stored in the system database. If the destination number is not registered, the system will send a warning and if the destination number is busy, the system will send a busy tone, as in Figure 6.

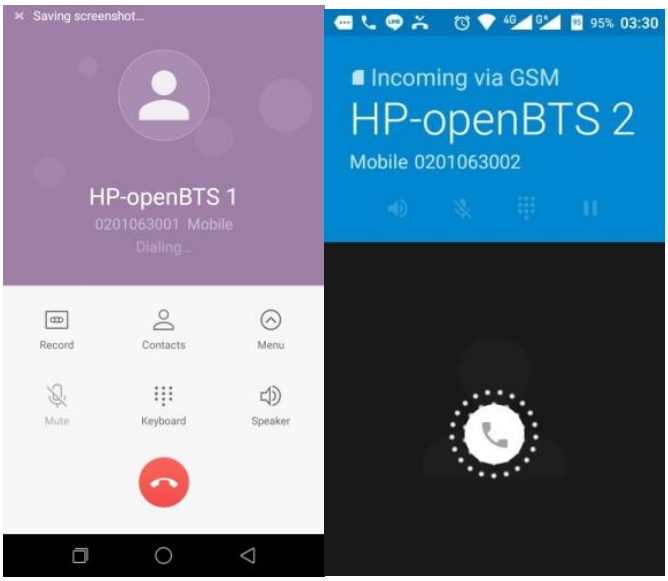



Fig. 6. Telephone Call

```
Feb 1 06:16:54 OpenBTS openbts: NOTICE 1952:2219 2018-02-01T06:16:54.1 TBF.cpp:1125:mtCancel:
06:16:54.1,513222: 06:16:54.1@@@failed tbf cause=3103 TBF#23 mtMS= MS#2,TLLI=80286f86
mtDir=RLCDir::Up#012#011 channels: down=( 0:1) up=( 0:1,usf=1)#012#011 mtState==TBFState::DataFinal mtAttached=1
mtTFI=23 mtTlli=0x80286f86#012#011 mtMsgExpectedBits=0x10 mtMsgAckBits=0x22
mtExpectedAckBSN[4]=513224#012#011 mtUnAckMode=0 OnCCCH=0 mtAssignCounter=1 N3103=9#012#011
msNumDataU5FGrants=46 msAckNackU5FGrant=45 Pacch= idle=0#012#011 TimingError=(-0.08 min=-1.27 max=1.30
avg=0.62 N=82) RSSI=(-63 min=-63 max=-57 avg=-61.50 N=82) CV=(63 min=56 max=63 avg=59.43 N=7) ILev=(0) RXQual=(0)
SigVar=(0) ChCoding=(0)#012#011 VR=6 VQ=6 stalled=0 mNumUpBlocksSinceAckNack=4 mtUpState=3#012#011 blocks:
total=14 unique=9 grants=46
Feb 1 06:17:01 OpenBTS CRON[2230]: (root) CMD ( cd / && run-parts --report /etc/cron.hourly)
Feb 1 06:17:14 OpenBTS asterisk[1188]: VERBOSE[1346][C-00000000]: netsock2.c:516 in ast_set_qos: == Using SIP RTP
TOS bits 184
Feb 1 06:17:14 OpenBTS asterisk[1188]: VERBOSE[1346][C-00000000]: netsock2.c:538 in ast_set_qos: == Using SIP RTP
CoS mark 5
Feb 1 06:17:14 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: pbx.c:4898 in pbx_extension_helper: -- Executing
[0201063001@from-openBTS:1] Verbose("SIP/00101100010-00000000", "Dialplan started") in new stack
Feb 1 06:17:14 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: app_verbose.c:104 in verbose_exec: Dialplan
started
Feb 1 06:17:14 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: pbx.c:4898 in pbx_extension_helper: -- Executing
[0201063001@from-openBTS:2] Set("SIP/00101100010-00000000", "CALLER_IMSI=IMSI510101421302212") in new stack
Feb 1 06:17:14 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: pbx.c:4898 in pbx_extension_helper: -- Executing
[0201063001@from-openBTS:3] Verbose("SIP/00101100010-00000000", "Get CID from CALLER_IMSI:
IMSI510101421302212") in new stack
Feb 1 06:17:14 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: app_verbose.c:104 in verbose_exec: Get CID from
CALLER_IMSI: IMSI510101421302212
Feb 1 06:17:14 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: pbx.c:4898 in pbx_extension_helper: -- Executing
[0201063001@from-openBTS:4] Set("SIP/00101100010-00000000", "CID=0201063002") in new stack
Feb 1 06:17:14 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: pbx.c:4898 in pbx_extension_helper: -- Executing
[0201063001@from-openBTS:5] Set("SIP/00101100010-00000000", "CALLERID(num)=0201063002") in new stack
Feb 1 06:17:14 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: pbx.c:4898 in pbx_extension_helper: -- Executing
[0201063001@from-openBTS:6] Verbose("SIP/00101100010-00000000", "Get IMSI from EXTEN: 0201063001") in new
stack
Feb 1 06:17:14 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: app_verbose.c:104 in verbose_exec: Get IMSI from
EXTEN: 0201063001
Feb 1 06:17:14 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: pbx.c:4898 in pbx_extension_helper: -- Executing
[0201063001@from-openBTS:7] Set("SIP/00101100010-00000000", "IMSI=IMSI510113157642463") in new stack
Feb 1 06:17:14 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: pbx.c:4898 in pbx_extension_helper: -- Executing
[0201063001@from-openBTS:8] Dial("SIP/00101100010-00000000", "SIP/00101100010/IMSI510113157642463") in new
stack
Feb 1 06:17:14 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: netsock2.c:516 in ast_set_qos: == Using SIP RTP
TOS bits 184
Feb 1 06:17:14 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: netsock2.c:538 in ast_set_qos: == Using SIP RTP
CoS mark 5
Feb 1 06:17:14 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: app_dial.c:2625 in dial_exec_full: -- Called
SIP/00101100010/IMSI510113157642463
Feb 1 06:17:19 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: app_dial.c:1400 in wait_for_answer: --
SIP/00101100010-00000001 is ringing
Feb 1 06:17:25 OpenBTS transceiver: NOTICE 1981:2103 2018-02-01T06:17:25.4 RAD1Device.cpp:454:readSamples:
updating timestamp offset to: 305
Feb 1 06:17:29 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: app_dial.c:1307 in wait_for_answer: --
SIP/00101100010-00000001 answered SIP/00101100010-00000000
Feb 1 06:17:29 OpenBTS asterisk[1188]: VERBOSE[2232][C-00000000]: rtp_engine.c:1591 in ast_rtp_instance_bridge: --
Locally bridging SIP/00101100010-00000000 and SIP/00101100010-00000001
```

Fig. 7. Telephone Call Debugging

The OpenBTS system debugging results in /var/log/syslog observations when used as a telephone service can be seen in Figure 7. It can be seen that the system runs the SIP protocol by calling IMSI, so it is not a call based on the telephone number, even though it is calling based on the telephone number, but actually the number is the card number, in this case, IMSI, while the telephone number is formed by the operator.

## C. SMS Service

The system starts by entering the message via the 1st cell phone then the message is sent to the 2nd cell phone. The OpenBTS system will check IMSI data, fill in the message and destination number before sending it to the destination number. If the destination number is registered, the SMS data will enter the SMS database and the SMS queue system will then be sent to the destination number registered on the system. The shipping process can be seen in Figure 8.

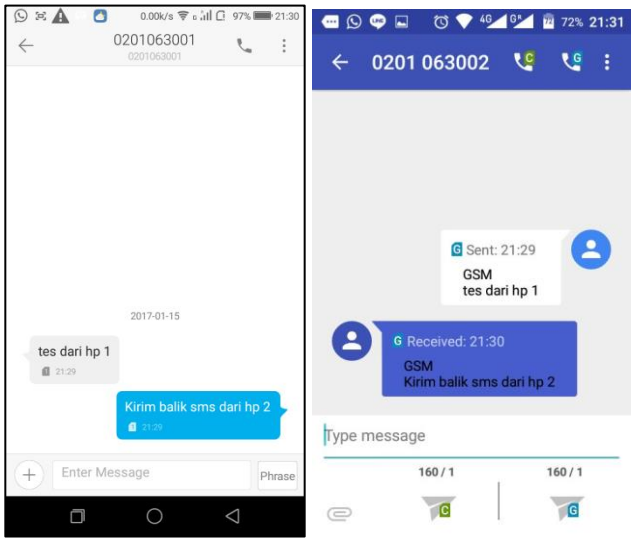



Fig. 8. SMS Between Cellphones

The OpenBTS system debugging results in /var/log/syslog observations when used as an SMS service can be seen in Figure 9. It appears that Smqueue happened to call -1st to the 2nd phone with the short message "test dari hp 1" and vice versa from the phone to-2 to the phone-1 with the contents of the short message "kirim balik sms dari hp 2".

```
Feb 1 06:14:25 OpenBTS transceiver: NOTICE 1981:2103 2018-02-01T06:14:25.4
RAD1Device.cpp:454:readSamples: updating timestamp offset to: 305
Feb 1 06:14:27 OpenBTS smqueue: NOTICE 1936:1966 2018-02-01T06:14:27.5
smqueue.cpp:2455:main_loop: Got SMS rqst qtag '700698--OBTSxhcthailugnhpcwe' from
IMSI510101421302212 for smsc
Feb 1 06:14:27 OpenBTS smqueue: NOTICE 1936:1967 2018-02-01T06:14:27.5
smqueue.h:505:get_text: Decoded text: tes dari hp 1
Feb 1 06:14:28 OpenBTS smqueue: NOTICE 1936:1967 2018-02-01T06:14:28.5
smqueue.h:505:get_text: Decoded text: tes dari hp 1
Feb 1 06:14:33 OpenBTS smqueue: NOTICE 1936:1967 2018-02-01T06:14:33.6
smqueue.cpp:318:handle_response: Got 200 response for sent msg '700698--
OBTSxhcthailugnhpcwe' in state 12
Feb 1 06:14:37 OpenBTS openbts: NOTICE 1952:2124 2018-02-01T06:14:37.4
GSML2LAPDm.cpp:277:unexpectedMessage: obj: LAPDm(C0T0 SDCCH/4-1(0xa0ff9f0) SAPI=SAPI0
State=LinkReleased C=1 R=0 V5=0 VA=0 VR=0 RC=0 EstablishmentInProgress=0 L1In.size()=0) from:5
Feb 1 06:14:59 OpenBTS openbts: NOTICE 1952:2110 2018-02-01T06:14:59.6
L3MobilityManagement.cpp:215:machineRunState: Machine=(L3IdentifyMachine tid=104 C0T0
SDCCH/4-0 CCState=null PopState=0)No IMSI or valid TMSI. Reqesting IMSI.
Feb 1 06:15:10 OpenBTS openbts: NOTICE 1952:2219 2018-02-01T06:15:10.8
TBF.cpp:1125:mtCancel: 06:15:10.8,508058: 06:15:10.8@@@failed tbf cause=3101 TBF#16
mtMS= MS#3,TLLI=c005e001,800c8001 mtDir=RLCDir::Up#012#011 channels: down=( 0:1 0:2) up=(
0:1,usf=1 0:2,usf=1)#012#011 mtState==TBFState::DataTransmit mtAttached=1 mtTFI=16
mtTlli=0xc005e001#012#011 mtMsgExpectedBits=0x20 mtMsgAckBits=0x2#012#011
mtUnAckMode=0 OnCCCH=0 mtAssignCounter=1 N3101=21 msT3193=21#012#011
msNumDataUSFGrants=89 msAckNackUSFGrant=67 msOldTlli=0x800c8001 Pacch= idle=0#012#011
TimingError=(1.84 min=-1.62 max=1.84 avg=0.72 N=92) RSSI=(-59 min=-63 max=-51 avg=-59.24
N=92) CV=(57 min=57 max=61 avg=59.40 N=5) ILev=(0) RXQual=(0) SigVar=(2 min=1 max=2 avg=1.67
N=3) ChCoding=(0)#012#011 VR=4 VQ=4 stalled=0 mNumUpBlocksSinceAckNack=0
mtUpState=1#012#011 blocks: total=29 unique=8 grants=89
Feb 1 06:15:11 OpenBTS openbts: NOTICE 1952:2118 2018-02-01T06:15:11.8
L3MobilityManagement.cpp:215:machineRunState: Machine=(L3IdentifyMachine tid=105 C0T0
SDCCH/4-1 CCState=null PopState=0)No IMSI or valid TMSI. Reqesting IMSI.
Feb 1 06:15:14 OpenBTS smqueue: NOTICE 1936:1966 2018-02-01T06:15:14.1
smqueue.cpp:2455:main_loop: Got SMS rqst qtag '700700--OBTSbncbvbnkzenknmkc' from
IMSI510113157642463 for smsc
Feb 1 06:15:14 OpenBTS smqueue: NOTICE 1936:1967 2018-02-01T06:15:14.1
smqueue.h:505:get_text: Decoded text: Kirim balik sms dari hp 2
Feb 1 06:15:15 OpenBTS smqueue: NOTICE 1936:1967 2018-02-01T06:15:15.1
smqueue.h:505:get_text: Decoded text: Kirim balik sms dari hp 2
Feb 1 06:15:23 OpenBTS smqueue: NOTICE 1936:1967 2018-02-01T06:15:23.1
smqueue.cpp:318:handle_response: Got 200 response for sent msg '700700--
OBTSbncbvbnkzenknmkc' in state 12
```

Fig. 9. SMS Debugging

### D. Signal Quality

Voice quality and SMS reception are very dependent on the condition of the signal strength emitted by the antenna. As an illustration of the results of sound quality testing in this study using indoor standard antennas can be seen in table 1.

TABLE I. SOUND QUALITY RECEIVED BASED ON DISTANCE AND SIGNAL STRENGTH

| Distance (m) | Signal Quality (dBm) | Signal Quality (mW) | Sound Quality |
|---|---|---|---|
| 3 | -53 | 0,0000050119 | Very good |
| 6 | -55 | 0,0000031623 | Very good |
| 9 | -57 | 0,0000019953 | Good |
| 12 | -57 | 0,0000019953 | Good |
| 15 | -59 | 0,0000012589 | Enough |
| 20 | -64 | 0,0000003981 | Not good |
| 25 | -67 | 0,0000001995 | Not good |
| 27 | -71 | 0,0000000794 | Bad |
| 30 | -81 | 0,0000000079 | Bad |

The next test is SMS, carried out on the message delay sent based on signal strength and distance from USRP, the results of the test can be seen in table 2.

TABLE II. SIGNAL QUALITY TESTING BASED ON DISTANCE AND QUALITY OF SMS

| Distance (m) | Signal Quality (dBm) | Signal Quality (mW) | SMS sending |
|---|---|---|---|
| 3 | -51 | 0,0000079433 | Sent |
| 6 | -53 | 0,0000050119 | Sent |
| 9 | -56 | 0,0000025119 | Sent |
| 12 | -56 | 0,0000025119 | Sent |
| 15 | -60 | 0,0000010000 | Sent |
| 20 | -64 | 0,0000003981 | Sent |
| 25 | -69 | 0,0000001259 | Sent |
| 27 | -71 | 0,0000000794 | Not sent |
| 30 | -85 | 0,0000000032 | Not sent |

From tables 1 and 2 it can be concluded that the quality of service from OpenBTS will depend on the distance and also the signal quality obtained by the user. The mathematical calculations to get signal quality (mW) are as follows:

$$P_{(mW)} = 1mW \cdot 10^{(P(dBm)/10)}$$

Example : 13dBm = ? Mh

$$P_{(mW)} = 1mW \cdot 10^{(13dBm/10)} = 19.95mW$$

So 13dBm is 19.95mW.

## V. CONCLUSION

From the results of research conducted, the design of mobile telephone communications without involving commercial cellular operators is very likely to be applied in remote areas, this is because without involving commercial cellular operators telephone and SMS services can be realized. In connection with sound quality and SMS reception, it can be improved by using an amplified outdoor antenna. The main problem in implementing the system is only on OpenBTS hardware that must be imported from outside because in Indonesia no one sells the equipment. But with the progress of the Online-based purchase system, the procurement of OpenBTS hardware is not an obstacle. From the results of the research achievements, further research that is currently being

developed is GPRS data service to provide information and post-disaster data to the SAR team regarding the location of the victim and can be seen in real-time through a web browser on a smartphone. This research can be realized because communication can be realized independently with smartphone communication devices that are commonly sold in the market but in realizing a communication system without involving commercial operators.

## REFERENCES

[1] Febrianty, O., "Definisi Komunikasi", 2015, (Online), (http://slideplayer.info/slide/2838078/, accessed 20-12-2017)

[2] Iedema, "Getting Started with OpenBTS", O'Reilly, Range Networks, 2015

[3] Juhana and Hatorangan, "Mobile phone auto registration to OpenBTS-based cellular network in disaster situation", 8th International Conference on Telecommunication Systems Services and Applications (TSSA), 2014

[4] Purbo, O., W., "Bongkar Rahasia openBTS untuk Jaringan Operator Selular", Andi Offset, Yogyakarta, 2013.

[5] Utama, "Perancangan Dan Instalasi SIP Sebagai Sarana Komunikasi Audio Visual Dengan Asterisk Sebagai IP PBX Server", Jurusan Teknik Elektro Fakultas Teknik Universitas Diponegoro, Semarang, 2012

[6] Wibowo, Rizal, and Gunawan, "Implementasi Voice Service Pada Jaringan GSM Menggunakan OpenBTS V5.0", e-Proceeding of Applied Science : Vol.1, No.2 Agustus 2015

[7] Jalendry, S. and Verma, S., "A Detail Review on Voice over Internet Protocol (VoIP)", Dept. of CSE., Manav Rachna International University,Faridabad, Haryana, India, 2015

[8] _____, "OpenBTS Application Suite Release 4.0", Range Networks, Inc., San Francisco, California 94107, United States of America, 2014

# Self-organized Population Segmentation for Geosocial Network Neighborhood

Low Shen Loong, Syarulnaziah Anawar, Zakiah Ayop, Mohd Rizuan Baharon, Erman Hamid
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

***Abstract*—Geosocial network neighborhood application allows user to share information and communicate with other people within a virtual neighborhood or community. A large and crowded neighbourhood will degrade social quality within the community. Therefore, optimal population segmentation is an essential part in a geosocial network neighborhood, to specify access rights and privileges to resources, and increase social connectivity. In this paper, we propose an extension of the density-based clustering method to allow self-organized segmentation for neighbourhood boundaries in a geosocial network. The objective of this paper is two-fold: First, to improve the distance calculation in population segmentation in a geosocial network neighbourhood. Second, to implement self-organized population segmentation algorithms using threshold value and Dunbar number. The effectiveness of the proposed algorithms is evaluated via experimental scenarios using GPS data. The proposed algorithms show improvement in segmenting large group size of cluster into smaller group size of cluster to maintain the stability of social relationship in the neighbourhood.**

***Keywords*—Segmentation; geosocial network; virtual neighbourhood; density-based clustering; dunbar's number**

## I. Introduction

Since Web 2.0 technology has gained its popularity, the use of online social networks (OSN) like Facebook and Instagram has increased to the point of becoming pervasive. With the introduction of social networking, people have more interaction in a neighbourhood level. Social networks tend to take over some of the functions of neighbourhood communities [1]. These virtual communities allow better quality of social interaction among neighbours in a social network. Geosocial network neighbourhood is one of the branches in social networking that allow user to share information and communicate with other people within a virtual neighbourhood or community. Typically, geosocial networking application uses location awareness to track geolocation information that consists of current user location coordinates; longitude and latitude. Location-aware features in users' mobile device will assist GPS self-check-in function to match users' house address and current location.

One of the important research areas in geosocial network is defining neighbourhood boundaries. Previously, several studies have proposed techniques for optimal definition of neighborhood boundaries in geographic information system (GIS) such as using collaborative tagging system [2], classification and regression trees [3], and clustering [4]. In the context of geosocial network neighbourhood, optimal neighbourhood boundaries can be achieved through population segmentation. Population segmentation is a key component of neighbourhood management strategy. Population segmentation is the process of dividing population into segments based on various characteristics. Segmentation is an essential part in a geosocial networking application for safer virtual neighborhood environment. Segment is utilized to determine residents in a neighbourhood and to differentiate whether a resident is staying within its own living areas. This is particularly important in order to specify access rights or privileges to applications' resources. In addition, social connection between each resident plays a vital role in the quality of geosocial network neighbourhood. A large and crowded neighbourhood will degrade the social quality within the community. Therefore, population segmentation will increase social connectivity and allows user to share information and communicate with other people effectively.

Population segmentation for defining neighbourhood boundaries have been first applied using clustering in geosocial network by [5]. Their work applies cluster technique using DBSCAN algorithms based on user-defined parameters, which is not suitable for population segmentation in a geosocial network neighbourhood. The study has been extended in [6], where the proposed solution considered user check-in time during the clustering process. In [7], EBSCAN algorithms is used to improve the shortcomings in [5] and [6], by removing the user-defined parameters thus improving system performance. However, several problems pertaining to defining neighborhood boundaries remains: (1) low accuracy for distance calculation, (2) inability for self-organized population segmentation, (3) less social connectivity within population. To address these challenges, we propose to design self-organized population segmentation for geosocial network neighbourhood. The improved population segmentation technique is important to provide better representation and accuracy in determining the residents of geosocial neighbourhood for better social connectivity within a neighbourhood segment.

The objective of this paper is two-fold: First, to improve the distance calculation in population segmentation in geosocial networking. Second, to implement self-organized population segmentation using threshold value and Dunbar number. The rest of this paper is organized as follows: Section 2 explain population segmentation in geosocial network neighborhood and review some related work. In Section 3, the basic concepts of DBSCAN and Dunbar Numbers used in the proposed work

are explained. In Section 4, the design of self-organized population segmentation is presented. Finally, we present the proof-of-concept implementation of our proposed algorithms in Section 5. This paper is summarized in the last section.

## II. Population Segmentation in Geosocial Network Neighbourhood

### A. Geosocial Networking

Geosocial network is a relatively new research field emerging as an integration of social network and location-based services. In recent years, some of the most prominent geosocial network applications are Waze, Foursquare and Facebook Places with millions of active users. Geosocial network applications use several techniques to provide its service, such as geocoding, geotagging, and geolocation (see Figure 1). Geocoding can be used to generate route in a traffic report application like Waze. On the other hand, geotagging is the process of adding geographical identification metadata.

Geolocation is a technique to estimate or identify a person or place geographical location within a set of geographic coordinates. With the use of mobile phone GPS to track user location, a geosocial neighborhood application will track user location through the GPS sensor with accuracy of 10 meter. Geolocation application such as Foursquare, Brightkite and others encourage user to provide details recent visited places, hometown or neighbourhood using GPS system. One example of geosocial network neighbourhood application that used geolocation technique is NextDoor.

Population segmentation is the process of dividing the population (users) in a geosocial network application into smaller group. In the process of dividing and segmenting, those who shared common characteristics such as common interest, common needs, similar lifestyles or even similar demographics profiles will be divided into segment. Population segmentation is important in understanding the distinctive needs of different parts of the population. Understanding the characteristics of population needs is important to identify services to be offered in neighborhood geosocial network. Tailoring services to specific segments is the best way of ensuring the most effective use of resources. The starting point for population segmentation strategy is identifying target populations.

Population segmentation in a geosocial network neighbourhood can be characterized physically and logically. The former is based on the distance and geolocation data, while the latter is based on the user behavior and activities within the neighbourhood. In physical segmentation, users must be divided into clusters to be more effectively targeted. One way to implement physical segmentation in geosocial network is through clustering techniques. Clustering algorithms have three basic categories that are hierarchical, partitioning and density-based [8]. Large number of data in huge databases can be deals by all these algorithms. Partitioning algorithm construct k clusters given n data object. Each cluster cannot have same common data object and can have as many group and object, where $k \leq n$ . Hierarchical algorithms create hierarchical decomposition presented in a dendrogram, where a tree splits set of data object into smaller subset until one data object represents a subset. Hierarchical algorithms can be classified into two types that is divisive and agglomerative. On the other hand, density-based algorithms are designed to discover arbitrary shape of cluster that has higher density than remainder data object. Low-density region of data object considers outlier or noise.

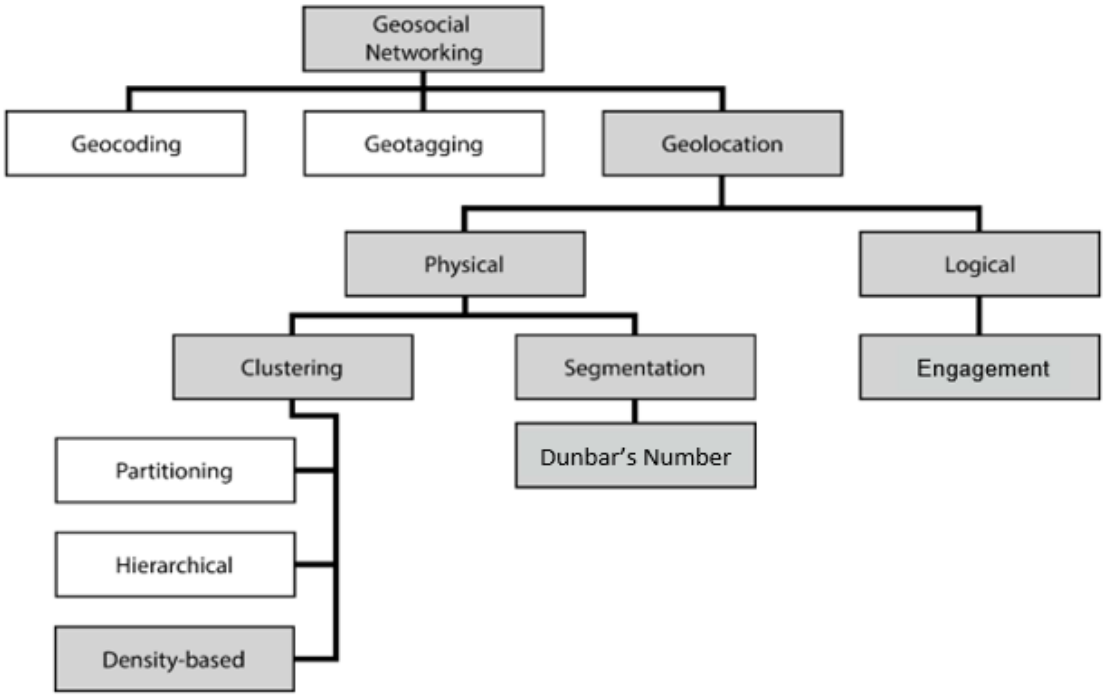



Fig. 1. Classification of Geosocial Networking.

### B. Density-Based Clustering

The density-based notion is a common approach for clustering. Density-based clustering algorithms are based on the idea that objects which form a dense region should be grouped together into one cluster. The algorithms use a fixed threshold value to determine dense regions. They search for regions of high density in a feature space that are separated by regions of lower density.

In this paper, we extend the work in [5] that used DBSCAN algorithms [9] because it has the ability in discovering clusters with arbitrary shape such as linear, concave, and oval. Furthermore, in contrast to some clustering algorithms, it does not require pre-determination of the number of clusters. This algorithm is a data clustering algorithm that given a set of points in some space, it groups together points that are closely packed together, marking as outlier points that lie alone in low-density regions. DBSCAN has been proven in its ability of processing very large databases [10],[11].

In the context of geosocial networking, the DBSCAN algorithms have been first applied for clustering in geosocial network by [5]. However, as the DBSCAN algorithms only has cluster technique that is based on user-defined parameters, the algorithms is less appropriate for population segmentation in geosocial network.

## III. Basic Concept used in the Proposed Work

In this section, two basic concepts used in the proposed algorithms, namely DBSCAN algorithms and Dunbar Numbers are explained.

### A. Understanding DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is designed to discover arbitrary-shaped clusters in any database, D, and at the same time can distinguish noise points. The main idea of DBSCAN is developing a cluster from each point with two parameters, Eps and MinPts. Eps (Epsilon) accepts a radius value based on a user defined distance measure and a value MinPts (Minimum Points) for the number of minimal points that should occur within Eps radius.

DBSCAN use distance function such as Euclidean Distance, Manhattan Distance, Haversine Formula for points $x_i$ and $x_j$ to determine as neighbourhood denoted by $\text{dist}(x_i, x_j)$. Eps is the maximum radius between points for them to be considered as in the same neighborhood. The Eps-neighbourhood is denoted as $\{x_j \in D \mid \text{dist}(x_i, x_j) \leq \text{Eps}\}$.

Number of objects within Eps-neighbourhood can be differentiate using three types of object that is core object, border object and noise object. These three objects will rely on the second parameter that is MinPts. MinPts specifies the minimum amount of points in a neighbourhood. A neighbourhood that contains more than, or minimum amount of points (MinPts) is define as core object. The core object will denote as $x_{core}$ derived in density $x_{core} \geq$ MinPts. Border object, $x_{border}$, is an object where the density is reachable from another core object, but it is not a core object. Border object belongs to a neighbourhood of core object, and the density that is less than MinPts is define as density $x_{border} <$ MinPts. Lastly, noise object is a point that fall within the neighbourhood radius, Eps, but less than the minimum amount of MinPts where no core object exists in it. Therefore, noise object does not belong to any clusters.

DBSCAN clustering have other important definition to define the relationship between objects, namely density reachable, density connected and cluster. Density reachable happen when two objects $x_1$ and $x_n$ are in a chain of objects $x_1, x_2, ..., x_n$ such as $x_{i+1}$ and $x_1$ are direct density reachable and $x_n$ as core object is density reachable to $x_1$, followed by the requirement of Eps and MinPts. Meanwhile, two objects $x_1$ and $x_2$ are considered as density connected when $x_1$ and $x_2$ are density reachable to $x_i$ with respect to MinPts and Eps. Lastly, a set of objects that density reachable to a core object will form cluster object. The relationship between the objects is illustrated in Figure 2.

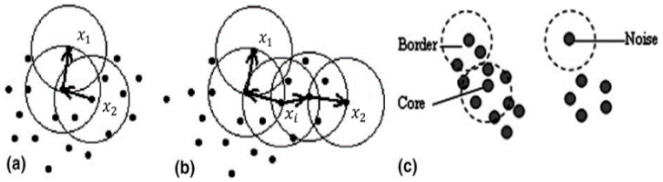



Fig. 2. Terms and Concept of DBSCAN a) Density-Reachable b) Density Connected c) Noise, Core Object and Border Object.

DBSCAN algorithms start from random points in a database, and retrieve all density-reachable of points within Eps radius. Points that are more than MinPts parameter will become core object and a new cluster will be formed. The chain of objects for the cluster will recursively detect all density-reachable objects. A seed list in DBSCAN algorithms is used to store all core objects in the chain and update newly discovered core objects into the list. dbscan and ExpandCluster are two major functions in DBSCAN algorithms as presented in [11].

### B. Dunbar's Number

The size of a neighbourhood segment has an impact on social interaction. With large amount of information in each neighbourhood, each user must make their own choices about the best way to handle and use the information given the priority of personal preferences, interests, and needs. In 1992, Dunbar measured the correlation between neocortical volume and typical social group size in a community. The limit imposed by neocortical processing capacity appears to define the number of individuals with whom it is possible to maintain stable interpersonal relationships in a group. The results indicate that humans' social network size is limited to between 100 and 200 individuals, i.e. Dunbar's number [12].

## IV. Design of Self Organized Population Segmentation

This study will extend previous effort by Shi et al. in [5], [6] for population segmentation in a geosocial network neighbourhood.

### A. Improving Distance Accuracy between Residents using Haversine Formula

This study proposed to improve the accuracy of distance calculation between residents in [5] and [6] using Haversine formula [13] due to its suitability in calculating the great-circle distance between two points on a sphere given their longitudes and latitudes. Current algorithms use Euclidean distance to calculate distance of the two points. However, Euclidean distance is only applicable on cartesian plane but cannot be applied on sphere shape. Since the Earth is nearly spherical, the great-circle distance formulas give the distance between points on the surface of the Earth with correctness within 0.5%. Thus, Haversine formula is chosen to replace Euclidean distance in the improved algorithms.

In the Haversine formula [13], $d$ is the distance between two points, $r$ is the radius of sphere 6371km, $\varphi_1, \varphi_2$ is latitude of point 1 and latitude of point 2, in radians, and $\lambda_1, \lambda_2$ is longitude of point 1 and longitude of point 2, in radians.

### B. Defining threshold Value to Self-Organize Segment Density

Next, the improved algorithms define a threshold value for both Eps and MinPts parameters to self-organize segment density, which are previously user defined. In [5], Eps parameter accepts radius value and MinPts accept number of minimal points in Eps radius. Therefore, both parameters should have a threshold value to achieve self-organize segment density without any user defined parameters. Thus, this study proposed that Eps and MinPts parameter is predefined with a threshold value.

To define the threshold value of Eps, the minimum distance between neighbours needs to be defined and understood. In real life society, a neighbour means a person living nearby or next door to the person referred to. Person who stay in front could also be considered as a neighbour. Thus, the threshold distance between both neighbours are the radius of house area and the width of roads. In Malaysia, there are few road design which the shortest width among others types with only 2.75 meter width length. The width length of road is the standard length that are set under Malaysian Public Works Department (JKR) which is responsible for construction and maintenance of public infrastructure in Malaysia.

Another threshold value for MinPts is defined by the minimum number of individuals to form a group. A group is a number of people that are located, gathered or classed together. The group size of people can vary from two persons to thousands of people. A German sociologist, Georg Simmel study the connections between group size and group actions, as well as the effect of the group size on social life. For MinPts parameter, a minimum value need to be assigned in the improved algorithms to group neighbours into a cluster. According to Simmel's studies of group size, dyad, or a group of two people, is the simplest group form that may exist between individuals [14]. Thus, 2 is selected as threshold value for MinPts parameter.

### C. Improving Social Connectivity through Re-Segmentation using Dunbar Number

In order to improve social connectivity, the improved algorithms implements cluster re-segmentation using Dunbar's number. Current cluster technique did not cluster based on the concern of social connectivity between neighbours. Therefore, this study proposed a method to determine group size of a cluster and re-segmentation of a cluster using Dunbar's number as shown in Figure 3.

The improved algorithms use Dunbar's Number as a threshold number for a cluster. Based on [15], a community should have a mean group size of 150 peoples to maintain a stable social relationship between each other. Thus, the improved algorithms should determine cluster group size that is larger than 150 peoples and save the cluster id into an array for later population segmentation.

**Algorithm 1**: Cluster_size (cluster_data)

**For** $C_i$ in clustered data set
  **If** $C_i$ size $C > 150$
    **add** $C_i$ to list
  **End**
**End**
**Return** list
**End** // cluster_size

Fig. 3. Determination of Cluster Group Size.

One major problem for this improvement is the algorithms will discover cluster in arbitrary shape which makes it difficult to equally segment the cluster. To solve this problem, we propose to use inverse Haversine formula to form a rectangle border to segment equally as seen in Equation 1 below:

$$\varphi_2 = a\sin\left(\sin\varphi_1 \cdot \cos\delta + \cos\varphi_1 \cdot \sin\delta \cdot \cos\theta\right)$$
$$\lambda_2 = \lambda_1 + a\tan 2\left(\sin\theta \cdot \sin\delta \cdot \cos\varphi_1, \cos\delta - \sin\varphi_1 \cdot \sin\varphi_2\right) \quad (1)$$

Where,

$\varphi$ is latitude,

$\lambda$ is longitude,

$\theta$ is the bearing (clockwise from north),

$\delta$ is the angular distance $d / R$; $d$ being the distance, $R$ the earth's radius.

The inverse Haversine formula is implemented in the improved segmentation algorithms as shown in Figure 4. In order to form a rectangle border for equal segmentation, the algorithms will first determine the distance of latitude and longitude by finding the cluster farthest north and south latitude, and east and west longitude points. Then, the number of slice is determined by dividing the total points for oversize segment with 150. The improved algorithms will then re-segment cluster group that have more than 150 people by creating a new cluster using the inverse Haversine formula.

**Algorithm 2**: Segmentation (over_size_cluster)

**If** $DLat > DLng$

Distance segmented area, $DSeg = DLat / S$

New Cluster List, $NList = inverse\_haver\sin e(NLat, SLat, DSeg)$

**Else**

Distance segmented area, $DSeg = DLng / S$

New Cluster List,

$NList = inverse\_haver\sin e(WLng, ELng, DSeg)$

**End**

**Return** *NList*

**End** // *segmentation*

Fig. 4. Determination of Rectangle Border for Arbitrary Shape Cluster.

Where,

$NLat$ = Most North of Latitude point

$SLat$ = Most South of Latitude point

$WLng$ = Most West of Longitude point

$ELng$ = Most East of Longitude point

Distance of Latitude, $DLat = NLat - SLat$

Distance of Longitude, $DLng = WLng - ELng$

Number of slice, $S$ = Total point for over-size / 150

## V. Implementation

This section discusses the implementation of the proposed population segmentation algorithms.

### A. Experimental Setup

The experiment is setup in Taman Bukit Melaka, Malaysia. Global Positioning System (GPS) coordinates are plotted based on Taman Bukit Melaka housing area as shown in Google Maps. All GPS coordinates are represented as the local residents. Two different scenarios of resident density are simulated in this experiment; low and high densities. All GPS coordinates data are recorded in JavaScript Object Notation (JSON) format for the evaluation of improved algorithms.

### B. Results and Discussions

Current work in progress is to identify low and high density residential areas. The improved algorithms consist of segmentation technique to control group size of a cluster based on pre-defined parameters and Dunbar number with average 150 persons per group. In this section, the comparison of current algorithms [5], and the improved algorithms are shown in Figure 5, Figure 6 and Figure 7. Both algorithms are implemented in Javascript programming language.

The first experimental scenario is to compare the population segmentation for a low-density neighbourhood. The result in Figure 5 shows that both current and improved algorithms produce the same segments. This indicate that both algorithms work well in a low-density neighbourhood.

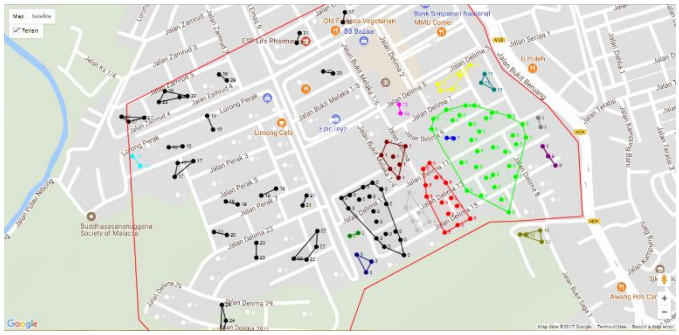


Fig. 5. Segmentation Algorithms for Low-Density Area using Current [5] and Improved Algorithms.

Fig. 6. Segmentation for High-Density Area using Current Algorithms [5].

The second experimental scenario is to compare the population segmentation for a high-density neighbourhood. Figure 6 shows the segmentation result for current algorithms [5] while Figure 7 shows the segmentation result for the proposed algorithms. The result in Figure 6 shows that current algorithms only cluster nearby GPS coordinates into few large segment. From the results, it can be seen that segmentation of the cluster group is not equally distributed and some of the segments are too crowded.

On the other hand, the result in Figure 6 shows that the proposed algorithms do not only cluster GPS coordinates, but also segment large group size of cluster into smaller group size of cluster with average of 150 people to maintain the stability of social relationship.

Fig. 7. Segmentation for High-Density Area using Improved Algorithms.

## VI. Conclusion and Future Work

In summary, this paper propose an improved population segmentation algorithms to provide better representation and accuracy in determining the residents of geosocial neighbourhood. In this paper, we show how the current population segmentation in geosocial network neighbourhood using density-based clustering method can be extended to improve the quality of social community. Three improvements are implemented in the proposed algorithms namely improving distance accuracy between residents using haversine formula, eliminating user-defined parameter by defining threshold value to self-organize segment density, and improving social connectivity through re-segmentation using Dunbar number. The improved algorithms have achieved a great result particularly on the segmentation for high-density neighbourhood and able to segment a crowded area into smaller group size of cluster to maintain the stability of social relationship.

In geosocial network neighbourhood, identification of demographic sub-group will help to better understand the needs and requirements for the people who are related to the segment. Therefore, future extension of our work may include and put an emphasis on integration of demographic information that may help to reveal patterns or differences between groups of people who may be similar in age, gender, race, religion or socioeconomic status. The addition of demographic variables will allow the most effective use of resources in a geosocial network neighbourhood application.

## Acknowledgment

A high appreciation to Center for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM) for supporting the work done in this paper.

## References

[1] K. Hampton, and B. Wellman, "Neighboring in Netville: How the Internet supports community and social capital in a wired suburb," in City & Community, vol. 2(4), pp. 277-311, 2003.

[2] F. Wilske, "Approximation of neighborhood boundaries using collaborative tagging systems," in GI-Days, vol. 32, pp. 179-187, 2008.

[3] R. Füss, and J.A. Koller, "The role of spatial and temporal structure for residential rent predictions," in International Journal of Forecasting, vol. 32(4), pp. 1352-1368, 2016.

[4] Y. van Gennip, B. Hunter, R. Ahn, P. Elliott, K. Luh, M. Halvorson, S. Reid, M. Valasik, J. Wo, G.E. Tita, and A.L. Bertozzi, "Community detection using spectral clustering on sparse geosocial data," in SIAM Journal on Applied Mathematics, vol. 73(1), pp. 67-83, 2013.

[5] J. Shi, N. Mamoulis, D. Wu, and D.W. Cheung, "Density-based place clustering in geo-social networks," in Proceedings of the 2014 ACM SIGMOD international conference on Management of data, pp. 99-110, June 2014,

[6] D. Wu, J. Shi, and N. Mamoulis, "Density-Based Place Clustering Using Geo-Social Network Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 30(5), pp. 838-851, 2018.

[7] S. Yokoyama, Á. Bogárdi-Mészöly, and H. Ishikawa, "EBSCAN: An entanglement-based algorithm for discovering dense regions in large geo-social data streams with noise," in Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, p. 7, November 2015.

[8] K. Mumtaz, and K. Duraiswamy, "A novel density based improved k-means clustering algorithm–Dbkmeans," in International Journal on computer science and Engineering, vol. 2(2), pp. 213-218, 2010.

[9] M. Ester, H.P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Kdd, vol. 96, no. 34, pp. 226-231. 1996.

[10] A. Zhou, S. Zhou, J. Cao, Y. Fan, and Y. Hu, "Approaches for scaling DBSCAN algorithm to large spatial databases," in Journal of computer science and technology, vol. 15, no. 6, pp. 509-526, 2000.

[11] M. Ester, H.P. Kriegel, J. Sander, and X. Xu, "Clustering for mining in large spatial databases," in KI, vol. 12, no. 1, pp. 18-24, 1998.

[12] R.I. Dunbar, "The social brain hypothesis," in Evolutionary Anthropology: Issues, News, and Reviews, vol. 6, no. 5, pp. 178-190, 1998.

[13] R.W. Sinnott, "Virtues of the Haversine," in Sky Telesc., vol. 68, p.159, 1984.

[14] R.L. Moreland, "Are dyads really groups?," in Small Group Research, vol. 41, no. 2, pp. 251-267, 2010.

[15] R. Dunbar, How many friends does one person need? Dunbar's number and other evolutionary quirks. London: Faber & Faber, 2010.

# A Serious Game for Healthcare Industry: Information Security Awareness Training Program for Hospital Universiti Kebangsaan Malaysia

Arash Ghazvini
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia,
43600 UKM, Bangi, Selangor, Malaysia

Zarina Shukur
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia,
43600 UKM, Bangi, Selangor, Malaysia

***Abstract*—This paper aims to develop an information security awareness training program for the healthcare industry to ensure the appropriate protection of electronic health systems. Serious games are primarily designed for training purposes rather than pure entertainment. Serious games are proven as an effective training approach for awareness programs. Serious games benefit learning as the games are fun to play and motivate learners to participate and interact with learning activities. Developing a serious game requires the revision of adequate guidelines that identify all characteristics to be incorporated in such games. Thus, this paper reviews serious game models that have been constructed as game development guidelines. To this end, a serious game is developed and implemented at a selected healthcare organization.**

***Keywords—Serious game; information security; awareness training program***

## I. Introduction

As a new approach in developing information security awareness training programs, computer game-based training is proven as an effective and engaging program [13][11][7][15]. Computer games provide engaging interfaces that enhance training, draw more trainees in, and simulate a variety of scenarios [16][12]. They provide a simulation environment with pre-defined scenarios that motivate players to try different approaches to run the system [1]. The objective for game players is to improve their ability in decision making and learn how to cope with challenging situations. Unlike traditional training programs that focus on the technical aspect of a system, this new approach improves the decision-making skill of game players for handling challenging situations [4]. Computer games give employees an experience by offering them a platform to apply their decisions before those decisions are actually implemented in the real place. For instance, computer games teach users about healthcare values, the information they have access to, the information they are allowed to share, as well as behaviors they are supposed to demonstrate to protect electronic health records [4]. Computer games are effective in the way that they create a platform for players to practice their behaviors and handle the consequences of their decisions in a virtual environment.

## II. Serious Games

Serious games (or applied games) are games primarily designed for training purposes rather than pure entertainment. When it comes to training purposes, computer games should be able to integrate educational elements, content, and multimedia while being playable to users with a pleasant interface [14]. References [5][10][19][3] emphasized the educational advantage of a serious game. "The serious games application is intended to help professionals, as well as enabling users to enjoy themselves through straightforward, real interaction while learning how to cope in several real social situations".

## III. Serious Game Models

This paper develops a serious game called InfoSecure as a training delivery method for Hospital Universiti Kebangsaan Malaysia (HUKM). The aims are to raise HUKM employees' awareness of information security and help them adhere to appropriate behaviors that do not compromise the security of information assets. However, the development process requires adequate guidelines that include all characteristics that should be incorporated in a serious game. Therefore, it is necessary to review available serious game models. The following part briefly discusses serious game models.

Reference [6] introduced a model that explains the relationship between learner and serious games. The model is too basic and designed for fundamental games. It does not provide a detailed design for optimal learning. Reference [8] proposed a framework that supports teachers' evaluation of serious games. The framework only discusses game evaluation but not game design [17]. Reference [9] developed a framework based on classroom teaching. This model imitates the textbook style and does not offer rooms for exploratory learning. Reference [2] introduced a model to identify interface requirements of a serious game. However, the model is too general with a little guideline for designing a serious game. Reference [17] offered a model that is highly dependent on a software called Emergo. His model remains at the theoretical stage and does not provide any design solution to use other game design tools.

Reference [18] proposed a serious game model that is the most efficient and effective model for serious games. Yusoff's

model is based on the review of learning and pedagogy perspectives in combination with the games. This conceptual model is developed to be used by game designers for efficient game development as well as the educational practitioner when designing serious games for effective learning. The conceptual framework proposed by Yusoff is the basis for InfoSecure serious games design (Figure 1). The gray boxes are entities added to the original model to explain how the model is used to develop a serious game for information security awareness training programs.

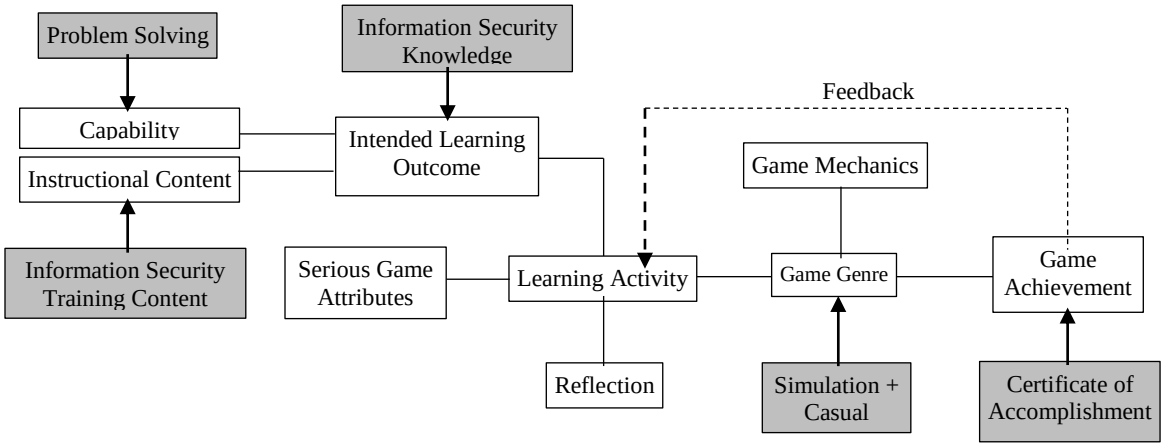



Fig. 1. Serious Game Model Adapted from Yusoff (2010)

### A. Capability

Capability refers to the cognitive skills to be developed by employees from playing a game. In training, conative problem-solving skills attempt to decrease inappropriate behaviors by developing a new perspective and enhancing knowledge about the subject matter. In the context of information security, the aim is to help employees adhere to appropriate behaviors that do not compromise the security of information assets.

### B. Instructional Content

Instructional content refers to the subject matters that must be learned by the employees. The InfoSecure game teaches employees how to protect healthcare information assets. The reason why InfoSecure is developed is due to the failure of the employees to understand and maintain the security of information assets.

### C. Learning Activities

The learning activity is the activity designed to keep the learner engaged and learn in the game world. InfoSecure games provide different activities to keep employees engaged and attracted. For example, the story of the InfoSecure topic is to feed a fish by answering all questions correctly and to avoid catching the bait. The most important fact is to ensure employees stay engaged without getting bored, and show interest to play the game for learning.

### D. Intended Learning Outcomes

Intended learning outcome refers to the objective achieved by the employees from playing the serious game. By the end of the game, employees will be able to use and apply the newly acquired knowledge in their daily work activities. The objective is to enhance the employees' knowledge about risks and hazards associated with information security.

### E. Reflection

Reflection is where the learner thinks about the purpose of the learning activities that have been undertaken. For healthcare employees, understanding the importance of information security is a motivational factor to participate in awareness training programs.

### F. Game Genre

Serious games can be of any genre with the ultimate objective of education. Each genre has its own characteristic that makes it attractive. A combination of simulation and casual genres is used to design InfoSecure. Simulation games are favored as they simulate the real work environment, which allows users to make mistakes and learn from those mistakes without worrying about the consequences of their actions as they would in real life. However, serious games also require the flexibility and fun aspects that usually appear in casual games.

### G. Game Mechanism

Game mechanics are the game operations, and the purpose of this notion is to enable the game to be more fun, enjoyable, and more engaging for the player. There should be at least one game mechanism for each genre. The simulation genre focuses on a single activity and attempts to replicate the real world experience. The outcomes can be highly consistent with a real-life experience or exaggerated. Simulation games place players as decision-makers to manage the simulated situations. Moreover, games with the casual genre are easy to play and master. The design of these games is adopted from conventional games such as chess and cards games. The interface of the game should be very simple. The common goal in casual games is to score.

### H. Serious Game Attributes

It refers to the attributes of a serious game that support learning and engagement. Yusoff (2010) identified twelve important attributes of a serious game. Table I explains how

these attributes are incorporated into the design of the InfoSecure game.

TABLE I. SERIOUS GAME ATTRIBUTES

| Attribute | Description | The InfoSecure Game |
|---|---|---|
| Incremental learning | The learning material is delivered in an incremental way | The target audience is general employees with basic knowledge of information security. Therefore, the training content is developed respectively. |
| Linearity | Learning is arranged sequentially | InfoSecure consists of eight topics; each covering one topic of the policy. |
| Attention span | Duration for learning concentration | InfoSecure breaks up<br>learning sessions into several intervals to produce effective learning. The time spent on each topic is about 5 minutes. |
| Scaffolding | Support and help during the learning process | Not applicable. |
| Transfer of learned skills | Applying skills to new learning based on previous learning | Each topic is designed to have one level to cover one distinct topic. |
| Interaction | Engagement in learning | InfoSecure is a combination of simulation and casual genre. |
| Learner control | Self-learning and active learning based on the learner's pace | There are eight topics. The player has the freedom to select and play any of the topics. |
| Practice and drill | Learning activities and exercises within the game | A topic can be played repeatedly until the full score is achieved. |
| Intermittent feedback | Inform learners of their progress in learning | Players' correct or wrong answers to information security questions are indicated by √ and × marks. |
| Reward | Incentives for the learner | A certificate of accomplishment is rewarded when full scores are achieved. |
| Situated and authentic learning | Placing the learner in an authentic learning environment | The themes and backgrounds of InfoSecure games are relevant to the healthcare work environment. |
| Accommodating the learners' style | Learning to suit learners' preferences | Serious games are suitable to be played by all types of learners. |

### I. *Game Achievement*

Game achievement is the level of the learners' achievement in playing these games. It can be indicated by the game scores, and it gives the pleasure of rewards. The game achievement or score indicates the level of learners' knowledge of the game. In the InfoSecure game, a certificate of accomplishment is rewarded when full scores are achieved for all the topics.

## IV. INFOSECURE CONCEPTUAL MODEL

Figure 2 is the conceptual model of the InfoSecure serious game developed in this paper. Because this is a conceptual model, it can be used as a framework that visually represents the arrangement of the InfoSecure game elements. The conceptual model consists of nine parts: 1) Capability, 2) Instructional Content, 3) Learning Activity, 4) Reflection, 5) Serious Game Attributes, 6) Game Genre, 7) Game Mechanism, 8) Evaluation, 9) Game Achievement.

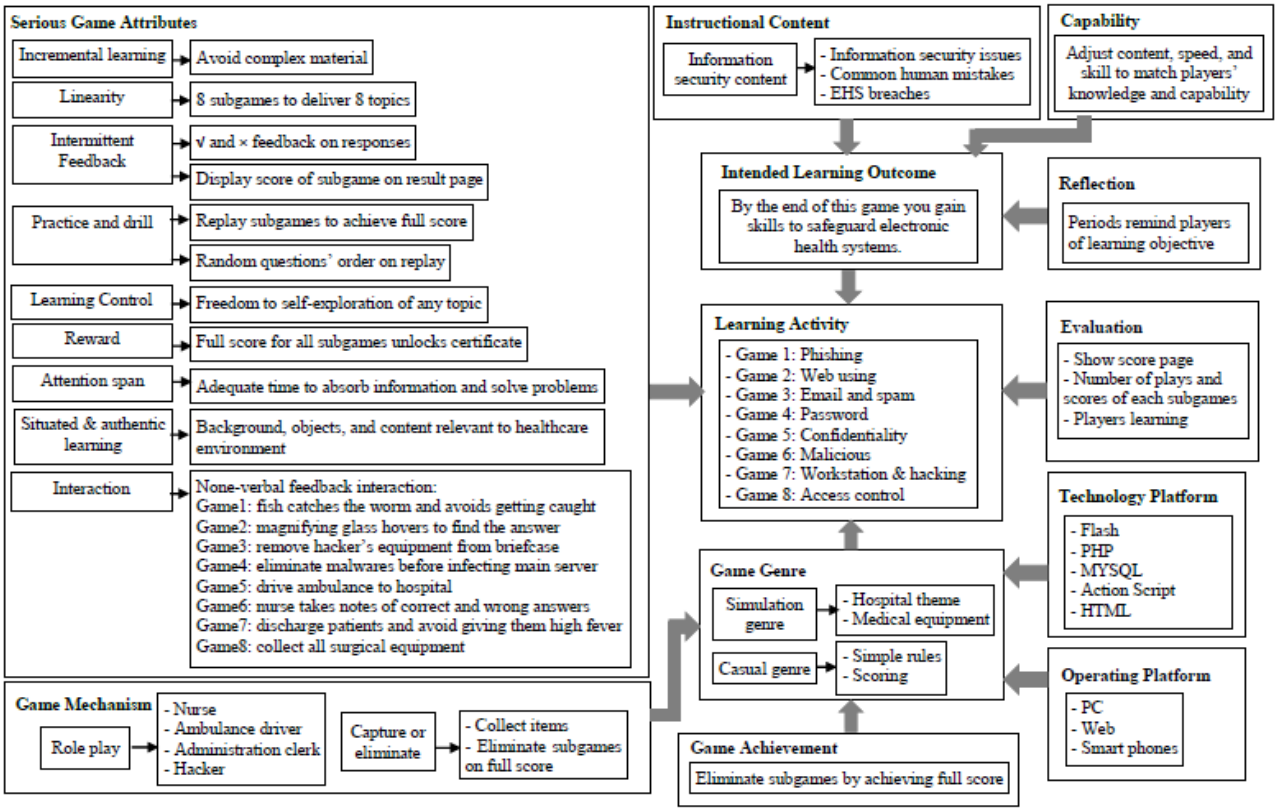



Fig. 2. InfoSecure conceptual model

Table II presents the implementation of the serious game model elements in the InfoSecure game. It explains how these elements are used to develop the InfoSecure game in this study.

TABLE II. IMPLEMENTATION OF THE SERIOUS GAME MODEL ELEMENTS IN THE INFOSECURE GAME

| **Serious Games Model Elements** | **Implementation in InfoSecure Game** |
|---|---|
| A. Intended Learning Outcomes | The intended learning outcome of playing the InfoSecure game is displayed on the homepage. The intended learning outcome is defined as follow: *By the end of this game, learners will be able to protect/ promote healthcare's information asset/security.* |
| B. Capability | InfoSecure is auto-adjusted to match player's knowledge and capability. Since this particular game is not an action-based genre but closed to a direct simulation genre and the target user capability is already known in advanced therefore it is not regulated to be a self-adjusted system to correct the content delivery in order to match to the learner capability. |
| C. Instructional Content | The InfoSecure game aims to teach learners about information security concepts. The instructional content that is intended for players to learn covers: 1) learning about information security issues 2) learning about human roles to maintain the security of information 3) learning how to prevent information breaches. |
| D. Learning Activities | The InfoSecure game consists of eight topics, each covering one information security issue in the healthcare. It provides different activity in each topic to keep employees engaged and attracted. For example, the story of game number 1 is to feed a fish by answering all questions correctly and avoid catching the bait. The important fact is to ensure employees stay engaged without getting bored and to ensure that they show interest in playing the game for learning. |
| E. Serious Game Attribute | |
| • Incremental Learning | The targeted audience is general employees with basic knowledge of information security. Therefore, training content is developed simple and easy to understand. InfoSecure breaks up learning sessions into several intervals to produce effective learning. |
| • Linearity | InfoSecure consists of eight topics, each covering one information security issue of the firm. |
| • Attention Span | The timing given to absorb information and solving the problems must be adequate enough. If the timing is too short, extend the timing. If the learning process too complicated, modify it to make it simple. The training time fluctuates depending on the number of questions set by the instructor. However, there is no time limit for players to finish the game. Adequate time to absorb information and solve problems. The player can leave and come back at any time to resume with training. It helps players digest new information and answer the questions. |
| • Transfer of Learnt Skills | Not applicable to InfoSecure. Each topic has a different scenario and covers different information security topics. Players do not need to acquire new skills to apply at the next level. |
| • Interaction | InfoSecure provides players with feedback interaction.<br>Game1: fish catches the worm and avoids getting caught on the hooks<br>Game2: magnifying glass determines if the selected answer is correct or wrong<br>Game3: remove hacker's equipment from the briefcase<br>Game4: eliminate malware before infecting the main server<br>Game5: drive the ambulance to the hospital<br>Game6: nurse checks if selected answer is correct or wrong<br>Game7: discharge patients and avoid giving them high fever<br>Game8: collect all surgical equipment |
| • Learner Control | There are eight topics on eight information security topics. Player has the freedom to select and play any of topics. InfoSecure provides self-learning to suit players' pace and experience. |
| • Practice and Drill | A topic must be replayed until the full score is achieved to eliminate the topic. Every time player replays a topic, questions are automatically reordered to avoid memorizing the pattern of correct answers. |
| • Intermittent Feedback | Fixed ratio feedback suits the strategy of InfoSecure. A feedback would be given after each response from a learner and at the end of each topic. Correct answers will be indicated by √ and wrong answers by × marks. After the player responded to all questions, the result page appears to show the topic of topic, username, and score. Once the full score is achieved for a topic, the topic will be eliminated from the homepage indicated by the √ mark. |
| • Reward | A certificate of accomplishment is rewarded when the full score is achieved for all topics. |
| • Situated and Authentic Learning | Incorporate familiar background, objects and common examples in games content, to relate it to the healthcare environment. |
| E. Game Genre | InfoSecure is developed based on characteristics of simulation and casual genres. Games with simulation genre are favored as they simulate real work environment which allows users to make mistakes and learn from those mistakes without worrying about the consequences of their actions as they would in the real life (Apperley, 2006). The serious game also requires the flexibility and fun appeared in casual games in order to release users' boredom |

| | |
|---|---|
| | and tiredness [13]. The design of these games is adopted from conventional games such as chess and cards games. The interface of the game should be very simple [13]. The common goal in casual games is to score. To fulfill the requirement of simulation genre, hospital theme is used as the main theme of InfoSecure, and medical objects are used in the game design. To fulfill the requirement of the casual genre, each topic is differently designed to avoid boredom and to make the game challenging and fun. Nevertheless, all topics hold the same goal of raising employees' awareness towards information security. InfoSecure is designed simple with familiar background and graphics and is easy to play and master. |
| F. Game Mechanism | Role-playing: learners get to play as a nurse, ambulance driver, administrative clerk, and computer hacker. Capture or eliminate: in order to achieve the certificate of accomplishment, the player must eliminate all eight topics by answering all questions correctly. |
| G. Technology Platform | The InfoSecure game is developed with Adobe Flash. Aside from being interactive, the flash timeline allows developers to create most of the animations without the need for coding. Games developed in Flash are standalone and could run independently unless it needs to be connected to the database. ActionScript 2 is used to code in Flash because it is easy to understand and develop. Flash is highly secure because the source code is not publicly accessible. MYSQL is used for the database which is a good platform for both basic and advanced projects. PHP connects Flash with MYSQL and enables developers to read, update, add, and remove data. |
| H. Operating Platform | InfoSecure can run on PC, smartphones, and browsers. |
| I. Game Evaluation | An advantage of InfoSecure over previously developed games is that every time a topic is played, questions are shuffled to prevent players from memorizing the pattern of correct answers. User progress can be viewed in the player's profile. The user and instructors can view the progress. They can keep track of employee's performance; how many time a topic is played, what are the scores achieved, what information security topics have been more difficult, what are employees' strength and weaknesses. It demonstrates employees' learning curve. Evaluating the recorded information help managers to monitor employees' performance and take necessary actions. |
| J. Game Achievement | A certificate of accomplishment is rewarded once the full score is achieved for all the topics. |

## V. INFOSECURE DEVELOPMENT

This section describes the development of the InfoSecure game. InfoSecure enhances previous approaches in a number of dimensions. It offers realism which is the balance between reality and fantasy. Simulation games are favored as they simulate real work environment which allows users to make mistakes and learn from those mistakes without worrying about the consequences of their actions as they would in the real life. The serious game also requires the flexibility and fun appeared in casual games in order to release users' boredom and tiredness.

With the said justification, InfoSecure is developed with a combination of simulation and casual genres. The game is characterized by the strength of simulation and casual genres. InfoSecure consists of eight topics. Every topic addresses one information security policy topic that is selected by HUKM management during semi-structured interviews. The information security topics include 1) phishing, 2) web using, 3) email and spam, 4) malicious code, 5) password protection, 6) privacy and confidentiality, 7) workstation and hacking, and 8) access control.

In order to fulfill the requirement of the simulation genre, hospital theme is used as the main theme of InfoSecure, and medical elements are used in the design of topics. In addition, to fulfill the requirement of the casual genre, each topic is designed differently to avoid boredom and to make the game challenging and fun. Nevertheless, all topics hold the same goal of raising employees' awareness towards information security.

An important advantage of InfoSecure is that the game is dynamic and flexible. As discussed before, the advantage of computer game training is that it can be repeated over time. However, a static game will get boring after repeated a few times. Therefore, to avoid reusing static game and to keep players motivated to participate, InfoSecure is developed dynamically. It allows IT managers to change and customize training content as well as the graphics. For instance, IT managers and instructors are able to change or update information security topics of the game as well as setting the number of questions. It is also possible to change the graphics and the overall look of the homepage and every individual topic. The aim is to keep gameplay more interesting. By changing graphics and content, the game will look fresh in the eyes of the player as if it is a different game. It is also worth noting that such flexibility transforms InfoSecure to a platform that can be used by other healthcare. Moreover, players are able to mute or unmute the background music as they desire.

InfoSecure is developed to satisfy several essential requirements obtained during semi-structured interviews with HUKM decision makers:

- Information security awareness: The overall aim is to design an effective information security awareness training program for HUKM. The training program attempts to raise employees' awareness toward information security and help them adhere to appropriate behaviors that do not compromise the security of information assets and with long-term impact.
- Large coverage: HUKM is categorized as a large organization in which all employees need to be trained toward information security. The advantage of InfoSecure is the capability to cover large population at lower cost.
- Accessibility: InfoSecure is easily accessible. The game will be sent to users via email. Employees receive a link that direct them to the game. This link

redirects users to the game page. It allows the game to run from a web browser without having a huge impact on computing power. Therefore, the game is accessible and runs smoothly even on old computers.

- Content Updatability: IT managers and instructors are able to change or update information security topics of the game as well as setting the number of questions. It is also possible to change the graphics and the overall look of the homepage and every individual topic.
- Fun and motivation: A well-developed serious game is fun, and it promotes employees engagement. InfoSecure is developed with a combination of simulation and casual genres that makes InfoSecure different from previously developed games.
- Performance evaluation: InfoSecure allows IT, managers and instructors, to record and keep track of employees' progress in the game. Instructors can review scores achieved by employees every time they play a topic.
- Repeat: InfoSecure can be played as many times as desired. A successful awareness program never ends and the awareness campaign must repeat its message to the employees. If the message is important, then it should be repeated more often in different manners each time. Since InfoSecure is dynamic, it inhibits higher and active learning with long-term impact.

## VI. INFOSECURE INTERFACE

Upon successful login user will be prompt with the welcome page (Figure 3). This page contains brief instruction of the game. By pressing the next button, welcome page disappears and the main page will be entirely visible. Eight linked objects are accessible within the homepage. Each object takes the user to a topic. The user is required to answer all questions correctly for each topic to obtain a full score, otherwise must replay. Once all questions are answered correctly, a green √ icon will appear on top of that particular topic to differentiate a successfully completed topic from others. Once a topic is completed, it will be deactivated and no longer can be played. This helps to push the user to move on and play incomplete topic.

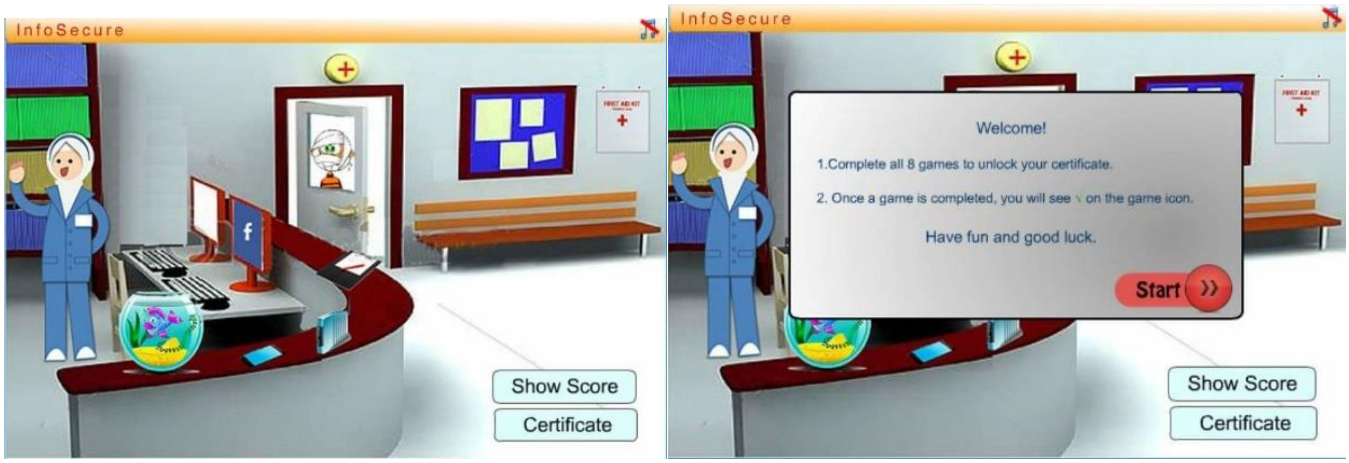



Fig. 3. The Welcome page of InfoSecure

Figure 4 is a screenshot of a game designed to cover a topic on information security, specifically on workstation and hacking. The story of the game is to cure and discharge all patients from the hospital by answering all questions correctly. The number of questions for this topic will be determined between one to ten by the trainer with administrative privilege. There are two icons on the top right corner including mute/unmute and home button. By clicking on the home button user will be redirected to the main page of the game to replay the game or play a different game. Below the top banner, you find a number of 10 beds displayed on the screen, and the number of patients on the beds is according to the number of questions determined by the trainer.

To display a question, the user must click on one of the patients laying on the bed. Once a question is answered correctly, a green color √ appears to confirm that the user has selected the correct answer. Therefore, the patient will be discharged and leave the bed vacant. If the user selects a wrong answer, a red color × appears to indicate that wrong answer is chosen and reveals the correct answer by a green √. Consequently, the patient will remain laying on the bed with a severe headache. After selecting an answer, either correct or wrong, the user can proceed to the next question by clicking on the next question button. All questions must be answered correctly otherwise; user needs to replay the game in order to mark the game as completed on the homepage. The order of questions changes randomly every time the game starts to prevent the user from memorizing the patterns of the correct answer. Once all questions are answered, the result page will come up that shows the topic of the game, the username, and the score.

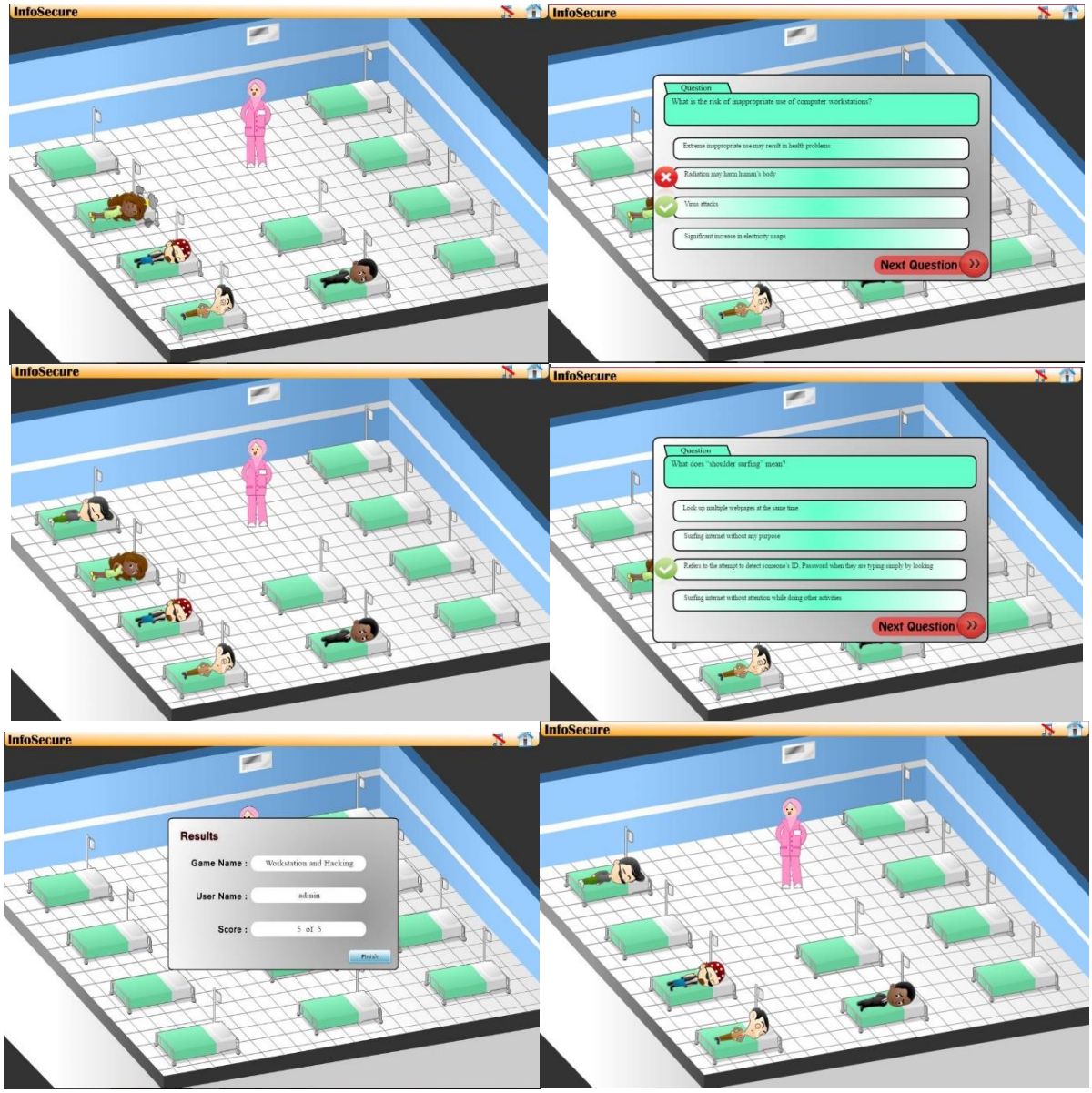



Fig. 4. The InfoSecure workstation and hacking topic

## VII. Implementation and Evaluation

Prior to pilot testing, the InfoSecure game was demonstrated for three full professors in computer science specialized in visual informatics, HCI and usability. They provided valuable feedbacks that are incorporated into the game. On the next phase, five students from the faculty of computer science, Universiti Kebangsaan Malaysia, randomly volunteered to play the game. A discussion was held with students and they provided corrective suggestions as well. The actual pilot test was conducted amongst HUKM employees to test the effectiveness and quality of the InfoSecuer game before handing over to IT department to reach out to all employees. A number of five employees randomly volunteered to play the InfoSecure. After playing the game, the participants showed good impressions of the gameplay experience, attraction, the graphics, and its idea.

When users play a topic for the first time, the answer to the information security questions based on their initial knowledge and understanding which might be wrong. Assuming that a player answers two questions correctly and scores 40% when playing a topic for the first time. The player will be prompted with a result page that three questions have been answered wrongly. The player needs to repeat the topic again in order to obtain a full score. In the second attempt, the player should select answers more carefully, knowing previously selected answers were wrong. Now assuming that the player manages to find the correct answer to four questions and scores 80%. There is still one more question, therefore, the player has to start the topic again from the beginning. The player cannot deactivate the topic and mark it as a completed game, before finding the correct answer to all five questions and scores 100%. Therefore, the user has to play the same topic over and over until he finds all the correct answers. Once

the user scores 100% the topic deactivates and is marked as completed.

In order to prevent players from memorizing the sequence and pattern of correct answers, the order of questions changes every time a topic is replayed. That means every time a topic restarts, questions are shuffled to display in random order. The InfoSecure game helps employees gain knowledge on information security and replace it with the wrong information they initially had in their minds. by playing the game, employees understand that they need to think carefully when dealing with electronic health systems. InfoSecure allows users to make mistakes and learn from those mistakes without worrying about the consequences of their actions as they would in the real life.

Every time an employee plays a topic, the scores are recorded in a database; from the first to the last attempt until player scores 100%. User progress can be viewed in the player's profile. The user and instructors can view the progress. They can keep track of employee's performance, the number of times a topic is played, obtained scores, the most challenging information security topics, and employees' strength and weaknesses. It demonstrates employees' learning curve. Evaluating the recorded information help managers to monitor employees' performance and take necessary actions.

A certificate of accomplishment is rewarded to players who scored 100% in all the topics. The certificated could be printed upon completion of the game. Nevertheless, obtaining a 100% score is not the end of the story. It is important to ensure that employees fully understood the topics and integrated them into their daily activities. As a result, the game must be played frequently, decided by hospital management. Therefore, to avoid reusing static game and to keep players motivated to participate, InfoSecure is developed to be dynamic. It allows IT managers to change and customize the training content as well as the graphics. The aim is to keep gameplay more interesting.

The main objective of asking computer science students to play is to get their feedback on the gameplay experience. It is not surprising that computer science students performed well and answered the most question correctly during the first play. However, HUKM employees did not perform well during the first play. They had to play a topic few times until score 100%. Table III shows the employees' records. For instance, the phishing topic was played four times by employee number 3. For the first play, only one question was answered correctly and scores 0%. The second and third play scores are 40% and 80%, respectively. During the fourth play, the player manages to select the correct answers to all questions and scores 100%.

TABLE III. EMPLOYEES' RECORD OF PLAYING INFOSECURE

| **Topic** | **Employee #1** | **Employee #2** | **Employee #3** | **Employee #4** | **Employee #5** |
|---|---|---|---|---|---|
| Phishing | 1st play: 60%<br>2nd play: 100% | 1st play: 80%<br>2nd play: 100% | 1st play: 20%<br>2nd play: 40%<br>3rd play: 80%<br>4th play: 100% | 1st play: 80%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 40%<br>2nd play: 80%<br>3rd play: 100% |
| Web using | 1st play: 40%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 60%<br>2nd play: 100% | 1st play: 20%<br>2nd play: 40%<br>3rd play: 80%<br>4th play: 80%<br>5th play: 100% | 1st play: 40%<br>2nd play: 60%<br>3rd play: 100% | 1st play: 20%<br>2nd play: 60%<br>3rd play: 100% |
| Email and spam | 1st play: 80%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 80%<br>2nd play: 100% | 1st play: 40%<br>2nd play: 60%<br>3rd play: 80%<br>4th play: 100% | 1st play: 40%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 80%<br>2nd play: 80%<br>3rd play: 100% |
| Malicious code | 1st play: 40%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 60%<br>2nd play: 100% | 1st play: 20%<br>2nd play: 60%<br>3rd play: 80%<br>4th play: 100% | 1st play: 40%<br>2nd play: 100% | 1st play: 40%<br>2nd play: 80%<br>3rd play: 100% |
| Password protection | 1st play: 60%<br>2nd play: 100% | 1st play: 60%<br>2nd play: 100% | 1st play: 40%<br>2nd play: 60%<br>3rd play: 100% | 1st play: 40%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 40%<br>2nd play: 80%<br>3rd play: 100% |
| Privacy and confidentiality | 1st play: 20%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 20%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 0%<br>2nd play: 40%<br>3rd play: 40%<br>4th play: 80%<br>5th play: 100% | 1st play: 40%<br>2nd play: 80%<br>3rd play: 80%<br>4th play: 100% | 1st play: 40%<br>2nd play: 60%<br>3rd play: 100% |
| Workstation and hacking | 1st play: 20%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 40%<br>2nd play: 60%<br>3rd play: 100% | 1st play: 0%<br>2nd play: 40%<br>3rd play: 60%<br>4th play: 80%<br>5th play: 100% | 1st play: 20%<br>2nd play: 40%<br>3rd play: 80%<br>4th play: 100% | 1st play: 20%<br>2nd play: 80%<br>3rd play: 100% |
| Access control | 1st play: 60%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 60%<br>2nd play: 100% | 1st play: 0%<br>2nd play: 40%<br>3rd play: 80%<br>4th play: 100% | 1st play: 40%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 80%<br>2nd play: 100% |

The records show that privacy and confidentiality, and workstation and hacking are more challenging topics compare to the others. Employee number 3 did not select any correct answer when played these two topics for the first time and scored 0%. He also scored 0% for the first time playing the topic on access control. Table IV shows employees total plays, the lowest and the highest score on first attempts. The two topics of privacy and confidentiality, and workstation and hacking were replayed more than the other topics, each for a total of 18 times in order to obtain a score of 100%. The lowest first play scores go to privacy and confidentiality (0%), workstation and hacking (0%), and access control (0%). The highest first play score goes to Phishing (80%), email and spam (80%), and access control (80%).

TABLE IV. TOTAL PLAY, LOWEST AND HIGHEST SCORES OF TOPICS

| Topic | Total Play | Lowest Score | Highest Score |
|---|---|---|---|
| Phishing | 14 | 20% | 80% |
| Web using | 16 | 20% | 60% |
| Email and spam | 15 | 40% | 80% |
| Malicious code | 14 | 20% | 60% |
| Password protection | 13 | 40% | 60% |
| Privacy and confidentiality | 18 | 0% | 40% |
| Workstation and hacking | 18 | 0% | 40% |
| Access control | 14 | 0% | 80% |

## VIII. INFOSECURE SECOND ROUND PILOT TEST

An awareness training program should not be a one-time session, whereas, it should be a regular program and reinforced periodically. However, many training programs have failed due to users' resistance to participate in the same program. The strength of computer games over other training programs is its capability to be repeated in different time intervals while promoting employees' engagement. InfoSecure is dynamic that allows IT managers to change and customize the training content as well as the graphics. The aim is to keep gameplay more interesting.

HUKM decision makers agreed that the training program should be repeated quarterly. Hence, a post-training program was conducted three months after the initial training. The post-training program aimed to evaluate and measure the success of the InfoSecure game as a training tool. The five employees who participated in the initial training participated in the post-training as well. Table V shows employees' records during the post-training. Looking at employee number 3, then he played Phishing four times during the initial training to score 100%, whereas, during the post-training, he only played the topic three times to score 100%.

TABLE V. EMPLOYEES' RECORD OF PLAYING INFOSECURE

| Topic | Employee #1 | Employee #2 | Employee #3 | Employee #4 | Employee #5 |
|---|---|---|---|---|---|
| Phishing | 1st play: 100% | 1st play: 100% | 1st play: 40%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 100% | 1st play: 80%<br>2nd play: 100% |
| Web using | 1st play: 60%<br>2nd play: 100% | 1st play: 100% | 1st play: 80%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 80%<br>2nd play: 100% | 1st play: 80%<br>2nd play: 100% |
| Email and spam | 1st play: 80%<br>2nd play: 100% | 1st play: 100% | 1st play: 60%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 80%<br>2nd play: 100% | 1st play: 80%<br>2nd play: 100% |
| Malicious code | 1st play: 40%<br>2nd play: 100% | 1st play: 80%<br>2nd play: 100% | 1st play: 60%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 100% | 1st play: 100% |
| Password protection | 1st play: 60%<br>2nd play: 100% | 1st play: 100% | 1st play: 40%<br>2nd play: 60%<br>3rd play: 100% | 1st play: 80%<br>2nd play: 100% | 2nd play: 100% |
| Privacy and confidentiality | 1st play: 40%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 60%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 60%<br>2nd play: 60%<br>3rd play: 100% | 1st play: 60%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 40%<br>2nd play: 60%<br>3rd play: 100% |
| Workstation and hacking | 1st play: 60%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 80%<br>2nd play: 100% | 1st play: 40%<br>2nd play: 80%<br>3rd play: 80%<br>4th play: 100% | 1st play: 40%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 60%<br>2nd play: 80%<br>3rd play: 100% |
| Access control | 1st play: 80%<br>2nd play: 100% | 1st play: 80%<br>2nd play: 100% | 1st play: 60%<br>2nd play: 80%<br>3rd play: 100% | 1st play: 80%<br>2nd play: 100% | 1st play: 80%<br>2nd play: 100% |

Similar to the previous records, privacy and confidentiality, and workstation and hacking are more challenging topics compare to the others. However, employees show improvement and obtained better results during the post-training. For instance, employee number 3 who did not select any correct answer for these two topics during initial training, scored 40% for privacy and confidentiality and 40% for workstation and hacking during the post-training. Table VI shows employees total plays, the lowest and the highest score on first attempts. The two topics of privacy and confidentiality, and workstation and hacking were replayed more than the other topics, each for a total of 15 times in order to obtain the score of 100%. Some of the employees scored 100% on the first attempt.

TABLE VI. LOWEST AND HIGHEST SCORES ON FIRST ATTEMPTS

| Topic | Total Play | Lowest Score | Highest Score |
|---|---|---|---|
| Phishing | 8 | 40% | 100% |
| Web using | 10 | 60% | 100% |
| Email and spam | 10 | 60% | 100% |
| Malicious code | 9 | 40% | 100% |
| Password protection | 9 | 40% | 100% |
| Privacy and confidentiality | 15 | 40% | 60% |
| Workstation and hacking | 15 | 40% | 80% |
| Access control | 11 | 80% | 80% |

The result of the post-training showed satisfactory outcome indicating that InfoSecure is an effective tool for the information security awareness training program, not only from the training result but also from the employees' perspective. The objective of this training program was to enhance employees' awareness towards information security. As the results show, there is a significant improvement in employees' performance. Moreover, employees have shown a willingness to participate in the program as they had a pleasant experience during the initial training and enjoyed playing the InfoSecure game.

## IX. CONCLUSION

The objective of this paper is to design a serious game for information security awareness training programs for the healthcare industry. To achieve the objective, this paper reviewed the serious game design models and adapted the most effective model. The serious game presented in this paper consists of eight topics; each addressing one information security issue selected by the healthcare IT managers. The topics include phishing, web using, email and spam, malicious code, password protection, privacy and confidentiality, workstation and hacking, and access control. The main advantages of InfoSecure include flexibility, content updatability, and accessibility. Such flexibilities enable organizations to conduct post-trainings without the need to purchase a new game for training. Moreover, not changing the graphics of the game makes it less interesting to play for post-training. Inflexibility in games increases the cost of training, which is a big disadvantage. According to the results, the employees showed good impressions of the game's attraction, graphics, and its idea.

## REFERENCE

[1] S. A. Alserri, N. A. M. Zin, and T. S. M. T. Wook, "Gender-based engagement model for designing serious games," in Electrical Engineering and Informatics (ICEEI), 2017 6th International Conference on, 2017, pp. 1–5.

[2] A. Amory, "Game object model version II: a theoretical framework for educational game development," Educ. Technol. Res. Dev., vol. 55, no. 1, pp. 51–77, 2007.

[3] N. A. Bartolome, A. M. Zorrilla, and B. G. Zapirain, "Can game-based therapies be trusted? Is game-based education effective? A systematic review of the Serious Games for health and education," in IEEE Intelligent Systems, 2011, pp. 275–282.

[4] R. C. Basole, D. A. Bodner, and W. B. Rouse, "Healthcare management through organizational simulation," Decis. Support Syst., vol. 55, no. 2, pp. 552–563, 2013.

[5] F. B. García, S. García-Martínez, E. M. Navarrete-Ibañez, and M. J. Cervelló-Donderis, "Designing Serious Games for getting transferable skills in training settings," in Interaction Design and Architecture (s), 2013, no. 19, pp. 47–62.

[6] D. Charles, T. Charles, and M. McNeill, "Using player and world representation techniques from computer games to improve student engagement," in Games and Virtual Worlds for Serious Applications, 2009. VS-GAMES'09. Conference in, 2009, pp. 36–42.

[7] B. D. Cone, C. E. Irvine, M. F. Thompson, and T. D. Nguyen, "A video game for cybersecurity training and awareness," Comput. Secur., vol. 26, no. 1, pp. 63–72, 2007.

[8] S. De Freitas and M. Oliver, "How can exploratory learning with games and simulations within the curriculum be most effectively evaluated?," Comput. Educ., vol. 46, no. 3, pp. 249–264, 2006.

[9] M. Hu and Y. Kuang, "Human-machine interface: Design principles of pagination navigation in web applications," in Computer Science & Education (ICCSE), 2014 9th International Conference on, 2014, pp. 1140–1143.

[10] W. Hu, "A reusable eduventure game framework," in Transactions on edutainment I, Springer, 2008, pp. 74–85.

[11] T. Monk, J. Van Niekerk, and R. Von Solms, "Concealing the Medicine: Information Security Education through Game Play.," in ISSA, 2009, pp. 467–478.

[12] M. Azwan and B. Ibrahim, "a Certification Criteria for Software of Measuring Instruments Based on Malaysia Environment," 2011.

[13] A. Nagarajan, J. M. Allbeck, A. Sood, and T. L. Janssen, "Exploring game design for cybersecurity training," in Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2012 IEEE International Conference on, 2012, pp. 256–262.

[14] H. M. Omar and A. Jaafar, "Usability of educational computer game (Usa_ECG): applying analytic hierarchy process," in International Visual Informatics Conference, 2011, pp. 147–156.

[15] M. Prensky, "J COMPUTER GAMES AND LEARN| NG: D 1 G ITAL GAME-BASED LEARN 1 NG," 2007.

[16] A. Shapi'i and S. Ghulam, "Model for educational game using natural user interface," Int. J. Comput. Games Technol., vol. 2016, 2016.

[17] W. Westera, R. J. Nadolski, H. G. K. Hummel, and I. G. J. H. Wopereis, "Serious games for higher education: a framework for reducing design complexity," J. Comput. Assist. Learn., vol. 24, no. 5, pp. 420–432, 2008.

[18] A. Yusoff, "A conceptual framework for serious games and its validation." University of Southampton, 2010.

[19] N. H. M. Zain, A. Jaafar, and F. H. A. Razak, "SGameFlow framework: how to experience enjoyment in serious game (SG) for motor impaired users (MIU)," in Computer & Information Science (ICCIS), 2012 International Conference on, 2012, vol. 2, pp. 1020–1024.

# Hashtag Generator and Content Authenticator

Kavinga Yapa Abeywardana[1], Ginige A.R.[2], Herath N.[3], Somarathne H.P.[4], Thennakoon T.M.N.S.[5]
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka

***Abstract*—In the recent past, Online Marketing applications have been a focus of research. But still there are enormous challenges on the accuracy and authenticity of the content posted through social media. And if the social media business platforms are considered, majority of the users who try to add a market value to their own product face the problem of not getting enough attention from their target audience. The purpose of this research is to develop a safe and efficient trending hashtag generating application solution for social media business users which generates trending and relevant hashtags for user content in order to get a broad reach of target audience, automatically generates a meaningful caption to their relevant posts and guarantees the authenticity of the product at the same time. The user content is analyzed and filters the important keywords, generates a meaningful caption, suggest related trending keywords and generates trending hashtags to get the required reach for online marketers. Additionally, the marketing products' content authentication is ensured. The application uses Natural Language Processing, Machine Learning, API technologies, Java and Python technologies. A unique database is assigned to users which contains rankings for each user. The target audience who engages in buying products get to know about the status of the sellers with respect to authenticity of the content. It is believed that the application provides a promising solution to existing audience reach problems of online marketers and buyers. The significance of this system is to help marketers and buyers to engage in online buying and selling with much effective, reliable and safer ways. This mitigate the vulnerability of bad social media marketing influences and helps to establish a safe and reliable online marketing practice to make both sellers and buyers happy. This paper provides a brief description on how to perform an organized online marketing discipline via the Trending Hashtag Generator & Image Authenticator application.**

***Keywords*—Hashtags; social media; NLP; machine learning; REST API; content authentication***

## I. Introduction

A hashtag is a name or an identifier that resolves to a description of its referent. In other words hashtag is a keyword or phrase preceded by the hash symbol, written within a post or comment to highlight it and facilitate a search for it. In the present day, hashtags are immensely used for brand promotion and social media discussions. In principle hashtags facilitate powerful identification functionality to any kind of HTTP based services (social media platforms etc.). Essentially, by including hash marks in your post, provided the appropriate privacy settings are in place it can be indexed by the social networks so that it is discoverable by everyone, even if the user is not subscribed to your account updates. There are countless number of individuals and merchants who intend to sell a genuine product by identifying and focusing on a targeted audience. Since hashtags play a vital role in online marketing, aforementioned people try to find the most relevant hashtags to increase the audience reach for their products. However most of them end up having a hard time to identify the most relevant and trending hashtags at particular given time. The major problem faced by the user is "unavailability of a platform that is made for clients which can analyze the post (photo/text content) and fetch trending hashtags specific for the relevant post, auto generate a meaningful caption, categorize the hashtags according to the popularity which can get a considerable awareness from target audience in a completely new and user-friendly way". Even though the content is published along with some hashtags manually, the content uploader does not have any prior knowledge whether the hashtags he/she manually searched are trending or not at the current time. On the other hand, if an uploader uploads a digitally captured image of the product, the audience does not have any proof whether the content is genuine or not. It is a commonly known fact that buyers are reluctant to purchase a product when they have a suspicion about the authenticity of the advertised content. Considering these facts the authors have introduced the "Hashtag Generator and Content Authenticator" as a web application which lets users easily upload their content, get a categorized trending hashtags list along with a caption, select the preferred hashtags according to his/her preferences and push the content to the relevant social media platform. The web app lets the consumer to generate the desired amount of reach via hashtags which could gain much required attention for commercial and noncommercial purposes. The authors have developed the app to authenticate the uploaded content with a specialized tag for the customers to identify pre authenticated content. This will prevent or hinder unauthorized copying and editing of the content which ensures authenticity. It will help uploaders to prove the authenticity of their content to their audience which is crucial from sales perspective. These processes are done using API technologies, Natural Language Processing (NLP), machine learning and data forensic technologies. The proposed features are unique compared to existing tools due to the unavailability of all these features in modern social medial platforms and related applications. The authors have provided an accurate, user friendly web application interface to ensure the authenticity of the products to be sold, making things easy for both the sellers and customers (target audience).

## II. Literature Review

### A. Analysis of the usage of Metadata

The literature conferred by Hazinah Kutty Mammi and Mohd Aliff Faiz bin Jeffry[1] explore the practicability of securing the images by using metadata with digital

watermarking. They have proposed to embed selected metadata to the image in a watermarking procedure. These embedded images were used in multiple social media platforms for analyzing the accomplishment of the project and proposed watermarking techniques. The literature discusses how metadata can be used to prove the originality of an image.

During the literature review, the priority was given to more modern approaches to establish a solution to prohibit users publishing downloaded images from the internet as their own .The research done by Walid Hussein, Osman Ibrahim and Mostafa A. Salama [2] on image processing using the signature verification techniques reviews how signature verification techniques can be done using an image.

*B. Analysis of Web Crawling Algorithms*

In the process of the literature survey, it was identified that the main web crawler approaches that can be used for comparison are based on the following characteristics.

- Resemblance and the relativity of the data that the crawler acquire
- Effective relevance forecasting to facilitate downloading the composition.
- Overall effectiveness of the crawler

In conclusion it was identified that the most beneficial algorithm presented is the 'Focused Crawling' algorithm due to its minimum response time. The research done by Handoko, et al. [3] on optimization of the focused crawler by using Genetic Algorithm has combined the focused crawler with genetic algorithm to resolve and fine tune web searching. It discusses about the complications caused by the local searching algorithms. Table 1 shows the comparison carried out on both approaches. The features used in the Focused crawler and the Genetic Algorithm will be used to achieve the goals in the proposed solution.

TABLE I. PRECISION OF THE WEB CRAWLERS' COMPARISON DONE IN THE RESEARCH[3]

| Category | BFS crawler's precision | GA crawler's precision |
|---|---|---|
| Education | 90% | 97% |
| Computer | 85% | 97% |
| Digital | 80% | 82% |
| Analog | 63% | 93% |
| Sport | 90% | 95% |

Researchers have studied and developed multiple NLP related solutions over time to solve problems in various domains. They can be categorized as follows.

- Lexical and morphological analysis, noun phrase generation, word segmentation, etc.
- Semantic and discourse analysis, word meaning and knowledge representation.
- Knowledge-based approaches and tools for NLP [4].

Noun phrasing is considered to be an important NLP technique used in information retrieval. One of the major goals of noun phrasing research is to investigate the possibility of combining traditional keyword and syntactic approaches with semantic approaches to text processing in order to improve the quality of information retrieval [4].

The growing technology of NLP suggests that there are two possible scenarios for the future interactions between computers and humans: in the user-friendliness scenario, computers become smart enough to communicate in natural language, and in the computer friendliness scenario humans adapt their practices in order to communicate with, and make use of, computers [4].

Topic modeling is a type of statistical modeling for discovering topics that occur in a collection of documents. Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

Recently, a number of studies have shown that the use of machine learning and text mining methods to automatically identify relevant studies has the potential to drastically decrease the workload [5].

Topic analysis is currently gaining popularity in both machine learning and text mining applications.

Automatic text classification for systematic reviews has been investigated by Bekhuis [6] who focused on using supervised machine learning to assist with the screening phase. Octaviano [6] combined two different features, i.e. content and citation relationship between the studies, to automate the selection phase as much as possible.

From a topic modelling perspective, Miwa firstly used LDA to automatically suggest topics for related keywords and reduce the difficulty of systematic reviews using an active learning strategy [6].

Since the emergence of topic models, researchers have introduced this approach into the fields of biological and medical document mining [7]. Such experiments proved LDA could be successfully applied to text classification. In the present day, LDA modeling is being developed for machine based communication purposes [7].

The trending hashtag recommendation problem addresses suggesting hashtags to explicitly tag a post made on a given social media platform, based upon the content and the context of the post. The issue of trending hashtag recommendation has emerged as a mainstream area of research overtime. "Hashtag recommendation for micro-blogs using topic specific translation" by *Ding, Q. Zhang, and X. Huang* [8][9] are researches based on NLP & Probability based algorithms by key phrase extraction and model them into topic specific translation. "Recommending #tags in twitter" by *E. Zangerle, W. Gassler, and G. Specht* [10] is also a research that targets microblogs but this is based on extracted hashtags ranking model. All these research efforts are specifically targeted at microblogs which is a highly specific area of content.

In the research "Semantic embedding from hashtags" by *Weston, S. Chopra, and K. Adams* [11] they used an NLP and ML based Convolutional Neural Network for hashtag recommendation with supervised word embedding. Comparatively this can be identified as a successful approach.

In present most of the research work utilize the advancements of ML to achieve their objectives. "User Conditional Hashtag prediction for Images" by *E. Denton, et al.* [12] is an approach that used ML along with the hashtags & contextual information about the user to perform hashtag prediction for user given image. Simply how user meta-data combined with images derived from a CNN can be used to predict hashtags. With the data, the researchers developed a user model which could be applied for a large dataset that is taken from Facebook. The user model primarily predicted hashtags, but the predicted hashtags were not "trending hashtags". In this approach a hashtag embedding model will be used that trains with the collected data. This method would be very practical because of the availability of data.

Data extraction from social media platforms comes under the categorization of social media data mining. Mainly there are three different ways to harvest data from social media platforms. Those are through APIs, personal archives and scraping. Since most of the social media platforms have updated their restrictions on data extraction due to various privacy related reasons, personal archiving method is not practical. Therefore, the preferred method for this proposed solution is scraping.

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability of a given sample belonging to a particular class. Bayesian classifier is based on Bayes' theorem. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computation involved and, in this sense, is considered "naive" [13]. Naïve Bayes method recommends hashtags by observing the content produced by the target user. In this paper it proposed to use Bayes model to estimate the probabilities of using different hashtags. Using this method, hashtags which are used by posts that has similar content can be identified [14].

The Natural Language Toolkit is a suite of program modules, data sets, tutorials and exercises, covering symbolic and statistical natural language processing. NL TK is written in Python and distributed under the GPL open source license. Over the past three years, NLTK has become popular in teaching and research [15]. In the proposed research application, Naïve Bayes classifier allows to classify the generated hashtags based on the analysis that given by the specific algorithms that trained using large training data-sets.

## III. Research Gap

Currently, there are few hashtag generating applications which can fetch hashtags only for a given input text. The current platforms do not facilitate users to generate hashtags based on images, paragraphs and/or URLs. Hence the user has to manually search for hashtags through the existing applications. Existing applications has a collection of hashtags which rarely updates with time. As a result, a precise decision cannot be made whether the fetched hashtags are trending or not. Therefore, the hashtags offered by these applications are mostly outdated and have a less tendency for the user to achieve the expected reach for his publications. Furthermore the user is unable to sort out the suggested tags according to the targeted audience. This is a major disadvantage for corporate users that use social media as a mode of advertising platform for their purposes and digital marketers as well as the personals that are involved in social commerce. The next identified gap is the current applications are unable to guarantee the authenticity of the seller uploaded content. Seekmetrics, All-hashtags and Hashtagify are few names of above mentioned applications which can generate hashtags only for a small text input. When the text input is heavy, the accuracy of the performance drops to a recognizable rate.

Trending Hashtag Generator and Content Authenticator covers all the gaps identified in the existing applications mentioned above.

## IV. Methodology

### A. Analysis and Requirement Gathering

The analysis phase was focused on gathering information about the existing systems and analyzing the weaknesses and strengths of the respective systems which lead to the concept of developing the new system. Requirements of the new system were clearly understood during the analysis phase. The research team identified main users of the 'Trending Hashtag Generator and Image Authenticator' are social media marketers, promoters and their target audience. After conducting a survey, the research team found that more than 70% of online marketers use hashtags to promote their content and majority of them were not satisfied with their audience reach even after using social media platform audience tools. 60% of online marketers and promoters were using existing systems to find matching hashtags for their content. Authors came into a conclusion that an efficient system which analyzes the content and generate "real time trending hashtags" would assist the online marketers and promoters to get the audience reach they expect.

### B. Implementation

The overall system was developed and built during this phase. The system architecture is mainly divided into four main components.

As depicted in Figure 1, authors are following a sequential approach to implement the proposed solution. Figure 1 also illustrates the type of output expected from the proposed solution.

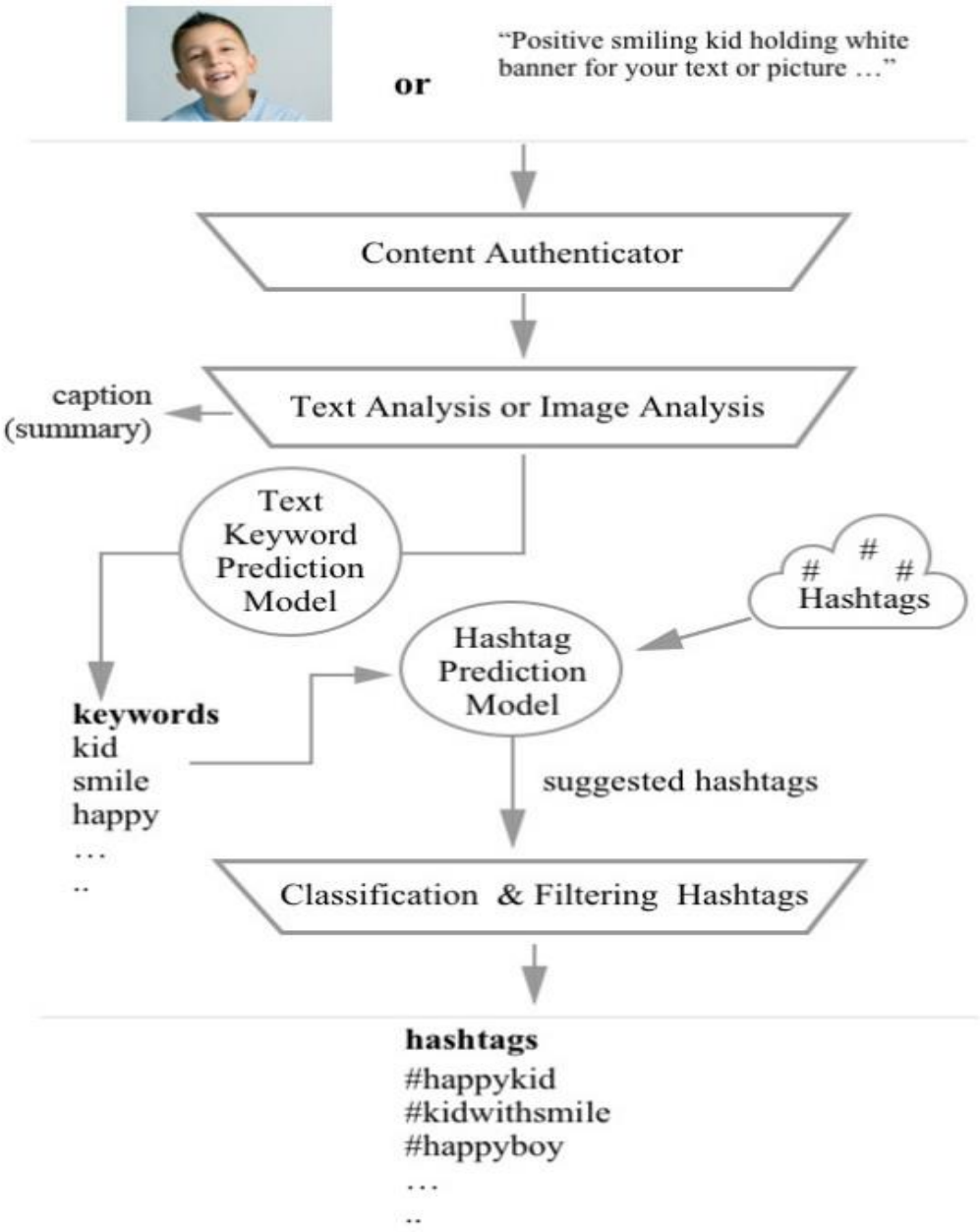



Fig. 1. System Architecture Diagram

### C. System Components

*1) Image Authentication using Metadata*

In this process, the content authentication will be ensured with the usage of metadata. The metadata will be extracted with the aid of java libraries and subjected for an appraisal. The main metadata tags that will be extracted will be the make of the camera, model of the camera and the tags that will aid in locating the geographical location and the serial number. The appraisal will be done based on a defined criteria that will aid in finding the authenticity of the uploaded image. This criteria will be mainly based on the quality of the Metadata. Based on the appraisal, a rating will be provided to the user and it will be provided based on predefined assumptions. These assumptions are created based on the targeted audience which are social media influencers and commercial users. Once the user gets the rating, the user will be provided with an option to accept the rating or to request for a re-assessment and provide any justification if required if the rating is not in a satisfactory level. Once the rating is accepted the user can move on to the next phase.

*2) Image Feature Extraction*

The main objective of this process is to identify if the user uploaded images are already existing in the internet. The image feature extraction process is done based on the scale-invariant feature transformation which is used to detect and describe local features in images. Correspondingly speeded up

robust feature techniques are used for object recognition, classification and image regression. Feature extraction is done using the images that are harvested from a web crawler and to the images that gets uploaded by the publisher. Once both features are extracted they will be compared to get a match. The feature extraction process will recognize interest points of the image and the RGB values will be compared. This matching process depends on the Euclidean distance between the interest points of the two images.

*3) Text Analysis*

This process is based on Natural Language Processing (NLP) and Machine Learning (ML). Python is used as the programming language. If the user inputs a URL which he needs to generate hashtags, the text content in the respective URL is fetched removing all the HTML markup elements and other unnecessary items. Using NLP libraries, stop words and punctuation marks are removed. The repetitive words are removed along with the canonical form of words. After above processes, a list of important keywords are generated. Nouns, verbs, pronouns etc. are separately identified. The nouns are then analyzed and automatically categorized into unique topics. This process is known as Topic Modeling or Latent Dirichlet Allocation (LDA). A unique ID is allocated to each topic and the relevant topic model is compared with a manually created dataset (1000 keywords and similar words related to each keyword). This process is carried out using cosine similarity algorithms which is commonly used in data science domain. The final result is a recommender system which is an application of machine learning which recommends related, most matching keywords for the generated words in LDA topic model. The generated keywords and recommended keywords are then passed in order to generate hashtags. The ultimate target in this phase is to automatically suggest many keywords as possible to user so that the system will generate various hashtags which are relevant to the user posted content. As a result, the user can get a broad and specific target audience's reach. Additionally, a meaningful summary is generated using term frequency–inverse document frequency (TF-IDF) which can be used as a caption when a user posts a post. If the user inputs a direct text, a similar process is carried out using NLP and ML techniques respectively. In this case, considering HTML elements is unnecessary.

*4) Recommending Hashtags*

The aim of this component is to recommend the best hashtags for the user content. A dataset which includes hashtags with popularity data is prepared for further processing. Social media public APIs and web scraping algorithms along with python is used to extract hashtags. After data extraction, data is cleaned and converted into the desired format. The respective dataset is divided into two datasets where the first dataset is to develop the solution and the other dataset is to test the model. With respect to image analysis process, since image classification is a broad research area that is not the scope of this research, Google Vision API is used to analyze images along with existing artificial intelligence and machine learning technologies. In this process, real time hashtags are taken from Instagram. In order to perform the operation, four factors are considered. Calculation of Cosine similarity of the hashtags and the keywords is one factor. To perform the task, a dataset has to be prepared manually to calculate the cosine similarity. The frequency of the hashtags are considered within a given period along with the like and comment count for the relevant hashtags. Then weights are assigned to indicate the relevancy of hashtags accordingly. Finally, hashtags are inserted into an equation to calculate the final score. For these processes, Gensim, Numpy, Tensorflow framework, Word2vec technologies are used. Since each iteration helps to perform the prediction more precisely, the training process is repeated.

*5) Classification and Filtering Hashtags*

This process is based on Naïve Bayes Classifier and Machine Learning (ML). The generated hashtags are filtered into classifications to reach the ideal target audience. The selected filter option will analyze each and every hashtag in order to find the ideal sections of the hashtags. The related keywords of each hashtag is analyzed using a specific algorithm which uses Naïve Bayes Classifier to generate the results. Each category has its own unique algorithm in order to provide the classifying results, based on the specific target selection of the content uploader. The keyword analyzing algorithms are based on Naïve Bayes Classifier which will provide a probabilistic outcome of the related section of each hashtag. The classification process of the system is developed using ML which uses scikit-learn and natural language toolkit (NLTK). The classifying algorithms are trained using specific data models which increases the quality of the classifying result. According to the selected filter option, a unique classifier algorithm analyzes the keywords in order to classify the generated hashtags.

## V. Result & Discussion

Hashtag Generator and Content Authenticator is a web based application that allows users to find the most popular hashtags for user specific content to get a recognizable amount of reach from a target audience with the content authentication service.

The images which go through the uploading process initially undergo the content authentication phase where the images will be authenticated with the aid of the metadata and feature extraction. These authentication processes are parallelly executing. Once the authentication is done, a tag will be shown to the audience indicating the authenticity of the images. Afterwards, the images will be forwarded to the image analysis component which generates a set of keywords relevant and unique to each uploaded image. If user enters a text content instead of an image, text will be analyzed and important keywords are analyzed along with a meaningful summary. Additionally, most related keywords are suggested to the user using the trained keywords model. In the next phase, keywords will be analyzed and relevant hashtags will be suggested using the trained hashtags model which consists of collected data of hashtags along with the popularity. Then the suggested hashtags will be classified and filtered in the last phase which allows the user to select his/her favorites based on the interest.


Fig. 2. System Output

In the content authentication process, the selection of relevant images which needs to be harvested and exclude unwanted images can be considered as a limitation as it is a technical challenge to achieve 100% accuracy. Furthermore, a high processing speed is required in order to compare the features which are extracted between the images harvested and the image uploaded, which can also be considered as a technical limitation.

In the text analysis process, a data model (LDA topic model) is implemented to automatically separate keywords into specific groups. In this process, the accuracy of topic filtering is not 100% as the current LDA topic modelling has lot of room for improvement which is an active area of research.

A set of hashtags with popularity domain was initially required for the hashtag suggesting process but it was difficult to find an up-to-date hashtag dataset with the popularity details. The training process of hashtag model needed existing trending hashtag dataset models. The retrieval of some datasets which has the popularity domain from social media is increasingly becoming difficult due to the recent social media privacy policy changes. General Data Protection Regulation (GDPR) changes in the European Union is a clear example for such rule change.

In the hashtag categorizing process, though there are infinite methods of categorizations, only a limited number of specific categorizations are considered. Each and every selected category requires specific algorithms which need to be trained with unique data sets.

## VI. Conclusion

The problem of accuracy and authenticity of the content published by online marketers is solved. A proper guidance can be provided to online sellers who don't have any idea about trending hashtags and reaching the proper target audience. The Metadata Extraction process successfully contributes towards the verification of the authenticity of the user uploaded content with the help of Image Feature Extraction process. Text Analysis phase generates a meaningful caption to user uploaded text content and will make sure the necessary keywords are generated and suggested at the same time which helps the seller to get more attention from the audience. The custom made dataset is built to suggest keywords which goes toe-to-toe with current marketing game on social media. Hashtag recommendation process identifies the image and/or the keywords supplied from the Text Analysis phase and generates relevant, real time trending hashtags. The generated hashtags are then categorized based on different attributes and the categorized hashtags are suggested to the user (online marketers).

## VII. Future Works

In the text analysis process, the recommender system uses a manually built dataset with 1000 keywords which are limited to the following domains; fashion, nature and travel. The dataset can be extended to several other domains as future works in order to cater user inputs from other domains.

General recommendations to those who are willing to develop this system further are as follows:

- Expand the API enabling the access to 3rd party users.
- Develop the accuracy of the topic model (LDA topic model) which suggests related keywords.
- Develop the advanced popularity prediction mechanism for the suggested hashtags.
- Expand the system in to mobile platforms as Android and IOS.
- Embed the personalization in to the system. Develop models to suggest hashtags depending on the person's past hashtag usage patterns.
- Expand the image authentication for wide range of images on the internet.

## References

[1] M. bin Jeffry and H. Mammi, "A study on image security in social media using digital watermarking with metadata - IEEE Conference Publication", Ieeexplore.ieee.org, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8270435/. [Accessed: 06- May- 2018].

[2] W. Hussein, M. Salama and O. Ibrahim, "Image Processing Based Signature Verification Technique to Reduce Fraud in Financial Institutions", Matec conferences.org, 2016. [Online]. Available: https://www.matec-conferences.org/articles/matecconf/pdf/2016/39/matecconf_cscc2016_05004.pdf. [Accessed: 10- May- 2018].

[3] B. Yohanes and H. Wardana, "Focused Crawler Optimization Using Genetic Algorithm", Telkomnika.ee.uad.ac.id, 2018. [Online]. Available: http://telkomnika.ee.uad.ac.id/n9/files/Vol.9No.3Des11/1SS-AI9.3.12.11.01.pdf. [Accessed: 12- May- 2018].

[4] Chowdhury, G. (2005). Natural language processing. [online] Available at: https://onlinelibrary.wiley.com/doi/full/10.1002/aris.1440370103 [Accessed 15 Mar. 2018].

[5] Li, S. (2018). Topic Modeling and Latent Dirichlet Allocation (LDA) in Python. [online] Towards Data Science. Available at: https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24 [Accessed 3 Aug. 2018].

[6] Mo, Y., Kontonatsios, G. and Ananiadou, S. (2015). Supporting systematic reviews using LDA-based document representations. [online] Available at:

https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-015-0117-0 [Accessed 4 Aug. 2018]

[7] Liu, L., Tang, L., Dong, W., Yao, S. and Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. [online] An overview of topic modeling and its current applications in bioinformatics. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5028368/ [Accessed 5 Aug. 2018].

[8] Z. Ding, Q. Zhang, and X. Huang, "Automatic Hashtag Recommendation for Microblogs using Topic-Specific Translation Model.," in COLING (Posters) , 2012, pp. 265–274

[9] Z. Ding, X. Qiu, Q. Zhang, and X. Huang, "Learning Topical Translation Model for Microblog Hashtag Suggestion.," in IJCAI , 2013, pp. 2078–2084.

[10] E. Zangerle, W. Gassler, and G. Specht, "Recommending\#-Tags in Twitter," in Proceedings of the Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR Workshop Proceedings , 2011, vol. 730, pp. 67–78.

[11] J. Weston, S. Chopra, and K. Adams, "#TagSpace: Semantic Embeddings from Hashtags.," in EMNLP , 2014, pp. 1822–1827.

[12] E. Denton, J. Weston, M. Paluri, L. D. Bourdev, and R. Fergus, "User Conditional Hashtag Prediction for Images.," in KDD , 2015, pp. 1731–1740.

[13] Ming K. 2007 Naive Bayesian Classifier [online] Available at: http://cis.poly.edu/~mleung/FRE7851/f07%%202FnaiveBayesianClassifier.pdf

[Accessed 04 Aug. 2018]

[14] Su Mon 2011 On Recommending Hashtags in Twitter Networks [online] Available at: https://link.springer.com/chapter/10.1007/978-3-642-35386-4_25 [Accessed 16 Mar. 2018]

[15] Steven Bird 2002 NLTK: The Natural Language Toolkit [online] Available at: https://dl.acm.org/citation.cfm?id=1118117 [Accessed 05 Aug. 2018]

# EEG Signals based Brain Source Localization Approaches

Anwar Ali Gaho[1], Sayed Hyder Abbas Musavi[2]
Department of Electrical Engineering, Faculty of Engineering Science and Technology, Indus University Karachi, Sindh, Pakistan

Munsif Ali Jatoi[3]
Department of Electrical Engineering Technology
Benazir Bhutto Shaheed University of Technology and Skill Development, Khairpur, Sindh, Pakistan

Muhammad Shafiq[4]
University of Innsbruck, IMGI, Austria
SUPARCO, Karachi, Pakistan

***Abstract*—This article is focused on the overview of functionality of the neurons and investigation of the current research and algorithms used for brain source localization. The human brain is made up of active neurons and continuously generates electrical impulses on scalp surface. The neurons transmit the message through the dendrites called pyramidal cells. The active parts of the brain are addressed and measured by various neuroimaging techniques such as electroencephalography (EEG), magnetoencephalography (MEG) etc. These techniques help to diagnose pathological, physiological, mental and functional abnormalities of the brain. EEG is a high temporal resolution and a low spatial resolution technique which yields the non-invasively potential difference measurements between pair of electrodes over the scalp. It is used in understanding behavior of brain which is further used to analyze various brain disorders. EEG brain source localization has remained an active area of research in neurophysiology since last couple of decades and still being investigated in terms of its processing time, resolution, localization error, free energy, integrated techniques and algorithms applied. In this paper, several approaches of forward problem, inverse problem and Bayesian framework have been explored to address the uncertainties and issues of localization of the neural activities incurring in the brain.**

***Keywords*—Electroencephalograph; brain source localization; forward problem; inverse problem; bayesian framework**

## I. INTRODUCTION

The human brain is made of $10^{10}$ neurons. The neurons pass out the message through the dendrites called pyramidal cells so called neural activity. This brain neural activity causes the generation of the electrical potentials over the scalp[1, 2]. The neurons' main function is to communicate chemically and electrically with other neurons[3]. They send the messages from one lobe to another lobe thereby causing the dipole currents inside the brain generated due to the stimuli or other brain disorders. and subsequently generate the synapse, pre-synaptic and post-synaptic actions in the brain [4]. This section consists of the following parts:

### *A. Brain Neurons*

The brain controls all main functions of the body. It is composed of cerebrum, cerebellum, and brainstem[5, 6]. The cerebrum is divided into left and right hemispheres. The fissures or grooves, divides cerebrum into four main parts i.e., frontal, temporal lobe, parietal lobe and occipital lobe. The mid-brain, pons, and medulla are interconnected with cerebrum with spinal cord[7]. Since the neuron is composed of cell body, nucleus, axons, nodes of ranvier, myelin sheath and dendrites as shown Fig.1. The neuron's structure physically change with the passage of the time and age of the subjects such as from infant to old [8].

### *B. EEG Current Dipoles*

On the application of the stimuli, electricity is generated by a particular group of neurons in the cortex so called dipoles as shown in Fig. 2. This dipole generates the electric field and creates potential field which is picked up by the scalp EEG electrodes.

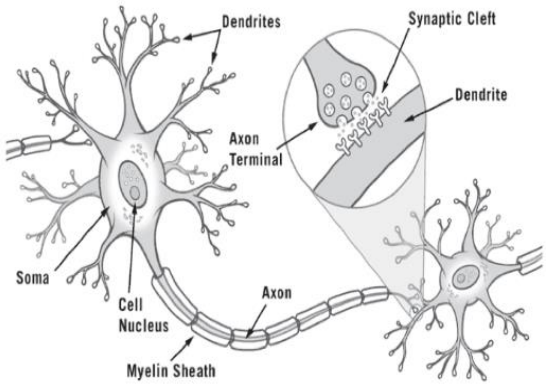



Fig. 1. Neuron structure[3]

EEG measures the current caused by the excitations of the dendrites of pyramidal neurons. The potential differences are caused by summed postsynaptic potentials from pyramidal cells that create diploes between soma and apical dendrites as shown in Fig. 3.

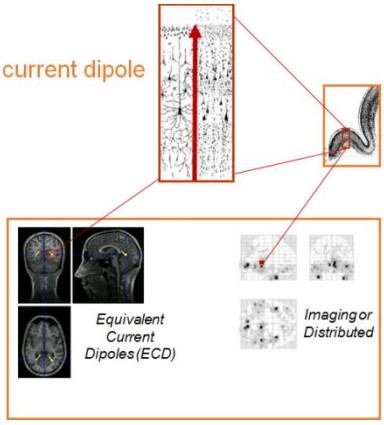



Fig. 2. Generation of current dipoles[6]

### C. EEG Generators

The pre-synaptic and post-synaptic are linked through the synapses. The neurons process the inhibitory action potentials and then convert it into post synaptic potentials for firing other neurons as shown in Fig. 3. The signals are superimposed synchronized and pass from grey matter and become normal to cortical surface thereby creating potentials on cortical surface.

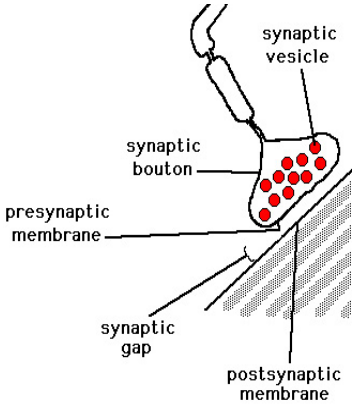



Fig. 3. Releasing ions [9]

The depolarization or excitatory postsynaptic potential (EPSP) reduces to -40 mV in the intracellular volume. The hyper polarization or inhibitory postsynaptic potential (IPSP) potential difference is subsequently increased. Due to the action of Na+, K+ and Cl- in the brain, potential difference of 70–110 mV for the period of 0.3 millisecond is generated as shown the Fig. 4.

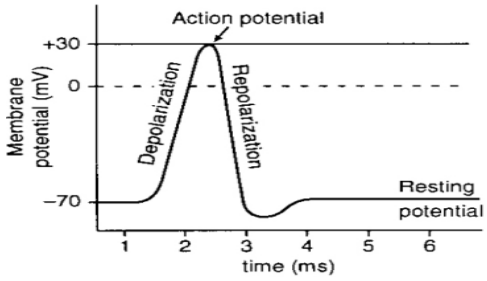



Fig. 4. Action potential [10]

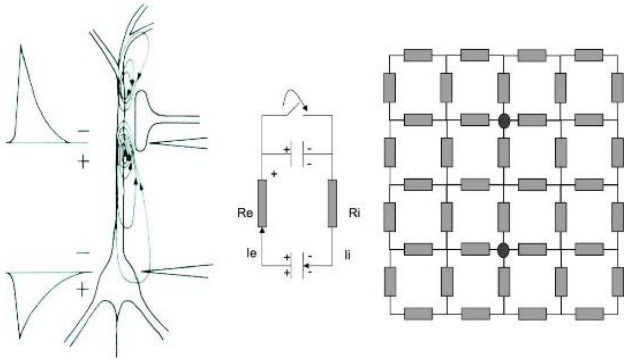



Fig. 5. Equivalent circuits for a neuron [12]

An equivalent circuit of capacitor, resistor and potential source is shown in Fig. 5. At rest, voltage stored between the intra-cellular and extracellular areas is denoted by two charged capacitors. One is assigned for potential difference between apical dendrites side and other for potential difference at the cell body for basal dendrite side. Action of charging and discharging takes place between capacitors with the on and off switch.

### D. EEG Applications

EEG being a revolutionized neuroimaging technique is not only used for helping in diagnosing pathological, physiological, mental and functional abnormalities of the brain but also for understanding epilepsy, schizophrenia, and Alzheimer's diseases. EEG is assisting in diagnosing disabilities, hyperactivity, sleep, awakening disorders, depression and surgery in patient with seizures. Moreover, EEG is useful for identifying coma, brain and death problems and locating injury, stroke, tumor, evoked potentials sources and brain cognitive behavior. EEG is also extensively used in investigating drug effects, aesthesia issues, brain growth and death. It is currently being integrated with other brain imaging modalities for better human life [11].

In the literature, there are several other neuroimaging techniques such as MEG, ECoG. Functional MRI (fMRI) and NIRS used in brain neurology. Their comparative analysis is given in TABLE I. where EEG mostly seems to be most favorable neuroimaging technique in given characteristics such as measurement, resolution and method.

TABLE I. COMPARISONS OF NEUROIMAGING TECHNIQUES[11]

| **Technique** | **Measurement** | **Temporal** | **Spatial** | **Method** |
|---|---|---|---|---|
| ***Type*** | ***Type*** | ***(m sec)*** | ***(mm)*** | ***Type*** |
| EEG | Electrical | 0.05 | 10 | Invasive |
| MEG | Magnetic | 0.05 | 5 | Invasive |
| ECoG | Electrical | 0.003 | 1 | Invasive |
| fMRI | Metabolic | 1 | 1 | Non-Invasive |
| NIRS | Metabolic | 1 | 5 | Non-Invasive |

This article is divided into three sessions. Session I is about the introduction of EEG source localization. Section II covers the EEG brain source localization approaches. Section III gives conclusion and remarks.

## II. EEG Source Localization Algorithims

The forward and inverse problems are used for locating the sources of the brain activity. The former problem measures potentials over the scalp of the skull whereas the later yields sources of the activity and localization[12]. Both Techniques are discussed in details in subsequent sections.

### A. Forward Problem

Forward problem is related to measuring of the potentials with the electrodes from the surface of the scalp and has a unique solution as shown in Fig. 6.

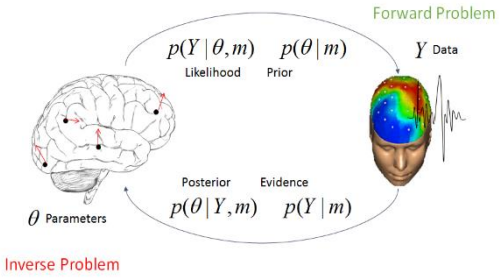



Fig. 6. Forward problem [12]

The grey matter as shown in Fig. 9. is composed of pyramidal cells which forward the currents to scalp surface through the apical dendrites with 2 milli-second duration. Orientation of the neurons causes the superimposing and cancelling out the electrical signals. Since neural activity is neither the static and nor the dynamic and therefore it is called quasi- statics conditions [13]. The Poisson's equation containing the divergence operator relates the potential difference, and current densities of the volume of the brain i.e.

$$\nabla.J = \lim_{G\to 0} \iint_{\partial G} JdS$$

This integral model represents the complete spherical shape of the surface and volume (G) where the flux in is negative and flux out is positive. Its current density form is represented by $\nabla.J = I_m$. Lines represent the flux caused by the currents and potentials generated in the volume[14] as shown in Fig. 7.

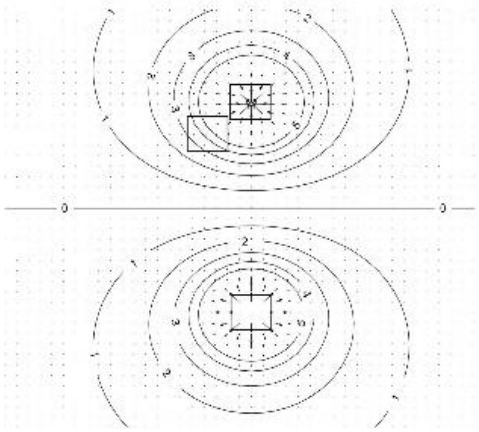


Fig. 7. Current density with equip-potential lines[12, 15]

The human head is composed of the isotropic and anisotropic tissues. The conductivity causes electric currents. The skull is made of spongiform layer, skull and scalp and composed of biological and chemical tissues [16] which affects the current flow from cortical surface to scalp. [17]. The skull consists of three layers such as a spongiform layer and two hard layers as shown in Fig.8. The conductivity tangential to the skull surface is 10 times larger than the radial conductivity.

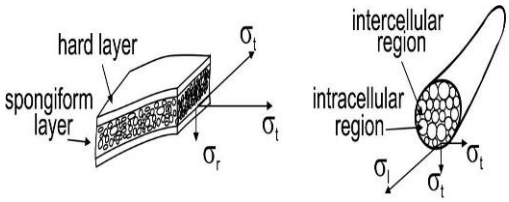



Fig. 8. Brain conductivity tissues [18].

The white matter has 9 times greater conductively than grey matter due to structure of the cells. The watery tissues have a greater conductivity. The several models of isotropic and anisotropic conductivities are incorporated in the forward problem to address the issues of the conductivities.

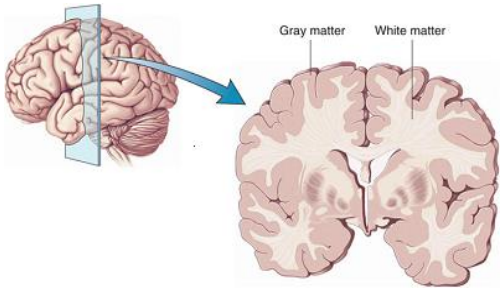



Fig. 9. Grey and white matter[7, 9]

Faraday's law gives zero results due to quasi-static conditions that is $\nabla \times E = 0$. Potential and electric fields with the gradient operator is given as $E= -\nabla V$. The negative sign reflects electric field direction. The human head is composed of the different types of the layers. As the outer layers of the human are scalp and air. As there is no current that flows in the outer layers due to air which is non-conducting material. Therefore, the current density inside the outer air layer is zero. The current dipole of pyramid cells has the same current flown in and flow out. The dipole has six parameters including three translational parameters and three spherical parameters with a unit vector at a pint in the source space between two monopoles as shown in Fig. 10.

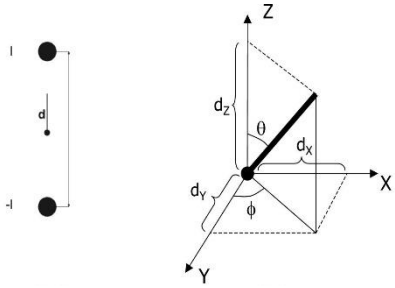



Fig. 10. Dipole parameters [12]

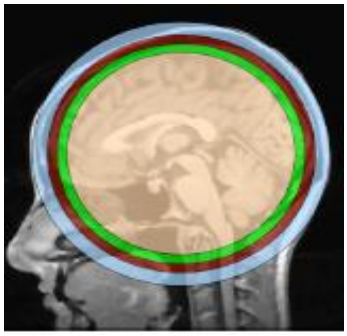


Fig. 11. Simple head model [9]

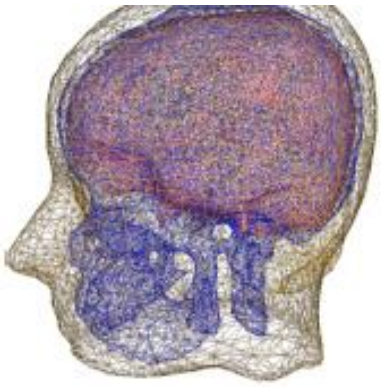


Fig. 12. Realistic head model[24]

The human head is a sphere and is composed of the conductivity tissues of skull and scalp. It is made of three concentric circles representing each circle as a layer [18]. Its conductivities are modeled in terms of radial and tangential layers[19]. The prolate and oblate spheroids or eccentric spheres are modeled and discussed in the literature. The Berg approximation [20] represents a single-sphere model for a three to four layer sphere model for increasing the computation and geometry accurateness [21, 22]. Simple head models are single layer spheroid models which are simple, fast and accurate one. The realistic head models with complexity are more numerical and realistic such as boundary element method (BEM)[19] and finite element method (FEM)[14, 23].

The realistic head models also called numerical models (Fig. 12) in computational time are higher than multi-layer spheres or analytical models (Fig. 11) due to formation of thousand pieces [25]. A good approach between spherical head model and realistic head one is counted a good quality of sensor-fitted approach i.e. exact on the location of the activity[26].

The BEM numerical technique calculates the surface potentials produced by current sources located in a piecewise uniform volume formed with isotropic conductivities structures [9, 27]. The pieces are made in such size that they are so small for yielding isotropic conductivities[28]. There are three boundaries with the three layers in spherical model. Each layer is put into pieces and is computed with small boundary elements. The distances of the layers from brain, skull and scalp to the center positions are 8 cm, 8.5 cm and 9.2 cm, respectively[29]. Each layer is digitized in triangles, for computing the potential at its center (see Fig 13 and 14).

The Green's first identity for integration is implemented to solve Poisson's equation in a realistic head model is the finite element method (FEM). The 3D volume conductor is digitized into small elements as shown in Fig 13 and Fig 14.

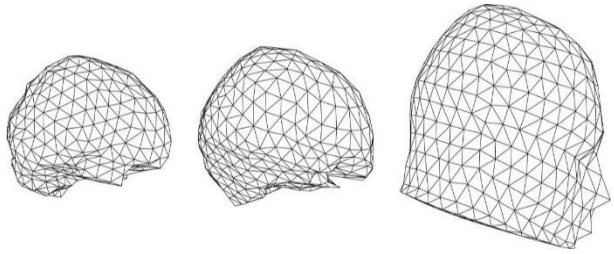


Fig. 13. Meshes of human head[14]

Fig 13, illustrates 2D volume conductor digitized with triangles and Fig.14 states the mesh in 2D coronal slice view only. In summary, forward problem for brain source localization covers all the algorithms and theory related to scalp, conductive medium and measuring potential differences techniques.

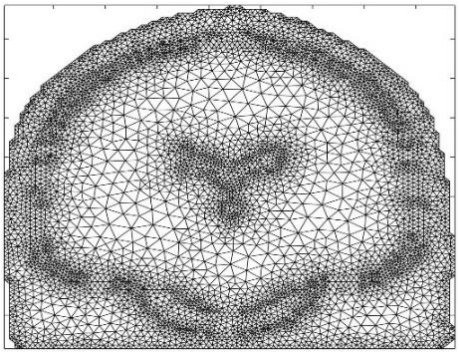


Fig. 14. Mesh in 2D coronal slice[12, 15]

### B. *Inverse Problem*

The inverse problem is an ill-posed and non-unique problem due to a reason that there are infinite dipoles inside the cortical conductor and limited number of the electrodes over scalp for fitting the data causing the uncertainties. Generally, true model in inversion problem consists of estimation problem and appraisal problem [15, 30] (see Fig 15).

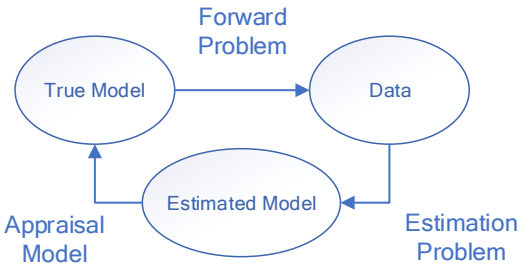



Fig. 15. Inversion problem of physical model [46]

Two methodologies such as equivalent current dipole (ECD) for limited areas for underlying neural activities and linear distributed for whole volume are assumed for current source densities. [31, 32]. Several algorithms have been developed with a minimum of localization error and high resolution and less computational time. The hierarchy of the inverse solutions and their history are shown in Fig.16 and Fig. 17. respectively.

The minimum norm estimation (MNE) estimates sources with 1 cm resolution [15]. It yields predicted information and provides better localization with increasing the number of magnetometers. It gives good resolution and current estimation. It provides improved localization error.

The low-resolution electromagnetic tomography (LORETA) estimates all active, neighboring and boundary areas. It not only yields good time resolution but also provides the poor spatial resolution of the blurring images. Further, it yields less errors and used for depth sources. LORETA and its family use the current source density for reconstruction.

The focal underdetermined system solution (FOCUSS) is a tomographic reconstruction scheme. It is high resolution non-parametric technique. It has a good spatial resolution. FOCUSS is used for deep source localizations. Recursive Multiple Signal Classification (Recursive MUSIC) [33] is a modified form of a RAP MUSIC. It yields less error of localization and is easy for processing of computation. It can face the situations where there are constraints or limits.

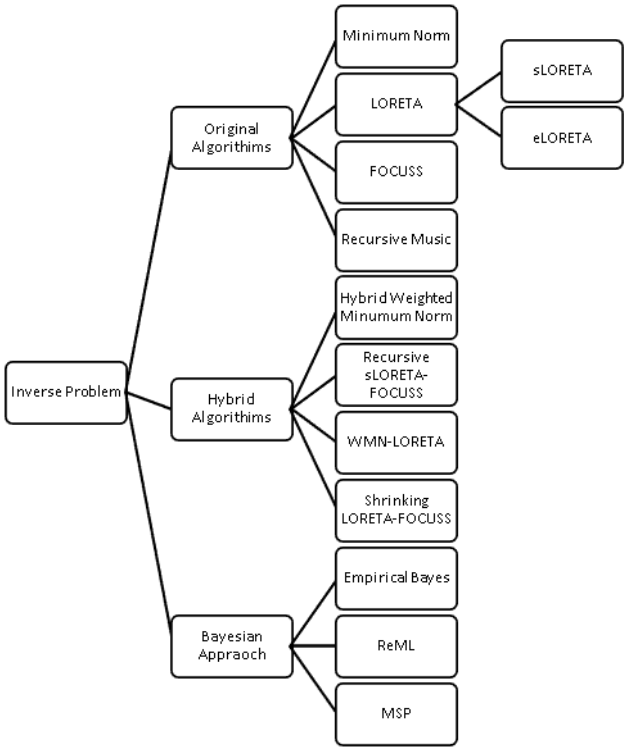



Fig. 16. Solutions of inverse problem [15]

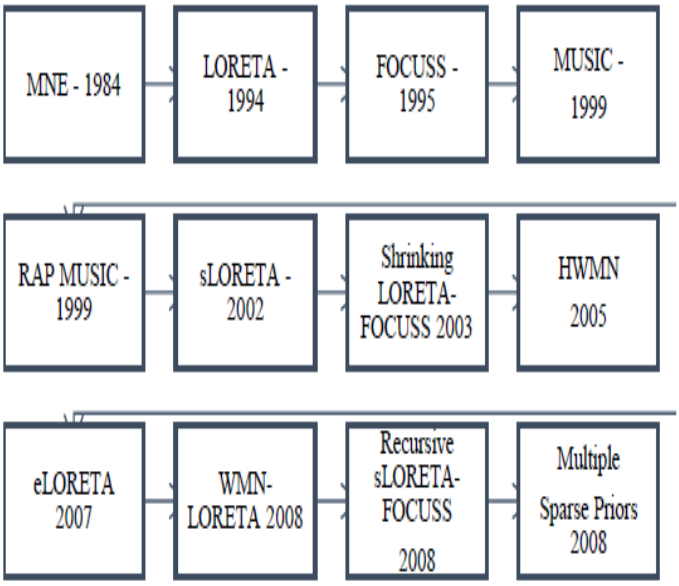



Fig. 17. History of solutions of the inverse problem [15]

The hybrid weighted minimum norm (HWMN) [15] is the modified form of LORETA [34, 39], FOCUSS [35] and WMN [15, 36], It yields the smoothness with a long processing time and therefore losses data while computing. The standard LORETA (sLORETA) [40] yields uniform variance across the whole brain volume and yields zero localization error. The results are verified on experimenting on 6430 voxels with having 5 mm spatial resolution. It has an exact localization with a zero error. Exact LORETA (eLORETA) [25] yields the variance of unity. All its simulations are based on LORETA software where zero-bias creates zero-localization error. Its weight matrix yields correct value with minimized error (12 to 7 mm). It has also the access of the touch to deep sources. There are some comparative studies for sLORETA and eLORETA[37]. WMN-LORETA experiments are carried out through by both simulation and experimented basis with 138 electrodes. Firstly, resolution matrix is designed. In these experiments such as conditions of identity matrix are achieved. However, error is subsequently reduced and accuracy is enhanced with less execution time. In Recursive sLORETA-FOCUSS, sLORETA gives smoothness whereas FOCUSS gives sparse solution and increase localization error with 494.03 mm when experiments are conducted. Its processing time is reduced to 330.45. Shrinking LORETA-FOCUSS a derived from LORETA and FOCUSS. It has the complexity of the weight matrix and iteration performance due to which it takes a time for processing. It provides good smoothness with low localization error. The sources construction with high spatial resolution is achieved. It uses LORETA and current densities for smoothness thereby subsequently weight matrix is computed. Summary of the inverse methods and their comparison are given in Table II and Table III.

TABLE II. SUMMARY OF INVERSE METHODS[10, 15]

| Method | Characteristics/Comparison |
|---|---|
| MNE[15, 38] | • It has good resolution and current estimation and yields good localization error as compared to LORETA, WMN etc.<br>• It fails for deep source localization and is incapable of localizing non-boundary sources. |
| LORETA [39] | • It is popular due to source localization and citations.<br>• However it is incapable of improving low spatial localization. |
| sLORETA [40] | • It yields poor stability and blurriness in the mage.<br>• It is a derived from LORETA and yields better localization.<br>• It has poor resolution and poor performance for recovering multiple sources |
| eLORETA[25] | • It is derived method of LORETA and gives authentic results in localization and handles the issue of variance very well.<br>• Its low resolution and blurring in images increases. |
| FOCUSS[35] | • Its possibility of sources localizing is on edge and deep areas.<br>• It yields spatial variations in good quantity.<br>• It is crucial to bring out the features of the brain activity due to its low resolutions and blurredness. |
| HWMN[15] | • It is a derived from MN and LORETA and provides better estimation.<br>• It uses LORETA and WMN for better estimation.<br>• It takes more iterations for more accurate, deep localization and less errors.<br>• It takes a large processing time.and these is a possibility of the loss of data due to continuous iterations and weight matrix involvement. |
| shrinking LORETA-FOCUSS[41] | • It is derived from LORETA and FOCUSS.<br>• Its results are not experimentally validated.<br>• It provides better minimized localization error. |
| MUSIC[33] | • It yields less localization error.<br>• However, it takes a large processing time but there is a chance of the loss of the data. |
| RAP-MUSIC [42] | • It yields good localization error but it takes a large processing time.<br>• There are chances of the loss of data.<br>• However, its random error and noise problem increase the issues of its processing. |
| WMN-LORETA [15] | • It yields better resolution if compared with LORETA and WMN separately. |
| Recursive sLORETA-FOCUSS [15] | • It is an efficient in terms of processing and less localization error. |

TABLE III. COMPARISONS OF INVERSE METHODS[15]

| Method | Resolution | Time | Validation |
|---|---|---|---|
| MNE | low | high | experiment |
| LORETA | low | high | experiment |
| FOCUSS | low | low | experiment |
| Recursive MUSIC | good | low | simulation |
| sLORETA | low | high | experiment |
| Shrinking LORETA-FOCUSS | low | low | simulation |
| Hybrid Weighted MN | low | low | simulation |
| eLORETA | low | high | experiment |
| WMN-LORETA | low | low | simulation |
| RecursivesLORETA-FOCUSS | low | high | simulation |

LORETA yields good localization error and estimated current density whereas the Shrinking LORETA- FOCUSS yields energy error, localization and maximum energy error. sLORETA- FOCUSS provides excellent localization error in simulations whereas recursive sLORETA-FOCUSS provides better processing time, accuracy and more accurate results. WMN-LORETA is a best method in terms of time processing and resolution.

This was the summary of the approaches and their comparative analysis used for ill-posed inverse problem.

### *C. Bayesian Approaches*

This approach has a vital role in source localization of the brain. Inversion scheme primarily consists of Bayesian approach. Much efforts are made to put the theoretical and mathematical approaches into empirical and experimental practices to locate sources of the underneath neural activities[43]. The optimized estimations are based on posterior and priori procedures called the Bayesian approach. EEG source localization depends on the estimation of priors or patches selected from parametric Bayes (PEB) so called empirical Bayes[44]. PEBs are simple parametric hierarchical linear models and parametric assumptions involve the randomly addition of Gaussian noise at each level[45]. The source space basically generates the data. The background of using this function is to limit on log-evidence or Likelihood by using Gaussian process priors on different models so called priors or covariance components of given measurements and best model automatically is selected from the hyperparameters by enforcing the conditional variances to minimum value [46]. The quality of using free energy and restricted maximum likelihood (ReML) are to get the grip on covariance components and hyperparameters [45]. These methods form an inversion scheme which helps in selection of the optimized priors either from sparse priors or distributed priors provided the nature of the data obtained. There are the definitely benefits for using inversion scheme such as it is fast and consists of the linear models for both said priors [47]. The empirical Bayes provide the best quantifying and relative information of spatial priors and accommodate multiple priors for the provision of more accurate source of the reconstruction of neural activity [48]. In empirical Bayes, different combinations of the priors so called sets of the models, are selected through the method of the model selection. The best utilization of ReML as log-likelihood is to bring improperness in using parameters of the hyperparameters [49]. Selected model is defined in terms of the covariance components.

Inversion scheme [50] is applied to the evoked highlighted neural activity by applying free energy bound or log-evidence of marginal likelihood or ReML objective function. The benefit of using ReML is that optimization can be carried out by selecting a specific model of covariance components from measurement of data irrespective of increasing the size and number of the resources. Empirical priors are the source estimated priors which can be observed during the desired pattern of the brain neural activity observed through EEG responses[51]. Its mathematical form is described as

$Q = \sum_{i=1}^{N_q} h_i C_i$ where priors are basically the components, partitions or the models which have large variances so called hyperparameters which are further spitted-up into a small group of the patterns called covariance components [52]. The evidences of the highly active regions of the neural activity with the local and compact support can be modelled as source priors. In Bayesian approach, the prior source covariance which are counted as weighted sum of multiple prior components are called as empirical Bayes as available in the literature as priors for source reconstruction[53]. The highly optimized combination of the compact and desired higher neural activity priors of the desired regions can be evaluated through evidences such as using Bayesian model approach. Hyperparameters have large variances. The variety of EEG responses can be observed though EEG responses. Such extreme situations can be addressed through minimum norm constraint and ARD solutions. The Bayesian model approach provides the log-evidences and Greedy search[54, 55] is iterated on evident provided patters or components for efficiently splitting the highest hyperparameters of the highest variances into two or more new components or partitions so called models[45]. Such iteration is repeated until the evidences are stopped for increasing and optimized results are obtained[56]. In a nutshell, both the greedy search and ARD schemes are used for optimizing the neural activity of EEG source localization.

In parametric empirical Bayes (PEB) [57], several models of the source neural activities are estimated. These are experimentally or empirically selected from the different regions of the brain. The empirical priors on the sources are given by [58, 59]: There is a variety of picking up the priori models of neural activity. The closed and compact neural activity priors are particularly paid attention These are also known as model correlated sources[60]. Any groups of priors can be enhanced or compared with one another's evidences so called Bayesian model comparison.

A greedy search (GS) that splits highest variance or hyperparameter into more than two components and subsequently optimized parameters. This action is iterated until its evidence stops increasing [49]. In Automatic relevance determination (ARD) one can start with one component per pattern and use ARD to eliminate other unused and unnecessary patterns. Inversion Scheme [50] is comprised of MN, IID, LOR and other MSPs. Specific designed M weight matrix which is also called mixture matrix as covariance components are used for optimization of estimating the reconstruction of the neural activity with the current density. In this scheme at sensors levels and noise levels optimized techniques are used for maximizing the current density [61]. Minimum norm (MN) Model is represented by $Q^{\varepsilon} = I$. This model asserts that all sources are active, with equal a priori probability and that none are correlated [50, 57]. Coherence (COH) model is a realistic model with two components modelling independent and coherent sources respectively. Multiple sparse prior model is represented by components such as $Q^{\varepsilon} = \left\{ q_1 q_1^T, ..., q_N q_N^T \right\}$ modeling activity in N patterns[57, 62].

## III. Conclusion and Remarks

In this article, several approaches of forward problem, inverse problem and Bayesian framework have been explored and discussed for EEG brain source localization in terms of their computational time, localization error, energy error, resolution etc. An EEG brain source localization not only yields the complete solution of understanding brain abnormalities, mental diseases and cognitive brain actions but it also provides the best spatiotemporal solution of diagnosing the disorders of the brain such as epilepsy and tumors etc. The forward yields only potential measurements on the scalp which are calculated through the electrodes and efforts are made to remove the noises from the signals due to uncertainties of the three layers or mediums, coupling of the cortical surfaces and environmental noises or stimuli and electrodes used for experimental purposes. EEG minimum norm and listed LORETA family algorithms used for ill-posed inverse problem subsequently yields the good results in terms of resolution localization error and computational time of the active brain sources. Bayesian framework involves the optimized selection of the patches of the surface or volume of the cortical surface or head. Multiple sparse priors (MSP) algorithm-based Bayesian algorithm is targeted for optimization of the source localization. ReML, Greedy search, ARD and Coherence approaches are used as optimization techniques in terms of giving out the good results for the maximum free energy. The uncertainties such as unnecessary active regions or patches, unnecessary equivalent current dipoles or unnecessary coupling of neighbor regions of the brain are surely removed to reach out at the exact active brain region for giving a maximum energy as discussed in Bayesian framework. However, there still more work is required on solving the forward problem and ill-posed inverse problems. There is a need of using exact analytical or numeral head models required. Conductivities and noise issues are yet to be improved by using the new techniques or algorithms of computational methods. The new integrated neuroimaging techniques with robust algorithms are yet to be worked out for giving better results in terms of spatiotemporal resolution, localization error, less processing time or even free energy. The multiple source prior technique seems to one of most advanced techniques which is capable for finding the most optimized source localization by removing unnecessary patches or their repetition.

This was an entire overview of EEG signals-based source localization. Future study will deeply be extended to cover their latest research and mathematical background which will surely will cover all the issues in optimizing the results of the brain source localization using EEG signals.

## References

[1] Q. Tran, T. Le, C. Le, and T. V. Van, "Electrophysiological modeling in generalized epilepsy using surface EEG and anatomical brain structures," in International Conference on the Development of Biomedical Engineering in Vietnam, 2017, pp. 699-704: Springer.

[2] A. Majkowski, Ł. Oskwarek, M. Kołodziej, and R. J. Rak, "An attempt to localize brain electrical activity sources using EEG with limited

number of electrodes," Biocybernetics and Biomedical Engineering, vol. 36, no. 4, pp. 686-696, 2016.

[3] R. Cook, G. Bird, C. Catmur, C. Press, and C. Heyes, "Mirror neurons: from origin to function," Behavioral and Brain Sciences, vol. 37, no. 2, pp. 177-192, 2014.

[4] M. A. Lancaster et al., "Cerebral organoids model human brain development and microcephaly," Nature, vol. 501, no. 7467, p. 373, 2013.

[5] M. Susta, H. Papezova, S. Petranek, and K. Sonka, "Brain activation sequences," Neuroendocrinology Letters, vol. 36, no. 8, 2015.

[6] P. Milz, K. Oberauer, P. D. D. C. Kiper, and D. Lehmann, "Brain Electric Mechanisms of Modalities of Thinking," Universität Zürich, 2016.

[7] M. J. Hawrylycz et al., "An anatomically comprehensive atlas of the adult human brain transcriptome," Nature, vol. 489, no. 7416, p. 391, 2012.

[8] L. G. Kiloh, A. J. McComas, and J. W. Osselton, Clinical electroencephalography. Butterworth-Heinemann, 2013.

[9] J. K. Mai, M. Majtanik, and G. Paxinos, Atlas of the human brain. Academic Press, 2015.

[10] W. Khalifa, A. Salem, M. Roushdy, and K. Revett, "A survey of EEG based user authentication schemes," in Informatics and Systems (INFOS), 2012 8th International Conference on, 2012, pp. BIO-55-BIO-60: IEEE.

[11] A. Chaturvedi, "Analysing the Use of EEG, fMRI, PET and Behavioural Studies of Brain Lesioned Patients Associated with Non-Literal Language eg Metaphors, Sarcasm, Some Types of Humour Etc," 2017.

[12] Hallez, Hans, et al. "Review on solving the forward problem in EEG source analysis." Journal of neuroengineering and rehabilitation 4.1 (2007): 46.

[13] R. Plonsey and D. B. Heppner, "Considerations of quasi-stationarity in electrophysiological systems," The Bulletin of mathematical biophysics, vol. 29, no. 4, pp. 657-664, 1967.

[14] M. R. Lakshmi, D. T. Prasad, and D. V. C. Prakash, "Survey on EEG signal processing methods," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 1, 2014.

[15] M. A. Jatoi, N. Kamel, A. S. Malik, I. Faye, and T. Begum, "A survey of methods used for source localization using EEG signals," Biomedical Signal Processing and Control, vol. 11, pp. 42-52, 2014.

[16] C. H. Wolters, "Influence of tissue conductivity inhomogeneity and anisotropy on EEG/MEG based source localization in the human brain," Max Planck Institute of Cognitive Neuroscience Leipzig, 2003.

[17] C. H. Wolters, A. Anwander, X. Tricoche, D. Weinstein, M. A. Koch, and R. MacLeod, "Influence of tissue conductivity anisotropy on EEG/MEG field and return current computation in a realistic head model: a simulation and visualization study using high-resolution finite element modeling," NeuroImage, vol. 30, no. 3, pp. 813-826, 2006.

[18] V. E. Montes Restrepo, "Determination of anisotropic ratio of the skull for EEG source localization in patients with epilepsy," in 11e FirW PhD Symposium, 2010, pp. 62-62: Universiteit Gent. Faculteit Ingenieurswetenschappen.

[19] M. A. Jatoi et al., "EEG-based brain source localization using visual stimuli," vol. 26, no. 1, pp. 55-64, 2016.

[20] P. Berg and M. Scherg, "A fast method for forward computation of multiple-shell spherical head models," Electroencephalography and clinical neurophysiology, vol. 90, no. 1, pp. 58-64, 1994.

[21] B. Cuffin, "Effects of local variations in skull and scalp thickness on EEG's and MEG's," IEEE Transactions on Biomedical Engineering, vol. 40, no. 1, pp. 42-48, 1993.

[22] N. Chauveau, X. Franceries, B. Doyon, B. Rigaud, J. P. Morucci, and P. Celsis, "Effects of skull thickness, anisotropy, and inhomogeneity on forward EEG/ERP computations using a spherical three-dimensional resistor mesh model," Human brain mapping, vol. 21, no. 2, pp. 86-97, 2004.

[23] M. A. Jatoi, N. Kamel, A. S. Malik, I. Faye, and T. Begum, "Representing EEG source localization using finite element method," in Control System, Computing and Engineering (ICCSCE), 2013 IEEE International Conference on, 2013, pp. 168-172: IEEE.

[24] T. Nyoni and W. G. Bonga, "Neuromarketing Methodologies: More Brain Scans or Brain Scams?," 2017.

[25] R. D. Pascual-Marqui, "Discrete, 3D distributed, linear imaging methods of electric neuronal activity. Part 1: exact, zero error localization," arXiv preprint arXiv:0710.3341, 2007.

[26] J. J. Ermer, J. C. Mosher, S. Baillet, and R. M. Leahy, "Rapidly recomputable EEG forward models for realistic head shapes," Physics in Medicine & Biology, vol. 46, no. 4, p. 1265, 2001.

[27] M. A. Jatoi, N. Kamel, I. Faye, A. S. Malik, J. M. Bornot, and T. Begum, "BEM based solution of forward problem for brain source estimation," in Signal and Image Processing Applications (ICSIPA), 2015 IEEE International Conference on, 2015, pp. 180-185: IEEE.

[28] F. Costa, H. Batatia, T. Oberlin, and J.-Y. Tourneret, "Skull Conductivity Estimation for EEG Source Localization," IEEE Signal Processing Letters, vol. 24, no. 4, pp. 422-426, 2017.

[29] B. Vanrumste, G. Van Hoey, R. Van de Walle, R. D. Michel, I. A. Lemahieu, and P. A. Boon, "Comparison of performance of spherical and realistic head models in dipole localization from noisy EEG," Medical Engineering and Physics, vol. 24, no. 6, pp. 403-418, 2002.

[30] R. Snieder and J. Trampert, "Inverse problems in geophysics," in Wavefield inversion: Springer, 1999, pp. 119-190.

[31] R. D. Pascual-Marqui, "Theory of the EEG inverse problem," Quantitative EEG analysis: methods and clinical applications, pp. 121-140, 2009.

[32] L. Duque-Muñoz, J. Martinez-Vargas, G. Castellanos-Dominguez, J. Vargas-Bonilla, and J. López, "Non-linear Covariance Estimation for Reconstructing Neural Activity with MEG/EEG Data," in International Work-Conference on the Interplay Between Natural and Artificial Computation, 2017, pp. 334-344: Springer.

[33] J. C. Mosher and R. M. Leahy, "Recursive MUSIC: a framework for EEG and MEG source localization," IEEE Transactions on Biomedical Engineering, vol. 45, no. 11, pp. 1342-1354, 1998.

[34] S. Ikeda, R. Ishii, L. Canuet, and R. D. Pascual-Marqui, "Source estimation of epileptic activity using eLORETA kurtosis analysis," BMJ case reports, vol. 2017, pp. bcr-2017-222123, 2017.

[35] I. F. Gorodnitsky, J. S. George, and B. D. Rao, "Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm," Electroencephalography and clinical Neurophysiology, vol. 95, no. 4, pp. 231-251, 1995.

[36] R. Khemakkhem, W. Zouch, A. Taleb-Ahmed, and A. B. Hamida, "A new combining approach to localizing the EEG activity in the brain: WMN and LORETA solution," in BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on, 2008, vol. 1, pp. 821-824: IEEE.

[37] M. A. Jatoi, N. Kamel, A. S. Malik, I. J. A. p. Faye, and e. s. i. medicine, "EEG based brain source localization comparison of sLORETA and eLORETA," vol. 37, no. 4, pp. 713-721, 2014.

[38] M. S. Hamalainen, "Interpreting measured magnetic fields of the brain: estimates of current distributions," Helsinki Univ. of Technol., Rep, 1984.

[39] R. D. Pascual-Marqui, C. M. Michel, and D. Lehmann, "Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain," International Journal of psychophysiology, vol. 18, no. 1, pp. 49-65, 1994.

[40] R. D. Pascual-Marqui, "Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details," Methods Find Exp Clin Pharmacol, vol. 24, no. Suppl D, pp. 5-12, 2002.

[41] H. S. Liu, F. Yang, X. Gao, and S. Gao, "Shrinking LORETA-FOCUSS: a recursive approach to estimating high spatial resolution electrical activity in the brain," in Neural Engineering, 2003. Conference Proceedings. First International IEEE EMBS Conference on, 2003, pp. 545-548: IEEE.

[42] J. C. Mosher and R. M. Leahy, "Source localization using recursively applied and projected (RAP) MUSIC," IEEE Transactions on signal processing, vol. 47, no. 2, pp. 332-340, 1999.

[43] A.-S. Hincapié et al., "The impact of MEG source reconstruction method on source-space connectivity estimation: a comparison between

minimum-norm solution and beamforming," NeuroImage, vol. 156, pp. 29-42, 2017.

[44] Y. Zhang, Y. Wang, J. Jin, and X. Wang, "Sparse Bayesian learning for obtaining sparsity of EEG frequency bands based feature vectors in motor imagery classification," International journal of neural systems, vol. 27, no. 02, p. 1650032, 2017.

[45] M. Dashti and A. M. Stuart, "The Bayesian approach to inverse problems," Handbook of Uncertainty Quantification, pp. 311-428, 2017.

[46] F. Costa, H. Batatia, T. Oberlin, C. D'giano, and J.-Y. Tourneret, "Bayesian EEG source localization using a structured sparsity prior," NeuroImage, vol. 144, pp. 142-152, 2017.

[47] Y. Bekhti, F. Lucka, J. Salmon, and A. Gramfort, "A hierarchical Bayesian perspective on majorization-minimization for non-convex sparse regression: application to M/EEG source imaging," arXiv preprint arXiv:1710.08747, 2017.

[48] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, Statistical parametric mapping: the analysis of functional brain images. Elsevier, 2011.

[49] J. López, V. Litvak, J. Espinosa, K. Friston, and G. R. Barnes, "Algorithmic procedures for Bayesian MEG/EEG source reconstruction in SPM," NeuroImage, vol. 84, pp. 476-487, 2014.

[50] K. Friston et al., "Multiple sparse priors for the M/EEG inverse problem," NeuroImage, vol. 39, no. 3, pp. 1104-1120, 2008.

[51] D. Wipf and S. Nagarajan, "A unified Bayesian framework for MEG/EEG source imaging," NeuroImage, vol. 44, no. 3, pp. 947-966, 2009.

[52] S. Hu, D. Yao, and P. A. Valdes-Sosa, "Unified Bayesian estimator of EEG reference at infinity: rREST," arXiv preprint arXiv:1802.02268, 2018.

[53] K. Sekihara and S. S. Nagarajan, "A unified Bayesian framework for MEG/EEG source imaging," in Electromagnetic Brain Imaging: Springer, 2015, pp. 119-137.

[54] F. Goksu, N. F. Ince, and A. H. Tewfik, "Sparse common spatial patterns in brain computer interface applications," in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, 2011, pp. 533-536: IEEE.

[55] D. Meunier, S. Achard, A. Morcom, and E. Bullmore, "Age-related changes in modular organization of human brain functional networks," Neuroimage, vol. 44, no. 3, pp. 715-723, 2009.

[56] K. Friston et al., "Bayesian decoding of brain images," Neuroimage, vol. 39, no. 1, pp. 181-205, 2008.

[57] J. Lopez, G. R. Barnes, and J. Espinosa Oviedo, Single MEG/EEG source reconstruction with multiple sparse priors and variable patches. 2012, pp. 136-144.

[58] M.-a. Sato et al., "Hierarchical Bayesian estimation for MEG inverse problem," NeuroImage, vol. 23, no. 3, pp. 806-826, 2004.

[59] Y. Yousra, F. Lucka, J. Salmon, and A. Gramfort, "A hierarchical Bayesian perspective on majorization-minimization for non-convex sparse regression: Application to M/EEG source imaging," arXiv. org e-Print archive, 2017.

[60] S. Villena, M. Vega, S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Bayesian combination of sparse and non-sparse priors in image super resolution," Digital Signal Processing, vol. 23, no. 2, pp. 530-541, 2013.

[61] J. López, F. Valencia, G. Flandin, W. Penny, and G. Barnes, "Reconstructing anatomy from electro-physiological data," NeuroImage, vol. 163, pp. 480-486, 2017.

[62] J. C. Ye, J. M. Kim, and Y. Bresler, "Subspace penalized sparse learning for joint sparse recovery," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, 2013, pp. 6039-6042: IEEE.

# An Efficient Protocol using Fuzzy Logic and Grids with Two-Dimensional Techniques for Saving Energy in WSN

Emad M Ibbini[†*1], Dr.Kweh Yeah Lun[2], Prof. Mohamed Othman[3], Dr. Zurina Mohd Hanapi[4]
[1]Department of Computer Science, University Putra Malaysia (UPM) Serdang, 43400 Malaysia

Amir Abbas Baradaran[5]
Payam Noor University,Tehran,
Department of Computer Engineering and Information Technology
IRAN

***Abstract*—This work proposes an energy-saving protocol for wireless sensor networks (WSNs) using fuzzy logic and grids with two-dimensional techniques, namely, gravity and energy centers, to address the pressing issue of energy efficiency in WSNs. The optimal cluster head is chosen in two stages of the proposed protocol to prolong the network lifetime and reduce the energy consumption. The proposed protocol evaluated the cluster-head radius according to the residual energy and distance to the base station(BS) parameters of the sensor nodes. The proposed scheme shows better improvements than other related protocols as it extends the lifetime of Two Dimensional Technique Based On Center of Gravity and Energy Center (TDTCGE) protocol by 54¥% and saves more energy. Fuzzy inference engine (Mamdani's rule) is used to elect the chance to be the best node. The results have been derived from matlab simulator which shows that the proposed protocol performs better than the TDTCGE protocol. Simulation results show also that our protocol offers a much better network lifetime and energy efficiency than other existing protocols.**

***Keywords*—Fuzzy logic; fuzzy inference engine; first node die; last node die; energy efficiency; lifetime***

## I. INTRODUCTION

Wireless sensor networks(WSNs) are generally utilized in diverse fields, which incorporate catastrophe aid projects and programs for agribusiness and social insurance [1]. In any case, its most predominant application is for getting data for savvy gadgets that make up inescapable sensor networks that are generally embraced in structures, homes, and transport and modern frameworks. A WSN is made of individual nodes/sensors that respond to physical parameters in its condition. Be that as it may, these nodes here and there rely upon batteries rather on a wired power supply [2]. Given the by and large long activity time in WSNs, such reliance represents an issue, especially as far as the energy productivity of correspondence protocols. In any case, such case isn't constantly watched, as different applications may organize the precision of results over is-sues in power supply [3].A WSN comprises of various arbitrarily or consistently circulated nodes in a sensor field. These nodes are utilized to screen physical conditions, for example, weight and weight on objects, since they can communicate with their condition by detecting outer physical parameters. Nodes and cluster heads (CHs) use Fuzzy rationale, which is a less perplexing legitimate framework, based on Fuzzy factors and if-then rules[6].The proposed protocol will be contrasted and the accompanying past protocols: TDTCGE [7] This protocol utilizes two dimensional techniques (Computes the Center of Gravity for every Grid and Computes the energy Center) to choose the ideal node as a Cluster head by which node is the closest to one of these Centers. The TDTCGE protocol takes care of the issue related with separation and how much more distant the cluster head from the Base station. In any case, it doesn't address the issue of sit without moving tuning in. The consequences of this protocol endorsed that this protocol enhanced the existence time and the energy utilization .In CRCWSN [8] this protocol utilizes two unique techniques for choosing cluster head (CH) that has been at first utilized by hereditary calculation and re-clustering strategy. This protocol thinks about separation and energy parameters. Contrasted with a few other proposed protocols, this protocol made a tar-get work that is more enhanced. It has a mix of chromosomes and timing of age rehash has been finished utilizing another strategy. The CRCWSN is productive and have more life time with diminishing age rehashes contrasted with past comparable techniques. Suitable cluster head (CH) race is one such issue, which can lessen the energy utilization dramatically[11]. In this paper we will utilize Fuzzy rationale clustering techniques in WSNs utilize Fuzzy rationale measure's for blending diverse clustering parameters to pick ideal cluster head to choose Cluster Heads[10].constrained energy is a conspicuous element for wireless sensor networks. Since the radio handset ordinarily expends a greater number of energies than some other equipment part on a sensor node, it is of awesome significance to plan energy advanced steering calculation to draw out network lifetime[12].The rest of this article is composed as take after. In Section 2, we present the Network model of wireless sensor networks. In Section 3, we

†The authors are with University Putra Malaysia (UPM) ,Malaysia.
* Corresponding author

present the star presented protocol in detail. In Section 4, we present our reenactment show and investigate the relative assessment after effects of the proposed plot through reproductions. At long last, a few conclusions are given..

## II. Literature Review

The proposed protocol is contrasted and the TDTCGE [5] protocol.

•Two-Dimensional Technique in light of Center of Gravity and Energy (TDTCGE) [5]

This protocol utilizes two-dimensional strategies. The focuses of gravity and energy for every framework are processed. The ideal node is chosen to be the cluster head (CH) in light of the fact that this node is the closest to one of the focuses. The TDTCGE protocol tends to the separation issue, especially the separation of the CH from the BS. In any case, the issue of sit out of gear listening is ignored. The aftereffects of this protocol demonstrate that both the lifetime and energy utilization are upgraded.

Fuzzy CLUSTERING ALGOROTHMS(FCA)[13] Some of the clustering calculations utilize Fuzzy rationale to deal with vulnerabilities in the WSNs. Fundamentally; FCAs utilize Fuzzy rationale for mixing distinctive clustering parameters to choose cluster-heads. They appoint opportunities to speculative cluster-heads as indicated by the defuzzified yield of Fuzzy if-then guidelines. The provisional cluster-head turns into a cluster-head in the event that it has the best shot in its region.

Energy aware distributed dynamic clustering protocol

Using fuzzy logic (ECPF) [9] is another fuzzy approach based algorithm for WSN. ECPF has two stages: setup and steady state stage. CH race and cluster development will occur. In setup stage. TDMA outline age and information gathering Occur amid the steady state stage. Node degree and node centrality are the two information variables and fuzzy expense is the yield variable. Every node in the system will sit tight for a postpone time which is contrarily relative to its lingering energy. After lapse of defer time, if a node does not get any provisional CH message from its adjacent nodes, at that point it communicates a provisional CH message containing its id, fuzzy cost, and its status as provisional CH. if a node does not receive any tentative CH message from its nearby nodes, then it broadcasts a tentative CH message containing its id, fuzzy cost, and its status as tentative CH. If there is no other node within its cluster range with lesser fuzzy cost value, it declares itself as the CH and broadcast final CH message within its cluster range.

This protocol enhances LEACH protocol utilizing Fuzzy Logic (LEACH-FL)[14], which takes battery level, separation and node density into thought. The proposed technique has been demonstrated improving a determination by examination reenactments utilizing Matlab contrasted and LEACH.

Fuzzy Logic Based Energy Efficient Clustering Hierarchy for No uniform Wireless Sensor Networks (FLECH) [15]This Protocol contributes a novel clustering calculation: Fuzzy Logic Based Energy Efficient Clustering Hierarchy (FLECH) for non-uniform WSN. The clusters in FLECH are made utilizing appropriate parameters which builds the lifetime of the WSN. Fuzzy rationale in FLECH is astutely used to join essential parameters like residual vitality, node centrality, and separation to BS for choosing best appropriate nodes as CH and builds the network lifetime. The reproduction results unmistakably show the lifetime increment by FLECH over different calculations and its energy preservation per round of information gathering in the network.

## III. Wireless Sensor Network Models

### A. Network Model

For this examination, we arbitrarily convey N sensor nodes in an observed zone and accept that the sensor network has the accompanying qualities:

*1) The situation of the BS in the sensor network is settled.*

*2) All nodes are heterogeneous and stationary and have diverse beginning supplies of vitality.*

*3) All the nodes are arbitrarily conveyed in the objective zone, and each can build up an association with the sink.*

### B. Energy Consumption:

Drain [4] incorporates a first-arrange radio model that can be used for computing equipment energy dissemination. For relative purposes, this paper utilizes a similar model. In this model, the energy consumptions of radios for sending and accepting information are both communicated as $E_{elect}$; the free space and the multi-path blurring channel models with particular intensifying lists ε fs and ε mp are utilized; the energy utilization of information combination is indicated by $E_{DA}$. The energy spent by a node that transmits a l-bit packet over separation d is figured utilizing the Heinzelman display. This model expresses that for every node to transmit L bits of information a separation d from itself, $E_t$ energy is expended:

*1) The energy required to get L bits of information squares with*

*2) The parameters are characterized as takes after:*

d0: hybrid separation

∈elect: energyimportant for actuating electronic circuits

∈mp, ∈fs: affectability and clamor in the collector, individually.

## IV. Limitation of the TDTCGE that FLG Proposed Protocol will Solve

TheFLG-proposed protocol will be compared with the following protocol: as TDTCGE. This protocol uses two-dimensional techniques (Computes the Center of Gravity for each Grid and Computes the energy Center) to select the optimal node as a Cluster head by which node is the nearest to one of these Centers. The TDTCGE protocol didn't solve the problem of the radius distance for each node but it minimize the distance only between the cluster head and the Centers which it's not enough. However, it is of great importance to design optimized routing algorithm to minimize the radius competition and prolong the lifetime of the FND (First node die).

## V. PROPOSED PROTOCOL

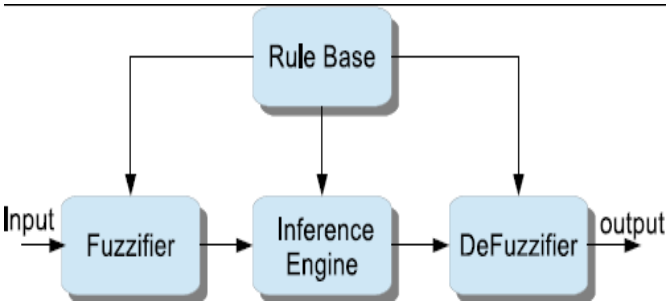



Fig. 1. Fuzzy logic process

The following four steps of fuzzy logic are used in the fuzzy inference system to compute chance values as shown in figure 1.

*1) In fuzzification, crisp input data are translated into the fuzzy set value required by the inference engine.*

*2) In rule evaluation, a set of fuzzy rules is defined to characterize the dynamic behavior of the system.*

*3) In aggregation, the inference system draws conclusions from the fuzzy rules and sends the output to the defuzzification unit.*

*4) In defuzzification, fuzzy sets are mapped into a space of crisp sets.*

In our study, we used fuzzy logic to elect CHs, which are selected from a nine-grid BS. In each round, the BS transmits its clustering node information to the member nodes in the nine grids. During CH election, the BS determines three fuzzy sets, namely, the distance between nodes, the distance between nodes and the BS, and the residual energy in each node.

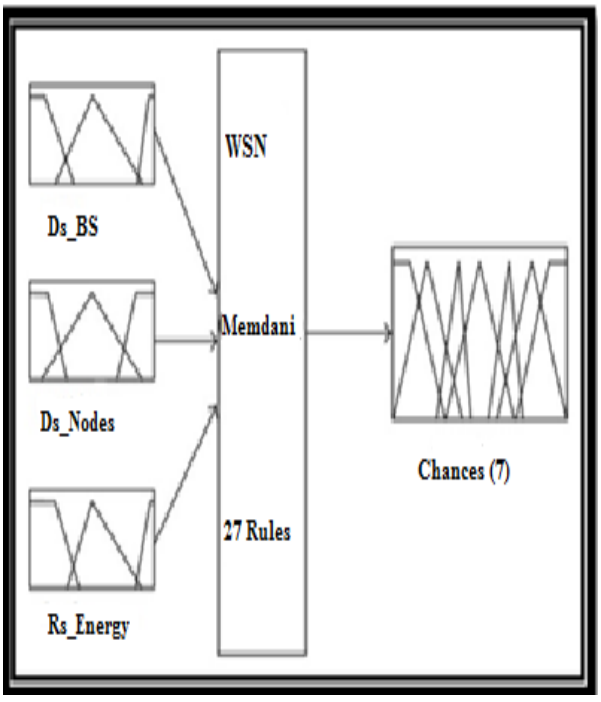



Fig. 2. Model of fuzzy system

Given that fuzzy logic is used, the BS elects CHs according to fuzzy rules. The BS, in which all clustering decisions are derived, then broadcasts the results to the entire network. Using the three criteria, the BS obtains highly precise CHs because its strength is greater than that of normal nodes and it has knowledge of the entire network.

Three criteria in fuzzy CH election (distance between nodes, distance between nodes and the BS, and residual energy in each node) as shown in figure 2, 3,4 and 5.

In the calculations of the radius between nodes, the FLG-proposed protocol utilizes residual energy, along with the parameter that measures the distance between nodes and BSs and the distance between two nodes. The residual energy of the CH may be reduced by minimizing the service area. This reduction in residual energy then prompts a change in competition radius; otherwise, the

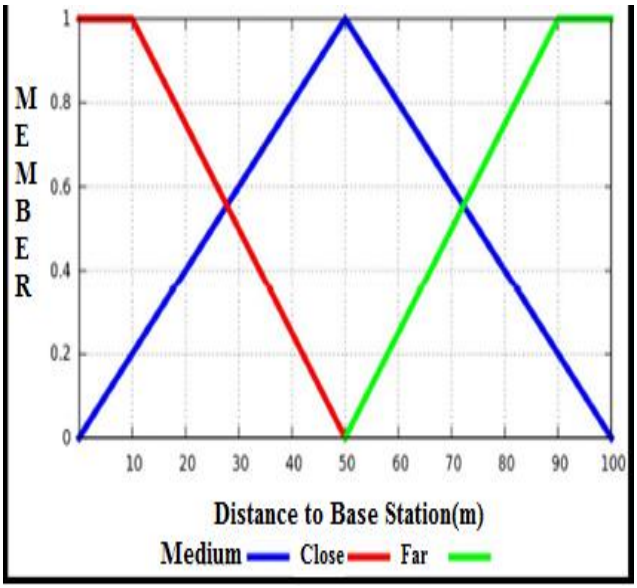



Fig. 3. Fuzzy input variable of distance between nodes

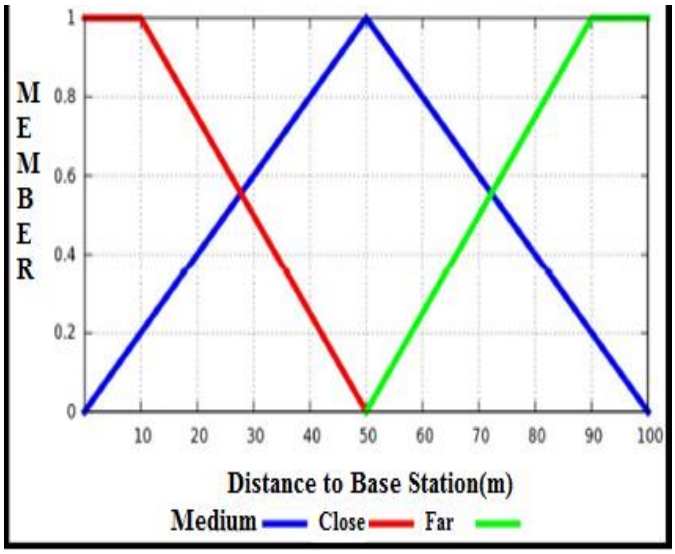



Fig. 4. Fuzzy input variable of distance between nodes and BS

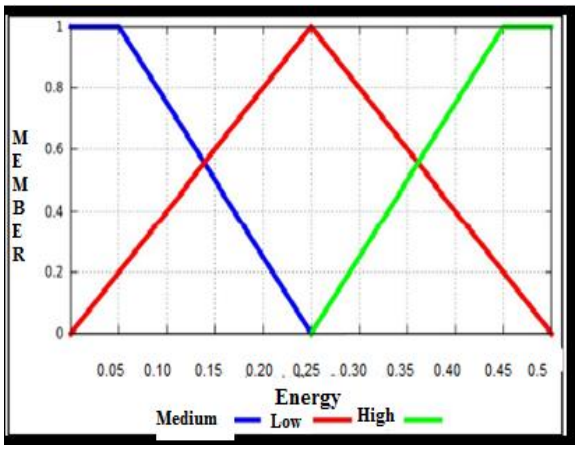



Fig. 5. Fuzzy input variable Residual energy for each node

They are added to each grid by calculating the formula for each center. The gravity center represents the average point of the object weight [7]. In figures 3,4 and 5 illustrates Fuzzy inputs variables which enter to the interference engine.

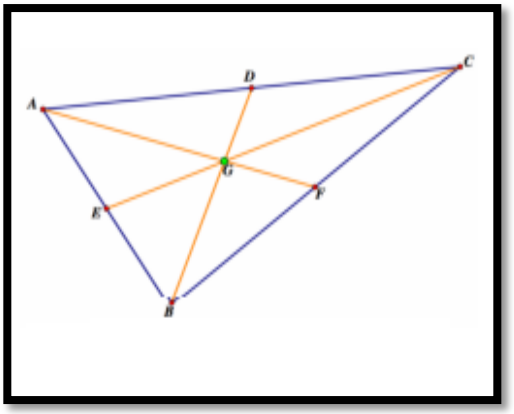



Fig. 6. Center of gravity

Figure. 6 shows the two-dimensional gravity and energy centers. The center of gravity for two points is determined as follows:

$Sum\ (X\_coordinate(node)\ *\ node\_mass)/node\_node)$ (8)

$Sum\ (Y\_coordinate(node)\ *\ node\_mass)/node\_mass)$ (9)

The proposed protocol comprises the setup and steady-state phases. In the setup phase, the network is separated into nine grids/clusters. Each grid/cluster comprises two centers (center of gravity and energy

Table 1. Decision-making criteria for optimal CH

center), one CH, and several

$$Xgc = (x1m1 + x2m2)/(m1 + m2) \quad (4)$$

$$Ygc = (y1m1 + y2m2)/(m1 + m2) \quad (5)$$

The following formula is used in the case of more than two object masses:

$$Xgc = Sum(X\_coordinate(node) * node\_mass)/AllMass \quad (6)$$

$$Ygc = Sum(Y\_coordinate(node) * node\_mass)/AllMass \quad (7)$$

To find the energy center for each grid, the energy center for two points or more is calculated as

$Sum\ (X\_coordinate(node)\ *\ node\_mass)/node\_node$ (8)

$Sum\ (Y\_coordinate(node)\ *\ node\_mass)/node\_mass$ (9)

### A. Fuzzy Ranking System

The proposed protocol comprises the setup and steady-state phases. In the setup phase, the network is separated into nine grids/clusters. Each grid/cluster comprises two centers (center of gravity and energy center), one CH, and several nodes with different energy levels. The setup phase comprises two stages for electing nine CHs: (i) initial fuzzy ranking system election and (ii) fuzzy logic with two-dimensional centers.

Using the fuzzy logic system, the BS ranks the nodes, measures the distances between nodes, and identifies the most energetic node.

As shown in figure 7, The BS selects the first nine nodes with the highest amount of remaining energy as CH candidates after using Table1 DecisionMaking.

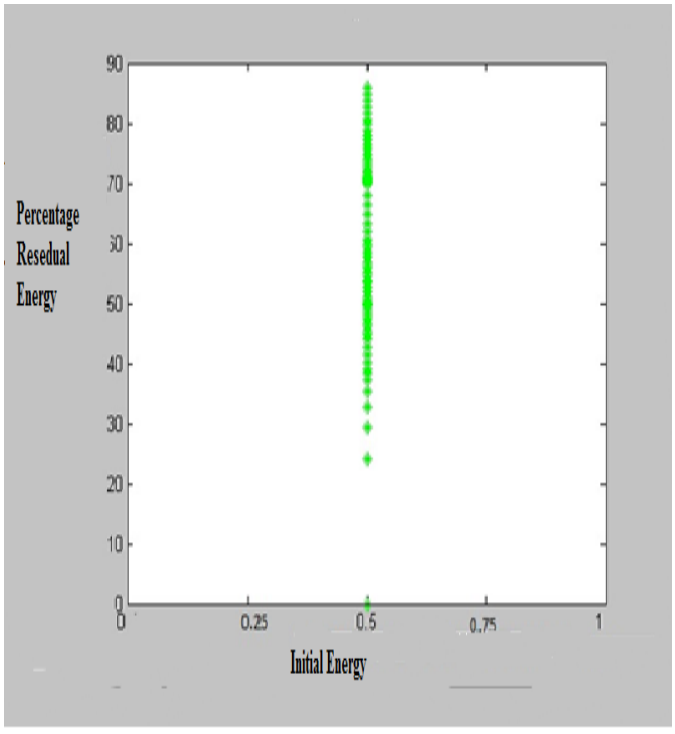



Fig. 7. Shows the nodes with the highest ranks based on the residual energy criteria; for 100 nodes, the ranking system ranges from 25% to 85%.

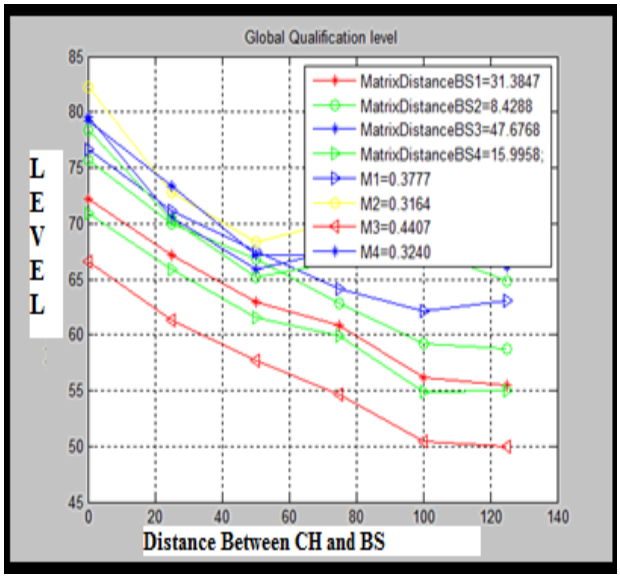



Fig. 8. Best CHs identified by the BS for the first stage.

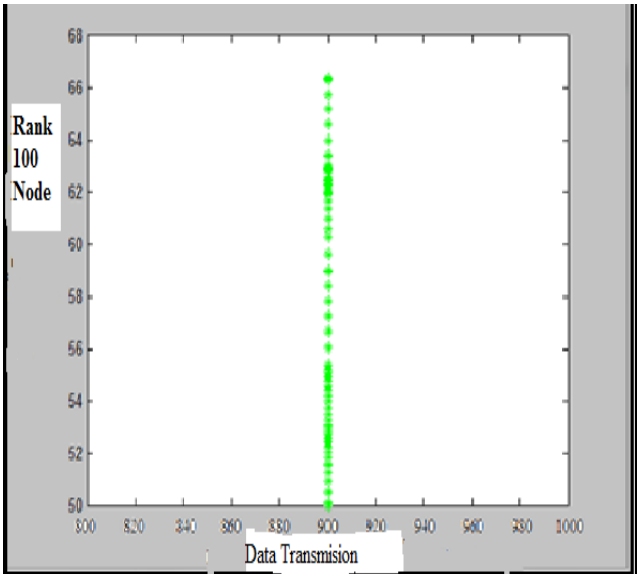



Fig. 9. Rank of nodes based on distance to the BS

The BS selects the optimal CH according to the three criteria: distance between nodes, the distance between nodes and the BS, and the residual energy in each node (Table 1).

As shown in Figure.8 shows the nodes with the highest ranks ranging from 50% to 67% based on their distance to the BS.

In each round, the proposed protocol involves 9 CHs, 100 nodes, and 1 BS. Thus, the distance from a node to the BS = distance from the node to the (candidate CHs) × distance from the candidate CH × BS. BS = (100 × 9 × 1) = 900.

Finally ,as in figure 9, for the first stage, the BS selects the top nine CHs with the highest percentages of 83%, 80%, 80%, 79%, 77%, 76%, 72%, 71%, and 67%. The selection is based on the ranking of residual energy and the distance between nodes and the BS.

### *B. 3Fuzzy logic with Gravity and Energy Centers for Other Rounds*

In the second stage, the nodes are randomly distributed within the grids. Then, CHs are elected by the BS on the basis of the two-dimensional centers. A node is selected as a CH if it has the shortest distance to the BS and the gravity center, which is near the BS. If a node is far from the gravity center but near the energy center, which is near the BS, this node with the highest energy is selected as the CH. A node is selected as a CH according to the following three criteria if its distances to the two centers are the same: (i) distance to the BS, (ii) distance to other nodes, and (iii) residual energy in the fuzzy logic system

### *C. 3Steady-State Phase*

In the steady-state phase, the proposed protocol adopts Time Division Multiple Access TDMA scheduling anddata transmission for each node. Specifically, the proposed protocol gathers data only from the nodes to the BS. The BS identifies the first nine nodes with the highest rank based on their distance to the BS as CH candidates. The setup phase and the steady state phase illustrated in the pseudo code (see figure 10).

TABLE I. MULTIPLE DECISION MAKING CRITERIA FOR FUZZY LOGIC

| **Energy_CH** | **Distance Between Nodes** | **Distance to BS** | **Chance of CH** |
|---|---|---|---|
| Low | Far | Far | V*week |
| Low | Far | Medium | Week |
| Low | Far | Close | L*week |
| Low | Medium | Far | Week |
| Low | Medium | Medium | L*W |
| Low | Medium | Close | Medium |
| Low | Close | Far | L*M |
| Low | Close | Med | Medium |
| Low | Close | Close | H*M |
| Medium | Far | Far | L*Strong |
| Medium | Far | Medium | L*M |
| Medium | Far | Close | Medium |
| Medium | Medium | Far | L*M |
| Medium | Medium | Medium | Medium |
| Medium | Medium | Close | H*M |
| Medium | Close | Far | Medium |
| Medium | Close | Medium | H*M |
| Medium | Close | Close | L*strong |
| High | Far | Far | Medium |
| High | Far | Medium | H*M |
| High | Far | Close | L*strong |
| High | Medium | Far | H*M |
| High | Medium | Medium | L*strong |
| High | Medium | Close | Strong |
| High | High | Far | L*strong |
| High | High | Medium | strong |

**Pseudo code:**

**Setup Phase:**

1 Divide the network into two separate grids.
2 Find the center of gravity for each grid.
3 **IF** (node count = 1)
4 The node nearest to the BS participates in grid computation.
5 **Else If** (node count > 1)
6 *AllMass = Calculate sum of all nodes' mass in grid*
*Sum(X_coordinate(node) × node_mass)/AllMass*
7 *Sum(Y_coordinate(node) × node_mass)/AllMass*
9 **End**
Find energy center for each grid.
10 **IF** (node count = 1)
11 The node nearest to the BS participates in grid computation.
12 **Else IF** (node count > 1)
13 *Sum(X_coordinate(node)×node_mass)/node_mass)/node_count*
*Sum(Y_coordinate(node)×node_mass)/node_mass)/node_count*
15 **IF** (Distance (center of gravity, BS) <= Distance (center of energy, BS))
Elect node nearest to the center of gravity as CH
17 **Else IF** Distance (center of gravity, BS) >= Distance (Center of energy,BS)
18 Elect most energetic node as CH
19 **IF** (Distance (center of gravity, BS) = Distance (center of energy, BS))
20 BS selects CH through fuzzy logic using three criteria (distance node to BS, distance between nodes, residual energy)
**End**
21
**Steady-state Phase:**
22 **IF** (node is normal = 0) (repeat)
23 node energy = node energy − consumed energy of sending a message
**Else**
24 node energy = node energy − aggregation energy − consumed energy of sending message
25 **End**
26 **End**

Fig. 10. Pseodecode of FLG-proposed

## VI. Result and Discussions

The simulation parameters in matlab simulator is shown in table2.

TABLE II. Simulation Parameters

| Parameter | Value |
|---|---|
| **Network size** | 100*100 m |
| **Ee** | 50 nJ/bit |
| **Tevent_all** | randi(9,1,m)+1)*1*$10^{-3}$ m |
| **T1** | 0 |
| **Pactive** | 6*$10^{-3}$ mw |
| **Tdown** | 1*$10^{-3}$ m |
| **Psleep** | 1*$10^{-3}$ mw |
| **L** | 1000 bit |
| **D0** | 87 m |
| **Grid Number** | 4 |
| **Mp** | 0.0013 *$10^{-9}$ |
| **Fs** | 10*$10^{-9}$ |
| **Number of nodes** | 100 |

### A. Performance Metrics

The performance of the FLG proposed protocol can be evaluated with a number of metrics.

- Network lifetime: The time interval from the beginning of operation (of a sensor network) until the death of the last alive sensor.
- First Dead Node (FDN): Number of rounds after the first sensor died. This parameter is directly related to the stability period parameter. A large FDN implies long stability period of the network.
- Last Dead Node (LDN): Number of rounds after all sensor nodes are dead.

As shown in figure 11, The FND occurs in round 2,100, and the LND occurs in round 5,000. This result indicates the prolonged network lifetime using the proposed protocol.

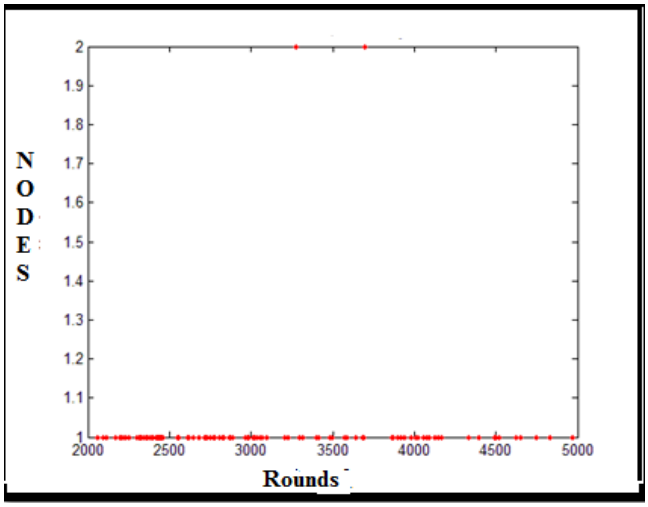



Fig. 11. Number of dead nodes in each round

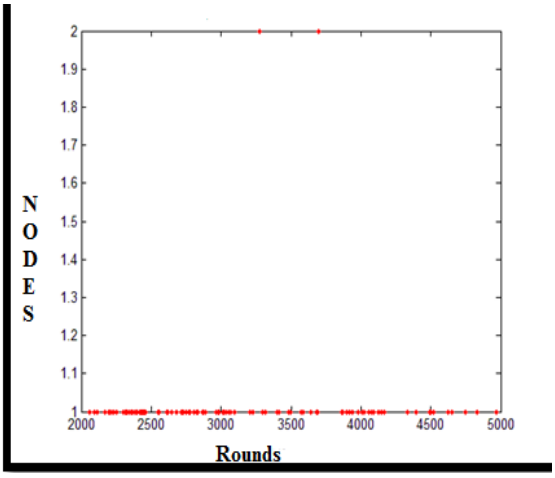



Fig. 12. Number of dead nodes in each round for the TDTCGE protocol

As shown in Figure 12, FLG-proposed protocol Compared with the TDTCGE protocol, our proposed protocol improved the FND by 50% and the LND by 2%.

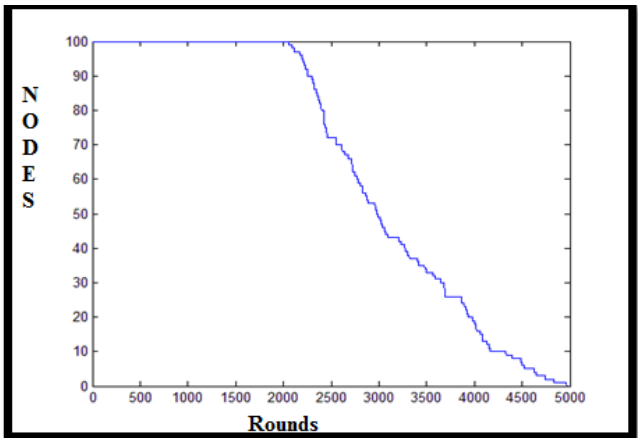



Fig. 13. Network lifetime improvement with the proposed protocol

As shown in figure 13, In comparison with the TDTCGE protocol, the fuzzy proposed protocol improved the network lifetime by 52%.

TABLE III. LIFETIME OF PROTOCOLS

| Protocol | FND | LND |
|---|---|---|
| **Proposed _FLG** | 2,100 | 5,000 |
| **TDTCGE** | 1,400 | 4,800 |
| **LEACH** | 780 | 1,100 |
| **LEACH-SWDN** | 1,100 | 1,490 |
| **EAERP** | 1,076 | 4,085 |
| **ERP** | 1,057 | 3,673 |
| **SEP** | 1,107 | 2,238 |
| **ECPF** | | |
| **CRCWSN** | 780 | 1,400 |

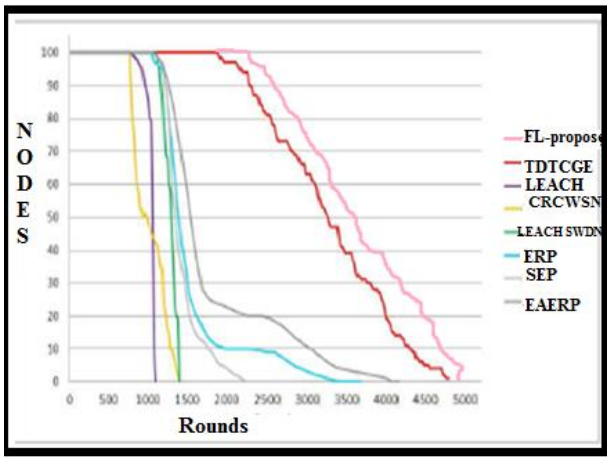



Fig. 14. Network lifetime improvement with various protocols

The FLG- proposed protocol is implemented using Matlab. In figure14 and Table 3. Illustratethe comparison of lifetime between FLG-Proposed and previous protocols.

The Simulation reveal that the lifetime extended by the FLG-proposed protocol is longer than that of the following protocols: TDTCGE,CRCWSN,LEACH, Low-energy Adaptive Clustering Hierarchy with Sliding Window and Dynamic Number of Nodes (LEACH-SWDN)[18], Clustered Routing Protocol (ERP)[19], Stable Election Protocol (SEP)[16],and Energy-aware Evolutionary Routing Protocol (EAERP )[17].

As shown in figure 14, an obvious improvement in network lifetime was achieved with the FLG-proposed protocol, the performance of which was superior to that of the TDTCGE, LEACH, CRCWSN, LEACH-SWDN, ERP, SEP, EAERP protocols.

As shown in figure. 15, the addition of fuzzy logic to the grids and the two centers yielded nine CHs near the centers, except for the eighth CH, which is far from the two centers by calculating the average of nearest distances nodes to the clusterheads (4+2+0+4+1+1+4+15+2) m / 9 cluster heads = 3.6 m.

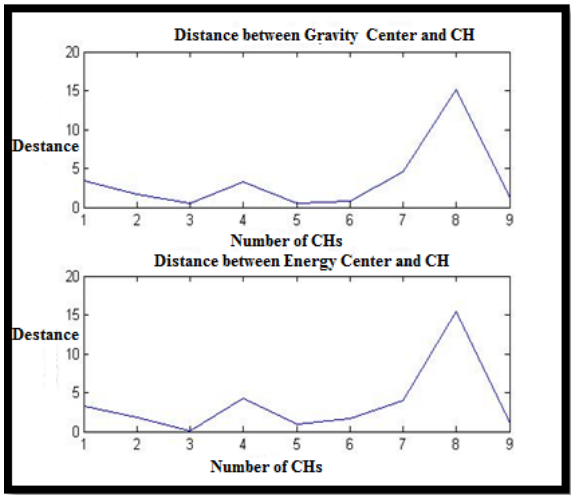



Fig. 15. Distance between centers and CH after adding fuzzy logic

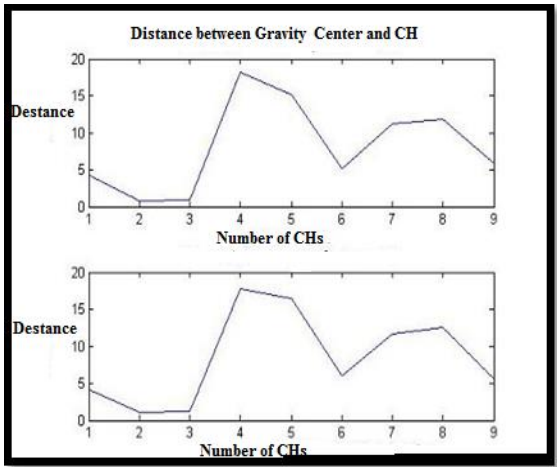



Fig. 16. Distance between centers and CH without fuzzy logic (TDTCGE protocol)

As shown in figure. 16, in the case without fuzzy logic, only CH numbers 1, 2, and 3 are observed near the centers; CH numbers 4, 5,6,7,8,9 and 8 are observed far from the two centers the percentage of nearest nodes is 33% . By adding Fuzzy logic Ranking system the percentage of the nearest nodes more than TDTCGE by by calculating the average of nearest distances nodes to the clusterheads (5+1+1+18+16+7+10+12+6) m / 9 cluster heads = 8.4 m. these calculations shows that FLGG will minimize distance radius competition 43%. than TDTCGE

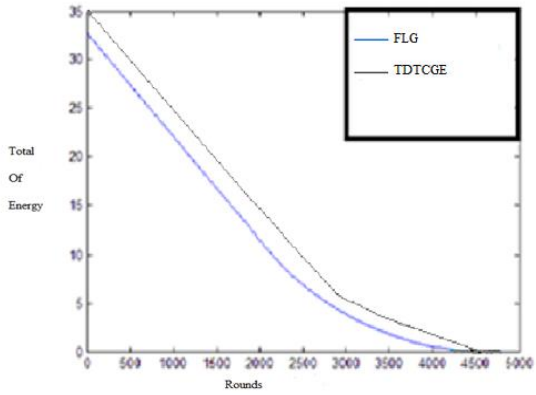



Fig. 17. Amount of energy consumption in each round

Fig. 17. shows that the proposed protocol consumes less energy than the TDTCGE protocol does, with its energy savings being 8% more than that of the TDTCGE protocol during round 5,000.

## VII. CONCLUSIONS

The addition of fuzzy logic to the grids obviously minimized the distance between nodes and the BS, as well as the distance between the CH to other nodes and the BS. The radius competition between nodes also decreased. The remaining node with the most energy is chosen by the BS as a CH. In this way, the proposed protocol based on fuzzy logic is able to extend network lifetime by nearly 52%. As revealed in the simulation, the proposed protocol also guarantees a balance in network load.

## REFERENCES

[1] AG Delavar, AA Baradaran. "CRCWSN: Presenting a routing algorithm by using re-clustering to reduce energy consumption in WSN", International Journal of Computers Communications & Control, 8.1. 61–69 (2012).

[2] IF Akyildiz, W Su, Y Sankarasubramaniam, E Cayirci, "A survey on sensor networks. Communications magazine", IEEE. 40(8), 102–114 (2002).

[3] EM Ibbini, KY Lun, M Othman, ZM Hanapi, MS Ibbini, "An efficient mathematical analysis for saving energy in WSN routing protocol."Journal of Theoretical and Applied Information Technology. 76(3), 314–321(2015).

[4] EM Ibbini, KY Lun, M Othman, ZM Hanapi, MS Ibbini,"A survey of routing MAC techniques for wireless sensor networks routing protocol."Journal of Theoretical and Applied Information Technology. 76(3), 322–332 (2015).

[5] V Godbole, "FCA-An approach on LEACH protocol of wireless sensor networks using fuzzy logic". International Journal of Computer Communications and Networks (IJCCN). 3.1, 1–13 (2012).

[6] Iancu I. "A Mamdani type fuzzy logic controller". INTECH , University Campus STeP Ri Slavka Krautzeka 83/A , Open Access Publisher; (2012).

[7] E Rezaei, AA Baradaran, A Heydariyan, "TDTCGE: Two-dimensional technique based on center of gravity and energy center in wireless sensor network", Journal of Basic and Applied Scientific Research (JBASR). 3(8), 194–201, (2013).

[8] Delavar, A.G. And A.A. Baradaran, "CRCWSN: Presenting A Routing Algorithm By Using Re-Clustering To Reduce Energy Consumption" , In WSN. International Journal Of Computers Communications & Control,(ISSN 1841-9836) 8(1): P. 61-69. , 102– 114, 2013.

[9] Aheri, Hoda, Peyman Neamatollahi, Ossama Mohamed Younis, Shahrzad Naghibzadeh, and Mohammad Hossein Yaghmaee. "An energy-aware distributed clustering protocol in wireless sensor networks using fuzzy logic." Ad Hoc Networks10, no. 7 (2012): 1469-1481.

[10] Taheri, Hoda, et al. "An energy-aware distributed clustering protocol in wireless sensor networks using fuzzy logic." Ad Hoc Networks 10.7 1469-1481(2012):.

[11] Jiang, Haifeng, Yanjing Sun, Renke Sun, and Hongli Xu. "Fuzzy-logic-based energy optimized routing for wireless sensor networks." International Journal of Distributed Sensor Networks 9, no. 8 (2013): 216561.

[12] Dadios Elmer, P., "Fuzzy logic–controls, concepts, theories and applications". Croatia, InTech, ISBN, pp.978-953,2012.

[13] D.Gupta R. I. and S. Sampalli, "Cluster-head election using fuzzy logic for wireless sensor networks," in 3rd Annual Conference on Communication Networks and Services Research, (2005) vol. 2.

[14] Ran, Ge, Huazhong Zhang, and Shulan Gong. "Improving on LEACH protocol of wireless sensor networks using fuzzy logic." Journal of Information &Computational Science 7, no. 3 (2010): 767-775.

[15] Balakrishnan, Baranidharan, and Santhi Balachandran. "FLECH: Fuzzy Logic Based Energy Efficient Clustering Hierarchy for Nonuniform Wireless Sensor Networks." Wireless Communications and Mobile Computing 2017 (2017).

[16] G. Smaragdakis, I. Matta, A. Bestavros, "SEP,a stable election protocol for clustered heterogeneous wireless sensor networks", in: Second International Workshop on Sensor and Actor Network Protocols and Applications (SANPA 2004), Boston, MA, 2004.

[17] Enan A. Khalil, Bara'a A. Attea,"Energy-aware evolutionary routing protocol for dynamic clustering of wireless sensor networks. Swarm and Evolutionary Computation", 1(4): p. 195-203,2011.

[18] Aimin Wang, Dailiang Yang , Dayang Sun :"A clustering algorithm based on energy information and cluster heads expectation for wireless sensor networks" . Computers and Electrical Engineering 38, 662–671, (2012).

[19] Bara'a, A. Attea, and Enan A. Khalil. "A new evolutionary based routing protocol for clustered heterogeneous wireless sensor networks." Applied Soft Computing 12.7,1950-1957, (2012).

# Design of Linear Phase High Pass FIR Filter using Weight Improved Particle Swarm Optimization

Adel Jalal Yousif, Ghazwan Jabbar Ahmed, Ali Subhi Abbood
Electronic Computing Center
University of Diyala
Diyala, Iraq

***Abstract*—The design of Finite Impulse Response (FIR) digital filter involves multi-parameter optimization, while the traditional gradient-based methods are not effective enough for precise design. The aim of this paper is to present a method of designing 24th order high pass FIR filter using an evolutionary heuristic search technique called Weight Improved Particle Swarm Optimization (WIPSO). A new function of the weight parameters is constructed for obtaining a better optimal solution with faster computation. The performance of the proposed algorithm is compared with two other search optimization algorithms namely standard Genetic Algorithm (GA) and conventional Particle Swarm Optimization (PSO). The simulation results show that the proposed WIPSO algorithm is better than GA and PSO in terms of the magnitude response accuracy and the convergence speed for the design of 24th order high pass FIR filter.**

***Keywords—Finite impulse response filter; evolutionary optimization; particle swarm optimization; fitness function; genetic algorithm; high pass filter; impulse response***

## I. Introduction

Digital filtering is one of the main fundamental aspect of digital signal processing, so digital filters are used in many applications such as seismic signal processing systems, biomedical applications systems, audio and video processing systems, and communication systems. In general, filters serve two purposes: the first is signal separation, which is applied when the desired signal is corrupted with another unwanted signal such as noise, and the second purpose is signal restoration, which is applied when the desired signal is distorted for some reason. A digital filter can be described as a mathematical algorithm which is implemented in hardware or software for achieving the objectives of the filtering process [1].

Digital filters can be utilized to process very low-frequency signals, such as those happen in seismic and biomedical applications in an effective manner. Moreover, the digital filters characteristics can be easily modified or altered through a control software that handles the content of the registers to meet the new specifications. Hence, a single programmable digital filter can be employed to implement multiple filtering functions without the need to add additional hardware components [2].

Based on the type of impulse response, the digital filters can be classified as Finite Impulse Response (FIR) filters and Infinite Impulse Response (IIR) filters. In case of an FIR digital filter, the impulse response is decay and settle to zero within a finite length of time. But, the impulse response for an IIR digital filter never dies out [3, 4]. FIR filter has a number of good features that make it an attractive choice for many researchers in various fields. FIR filters are inherently stable, and usually guarantee a linear phase response. This is due to the fact that FIR filter requires no feedback and all the poles are located within the unit circle [5, 6] Hence, digital FIR filters are easily designed and simply implemented. FIR filters are also referred to as feed-forward or non-recursive filters.

The recent approaches used in the design of FIR filters utilize evolutionary techniques that have been proven to be efficient in multidimensional nonlinear environment [7] such as Genetic Algorithm (GA) [8], Differential Evolution (DE) [8], Simulated Annealing (SA) [9], Particle Swarm Optimization (PSO) [10, 11] and so on. In this work, the conventional PSO algorithm has been improved to overcome the constraints encountered in the filter design problem. This paper presents a new method for designing the HP FIR digital filter using Weight Improved Particle Swarm Optimization (WIPSO) approach.

The rest of the paper is organized as follows. The HP FIR filter design problem is described in section 2. In section 3, Different optimization algorithms namely, GA, PSO, and WIPSO are briefly discussed. Simulation results and comparative analysis of HP FIR filter design problem using GA, PSO, and WIPSO are presented in section 4. Finally, section 5 contains the conclusions drawn from the results evaluation..

## II. FIR Filter Design Problem

The design of digital FIR filter of length N with an input-output relationship can be described by the following difference equation [12]:

$$y(n) = \sum_{k=0}^{N} b_k\, x(n-k) \tag{1}$$

$$= b_0 x(n) + b_1 x(n-1) + \ldots\ldots + b_N x(n-N+1)$$

Where ($b_k$) represents the filter coefficients set. The output y(n) is a function only of the input signal x(n). FIR filter can also be characterized by its transfer function as follows:.

$$H(z) = \sum_{n=0}^{N} h(n) z^{-n} \quad n=0,1,\ldots,N \tag{2}$$

Where the coefficients h(n) represent the impulse response of finite length and N represents the filter order. Thus, the

number of coefficients will be (N +1). The type of the filter for a pass or attenuate e.g. high pass, low pass, band pass, etc. is determined by the filter coefficients h(n) that are to be specified in the design steps. When the coefficients of FIR filter are symmetrical around the center coefficient, the FIR filter is linear-phase In another word, the coefficients number that is actually optimized is (N/2+1), which is equal to 13 in this work. In (2), the coefficient vector $\{h_0,h_1,...,h_N\}$ is represented by the particles positions in (N+1) dimensional search space of PSO algorithm. Hence, in each cycle of the evolutionary algorithm, these particles find new positions to be the new coefficient vector of the transfer function. The following equation describes the FIR filter frequency response [13, 14]:

$$H_d(e^{jw}) = \sum_{n=0}^{N} h(n)e^{-jwn} \quad (3)$$

Where $w = \frac{2\pi}{N}k$ , $K = 0,1,\dots N-1$.

For ideal HP filter, the following equation defines the filter response:

$$H_i(e^{j\omega}) = \begin{cases} 0 & 0 \le \omega \le \omega_c \\ 1 & otherwise \end{cases} \quad (4)$$

Where $\omega_c$ is the cut-off frequency as shown in figure 1.

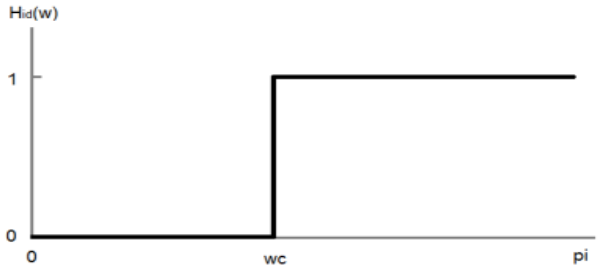



Fig. 1. Ideal HP Filter Frequency Response.

By sampling the frequency in the range [0, π] with L points, we get the following equations:

$$H_d(\omega) = [\, H_d(\omega_1), H_d(\omega_2), H_d(\omega_3), \dots, H_d(\omega_L)]^{\mathrm{T}} \quad (5)$$

$$H_i(\omega) = [\, H_i(\omega_1), H_i(\omega_2), H_i(\omega_3), \dots, H_i(\omega_L)]^{\mathrm{T}} \quad (6)$$

Where $H_d(\omega)$ and $H_i(\omega)$ are the frequency response of the designed and ideal filter respectively. So, the error function

$E(\omega)$ is defined by the following equation:

$$E(\omega) = [\, H_d(\omega) - H_i(\omega) \,] \quad (7)$$

Then, the error function above has been utilized for obtaining the fitness function as follows:

$$\text{fitness} = \sum_1^L \left|\left(E(\omega)\right)\right| \quad (8)$$

## III. Evolutionary Algorithms Utilized

### A. Genetic Algorithm (GA)

Genetic algorithm is basically a heuristic search method that can be employed to find an optimum solution for optimization problems depending on evolution and natural selection principles. Genetic algorithm manipulates a population of individuals at each iteration (cycle) where each individual, which is known as a chromosome, is a coded string of a probable solution of the optimization problem. Chromosomes are usually of a fixed length and constructed over some particular alphabet. Each chromosome of the population represents one candidate solution to the evolution function of the optimization problem according to its fitness value, which is calculated by a function called fitness function. A standard GA can be summarized by the following steps [15]:

- Randomly create an initial population of chromosome strings
- Evaluating fitness values of population strings
- Selection of elite strings
- Copying process of the elite strings is applied to the non-chosen chromosome strings
- Producing the off-springs by applying crossover and mutation operators
- Updating of the genetic cycle starting with the second step
- Stopping the iteration when the termination condition is met

### B. Particle Swarm Optimization (PSO)

Particle swarm optimization (PSO) is an evolutionary computation optimization algorithm inspired by social behavior of a swarm of bird in searching of food. The PSO technique was originally introduced by Kennedy and Eberhart in 1995 [16]. The key features of PSO are the ease of implementation, robustness to control parameters and effectiveness of computation compared with other existing probabilistic search algorithms. Additionally, PSO has the ability to handle problems with vast search space and non-linear objective function and giving better results within a reasonable amount of time.

PSO algorithm operates with a random population of individuals called a swarm, where each individual in the swarm is called a particle. The fitness value of each particle in the swarm is calculated iteratively through the search space at different locations by using a pre-defined fitness function. Hence, each particle in the search space has the experience to know its best value so far (pbest) or local best. Moreover, each particle vector, which is represent the coefficient vector $\{h_0,h_1,...,h_N\}$ in FIR filter design problem, knows the best value so far in the group (gbest) among pbests. As a result, the best solution is improved by particle's movement in the search space through the generations. Each particle tries to adjust their velocity and position based on the best encountered positions through the search space by using these two equations[17, 18]:

$$v_i^{k+1} = w * v_i^k + c_1 * rand_1 * \left(pbest - x_i^k\right) + c_2 * rand_2 * \left(gbest - x_i^k\right) \quad (9)$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \quad (10)$$

Where $v_i^k$ and $x_i^k$ are the velocity and position of the *i*th particle at the *k*th iteration respectively; *w* is the inertia weight;

$rand_1$ and $rand_2$ are random numbers in the interval [0,1]; *c1* and *c2* are constants usually having the same value, known as cognitive and social acceleration factors respectively. The factors $c1$ and $c2$ help the particles to reach the gbest, while the inertia weight determines the exploration size through the search space so that a large inertia weight causes large exploration and vice versa [19-21].

### C. Weight-Improved Particle Swarm Optimization (WIPSO)

The capability of the global search of conventional PSO has been improved by modifying the weight parameter, cognitive factor, and social factor. This improved PSO is called as Weight-Improved Particle Swarm Optimization (WIPSO). Hence, the velocity equation that is described in (9) has been rewritten as [22, 23]:

$$v_i^{k+1} = w_{new} * v_i^k + c_1 * rand_1 * \left(pbest - x_i^k\right) + c_2 * rand_2 * \left(gbest - x_i^k\right) \quad (11)$$

Where:

$$w_{new} = w_{min} + w * rand_3 \quad (12)$$

$$w = w_{max} - \frac{w_{max} - w_{min}}{Iter_{max}} * iter \quad (13)$$

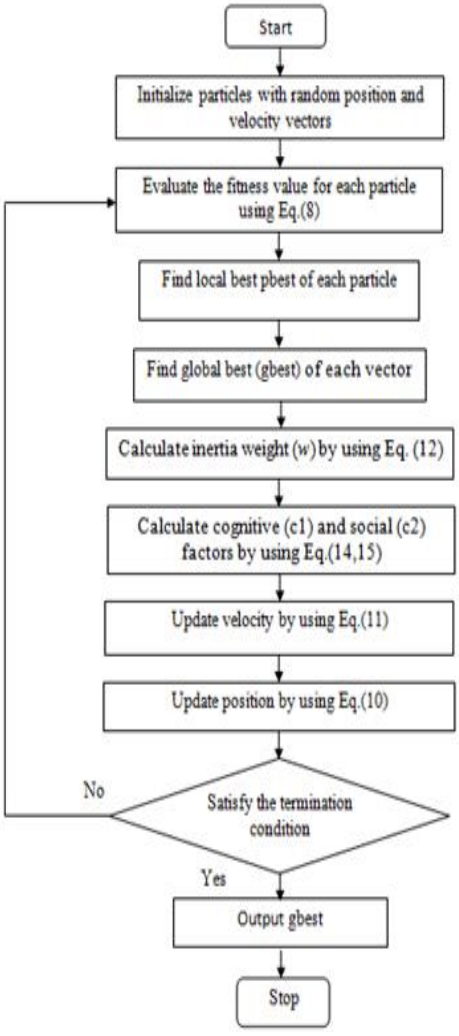



Fig. 2. Flow Chart of the WIPSO-based HP FIR Filter Design.

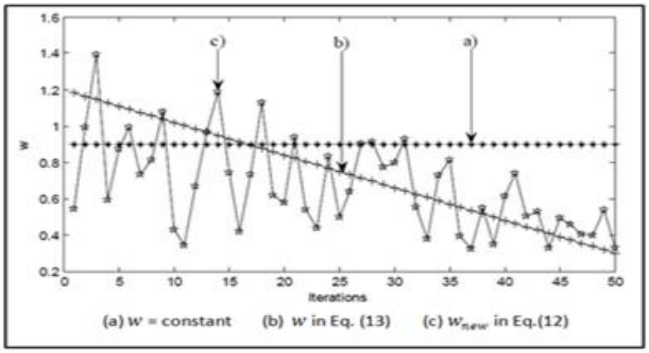



Fig. 3. Comparison of Weight Parameters Characteristics.

Here, $w_{min}$ and $w_{max}$ are the initial and final values of inertia weight respectively, *iter* represents the current number of iterations, $iter_{max}$ represents the maximum number of iterations. The factors *c1* and *c2* are modified as follows [24-26]:

$$c_1 = c_{1max} - \frac{c_{1max} - c_{1min}}{Iter_{max}} * iter \quad (14)$$

$$c_2 = c_{2max} - \frac{c_{2max} - c_{2min}}{Iter_{max}} * iter \quad (15)$$

Where;

$c1_{max}$ , $c1_{min}$ represent the initial and final values of cognitive factors respectively.

$c2_{max}$ , $c2_{min}$ represent the initial and final values of social factors respectively.

The flow chart of the proposed WIPSO-based HP FIR filter design is depicted in figure 2.

Figure 3 shows characteristics of three weight parameters, where, line *a* represents a constant weight, line *b* represents the weight parameter of (13) and line *c* describes our proposed weight parameter expressed by (12).

## IV. Results and Comparison

For evaluating the performance of the proposed method for FIR filter design, two optimization algorithms namely GA and PSO in addition to the proposed WIPSO have been carried out using MATLAB simulation. The proposed WIPSO algorithm has been applied to the design of 24*th* order (25-tap) high pass FIR filter with cut-off frequency $w_c = 0.5\pi$. Hence, the coefficient vector length will be 25. The control parameters values are chosen for GA, PSO, and WIPSO algorithms are shown in table I.

The main objective of the proposed modification of the conventional PSO is to minimize the fitness function in (8). A wide range of control parameters values of the PSO algorithm have been chosen to show the effect of changing these parameters on the optimization process, and to choose the best parameters values for WIPSO algorithm. Figure 4 shows the effect of inertia weight (w) changing on the fitness function value. Figure 5 displays the relation between the fitness function value and the (cognitive (c1) & social (c2)) learning parameters. Based on Fig.4 and Fig.5, it can be seen that the range (from 0.6 to 0.9) is a good choice for w and the range (from 0.1 to 1.3) is a good choice for c1 and c2 for the linear

phase HP FIR filter design. With these above ranges of w,c1 and c2, the PSO will have higher chance to find the optimal filter coefficients within a reasonable time. Depending on Fig.4 and Fig.5, the control parameters values (w,c1,and c2) of the conventional PSO algorithm have been chosen to be (0.7, 0.4, and 0.4) respectively and the control parameters values ($w_{min}$, $w_{max}$, $c1_{min}$, $c1_{max}$, $c2_{min}$, and $c2_{max}$) of the proposed WIPSO algorithm have been chosen to be (0.6, 0.9, 0.1, 1.3, 0.1, and 1.3) respectively**.**

TABLE I. CONTROL PARAMETERS FOR GA, PSO AND WIPSO

| **Parameters** | **GA** | **PSO** | **WIPSO** |
|---|---|---|---|
| Nvar (h(n)) | 13 | 13 | 13 |
| Population size | 1000 | 1000 | 1000 |
| Iteration Cycle | 300 | 300 | 300 |
| Selection | Roulette Wheel | - | - |
| Mutation | Gaussian Mutation | - | - |
| Mutation rate | 0.025 | - | - |
| Crossover | Two Point Crossover | - | - |
| Crossover rate | 0.8 | - | - |
| Inertial weight (w) | - | 0.7 | - |
| Minimum inertia weight (wmin) | - | - | 0.6 |
| Maximum inertia weight (wmax) | - | - | 0.9 |
| c1 | - | 0.4 | - |
| c2 | - | 0.4 | - |
| Minimum cognitive factor (c1min) | - | - | 0.1 |
| Maximum cognitive factor (c1max) | - | - | 1.3 |
| Minimum social factor (c2min) | - | - | 0.1 |
| Maximum social factor (c2max) | - | - | 1.3 |

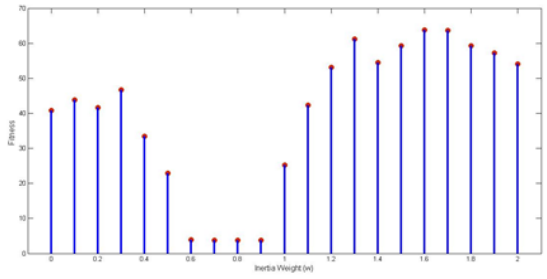



Fig. 4. The Fitness for Different Inertia Weights (w).

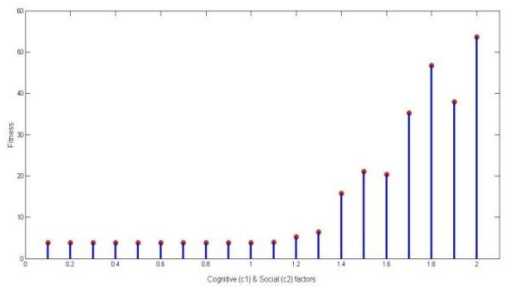



Fig. 5. The Fitness for Different Cognitive (c1) and Social (c2)) Learning Parameters.

Figures 6, 7, 8 and 9 represent the convergence behavior, amplitude response, filter coefficients & Impulse Response, and magnitude & phase response of the Weight Improved Particle Swarm Optimization (WIPSO) based 25-tap linear phase HP FIR filter.

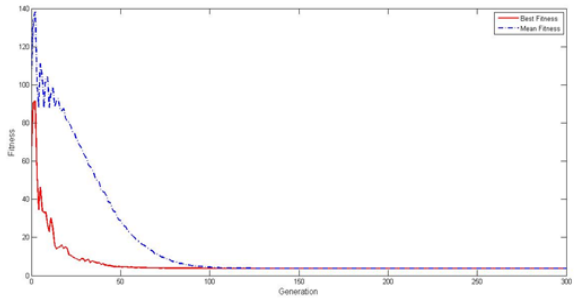



Fig. 6. Convergence Behavior of WIPSO in the Design of the 25-Tap HP FIR Filter.

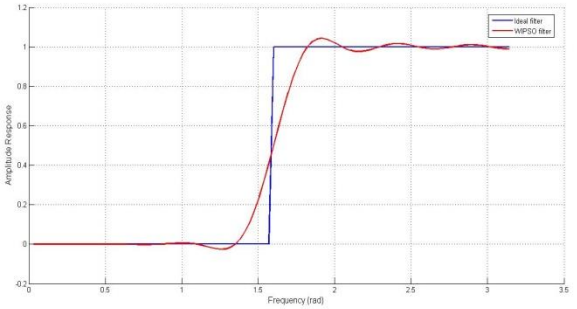



Fig. 7. Amplitude Response for the 25-tap WIPSO-based HP FIR filter.

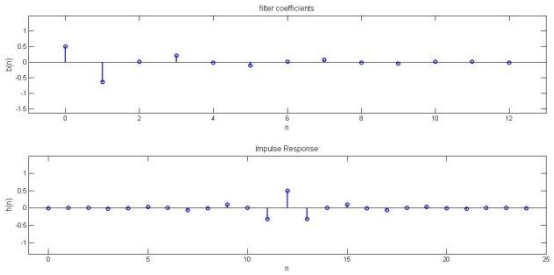



Fig. 8. Filter Coefficients and Impulse Response for the 25-tap WIPSO-based HP FIR filter.

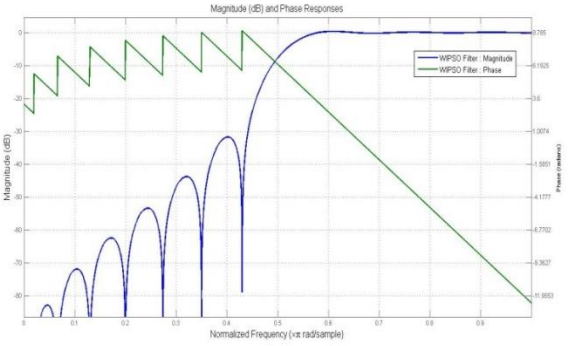



Fig. 9. Magnitude and Phase Responses for the 25-tap WIPSO-based HP FIR filter.

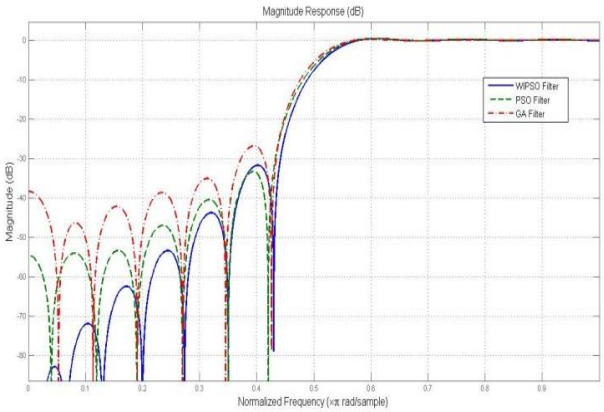



Fig. 10. Magnitude Response for the 25-tap HP FIR filter using WIPSO, PSO, and GA.

TABLE II. OPTIMIZED HP FIR FILTER COEFFICIENTS OF ORDER 24

| h(n) | WIPSO-based High Pass FIR Filter coefficients | PSO-based High Pass FIR Filter coefficients | GA-based High Pass FIR Filter coefficients |
|---|---|---|---|
| h(1) = h(25) | -0.0052 | -0.0031 | 0.0014 |
| h(2) = h(24) | 0.0099 | 0.0096 | 0.0110 |
| h(3) = h(23) | 0.0057 | 0.0031 | -0.0025 |
| h(4) = h(22) | -0.0206 | -0.0205 | -0.0214 |
| h(5) = h(21) | -0.0063 | -0.0025 | 0.0028 |
| h(6) = h(20) | 0.0343 | 0.0341 | 0.0354 |
| h(7) = h(19) | 0.0069 | 0.0023 | -0.0025 |
| h(8) = h(18) | -0.0555 | -0.0556 | -0.0559 |
| h(9) = h(17) | -0.0080 | -0.0016 | 0.0029 |
| h(10) = h(16) | 0.1013 | 0.1011 | 0.1012 |
| h(11) = h(15) | 0.0085 | 0.0016 | -7.2773e-04 |
| h(12) = h(14) | -0.3167 | -0.3170 | -0.3160 |
| h(13) | 0.4914 | 0.4991 | 0.5006 |

Figure 10 represents the magnitude response comparison for the conventional PSO, the standard GA and the proposed WIPSO drawn from the optimized coefficients for the designed 24*th* order high pass FIR filters. Depending on Fig.10, the magnitude response is much better achieved by WIPSO in terms of transition width, pass-band ripple and stop-band attenuation than by GA and PSO. Table II shows the best-optimized HP FIR filter coefficients of order 24 obtained by GA, PSO, and WIPSO.

## V. CONCLUSIONS

In this work, the proposed weight improved particle swarm optimization (WIPSO) algorithm is applied to solve the design problem of the high pass FIR filter. The WIPSO performance is compared with two other algorithms namely standard GA and conventional PSO. The experiments clarify that the proposed WIPSO method outperforms the GA and the PSO in terms of the magnitude response accuracy and the convergence speed. Additionally, simulations have been performed to demonstrate the influence of changing the PSO design parameters to determine the best parameters values for the proposed WIPSO method. It is concluded that the proposed WIPSO method with the inertia weight (w) in the range from 0.6 to 0,9 and the (cognitive (c1) & social (c2)) in the range from 0.1 to 1.3 will have a better performance in HP FIR filter design problem.

## REFERENCES

[1] Eleti, A.A. and A.R. Zerek. FIR digital filter design by using windows method with MATLAB. in Sciences and Techniques of Automatic Control and Computer Engineering (STA), 2013 14th International Conference on. 2013. IEEE.

[2] Mondal, S., et al., Linear phase high pass FIR filter design using improved particle swarm optimization. World academy of science, engineering and technology, 2011. 60: p. 1620-1627.

[3] Hammid, A.T., Applications of Tuning Control Actions for the Efficient Load/frequency Control in Steam Turbine. International Journal of Current Engineering and Technology, 2013. 3(5): p. 1895-1898.

[4] Hammid, A.T., Direct on Line Starter Motor and Reverse System in Allen-Bradley PLC. Diyala Journal for Pure Sciences (DJPS), 2016. 12(4): p. 132-148.

[5] Hammid, A.T., et al., Load Frequency Control for Hydropower Plants using PID Controller. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 2016. 8(10): p. 47-51.

[6] Hammid, A.T., M.H. Sulaiman, and H.J. Kareem, Control on Hydropower Plant using Fuzzy Neural Network based on Right-Angle Triangle Membership. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 2018.

[7] Ghosh, A., et al. Linear phase low pass FIR filter design using Genetic Particle Swarm Optimization with dynamically varying neighbourhood technique. in Evolutionary Computation (CEC), 2012 IEEE Congress on. 2012. IEEE.

[8] Zhao, A.-X., et al. The optimal design method of FIR filter using the improved genetic algorithm. in Industrial Electronics and Applications (ICIEA), 2014 IEEE 9th Conference on. 2014. IEEE.

[9] Chen, S., R. Istepanian, and B.L. Luk, Digital IIR filter design using adaptive simulated annealing. Digital Signal Processing, 2001. 11(3): p. 241-251.

[10] Kar, R., et al., FIR Filter Design using Particle Swarm Optimization with Constriction Factor and Inertia Weight Approach 1. 2011.

[11] Hammid, A.T. and M.H. Sulaiman. Optimal Long-Term Hydro Generation Scheduling of Small Hydropower Plant (SHP) using Metaheuristic Algorithm in Himreen Lake Dam. in MATEC Web of Conferences. 2017. EDP Sciences.

[12] Mandal, S., et al., Design of optimal linear phase FIR high pass filter using craziness based particle swarm optimization technique. Journal of King Saud University-Computer and Information Sciences, 2012. 24(1): p. 83-92.

[13] Hammid, A.T., M.H.B. Sulaiman, and A.A. Kadhim, Optimum Power Production of Small Hydropower Plant (SHP) using Firefly Algorithm (FA) in Himreen Lake Dam (HLD), Eastern Iraq. Research Journal of Applied Sciences, 2017. 12(10): p. 455-466.

[14] 1Hammid, A.T. and M.H. Sulaiman, Series Division Method based on PSO and FA to Optimize Long-Term Hydro Generation Scheduling. Sustainable Energy Technologies and Assessments, 2018. 29.

[15] Abbood, A.S., DESIGN OF LINEAR PHASE LOW PASS FINITE IMPULSE RESPONSE FILTER BASED ON PARTICLE SWARM OPTIMIZATION ALGORITHM. DIYALA JOURNAL OF ENGINEERING SCIENCES, 2015. 8(2): p. 134-150.

[16] Hammid, A.T., M.H. Sulaiman, and T.A. Abdulrazzaq, Control of Hydropower Plant Modeling using Fuzzy Neural Network based on Tuning Membership Function. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 2018.

[17] Hammid, A.T., M.H. Sulaiman, and Y.K. Hameed, Control on Hydropower Plant Modeling using Fuzzy Neural Network based on Normalized Firefly Algorithm. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 2018.

[18] Hammid, A.T., M.H.B. Sulaiman, and A.N. Abdalla, Prediction of small hydropower plant power production in Himreen Lake dam (HLD) using

artificial neural network. Alexandria Engineering Journal, 2018. 57(1): p. 211-221.

[19] Hammid, A.T., M.H. Sulaiman, and A.A. Jasim, Modeling of Hydropower Plant Production using Artificial Neural Network. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 2018.

[20] Hammid, A.T., et al., Priority of Kaplan Turbine and Small Hydropower Plants over Other Resources: An Overview. Journal of Engineering and Applied Sciences, 2017. 12(9 SI): p. 8695-8705.

[21] Hammid, A.T., A. Bhardwaj, and S. Prakash, Design Remote Power Control I/O Data Acquisition System and Control on Home Automation. IJECCE, 2013. 4(2): p. 528-535.

[22] Vu, P., et al. A novel weight-improved particle swarm optimization algorithm for optimal power flow and economic load dispatch problems. in Transmission and Distribution Conference and Exposition, 2010 IEEE PES. 2010. IEEE.

[23] Hammid, A.T., Optimization and Control of Hydro Generation Scheduling using Hybrid Firefly Algorithm and Particle Swarm Optimization Techniques. 2018, Universiti Malaysia Pahang.

[24] Hammid, A.T., M.H.B. Sulaiman, and O.I. Awad, A robust firefly algorithm with backpropagation neural networks for solving hydrogeneration prediction. Electrical Engineering, 2018.

[25] Hammid, A.T., O.A. Imran, and Y.K. Hameed, Executing Work Principle Star/Delta Starter in Motor Systems using Allen-Bradley PLC. Journal of Engineering and Applied Sciences, 2018.

[26] Hammid, A. T., & Kareem, M. H. B. S. J. (2018). Control on Hydropower Plant using Fuzzy Neural Network based on Right-Angle Triangle Membership.

# The Designing of Adaptive Self-Assessment Activities in Second Language Learning using Massive Open Online Courses (MOOCs)

Hasmaini Hashim[1], Sazilah Salam[2], Siti Nurul Mahfuzah Mohamad[3], Nur Syafiatun Safwana Sazali[4]
Fakulti Teknologi Maklumat dan Komunikasi
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

***Abstract*—Massive Open Online Courses (MOOCs) provides an effective learning platform with various high-quality educational materials accessible to learners from all over the world. In this paper, the types of learner characteristics in MOOCs second language learning are discussed. However, there are still problems and challenges including assessment. A quantitative research method approach has been utilized in this study. Results of the study are then used for implementing suitable adaptive self-assessment activities in MOOCs learning. Findings of this study are two folds: (1) The dimension of learner characteristics (learning styles and cognitive style) for improving student performance in MOOCs learning and (2) suitable self-assessment activities that consider learners requirement or adaptive to learner characteristics for improving MOOCs learning performance. Based on the findings, the data indicate that visual, active, thinking and intuitive learner is the proposed dimension used in this study. In this study, our aim is to propose adaptive self-assessment activities for improving MOOCs learning in the second language course. In the future study, students will be investigated about their engagement using MOOC assessment in the second language.**

***Keywords*—MOOCs; adaptive self-assessment; learning styles; cognitive styles; second language**

## I. Introduction

Massively Open Online Courses (MOOCs) have the potential to reach hundreds of thousands of learners enrolled for a variety of reasons [1]. Authors highlighted three controversies for learning in MOOCs in for higher education; (1) measurement, (2) assessment and (3) accountability. Wong [2] highlighted that in MOOC assessment is a big challenge for a large number of students to get detailed and timely feedback. However, the assessment can increase the student performance and at the same time, the teacher brings benefits [3]. Current MOOCs are (1) lacking personalized to learning guidance and (2) intelligent assessment for learners [4]. In MOOC assessment, the issue is, there was a reduction of interest and activity of students during the session of the course [5]. Three requirements to designing effective online course are (1) learners characteristics, (2) context and (3) activities [6]. These authors suggest a reconceptualization of (1) curriculum activities and (2) student achievement based in various indicators of performance. The previous researcher mentioned that assessment is an important part to test and control the learning process of learners [7]. The author stated that in adaptive learning, the suitable adaptive assessment methods are needed to provide the learners with a consistent learning process.

This study aimed to propose adaptive self-assessment activities for improving Massive Open Online Courses (MOOCs) learning performance in the second language and the research questions (RQ) were constructed as follows:

RQ 1: What are the learning styles of learners that used the second language in MOOCs learning?

RQ 2: What are the cognitive styles of learners that used the second language in MOOCs learning?

RQ 3: What are the suitable adaptive self-assessment activities element designs for the student in second language MOOCs learning?

This paper is organized as follows: Section II is a literature review. Section III describes the methodology and method. In Section IV, we review the results, discuss the designs and propose an adaptive self-assessment based on elements of learner characteristics in MOOC assessment. Finally, Section V gives a conclusion of the main finding in this paper.

## II. Literature Review

Learner characteristics are important in all forms of online learning [8]. Most researchers focused predominantly on user perception of MOOC features, rather than individual learner characteristics such as exploring the factors which affect MOOC completion or learner retention [9]. However, Judy and Jean [10] listed three types of learner characteristics that are effective factors for student learning: (1) learning style, (2) cognitive style and (3) multiple intelligence.

### *A. Learning Styles*

According to Ali [11], learning styles are the ways of receiving and responding to a learning stimulus with (1) unique psychological, (2) affective and (3) cognitive composition. Learning styles refer to the variations in an individual's ability to accumulate and assimilate information, sensory preferences that have the impact on learning and related to personality [10], [11], [12]. Previous researchers mentioned that learning styles refer to the way learners receive and perceive the information [13]. Learning styles are authentic as they are the appropriate techniques or methods in which learners learn, comprehend and get information [14].

TABLE I. ELEMENTS FOR THE DIMENSION OF LEARNING STYLES

| Dimension | Element | Author | | | | |
|---|---|---|---|---|---|---|
| | | [14] | [13] | [15] | [16] | [17] |
| Active | Learn by trying things | | | X | | |
| | Can be impulsive | X | | | | |
| | Risk-takers | X | | | | |
| | Do not prefer lectures | X | | | | |
| | Interpersonal | X | | | | |
| | Prefer group work | X | X | X | X | X |
| | Participate in learning tasks immediately | | | | X | X |
| Reflective | Think about concepts before action | X | | | | |
| | Like writing | X | | | | |
| | Not inclined to too much note-taking | X | | | | |
| | Learn by thinking | X | X | X | | |
| | Intrapersonal & introspective | X | | | | |
| | Adopt analytical approach | | | | X | X |
| | Working alone | | X | X | X | X |
| Sensing | Learn from concrete material | X | X | X | | |
| | Prefer facts | | | | X | X |
| | Follow tutors' approaches in problem-solving | | | | X | X |
| Intuitive | Prefer to learn abstract material. | X | X | X | X | X |
| | Learn and tend to apply their own innovation | | | | X | X |
| Visual | Learn best from what they see | | | X | | |
| | Prefer pictorial materials | X | X | | X | X |
| Verbal | Prefer to learn from words | X | X | X | | |
| | Prefer written and listen | | | | X | X |
| Sequential | Prefer step in a linear way | | | X | X | X |
| | Logically sequenced steps | X | | | | |
| | Focus on details | | | | | X |
| | Learn in continual small steps | | X | | | |
| Global | Prefer to learn in large leaps, skipping, understand and look at detail | | | X | X | X |
| | Learn holistically in large jumps | X | X | | | |

As mentioned by Mohamad [18], each student has his or her own learning style to be considered during the learning process. Ali [11] listed three important elements for learning styles: (1) academic achievements, (2) attitudes towards learning and (3) multimedia technology. Sadhasivam and Babu [14] listed a few learning style models to build up their pedagogical hypothesis: (1) Kolb Experiential Learning Theory; (2) VARK Model, (3) Felder & Silverman Learning/Teaching Style Model and 4) Dunn and Dunn Learning Style Mode. The previous researcher mentioned that the most appropriate model for open learning is Felder and Silverman Learning Style Model [13].

Previous studies have found four different dimensions of learning styles: (1) processing (active/reflective), (2) perception (sensory/intuitive), (3) input (visual/verbal) and (4) understand (sequential/global) [13], [19], [15]. Rohaniyah [20] stated the major differences in learning styles are the (1) the way a people perceive (sensation versus intuition), (2) the way they made a decision (logical thinking versus imaginative feelings) and (3) how active or reflective for interacting (extroversion versus introversion). Table I shows elements for the dimension of learning styles. Researchers highlighted that there is a lack of studies investigating how learning styles affect to the students' reactions in the context of assessment [21]. This finding shows visual learning styles was the control item and suggested making use of different learning styles as an independent variable.

### *B. Cognitive Styles*

TABLE II. ELEMENTS FOR THE DIMENSION OF COGNITIVE STYLES

| Dimension | Elements | Author | | |
|---|---|---|---|---|
| | | [10] | [22] | [23] |
| Extrovert | Outer world | X | | X |
| | Collaborate with others | | X | |
| | Try things out for himself | | | X |
| Introvert | Inner world | X | | X |
| | Independent | | X | X |
| | Reflect on thoughts and ideas | | | X |
| Sensing | Real objects and solid facts. | X | | X |
| | Physical qualities and affection by other information | | X | |
| | Rely on past experiences | | | X |
| Intuitive | Possibilities and personal meaning | X | | |
| | Intuitive types | | X | |
| | Speculations | | | X |
| Thinking | Analyzing fact | X | | X |
| | Structure and function | | X | |
| | Logical and rational decisions | | | X |
| Feeling | Subjective values and views | X | | X |
| | Initial energetic condition and interactions | | X | |
| | Decisions based on the people and their actions | | | X |
| Judging | Planned | X | | |
| | Organized way | X | X | X |
| | Prefers control | X | | |
| | Seek closure | | X | |
| | Think sequentially | | X | |
| | Orderly | | | X |
| Perceiving | Flexible | X | X | X |
| | Spontaneous way of life | X | | X |
| | To keep things open ended | | X | |

Previous research listed five definition for cognitive styles, namely (1) the way people think, (2) the accuracy of their perception, (3) how they process and remember information, (4) how they use the information in problem-solving or (5) how they organize and process information [18] and [19]. The

previous study listed three parts of cognitive style: (1) an individual's stability, (2) characteristic of the mental approach, or (3) cognitive preferences [20]. The author has proposed three elements of the cognitive style dimension: (1) processing information, (2) solving problems and (3) making decisions.

Cognitive styles are preferences or strategies used by individuals that influence functions such as (1) perceiving, (2) remembering, (3) thinking and (4) problem solving [24]. Previous researcher has used a cognitive style questionnaire proposed by Ancona [25], and the cognitive style dimensions are: (1) energy (extroversion/introversion), (2) information (sensing/ intuitive), (3) decisions (thinking/feeling) and (4) lifestyle (judging/perceiving) [26]. Researchers conclude that the student's cognitive style impacts their performance via e-activities. Therefore, in this study, we focus this questionnaire to investigate the type of learner dimension for the student using MOOCs learning and proposed the suitable self-assessment activities. Table II shows elements for the dimension of cognitive styles.

### C. Assessment

MOOC in education must have three requirements; (1) assessment (2) instructor and (3) model [27]. The researchers list two factors caused a lack of assessment to student achieved learning outcomes in higher education, which are (1) limitation of experiential learning and (2) many high-impact practices [28]. Gikandi et. al. [29] stated the term of assessment refer to the measurement of the learner's achievement and progress in a learning process. Two major types of assessment exist, which are (1) formative and (2) summative assessments. Formative assessment refers to the continuing feedback aiming to advance teaching and learning [30], while summative assessments, on the others hands refers to measure students' learning progress at the end of learning [31]. Previous researchers stated that the quality of both assessments was moderate [32]. In MOOC assessment, the issue is, there was a reduction of interest and activity of students during the session of the course [5].

Chan and King [33] mentioned, one of the most challenging problems in MOOCs is that it is infeasible for the teaching staffs to grade all the assignments on such a large scale. The design of engaging and challenging assessment tasks is one of the most important elements of planning a course [34]. The is a limitation of experiential learning and many high-impact practices in higher education are lacking assessment for embedded learning outcomes at the individual learner level [28]. Based on the previous study, in an attempt to avoid focusing on a broad assessment issue in of MOOC as a phenomenon, the researchers limited each case study by time and course design and activities [1]. Authors listed that common activities amongst all courses were: (1) Video lectures, (2) demonstrations, (3) discussion forums, (4) quizzes, (5) projects, (6) PowerPoint slides, (7) reading the material and (8) online learning resources. In this study, researchers focus on five types of activities designed: (1) quiz, (2) listening assessment, (3) forum, (4) mid-term test and (5) project.

| Assessment features/MOOC Platform | edX | coursera | miríada | KHAN ACADEMY | openlearning |
|---|---|---|---|---|---|
| Hints, adaptive exercises, etc. | ↑ | - | - | ↑ | ↓ |
| Automatic correction of exercises with artificial intelligence | → | - | - | - | ↓ |
| Interactive exercises | ↑ | - | - | → | → |
| Peer review | ↑ | ↑ | ↑ | - | → |
| Gamification | - | - | - | → | ↓ |
| Adaptation of the learning process and recommendations | - | - | - | ↓ | ↓ |
| Authoring tools for complex assessment activities | ↑ | - | - | - | ↓ |
| Learning analytics | ↓ | ↓ | ↓ | ↑ | ↑ |
| Social interaction functionality | ↑ | ↑ | → | ↓ | ↑ |
| Setting students own learning objectives and goals | - | - | - | → | - |

Adapted from (Muñoz-Merino et al., 2015)

Fig. 1. Comparison of the assessment functionality for MOOCs

All of the MOOC platforms can greatly improve the automatic correction of exercises using artificial intelligence techniques, the adaptation of assessment activities or the gamification features, although some platforms have already done some initial steps in these directions. The challenges when applying these online assessment methods are (1) student's engagement and (2) motivation to complete the e-activities. The researcher suggested two requirements to improved assessment in MOOC platform (1) new developments must make a perfect solution for assessment and (2) need to improve in the different aspects and features [35]. Fig. 1 shows the comparison of the assessment functionality for MOOCs between popular MOOC platforms.

Taras [36] also highlighted that self-assessment has been shown to support student learning. The author mentioned that the term 'self-assessment' is used to cover all judgments by learners of their work which subsumes terms such as (1) 'self-evaluation' and (2) 'self-appraisal'. However, models used for adaptive assessment have been mostly summative: they measure or rank effectively examinees, but do not provide any other feedback [37]. Gohokar [38] listed three requirements for improving learning performance of the course (1) adaptive, (2) innovative and cognitive tools and (3) techniques. In an assessment, adaptive can lead to improved personalization, by organizing learning resources [37]. The authors mentioned adaptive assessment is more and more useful in the current age of MOOCs, where motivation plays an important role. In this study, the aim is to propose adaptive self-assessment activities in MOOCs learning in the second language course based on learner characteristics to improving student performance.

## III. Materials and Methods

This part described the research methodology and the research design of this study. This part also addressed aspects of research design such as theoretical framework and proposed assessment activities.

### A. Questionnaire

The questionnaire consists of 2 part and 24 survey items measuring the different types of learner characteristics and 50 students participated as the respondents for the study. The survey items were adapted from several research projects with similar research scope [25] and [39]. The questionnaires were distributed and instructions were given to the respondents. Participation was voluntary. All participants were given the survey via by hand.

### B. Research Instrument

There are many ways to determine learner styles strengths dimension. The survey items were adapted from several research projects with similar research scope [25] and [39] had been created for this purpose. Questionnaire method was used as the research instrument in this study. This questionnaire is used to determine which learning styles and cognitive styles dimension are the strongest for technical student. Once the items of the survey instrument were scored, the points for each of the learning styles and cognitive styles dimension were totaled for each student using the Ms. Excel.

### C. Theoretical Framework

Fig. 2 illustrated the theoretical framework of this paper. Exclusively, the theoretical framework defined the learner characteristics as the independent variable and dependent variable are student performance. The learner characteristics consist of two construct: (1) learning styles and (2) cognitive styles. The student performance are measured the mark for each MOOC assessment such as mid-term test.

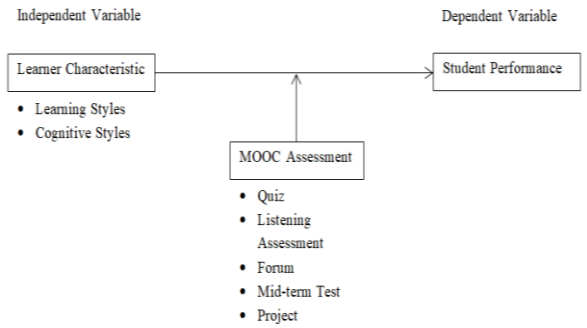



Fig. 2. The theoretical framework

## IV. Results and Discussion

In this part, the results of data analysis are presented from both methods. The findings are presented, analyzed and discussed based on the research question.

### A. Research Question 1: What are the Learning Styles of Learners that used Second Language in MOOCs Learning?

The findings of the questionnaires indicated visual as the highest percentage of learning styles for students based on the Felder and Silverman Learning Style Model in Mandarin MOOC. According to the results of the learning styles questionnaire, Fig. 3 shows the dimension of learning styles by percentages in MOOCs learning. Visual and active learner dimension has the highest percentage, values of 76.00% and 74%, respectively.

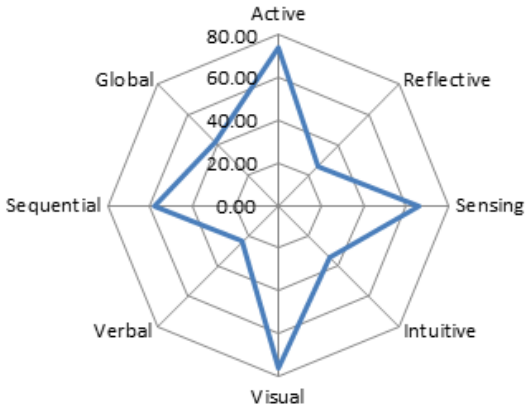



Fig. 3. Dimensions of learning styles

The previous researcher stated that the stronger the visual style of learning was, the more time the learners spent on pictorial content [24]. Most researchers mentioned that visual learners outperformed verbal learners such as focused on pictorial learning objects and learn best from what they see [15] and [24]. Visual learners almost prefer video, demonstrations, pictures and graphs [19]. The visual persons like to receive information through pictures and diagrams [40], [41]. Therefore these learners prefer the use of pictures, imageries and spatial perception [40]. Zhong [42] stated that online students are mostly visual learners while the traditional student are mostly auditory or kinesthetic learners. For example, almost all students from Asian cultures are highly visual learner [40].

Active learners are defined the as learners who prefer to process information actively by doing something with the learned material, for example discussing, explaining or testing [43]. Active learners acquire knowledge best by trying things out and working with others in groups [41]. Active learners prefer studying in groups, immediately doing learning tasks, working with others, learn by doing something with information and they prefer to process information by talking about it and trying it out [13], [14], [17]. The previous researcher found that active learners possessed significantly higher intention towards an interactive learning technology (wiki) than the reflective group [17]. However, active learners did not like to be passive participants in educational activities. The researcher listed two requirements needed to support student engagement and guide the student to become an active learner, namely: (1) course tools and (2) media [44].

Based on the literature review, two elements in the visual learner dimension were identified: (1) learn best from what they see, and (2) learners priorities pictorial materials. While seven elements in the active learner dimension were identified: (1) learn by trying things, (2) can be impulsive, (3) risk-takers, (4) do not prefer lectures, (5) interpersonal, (6) prefer group work and (7) participate in learning tasks immediately. Based on findings from RQ1, the researchers proposed an existing element learning styles (visual and active) in adaptive self-assessment activities for improving student performance using MOOCs learning for the second language course.

### B. Research Question 2: What are the Cognitive Styles of Learners that used Second Language in MOOCs Learning?

Based on the results of the learner characteristics questionnaire, the dimension of cognitive style by percentages in MOOCs learning is shown in Fig. 4 below. Thinking and intuitive learner dimension have the highest percentage, values of 70.00% and 64%, respectively. Ghaedi and Bashir [45], mentioned thinking learner tends to complete their work in an organized and efficient manner. Intuitive learners often prefer discovering possibilities and relationships [46].

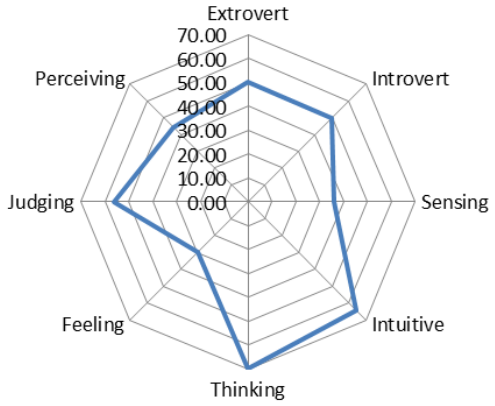



Fig. 4. Dimensions of cognitive styles

Based on the literature review, three elements in the thinking style were identified: (1) analyzing fact, (2) structure and function and (3) logical and rational decisions. While three elements in the active learner dimension were also identified: (1) possibilities and personal meaning, (2) intuitive types and (3) speculations. Based on findings from RQ2, the researchers proposed an existing element cognitive style (thinking and intuitive) in adaptive self-assessment activities for improving student performance using MOOCs learning for the second language course.

### C. Research Question 3: What are the Suitable Adaptive self-Assessment Activities Element Designs for the Student in Second Language MOOCs Learning?

To answer RQ3, the researchers focused on designing MOOC self-assessment activities for two selected dimensions of learning styles and two selected dimensions of cognitive styles. Based on the findings, the data indicate that visual, active, thinking and intuitive learner is the highest dimension for both learner characteristics. Once the dimensions have been determined, the MOOC topic learning outcomes were used as guidance in selecting which types of MOOC self-assessment activities can be incorporated with which learning styles and cognitive styles dimensions. Table III shows the dimensions of both styles that can be adapted into potential MOOC self-assessment activities to improve student performance in MOOCs learning. In this table, the design of a question for this self-assessment activity (e.g. Quiz) will use the dimension of learning styles (e.g. Visual).

TABLE III. DIMENSIONS OF BOTH STYLES

| Learner Characteristics | Dimension | MOOC Self-Assessment Activities |
|---|---|---|
| Learning Styles | Visual | Quiz: Multiple Choice (MC), True/False (T/F), Drag & Drop, Match |
| | Active | Quiz: Multiple Choice (MC), True/False (T/F), Drag & Drop, Match<br>Listening Assessment<br>Forum<br>Mid-Term Exam<br>Project: Writing, Presentation |
| Cognitive Styles | Thinking | |
| | Intuitive | |

Results of further analysis on each question/item in self-assessment activities for each element in learning styles and cognitive styles dimensions are presented in Table IV. In this table, a shaded box means that the design of a question for this self-assessment activity (e.g. T/F) will incorporate the element of its associated learning styles (e.g. Trying things).

TABLE IV. SELF-ASSESSMENT ACTIVITIES VERSUS ELEMENT OF DIMENSION IN MOOC PLATFORM

| Dimension | Element | Quiz | | | | Listening Assessment | Forum | Mid-Term | Project | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MC | T/F | Drag & Drop | Match | | | | Writing | Presentation |
| Visual | See | | | | | | | | | |
| | Pictorial materials | | | | | | | | | |
| Active | Trying things | | | | | | | | | |
| | Impulsive | | | | | | | | | |
| | Risk-takers | | | | | | | | | |
| | Do not prefer lectures | | | | | | | | | |
| | Interpersonal | | | | | | | | | |
| | Group work | | | | | | | | | |
| | Tasks immediately | | | | | | | | | |
| Thinking | Analysing fact | | | | | | | | | |
| | Structure and function | | | | | | | | | |
| | Logical and rational decisions | | | | | | | | | |
| Intuitive | Possibilities and personal meaning | | | | | | | | | |
| | Existence intuitive types | | | | | | | | | |
| | Speculations | | | | | | | | | |

| Color code | Activities | Color code | Activities | Color code | Activities |
|---|---|---|---|---|---|
| | Multiple Choice (MC) | | Match | | Mid-Term |
| | True/False (T/F) | | Listening Assessment | | Writing |
| | Drag & Drop | | Forum | | Presentation |

Visual learner dimensions divided by four of the adaptive self-assessment activities for MOOCs: (1) Multiple Choice (MC), (2) True/False (T/F), (3) Drag & Drop and (4) Match. Meanwhile, active , thinking and intuitive learner dimensions divided by nine of the adaptive self-assessment activities for MOOCs: (1) Multiple Choice (MC), (2) True/False (T/F), (3)

Drag & Drop, (4) Match (5) Listening Assessment, (6) Forum, (7) Mid-Term, (8) Writing and (9) Presentation. Further analysis, the researchers focused on designing MOOC self-assessment activities using selected MOOC functional features that are suitable for a particular learner characteristic dimension such as hints and interactive. The design of a question for this self-assessment activity (e.g. MC) will incorporate the feature of its associated learner characteristic (e.g. 'Hints' in 'Graphics' form).

## V. Conclusion

This initial study identified the learner characteristics of learners using MOOC assessment. The findings of the questionnaire indicated visual and active as the learner dimension for learning styles. While the learner dimension for cognitive styles is thinking and intuitive. The results from the percentage analysis revealed that out of the eight dimensions of learning styles, the preferred dimensions were visual and active learner (76% and 74%). As for cognitive styles, the preferred dimensions were thinking and intuitive learner (70% and 64%). The suitable adaptive self-assessment activities design was also proposed for improving student performance in MOOCs learning for the second language course. In the future study, students will be investigated about their engagement using MOOC assessment in the second language course.

## Acknowledgment

This research was conducted by the Pervasive Computing & Educational Technology Research Group, C-ACT, Universiti Teknikal Malaysia Melaka (UTeM), and supported by the Ministry of Higher Education (MOE). FRGS grant: FRGS/1/2016/ICT01/FTMK-CACT/F00327.

## References

[1] A. Stanley and M A. Stanley and M. Moallem, "Assessment in MOOCs: A Comparative Analysis," in Proceedings of Society for Information Technology & Teacher Education International Conference 2016, 2016, pp. 394–400.

[2] B. T. Wong, "Factors leading to effective teaching of MOOCs," Asian Assoc. Open Univ. J., vol. 11, no. 1, pp. 105–118, 2016.

[3] T. Tenório, I. I. Bittencourt, S. Isotani, and A. P. Silva, "Does peer assessment in on-line learning environments work? A systematic review of the literature," Comput. Human Behav., vol. 64, pp. 94–107, 2016.

[4] Z. Wang, "Structured Knowledge Tracing Models for Student Assessment on Coursera," in Proceedings of the Third (2016) ACM Conference on Learning @ Scale, 2016, pp. 209–212.

[5] L. S. Lisitsyna and A. E. Evgenii, "Making MOOCs More Effective and Adaptive on Basis of SAT and Game Mechanics," Smart Educ. e-Learning 2017, Smart Innov. Syst. Technol. 75, vol. 75, 2017.

[6] O. Pilli and W. Admiraal, "Students' Learning Outcomes in Massive Open Online Courses (MOOCs): Some Suggestions for Course Design," J. Higher Educ., vol. 7, no. 1, pp. 46–71, 2017.

[7] F. E. Louhab, A. Bahnasse, and M. Talea, "Towards an Adaptive Formative Assessment in Context-Aware Mobile Learning," in Procedia Computer Science, 2018, vol. 135, pp. 441–448.

[8] Y. Park, I. Jung, and T. C. Reeves, "Learning from MOOCs: a qualitative case study from the learners' perspectives," EMI. Educ. Media Int., vol. 52, no. 2, pp. 72–87, 2015.

[9] K. S. Hone and G. R. El Said, "Exploring the factors affecting MOOC retention: A survey study," Comput. Educ., vol. 98, pp. 157–168, 2016.

[10] J. Lever–Duffy and J. B.McDonald, Teaching and Learning With Technology, 3rd ed. Boston: Pearson, 2009.

[11] R. Ali, "Learning Style Construct in Student ' s Learning," J. Indones. untuk Kaji. Pendidik., vol. 1(2), no. November, pp. 213–222, 2016.

[12] M. E. R. Abante, B. C. Almendral, J. E. Manansala, and J. Mañibo, "Learning Styles and Factors Affecting the Learning of General Engineering Students," Int. J. Acad. Res. Progress. Educ. Dev., vol. 3, no. 1, pp. 16–27, 2014.

[13] H. Fasihuddin, G. Skinner, and R. Athauda, "Towards adaptive open learning environments: Evaluating the precision of identifying learning styles by tracking learners' behaviours," Educ. Inf. Technol., vol. 22, no. 3, pp. 807–825, 2017.

[14] J. Sadhasivam and R. Babu, "MOOC : A Framework for Learners Using Learning Style," Int. Educ. Res. J., vol. 3, no. 2, pp. 21–24, 2017.

[15] B. Hmedna, A. El Mezouary, O. Baz, and D. Mammass, "A Machine Learning Approach to Identify and Track Learning Styles in MOOCs," in International Conference on Multimedia Computing and Systems (ICMCS'16), 2016.

[16] A. Al-Azawei, A. Al-bermani, and K. Lundqvist, "Evaluating the Effect of Arabic Engineering Students ' Learning Styles in Blended Programming Courses," J. Inf. Technol. Educ. Res., vol. 15, no. April, pp. 109–130, 2016.

[17] A. Al-azawei, P. Parslow, and K. Lundqvist, "Investigating the effect of learning styles in a blended e-learning system : An extension of the technology acceptance model ( TAM )," Australas. J. Educ. Technol., vol. 33, no. 2, pp. 1–23, 2017.

[18] S. N. M. Mohamad, "Model for Online Teaching Tools Based on Interpersonal, Visual and Verbal Intelligence," Universiti Teknikal Malaysia Melaka, 2014.

[19] A. Al-Azawei, P. Parslow, and K. Lundqvist, "Investigating the effect of learning styles in a blended e-learning system: An extension of the technology acceptance model (TAM)," Australas. J. Educ. Technol., vol. 33, no. 2, pp. 1–23, 2017.

[20] J. Rohaniyah, "Integrating Learning Style and Multiple Intelligences in Teaching and Learning Process," J. Pemikir. Penelit. Pendidik. dan Sains, vol. 5, no. 1, pp. 19–27, 2017.

[21] T. C. Hsu, S. C. Chang, and N. C. Liu, "Peer assessment of webpage design: Behavioral sequential analysis based on eye tracking evidence," Educ. Technol. Soc., vol. 21, no. 2, pp. 305–321, 2018.

[22] S. Mukherjee, "Learning Style of Humanities , Commerce and Science Students : A Study on Higher Secondary Students from West Bengal," Int. J. Indian Psychol., vol. 3, no. 3, pp. 15–22, 2016.

[23] L. D. Garner-O'Neale and S. Harrison, "An Investigation of the Learning styles and Study Habits of Chemistry Undergraduates in Barbados and their Effect as Predictors of Academic Achievement in Chemical Group Theory," J. Educ. Soc. Res., vol. 3, no. May, pp. 107–122, 2013.

[24] M. Koc-Januchta, T. Hoffler, G.-B. Thoma, H. Prechtl, and D. Leutner, "Visualizers versus verbalizers: Effects of cognitive style on learning with texts and pictures - An eye-tracking study," Comput. Human Behav., vol. 68, pp. 170–179, 2017.

[25] Ancona, Kochan, Scully, V. Maanen, and Westney, "Managing For the Future," 1997. [Online]. Available: http://www.analytictech.com/mb021/cogquest.htm. [Accessed: 30-Apr-2017].

[26] J. Simuth and I. Sarmany-Schuller, "Cognitive Style Variable in E-learning," in Procedia - Social and Behavioral Sciences, 2014, vol. 116, no. January 2013, pp. 1464–1467.

[27] D. Baneres, S. Caballé, and R. Clarisó, "Towards a Learning Analytics Support for Intelligent Tutoring Systems on MOOC Platforms," in Proceedings - 2016 10th International Conference on Complex, Intelligent, and Software Intensive Systems, CISIS 2016, 2016, no. October, pp. 103–110.

[28] W. F. Heinrich and J. E. Rivera, "A Framework for PLA in Traditional Higher Education : Experiential Learning Assessment for Embedded Outcomes," Prior Learning Assessment Inside Out, vol. 5, no. 5, pp. 1–5, 2016.

[29] J. W. Gikandi, D. Morrow, and N. E. Davis, "Online formative assessment in higher education: A review of the literature," Comput. Educ., vol. 57, no. 4, pp. 2333–2351, 2011.

[30] Z. G. Baleni, "Online Formative Assessment in Higher Education: Its Pros and Cons.," Proc. Eur. Conf. e-Learning, vol. 13, no. 4, pp. 43–47, 2014.

[31] OECD, "Assessment For Learning - The Case for Formative Assessment," in OECD/CERI International Conference - Learning in the 21st Century: Research, Innovation and Policy, 2008, pp. 1–24.

[32] W. Admiraal, B. Huisman, and M. Van De Ven, "Self- and Peer Assessment in Massive Open Online Courses," Int. J. High. Educ., vol. 3, no. 3, pp. 119–128, 2014.

[33] H. P. Chan and I. King, "Leveraging Social Connections to Improve Peer Assessment in MOOCs.," WWW (Companion Vol., pp. 341–349, 2017.

[34] M. Bearman, P. Dawson, D. Boud, M. Hall, S. Bennett, E. Molloy, and G. Joughin, "Guide to the Assessment Design Decisions Framework," 2014.

[35] P. J. Muñoz-Merino, J. A. Ruipérez-Valiente, J. L. Sanz, and C. Delgado Kloos, "Assessment Activities in MOOCs," 2015.

[36] M. Taras, "Student self-assessment: Processes and consequences," Teach. High. Educ., vol. 15, no. 2, pp. 199–209, 2010.

[37] J.-J. Vie, F. Popineau, É. Bruillard, Y. Bourda, and Jill-Jênn, "A Review of Recent Advances in Adaptive Assessment," Learn. Anal. Fundam. Appl. Trends, vol. 94, pp. 113–142, 2017.

[38] V. N. Gohokar, "Adaptive , Cognitive and Innovative Tools & Techniques for improving Learning Performance : A Case Study," in International Conference on MOOC, Innovation and Technology in Education (MITE), 2014, pp. 47–52.

[39] T. Kintakaningrum, "A Study Of Consequences On Individual And Group Learning Performance Using A Web-Based And Mobile Supported Learning Management System," Universiti Teknikal Malaysia Melaka. Malaysia, 2012.

[40] K. Karthigeyan and K. Nirmala, "Learning Style Preference of English Language Learners," Educ. Confab J., vol. 2, no. 1, pp. 134–140, 2013.

[41] A. E. Labib, J. H. Canós, and M. C. Penadés, "On the way to learning style models integration: a Learner's Characteristics Ontology," Comput. Human Behav., vol. 73, pp. 433–445, 2017.

[42] S. H. Zhong, Y. Li, Y. Liu, and Z. Wang, "A computational investigation of learning behaviors in MOOCs," in Computer Applications in Engineering Education, 2017, no. November 2016, pp. 1–13.

[43] S. Graf, Kinshuk, and T. C. Liu, "Identifying learning styles in learning management systems by using indications from students' behaviour," in Proceedings - The 8th IEEE International Conference on Advanced Learning Technologies, ICALT 2008, 2008, pp. 482–486.

[44] P. R. Lowenthal and C. B. Hodges, "In search of quality: Using Quality Matters to analyze the quality of massive, open, online courses (MOOCs)," Int. Rev. Res. Open Distrib. Learn., vol. 16, no. 5, pp. 83–101, 2015.

[45] Z. Ghaedi and B. Jam, "Relationship between Learning Styles and Motivation for Higher Education in EFL Students," in Theory and Practice in Language Studies, 2014, vol. 4, no. 6, pp. 1232–1237.

[46] P. Tulsi, M. Poonia, and P. Anu, "Learning Styles of Engineering Students," J. Eng. Educ. Transform. , vol. 30, no. 2, pp. 44–49, 2016

# Consequences of Customer Engagement in Social Networking Sites: Employing Fuzzy Delphi Technique for Validation

Noraniza Md Jani[1], Mohd Hafiz Zakaria[2], Zulisman Maksom[3]
Fakulti Teknologi Maklumat dan Komunikasi,
Universiti Teknikal Malaysia Melaka, Malaysia

Md. Shariff M. Haniff[4]
Faculty of Business and Management,
Universiti Teknologi MARA
(Alor Gajah Campus), Melaka, Malaysia

Ramlan Mustapha[5]
Institute of Teacher Education, Tengku Ampuan Afzan Campus, Kuala Lipis, Pahang,
Malaysia

***Abstract*—The consequences of the customer engagement in the Social Networking Sites (SNS) community have direct impact on the brand. This present research was conducted to examine the cohesive mechanisms for item verification on the most influential consequences of participating the brand community and joining the electronic Word-of-Mouth (eWOM) as the manifestation of the behavior of such communities. Using Fuzzy Delphi techniques, a total of 12 heterogeneous experts are involved in the verification process through a 7-point linguistic scale of the questionnaire survey. The results show good evidence of expert consensus by reaching 75% for each consequence of the engaged customers. On the SNS platform, further aspects of the inspected effects can be expanded to be studied on relevant domains. Practitioners will be more strategic in maintaining and fostering customer relationships, and consistently influencing new customers when interacting actively through SNS brand pages.**

***Keywords*—Customer engagement; fuzzy delphi; SNS; consequences; brand page***

## I. Introduction

The explosive growth of Social Networking Sites (SNS) is the basis for firms to open space for customer engagement in the proliferation of brand-related ideas, information, and thought surrounding their network venture. In the circle of electronic Word-of-Mouth (eWOM) activities, customers and prospects use SNS not just to acquire information but to engage with brand and products [1]-[2]. Customers are equipped with the brand pages that are created by firms for marketing and communication activities related to brand and products. These pages are able to tie existing customers and catch the prospect's attention due to the firms' series of repeated and effective strategies, for an instance, posting the good copywriting to advertise a newly-launched product.

Notwithstanding the evidence, engaging customers and ensuring their commitment to the online community is a critical task for the brand to succeed. The online marketing initiatives has put a big challenge on most brands in the aspects of broadcast advertisements, messages, and campaigns by firm marketers that are usually influenced by community partners and ultimately alter the customer's decision beyond the transaction [3]-[5]. The postal content generated by marketers via the brand pages are greatly affected by various types of data manipulated through User-Generated Content (UGC) in the form of audio, video, images, animation and text [6]. Such activities are the reaction of customers that occur when they are stimulated by the factors around their communities.

The core of the relationship between the customer, the firm, and the brand is through the virtual customer engagement [7]-[8]. In particular, there is a demand to provide a broader view of the online community in the social media landscape [9]-[11]. The concept of customer engagement is the need for a brand to handle online affairs more effectively and overcome the firm's challenge to understand the exact mechanism to measure online customer preferences and reactions.

Customer engagement is an evolving concept of relationship marketing theory that stands as a key role motivated by various drivers, thereby delivering impact on other consequences (see [12]-[16]). In consumer behaviour research, this principle theory is anchored with service-dominant logic and social exchange theories in which are mutually complementing in terms of the body knowledge contribution. The lens of service-dominant logic has centralized the notion of customer engagement on the value co-creation of firm-customer exchange in the network of interactive service system [17]-[18]. According to [19]'s social exchange theory, customers are predicted to respond positively to their thoughts, feelings, and behaviors towards the brand as they are motivated by certain future returns from established obligations [20]. Communities interacting with SNS have acknowledged the tendency of customers to be grateful and reciprocate by repaying the benefits they acquired in the process of exchanging with the brand.

Dominantly, online customer engagement can be assessed through three main dimensions - cognitive, affective and

behavioural – as treated by various researchers in different field areas [10], [21]. This study, however, argues about the importance of customer behavior within the SNS community. The behavioral aspect profoundly seen as priority since it serves as a strong indicator of customer engagement in social media and supported by many claims in previous studies [8], [22]-[24]. Manifesting behavior from such customers includes all kinds of behaviours that they translate beyond purchase arrangements; such as suggestions for service improvements, product purchasing insights, search of brand and product information, and sharing of usage experience with other customers. Each of their behaviours in the app interface panel greatly determines the involvement of other community members to the brand.

It is clear that the study referred to customer engagement within social media community is increasingly important [2], [25]-[26]. These scholars focus on examining the intangible indicators that pose impacts to the engagement domain, for examples reputation, advocacy, and others. In spite of this, literature recognizes the lack of agreement on research of the consequences of customer engagement [21]. While the impact of customer engagement in the communities studied is more important as it determines the restructuring of a more effective strategy by the brand for achieving better business objectives, scientific research investigating the behavior of customer engagement in the SNS is rarely implemented [27]. In the similar consonant, authors call for more research to gain insights on social media use as an effective integrated marketing tools [28]-[30]. The development and verification of indicators to assess the involvement of the community in SNS activities can meet the claims of those researchers.

Therefore, to encourage future research on the consequences of customer engagement in behavioural dimension within the SNS community, it is essential to have a systematic approach to validate the indicators. The study seeks to obtain a consensus of experts relating to the development and validity of the proposed consequences of customer engagement due to the scarcity of works conducted to this date. The effective tool that will be used in this case is the Delphi Fuzzy technique in which is expected to work very well as the validity measurement of the variables studied. Thus, the objective of this study is:

*What is the level of agreement among experts towards the indicators for evaluating the behavioral manifestation from the engaged customer (consequences) via the SNS platform?*

The rest of the article presents the conceptualization of customer engagement in SNS held by this study, followed by the methodology section to elucidate the implementation of Fuzzy Delphi approach to achieve the above objective.

## II. Conceptual Framework

Firms must accommodate opportunities to encourage customers in the participation of product review, exchange information, development, and innovations [31]-[32]. These activities can be reached when members of the online community continue to communicate in the direction of bilateral in their post-engagement in SNS through the eWOM network. Brand proactive activities will increase the participation of more customers in the community, in turn will add the concept of sharing and dissemination of brand information. Subsequently, customers involved in the SNS platform are believed to have shown certain reactions of behavior as a result of intent to show their involvement.

This study put a debate on the increases of eWOM and the active participation in brand community as the prominent consequences that symbolizes the behavioral of engaged customers in SNS. Furthermore, knowing the consequences are even more important for firms to get valuable insights of what outcomes after customers are considered as engaged person, hence giving input to their online marketing strategies to be reviewed and optimized as to making them parallel with the business goals. The conceptualization of customer engagement and it's consequences in the SNS brand page is illustrated in Fig. 1 below.

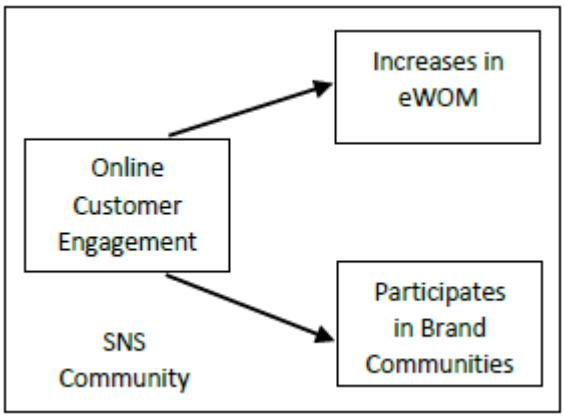



Fig. 1. Conceptual Framework of Customer Engagement in SNS

### A. Increases in Electronic Word-of-Mouth (eWOM)

Positively engaged customers are likely communicating in positive eWOM about firms, brands and products. In adjustment with the context of this study, eWOM was labelled as "SNS environment which create new forms of group conversations in communities, with large numbers of participants participate in different roles and stages, allowing potential and existing customers to positively express statement and exchange information about a brand, product or a firm" [33]-[35].

In the literature of marketing and communication, two important dimensions that conceptualize social influences of customers' attitudes and behaviors in the eWOM process are opinion leadership and opinion seeking [35]-[36]. The exceptional growth of SNS offers opinion leaders with the ability to strengthen their interpersonal characteristics by motivating other community fellows, hence will encourage the eWOM development [37]. Relatively posses lower product knowledge, opinion seekers actively search information and advice in a given category [35]. Another important dimension in eWOM is online pass-along behavior [38] in an attention to provide the community SNS members with a useful tool to exchange meaningful information and past worth experiences about brand and product [37].

### B. Participates in Brand Community

The first concept of brand community introduced to the marketing literature is proposed by [39], defined as "a specialized, non-geographically bound community, based on a structured set of social relationships among admirers of a brand". The engaged customers are experienced customers with the focal brand, hence, they highly possible to deliver some valuable knowledge-sharing information to advise and to influence community peers essentially by helping each other and fixing each other's problems with the brand [40]. The engaged customers' activities will attract and influence enormous potential customers who interest about the brand to participate and constructing larger groups, and performing many important tasks on behalf of the brand [18], [25].

As a quest to universally connecting all benefits of brand communities announced by prior studies, [26] had identified a diverse set of dimensions which is empirically tested to gain insight of brand communities' motivational drivers – brand influence, connecting, helping, like-minded discussion, seeking assistance, self-expression and up-to-date information. Adapted from these empirically tested measures, this study aims to test the relationship between customers engaged with brands with the SNS community in which they are located.

TABLE I. CONSEQUENCES OF CUSTOMER ENGAGEMENT IN SNS, CONCEPTUAL FRAMEWORK OF CUSTOMER ENGAGEMENT IN SNS

| Author(s) | Consequences of customer engagement | | | | | | | | Context, research type |
|---|---|---|---|---|---|---|---|---|---|
| | ** Participates in Brand Community* | ** electronic Word of Mouth (eWOM)* | *Brand Loyalty* | *Brand Satisfaction* | *Advocacy* | *Firm's performance* | *Online Review Intention* | *Brand Usage Intention* | |
| [12] | | √ | | | | | √ | | Offline, conceptual |
| [41] | √ | √ | | | | | | | Online social platform, conceptual |
| [7] | | | √ | √ | | | | | SNS, empirical |
| [42] | √ | √ | √ | | | | | | Offline, conceptual |
| [1] | √ | | | | | | | | Online community, conceptual |
| [14] | | | | | | | | √ | Social media, empirical |
| [27] | | | √ | | | | | | SNS, empirical |
| [8] | | | √ | | | | | | Social media, empirical |
| [16] | | | | | | | √ | | Mobile devices, empirical |
| [43] | | | | | √ | √ | | | Offline, conceptual |

In this study, the manifestations of customers involved in SNS is confined within a one-dimensional outcome - which is behaviour. Following the exploration of previous works (refer Table I where the constructs of behaviour dimension are marked with *), it is strongly believed that these two consequences could result in significant relationships with the domain of customer engagement using the Delphi Fuzzy method. To achieve results, the manuscript on hand has adjusted 9 items to *Increases in eWOM* from [37] and [44], as well as 24 items were drawn and coordinated from [26] to *Participates in Brand Communities* to measure the scale of customer engagement behavior as conceptualized in Fig. 1.

## III. METHODOLOGY

Fuzzy Delphi technique has been utilized by this study to validate the conceptual framework of customer engagement in SNS as well as adapted measurement scales. Inspiration of the use of Fuzzy Delphi is derived from the effectiveness of the results in various types of IT/IS research variations presented by [45]. This tool is a very helpful method when the level of research requires approval from a group of experts. In addition, the Fuzzy Delphi technique is an interesting method of collective decision-making about the blurred ideas of the alignment of experts' opinions [46]. Their responses can be implemented in a short time because of the use of survey methods in parallel with the lower cost and experts can fully express their true response without misinterpretation to ensure perfection and consistency of opinion [47].

The whole methodology is done by dividing the processes into four major stages. The first stage is prior to the Delphi Fuzzy technique is performed where the items to be certified by experts are developed. The second stage refers to the selection of 12 appropriate experts from the background of the academic and business fields. The third stage is related to data collection activities from designated experts within a certain period of time. The final stage is definitely an analysis process in terms of generating significant relationships between the latent variable for review, customer engagement and the two identified consequences. Table II(A) and II(B) illustrates the four levels mentioned.

### A. Instrument Development

The formation of a questionnaire instrument was conducted after the identification of two of the leading consequences of SNS's customer engagement (as presented in Table I). In order to meet the criteria of measurement items that are expected to yield significant results, these items are withdrawn from the relevant prior research. The statement of these questions has been modified to conform to the context and environment of the SNS medium to be studied. These 42 questions were reviewed in terms of content and format at the beginning by 3 academicians in the relevant field of social media and business. In conjunction with this, two language specialists, an English language lecturer and another Malay language lecturer were also invited to evaluate the suitability of the item presented. Some improvements have been made to the content of the question with the aim of enhancing the clarity and ease of formatting. Adaptive questions from previous studies with different mediums and samples can affect the reliability and validity of items to be tested in real surveys. Therefore, this present study has taken into account the views of practitioners (in this case the Small Medium Enterprises (SME)) through small-scale structured interview sessions to validate items captured from literature [45]. The examined set of items is then properly formatted into the questionnaire as outlined in Table II(A).

### B. Sampling and Selection of Respondents

The expert group required for this study is from those who meet the following criteria: knowledge and experience with issues being examined, ability and willingness to participate, qualification, individual character, comparative ability, communication skills as well as consistent and credible within their respective fields [46], [48]. Hence, this study has chosen the purposive sampling through the identification of 12 skilled experts in one or more of the following areas: IT, social media, online governance, marketing, and consumer behaviour.

TABLE II. (A).FUZZY DELPHY – STAGES OF INSTRUMENT DEVELOPMENT, SELECTION OF EXPERTS AND DATA COLLECTION, CONCEPTUAL FRAMEWORK OF CUSTOMER ENGAGEMENT IN SNS

**Instrument development**

| Content | Approach | Measurement items |
|---|---|---|
| A questionnaire form with 42 items and was divided into four sections:<br>1) Cover page<br>2) Demographic of respondents<br>3) Item questions in 7-Lickert scale<br>4) Spaces for comment | Adapted scales from relevant literature (refer Table I) | See nested table below |

| Latent variable | Adapted Scales | Number of item scales |
|---|---|---|
| Online Customer engagement | [14], [54], [55] | 9 items |
| **Consequences** | **Adapted Scales** | **Number of item scales** |
| Participates in Brand Communities | [26] | 24 items |
| Increases in WOM | [37], [44] | 9 items |
| **Total** | | **42 items** |

**Selection of experts and data collection**

| Technique | Approach | Classification of experts |
|---|---|---|
| Purposive sampling by identifying, contacting and confirming the experts to be participants | Form delivery<br>A scheduled and face-to-face meeting based on experts' availability<br><br>Form collection<br>Returned forms either by face-to-face or submission via email | See nested table below |

| Experts | Institution/Sector | Number of participants |
|---|---|---|
| Academicians | UTeM (IT domain) | 2 |
| | UTeM (Business domain) | 1 |
| | UITM (IT domain) | 1 |
| | UITM (Business domain) | 1 |
| | UPM (IT domain) | 1 |
| | Polytechnic (IT domain) | 1 |
| Practitioners (SMEs) | Manufacturing | 2 |
| | Services | 3 |
| **Total** | | **12** |

A total of 12 experts were taken from the suggestion of [49] which states that heterogeneous sampling requires between 10 and 35 experts. The experts who come from different profiles and backgrounds are made up of 7 academics and 5 practitioners (refer Table II(A) for the classification). By using email and follow-up via phone calls, they have been invited to take part in the survey and agreement have been dealt with as soon as they give consent. By scheduling different meetings, a face-to-face approach with these 12 designated experts are carried out to provide an explanation of the items reviewed if any issues arise. The questionnaires are then distributed and the item verification process from each expert is conducted based on the procedures outlined. This level ends with the collection of forms at different times where the items being evaluated are ready to be analyzed using the Fuzzy Delphi method in Microsoft Excel.

TABLE II. (B). FUZZY DELPHY – STAGES OF DATA ANALYSIS

<table>
<tr><th colspan="2">Data analysis</th></tr>
<tr><td>Step 1<br>The determination of triangular fuzzy numbers</td><td>
<table>
<tr><th>Linguistic variables</th><th colspan="3">Fuzzy scales</th></tr>
<tr><td>Strongly Disagree</td><td>0.0</td><td>0.0</td><td>0.1</td></tr>
<tr><td>Disagree</td><td>0.0</td><td>0.1</td><td>0.3</td></tr>
<tr><td>Somewhat Disagree</td><td>0.1</td><td>0.3</td><td>0.5</td></tr>
<tr><td>Neither agree or disagree</td><td>0.3</td><td>0.5</td><td>0.7</td></tr>
<tr><td>Somewhat agree</td><td>0.5</td><td>0.7</td><td>0.9</td></tr>
<tr><td>Agree</td><td>0.7</td><td>0.9</td><td>1.0</td></tr>
<tr><td>Strongly agree</td><td>0.9</td><td>1.0</td><td>1.0</td></tr>
</table>
All linguistic variables used in the questionnaire are transformed to triangular fuzzy numbers [50].</td></tr>
<tr><td>Step 2<br>The determination of average responses</td><td>Next, the average value of each expert's response will be calculated with the following equation (1):<br>$$m = \frac{\sum_1^n m_i}{n} \quad (1)$$</td></tr>
<tr><td>Step 3<br>The determination of threshold value (“d”)</td><td>The threshold value (“d”) between two Fuzzy numbers (m= $m_1$, $m_2$, $m_3$) dan (n= $n_1$, $n_2$, $n_3$) is then computed using the equation (2):<br>$$d(\overline{m}, \overline{n}) = \sqrt{\frac{1}{3}[(m_1 - n_1)^2 + (m_2 - n_2)^2 + (m_3 - n_3)^2]} \quad (2)$$<br>This “d” value must be <= 0.2 to reach an agreement of all experts, based on [50]-[51].</td></tr>
<tr><td>Step 4<br>The determination of group consensus</td><td>The overall value of the items in each construct is assessed to exceed 75% in order to reach the group agreement [50], by using the formula (3) as the following:<br>$$\frac{Total\ of\ \text{"d"} < 0.2\ X\ 100}{Total\ of\ experts} \quad (3)$$</td></tr>
<tr><td>Step 5<br>The identification of defuzzification and ranking of items</td><td>To calculate the defuzzification, either of these formulas (i or ii or iii) can be used:
<table>
<tr><td>i</td><td>$A_{\max} = \frac{m_1 + m_2 + m_3}{3}$</td></tr>
<tr><td>ii</td><td>$A_{\max} = \frac{m_1 + 2m_2 + m_3}{4}$</td></tr>
<tr><td>iii</td><td>$A_{\max} = \frac{m_1 + 4m_2 + m_3}{6}$</td></tr>
</table>
The resulting score numbers (defuzzification) must be in the range 0 to 1 [56] and the value of alpha-cut must be exceeding 0.5 [52]-[53] to indicate the expert consensus.</td></tr>
</table>

### C. Data Analysis using Fuzzy Delphi

The analysis was conducted with the aim to prove that the individual item and the overall examined constructs - *Customer Engagement* and the two consequences namely *Increases in eWOM* and *Participates in Brand Communities* - are the valid model and will be eligible as the measurement items for the actual survey in the following research phase. Overall, the Fuzzy Delphi systematic analysis needs to be through 5 steps to obtain the final decision.

As summarized in Table II(B), data collected in the form of linguistic scales should first be altered to fuzzy triangular numbers [50]. This is done for each expert before following the next procedure, determining average responses. Therefore, in the third step, the calculation of the threshold value (“d”) for each item and construct is calculated to ensure that the decision does not exceed 0.2 to reach an expert agreement [50]-[51]. The overall value of items in each construct is then assessed to obtain a minimum of 75% in order to reach the group's consent [50]. This analysis ends by calculating the defuzzification value for each item in prioritizing its position according to the construct [52]-[53]. Each of the above steps has been presented distinctly in Table II(B).

TABLE III. VALUE "D" OF ITEM, VALUE "D" OF CONSTRUCT, % OF EXPERTS CONSENSUS, DEFUZZIFICATION AND ITEM RANKING FOR *ONLINE CUSTOMER ENGAGEMENT* (Q1 – Q9)

| Expert | ONLINE CUSTOMER ENGAGEMENT | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 |
| 1 | 0.22 | 0.03 | 0.06 | 0.06 | 0.04 | 0.09 | 0.05 | 0.03 | 0.11 |
| 2 | 0.05 | 0.03 | 0.06 | 0.06 | 0.04 | 1.26 | 0.05 | 0.03 | 1.24 |
| 3 | 0.17 | 0.03 | 0.10 | 0.06 | 0.14 | 0.19 | 0.11 | 0.03 | 0.11 |
| 4 | 0.05 | 0.03 | 0.06 | 0.06 | 0.04 | 0.09 | 0.28 | 0.27 | 0.19 |
| 5 | 0.17 | 0.14 | 0.10 | 0.09 | 0.14 | 0.19 | 0.11 | 0.13 | 0.21 |
| 6 | 0.17 | 0.14 | 0.10 | 0.09 | 0.14 | 0.19 | 0.11 | 0.13 | 0.21 |
| 7 | 0.22 | 0.26 | 0.30 | 0.06 | 0.26 | 0.21 | 0.28 | 0.27 | 0.19 |
| 8 | 0.17 | 0.14 | 0.10 | 0.09 | 0.14 | 0.19 | 0.11 | 0.13 | 0.21 |
| 9 | 0.05 | 0.03 | 0.06 | 0.06 | 0.55 | 0.09 | 0.05 | 0.03 | 0.11 |
| 10 | 0.52 | 0.26 | 0.10 | 0.09 | 0.14 | 0.19 | 0.11 | 0.13 | 0.21 |
| 11 | 0.17 | 0.14 | 0.10 | 0.09 | 0.14 | 0.19 | 0.11 | 0.13 | 0.21 |
| 12 | 0.05 | 0.03 | 0.06 | 0.06 | 0.04 | 0.09 | 0.05 | 0.03 | 0.21 |
| Value "d" of each item | 0.17 | 0.10 | 0.10 | 0.07 | 0.15 | 0.25 | 0.12 | 0.11 | 0.27 |
| Value "d" construct | **0.15** | | | | | | | | |
| Percentage (%) of expert consensus for each item | 75% | 83% | 92% | 100% | 83% | 83% | 83% | 83% | 42% |
| Percentage (%) of expert consensus for construct | **81%** | | | | | | | | |
| Defuzzification | 0.84 | 0.86 | 0.90 | 0.91 | 0.86 | 0.82 | 0.88 | 0.87 | 0.79 |
| Item ranking in each construct | **7** | **5** | **2** | **1** | **6** | **8** | **3** | **4** | **9** |
| Item content | Spend a lot of time using [brand] | Usually use [brand] when use [category] | Inform positive things about [brand] | Recommend [brand] to other people | Read messages on [brand] post | Like messages on [brand] post | Write comments on [brand] post | Share messages on [brand] post | Follow the [brand] page |

## IV. RESULT

Fuzzy Delphi analysis that has been implemented on 42 items has met the objective with compelling results. Available in Table III are the results obtained for *Online Client Engagement* in the dimensions of the behaviour as a latent construct under investigation. Boxes marked with colour display values of "d" above 0.2, indicating that some experts disagree with the items reviewed.

However, unmarked boxes have a much larger number, concluding that most items are relevant in their view. It has reached a member consensus percentage of more than 75%, and defuzzification score of individual items is greater than 0.5. With the defuzzification scores, the item ranking can be calculated to know the priorities. Counting scores displaying two items that tend to be an expert choice are the similarity of questions that put forward suggestions and positive feedback about the brand initiated by a customer to other prospects or customers within the community they are interacting with (refer Q3 and Q4).

TABLE IV. VALUE "D" OF ITEM, VALUE "D" OF CONSTRUCT, % OF EXPERTS CONSENSUS, DEFUZZIFICATION AND ITEM RANKING FOR *INCREASE IN ELECTRONIC WORD-OF-MOUTH (EWOM)* (Q10 – Q14)

| Expert | INCREASES IN EWOM | | | | |
|---|---|---|---|---|---|
| | **Q10** | **Q11** | **Q12** | **Q13** | **Q14** |
| **1** | 0.07 | 0.05 | 0.05 | 0.07 | 0.16 |
| **2** | 0.07 | 0.05 | 0.05 | 0.07 | 0.05 |
| **3** | 0.19 | 0.17 | 0.05 | 0.19 | 0.16 |
| **4** | 0.49 | 0.52 | 0.21 | 0.20 | 0.23 |
| **5** | 0.19 | 0.17 | 0.18 | 0.19 | 0.16 |
| **6** | 0.19 | 0.17 | 0.18 | 0.19 | 0.16 |
| **7** | 0.07 | 0.05 | 0.05 | 0.07 | 0.05 |
| **8** | 0.19 | 0.17 | 0.18 | 0.19 | 0.16 |
| **9** | 0.49 | 0.22 | 0.21 | 0.79 | 0.23 |
| **10** | 0.20 | 0.22 | 0.50 | 0.20 | 0.53 |
| **11** | 0.19 | 0.17 | 0.18 | 0.19 | 0.16 |
| **12** | 0.07 | 0.05 | 0.05 | 0.07 | 0.05 |
| Value "d" of each item | 0.20 | 0.17 | 0.16 | 0.20 | 0.17 |
| Value "d" construct | | | **0.18** | | |
| Percentage (%) of expert consensus for each item | 83% | 75% | 75% | 92% | 75% |
| Percentage (%) of expert consensus for construct | | | **80%** | | |
| Defuzzification | 0.82 | 0.84 | 0.83 | 0.82 | 0.85 |
| Item ranking in each construct | **5** | **4** | **2** | **3** | **1** |
| Item content | Persuade contacts in the SNS to buy products of [brand] | Influence the contacts' opinions about [brand] and its products | Ask contacts in the SNS for advice in considering new products | Get contacts opinions in the SNS before buying new products of [brand] | Pass along the [brand] related information or opinion to other contacts in the SNS |

On the other hand, an assessment of the 9 items in the eWOM construct returned the unpredictable results initially. This is because nearly half of the threshold item value ("d") has exceeded 0.2 indicating the disagreement of experts on some adapted items from the previous context and research setting. To achieve a good overall agreement, this study has released 4 individual items with the lowest percentage of "d" value (not reaching 75%). Consequently, the scale supporting the *Increase in eWOM* measurements is comprised of only 5 items left as these have the percentage of expert consensus for more than 75%, contributing the total consensus of 80%. Refer Table IV.

According to the same table, score numbers (defuzzification) lie between 0 and 1 [56]. The value of alpha-cut is calculated to be greater than 0.5 [52]-[53] to support the consensus of experts on 5 items for the eWOM consequence. Attitudes to conveying brand-related information to their contact network at SNS are the most accepted items by experts who evaluate the question. Online subscribers are also interested in soliciting opinions or suggestions of members of the branded site community about their new favorite product release.

TABLE V. (A). VALUE "D" OF ITEM, VALUE "D" OF CONSTRUCT, % OF EXPERTS CONSENSUS, DEFUZZIFICATION AND ITEM RANKING FOR *PARTICIPATES IN BRAND COMMUNITIES* (Q15 – Q24)

| Expert | PARTICIPATES IN BRAND COMMUNITIES | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Q15** | **Q16** | **Q17** | **Q18** | **Q19** | **Q20** | **Q21** | **Q22** | **Q23** | **Q24** |
| **1** | 0.03 | 0.03 | 0.11 | 0.09 | 0.06 | 0.08 | 0.06 | 0.03 | 0.10 | 0.10 |
| **2** | 0.03 | 0.03 | 0.05 | 0.06 | 0.06 | 0.08 | 0.09 | 0.03 | 0.06 | 0.06 |
| **3** | 0.03 | 0.03 | 0.05 | 0.06 | 0.09 | 0.08 | 0.06 | 0.03 | 0.06 | 0.06 |
| **4** | 0.26 | 0.28 | 0.29 | 0.06 | 0.06 | 0.08 | 0.09 | 0.03 | 0.06 | 0.06 |
| **5** | 0.14 | 0.12 | 0.11 | 0.09 | 0.09 | 0.08 | 0.06 | 0.12 | 0.10 | 0.10 |
| **6** | 0.14 | 0.12 | 0.11 | 0.09 | 0.09 | 0.08 | 0.06 | 0.12 | 0.10 | 0.10 |
| **7** | 0.03 | 0.03 | 0.05 | 0.06 | 0.06 | 0.08 | 0.09 | 0.28 | 0.30 | 0.30 |
| **8** | 0.14 | 0.12 | 0.11 | 0.09 | 0.09 | 0.08 | 0.06 | 0.12 | 0.10 | 0.10 |
| **9** | 0.03 | 0.03 | 0.05 | 0.06 | 0.06 | 0.08 | 0.09 | 0.03 | 0.06 | 0.06 |
| **10** | 0.26 | 0.03 | 0.05 | 0.06 | 0.06 | 0.08 | 0.06 | 0.03 | 0.10 | 0.10 |
| **11** | 0.14 | 0.12 | 0.11 | 0.09 | 0.09 | 0.08 | 0.06 | 0.12 | 0.10 | 0.10 |
| **12** | 0.03 | 0.03 | 0.05 | 0.06 | 0.06 | 0.08 | 0.09 | 0.03 | 0.06 | 0.06 |
| Value "d" of each item | 0.10 | 0.08 | 0.09 | 0.07 | 0.07 | 0.08 | 0.07 | 0.08 | 0.10 | 0.10 |
| Value "d" construct | **0.11** | | | | | | | | | |
| Percentage (%) of expert consensus for each item | 83% | 92% | 92% | 100% | 100% | 100% | 100% | 92% | 92% | 92% |
| Percentage (%) of expert consensus for construct | **86%** | | | | | | | | | |
| Defuzzification | 0.86 | 0.88 | 0.89 | 0.91 | 0.91 | 0.92 | 0.93 | 0.88 | 0.90 | 0.90 |
| Item ranking in each construct | **15** | **12** | **9** | **4** | **3** | **2** | **1** | **11** | **5** | **6** |
| Item content | Can help improve [brand] and its products | Increasing the strength of relationships and the brand community | Feel more connected to the [brand] | Feel more connected to other customers of [brand] | Use experience to help other customers | Share experience and knowledge with others so as to help them be more educated about [brand] | Feel good to help answering member's questions | Look forward to discussing opinions about [brand] and its products with other people | Enjoy conversing with similar people | It is important to have conversations with people who share the same views about [brand] |

The second consequence of *Online Customer Engagement*, namely *Participates in Brand Communities* was captured as significance by the findings (refer Table V(A) and Table V(B)). Examined items were calculated with the value of "d" ≤ 0.2, proven that all questions are relevant for reaching an agreement 86%. When the selected group of experts evaluates these items, they have confirmed that they are suitable as measuring scales to the studied constructs.

According to the results of defuzzification in both tables, the scores obtained for all items are within the range of 0 to 1 and the alpha-cut values have exceeded 0.5 [52]-[53], [56]. The achievement of these values demonstrates the success of obtaining the experts' consent on each item submitted after the analysis. As mentioned earlier, *Participates in Brand Communities* has adopted the most number of items from literature. Following that, 9 out of 24 questions (leaving 15 items) have been excluded from this construct based on the ranking to minimize and facilitate respondents in answering the questionnaire in the actual survey soon.

TABLE V. (B). VALUE "D" OF ITEM, VALUE "D" OF CONSTRUCT, % OF EXPERTS CONSENSUS, DEFUZZIFICATION AND ITEM RANKING FOR *PARTICIPATES IN BRAND COMMUNITIES* (Q25 – Q29)

| Expert | PARTICIPATES IN BRAND COMMUNITIES | | | | |
|---|---|---|---|---|---|
| | Q25 | Q26 | Q27 | Q28 | Q29 |
| 1 | 0.03 | 0.11 | 0.10 | 0.11 | 0.04 |
| 2 | 0.03 | 0.05 | 0.06 | 0.05 | 0.04 |
| 3 | 0.13 | 0.05 | 0.10 | 0.05 | 0.13 |
| 4 | 0.03 | 0.05 | 0.06 | 0.05 | 0.26 |
| 5 | 0.13 | 0.11 | 0.10 | 0.11 | 0.13 |
| 6 | 0.13 | 0.11 | 0.10 | 0.11 | 0.13 |
| 7 | 0.27 | 0.29 | 0.30 | 0.29 | 0.26 |
| 8 | 0.13 | 0.11 | 0.10 | 0.11 | 0.13 |
| 9 | 0.03 | 0.05 | 0.06 | 0.05 | 0.04 |
| 10 | 0.27 | 0.05 | 0.06 | 0.05 | 0.26 |
| 11 | 0.13 | 0.11 | 0.10 | 0.11 | 0.13 |
| 12 | 0.03 | 0.05 | 0.06 | 0.05 | 0.13 |
| Value "d" of each item | 0.11 | 0.09 | 0.10 | 0.09 | 0.14 |
| Value "d" construct | | | **0.11** | | |
| Percentage (%) of expert consensus for each item | 83% | 92% | 92% | 92% | 75% |
| Percentage (%) of expert consensus for construct | | | **86%** | | |
| Defuzzification | 0.87 | 0.89 | 0.90 | 0.89 | 0.86 |
| Item ranking in each construct | **13** | **8** | **7** | **10** | **14** |
| Item content | Derives enjoyment from problem solving and idea generation | Community members can use their knowledge to help | Gives an opportunity to receive help from other community members | Able to use the community to find answers about [brand] | Freely share interest in the brand community |

After the analysis and modification process, the number of items retained is 29 against 42 original items reviewed in the provided questionnaire by the expert group. Table VI presents the summary of the items investigated after the Fuzzy Delphi analysis by the experts. The findings demonstrate that items obtained and modified from literature has been validated to be tested with the actual sample of brand communities in SNS environment.

TABLE VI. THE MODIFIED NUMBER OF ITEMS AFTER FUZZY DELPHI ANALYSIS

| Construct | Number of initial items | Number of items (after analysis) | Labelled of questions |
|---|---|---|---|
| Online Customer Engagement | 9 items | 9 items | Q1 – Q9 |
| eWOM | 9 items | 5 items | Q10 – Q14 |
| Participate in Brand Communities | 24 items | 15 items | Q15 – Q29 |
| **Total** | **42 items** | **29 items** | |

## V. DISCUSSION AND CONCLUSION

Taking the example of the analysis has been discussed, Fuzzy Delphi method proved very effective to confirm the modified set of items of literature into a different dimension of the study in terms of platform, object and subject of customer engagement. Indeed, it shows that *Increases in eWOM* and *Participates in Brand Communities* is a relevant result of customers engaged in behavioural aspects of the SNS community network. It coincides with the success of previous researcher studies that use Fuzzy Delphi for verification thus determining the priorities of factors or elements on a subject being examined [46]-[47], [56]-[57].

The analyzed items were well-calculated, as evidenced by the average threshold "d" ≤ 0.2 [50]-[51] and the average achievement of specialist consensus for proximal customer engagement variables and both consequences, which exceeded 75% [50]. The defuzzification scores obtained for all items is in the range of 0 to 1 and alpha-cut values have exceeded 0.5 as mentioned by [52] and [53] that confirms items analyzed are acceptable and contribute to the body of knowledge. The

full analysis results conclude that it has met the objective of this study to validate the measurement items for customer engagement and its consequences from the manifestation of their behaviour on the SNS avenue.

The outcome of this study offers two implications via the side of practical and social. First, by knowing the scales of the customer's engagement in the SNS interface such as indicators like, share and comment, digital marketers can establish effective strategies to drive customer or prospect towards promotional activities, campaigns, and brand-run contests. The items agreed by experts to evaluate the increase of customers through eWOM and the active participation in brand communities should be noted by the firms so that their brand can be recognized by the prospect and maintained by existing customers. For example, customers who continuously react to specific product or brand; by liking or sharing tips and uses of a brand are loyal customers that must be preserved by firms because they are valuable assets for brand life sustainability. This situation dictates the concept of relationship marketing that emphasizes those customers who have interacted with and are familiar with the brand become the source of informers to members within the brand's community site. Other theoretical support for customer engagement in brand community activities is derived from social exchange theory, where customers who are satisfied with the quality of their ties from the brand will resonate the sentiment, for example, through sharing of ideas to advance future brand products.

In terms of social implications, SNS software providers can offer a non-charge indicator within a wider radius to firms, capable of capturing the number of active loyal customers who often contribute brand-related information to other customers. This view aligns with the tendency of firms that often use built-in dashboards as they are cheaper and easier to learn on their own [58]-[59]. Programmers and social media administrators are also encouraged to provide similar approaches to these firms. With that, virtual marketers who are mostly small-scale businesses can leverage it to measure the effectiveness of their brand marketing efforts in the community.

## ACKNOWLEDGEMENT

The author would like to thank Universiti Teknikal Malaysia Melaka (UTeM) for providing research facilities at the time of conduction of this study and for the financial incentive for publishing this paper.

## REFERENCES

[1] J. Wirtz et al., "Managing brands and customer engagement in online brand communities," *Journal of Service Management*, vol. 24, no. 3, pp. 223–244, Jun. 2013.

[2] F. Kujur and S. Singh, "Engaging customers through online participation in social networking sites," *Asia Pacific Management Review*, vol. 22, no. 1, pp. 16–24, Mar. 2017.

[3] V. Blazevic *et al.*, "Beyond traditional word-of-mouth," *Journal of Service Management*, vol. 24, no. 3, pp. 294–313, 2013.

[4] "2016-2018 Research Priorities", *Marketing Science Institute*, 2016. [Online]. Available: http://www.msi.org/research/2016-2018-research-priorities. [Accessed: Nov. 07, 2017].

[5] T. Hennig-Thurau, C. F. Hofacker, and B. Bloching, "Marketing the pinball way: understanding how social media change the generation of value for consumers and companies," *Journal of Interactive Marketing*, vol. 27, no. 4, pp.237-241, 2013.

[6] T. Aichner and F. Jacob, "Measuring the degree of corporate social media use," *International Journal of Market Research*, vol. 57, no. 2, pp. 257–275, 2015.

[7] J. Gummerus, V. Liljander, E. Weman, and M. Pihlström, "Customer engagement in a Facebook brand community," *Management Research Review*, vol. 35, no. 9, pp. 857–877, Aug. 2012.

[8] L. Dessart, C. Veloutsou, and A. Morgan-Thomas, "Consumer engagement in online brand communities: a social media perspective," *Journal of Product & Brand Management*, vol. 24, no. 1, pp. 28–42, Mar. 2015.

[9] C. M. Sashi, "Customer engagement, buyer-seller relationships, and social media." *Management Decision*, vol. 50, no. 2, pp. 253-272, 2012.

[10] R. J. Brodie *et al.*, "Consumer engagement in a virtual brand community : an exploratory analysis consumer engagement in a virtual brand community : An exploratory analysis," *Journal of Business Research*, vol. 66, no. 1, 2013.

[11] D. E. Schultz and J. Peltier, "Social media's slippery slope: challenges, opportunities and future research directions," *Journal of Research in Interactive Marketing*, vol. 7, no. 2, pp. 86–99, May 2013.

[12] J. Van Doorn *et al.*, "Customer engagement behavior: theoretical foundations and research directions." *Journal of Service Research*, vol. 13, no. 3, pp. 253-266, 2010.

[13] R. C. Gambetti, G. Graffigna, and S. Biraghi, "The Grounded Theory approach to consumer–brand engagement," *International Journal of Market Research*, vol. 54, no. 5, pp. 659-687, 2012.

[14] L. D. Hollebeek, M. S. Glynn, and R. J. Brodie, "Consumer brand engagement in social media: conceptualization, scale development and validation," *Journal of Interactive Marketing*, vol. 28, no. 2, pp. 149–165, May 2014.

[15] F. J. Martínez-López, R. Anaya-Sánchez, S. Molinillo, R. Aguilar-Illescas, and I. Esteban-Millat, "Consumer engagement in an online brand community," *Electronic Commerce Research and Applications*, vol. 23, pp. 24–37, 2017.

[16] R. Thakur and S. P. Jain, "Customer engagement and online reviews," *Journal of Retailing and Consumer Services*, *41*, pp.48-59, 2017.

[17] S. L. Vargo and R. F. Lusch, "Service-dominant logic: continuing the evolution." *Journal of the Academy of Marketing Science*, vol. 36, no. 1, pp. 1-10, 2008.

[18] M. Clark, H. G. Black, and K. Judson, "Brand community integration and satisfaction with social media sites: a comparative study," *Journal of Research in Interactive Marketing*, vol. 11, no. 1, pp. 39–55, 2017.

[19] P. M. Blau, *Exchange and Power in Social Life*. Transaction Publishers, 1964.

[20] S. J. Pervan, L. L. Bove and L. W. Johnson, "Reciprocity as a key stabilizing norm of interpersonal marketing relationships: Scale development and validation." *Industrial Marketing Management*, vol. 38, no. 1, pp. 60-70, 2009.

[21] A. Dovaliene, A. Masiulyte, and Z. Piligrimiene, "The relations between customer engagement, perceived value and satisfaction: the case of mobile applications," *Procedia-Social and Behavioral Sciences*, vol. 213, pp. 659–664, Dec. 2015.

[22] T. H. Bijmolt, M. Eisenbeiss, B. G. Hardie, A. Lemmens and P. Saffert, "Analytics for customer engagement." *Journal of Service Research*, vol. 13, no. 3, pp. 341-356, 2010.

[23] A. Javornik and A. Mandelli, "Behavioral Perspectives of Customer Engagement: An exploratory study of customer engagement with three Swiss FMCG brands," *Journal of Database Marketing & Customer Strategy Management*, vol. 19, no. 4, pp. 300–310, 2012.

[24] W. Kunz, L. Aksoy, H. Hall, Y. Bart, K. Heinonen, and R. Marketing, "Customer engagement in a big data world," *Journal of Services Marketing*, vol. 31, no. 2, pp. 1–39, 2017.

[25] M. R. Habibi, M. Laroche, and M.-O. Richard, "The roles of brand community and community engagement in building brand trust on social media," *Computers in Human Behavior*, vol. 37, pp. 152-161, 2014.

[26] B. J. Baldus, C. Voorhees, and R. Calantone, "Online brand community engagement: Scale development and validation," *Journal of Business Research*, vol. 68, no. 5, pp. 978–985, 2015.

[27] G. Greve, "ScienceDirect The moderating effect of customer engagement on the brand image – brand loyalty relationship," *Procedia -Social Behavior*. Science, vol. 148, no. 148, pp. 203–210, 2014.

[28] G. Broekemier, N. N. Chau, and S. Seshadri, "Social Media Practices Among Small Business-to-Business Enterprises," *Small Business Institute® Journal Small Business Institute®*, vol. 11, no. 1, pp. 37–48, 2015.

[29] S. Hassan, S.Z.A. Nadzim and N. Shiratuddin, "Strategic use of social media for small business based on the AIDA model." *Procedia-Social and Behavioral Sciences*, vol. 172, pp. 262-269, 2015.

[30] H.M. Taiminen and H. Karjaluoto, "The usage of digital marketing channels in SMEs." *Journal of Small Business and Enterprise Development*, vol. 22. no. 4, pp. 633-651, 2015.

[31] C. K. Prahalad and V. Ramaswamy, "Co-creating unique value with customers." *Strategy & leadership*, vol. 32, no. 3, pp. 4-9, 2004.

[32] E. Jaakkola and M. Alexander, "The role of customer engagement behavior in value co-creation : a service system perspective," *Journal of Service Research*, vol. 17, no. 3, pp. 247–261, 2014.

[33] T. Hennig-Thurau, K. P. Gwinner, G. Walsh, and D. D. Gremler, "Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the Internet?," *Journal of Interactive Marketing*, vol. 18, no. 1, pp. 38–52, Jan. 2004.

[34] J. Brown, A. J. Broderick and N. Lee, "Word of mouth communication within online communities: Conceptualizing the online social network." *Journal of Interactive Marketing*, *21*(3), pp.2-20, 2007.

[35] R. E. Goldsmith and R. A. Clark, "An analysis of factors affecting fashion opinion leadership and fashion opinion seeking." *Journal of Fashion Marketing and Management*, vol. 12, no. 3, pp. 308-322, 2008.

[36] A. Shoham and A. Ruvio, "Opinion leaders and followers: A replication and extension." *Psychology & Marketing*, vol. 25, no. 3, pp. 280-297, 2008.

[37] S. Chu and Y. Kim, "Determinants of consumer engagement in Electronic word-of-mouth (eWOM) in Social Networking Sites", *International Journal of Advertising*, vol. 30, no. 1, pp. 47-75, 2011

[38] A. T. Norman and C. A. Russell, "The pass-along effect: Investigating word-of-mouth effects on online survey procedures." *Journal of Computer-Mediated Communication*, vol. 11, no. 4, pp. 1085-1103, 2006.

[39] A. M. Muniz and T. C. O'Guinn, "Brand Community." *Journal of Consumer Research*, vol. 27, no. 4, pp. 412–32, 2001.

[40] V. A. Barger and L. Labrecque, "An integrated marketing communications perspective on social media metrics." International *Journal of Integrated Marketing Communications*, pp. 64-76, 2013.

[41] C. M. K. Cheung, M. K. O. Lee, and X.-L. Jin, "Customer Engagement in an Online Social Platform: A Conceptual Model and Scale Development." in *Thirty Second International Conference on Information Systems*, 2011.

[42] S. D. Vivek, S. E. Beatty and R. M. Morgan, "Customer Engagement: Exploring customer relationships beyond purchase." *Journal of Marketing Theory and Practice*, vol. 20, no. 2, pp. 122-146, 2012.

[43] M. Á. Moliner, D. Monferrer, and M. Estrada, "Consequences of customer engagement and customer self-brand connection," *Journal of Services Marketing*, vol. 32, no. 19, pp. 8357–25, 2018.

[44] T. Wang, R. K. Yeh, C. Chen and Z. Tsydypov, "What drives electronic word-of-mouth on social networking sites? Perspectives of social capital and self-determination." *Telematics and Informatics*, vol. 33, pp. 1034–1047, 2016.

[45] S. Gregory, J, F. T. Hartman and J. Krahn, "The Delphi method for graduate research.", *Journal of Information Technology Education: Research*, vol. 6, pp. 1-21, 2007.

[46] R. Mustapha, Z. Hussin and S. Siraj, "Analisis faktor penyebab ketidakjujuran akademik dalam kalangan mahasiswa: aplikasi teknik Fuzzy Delphi.", *Jurnal Kurikulum & Pengajaran Asia Pasifik*, vol. 2, no. 2, pp. 1-17, 2017.

[47] N. Mat Noh, S. Abd Razak, N. Alias, S. Siraj, M. Mohd Jamil and Z. Hussin, "Usage of Facebook : the future impact of curriculum implementation on students in Malaysia", *Procedia-Social and Behavioral Sciences*, vol. 103, pp. 1261-1270, 2013.

[48] M. Adler and E. Ziglio, "Gazing into the oracle: the Delphi Method and its application to social policy and public health." London: Jessica Kingsley Publishers, 1996.

[49] T. J. Gordon, *The Delphi Method*. Future Research Methodology, 1994.

[50] P. L. Chang, C. W. Hsu and P. C. Chang, "Fuzzy Delphi Method for evaluating hydrogen production technologies." *International Journal of Hydrogen Energy*. pp. 14172 – 14179, 2011.

[51] C.H. Cheng and Y. Lin, "Evaluating the best main battle tank using Fuzzy Decision Theory with linguistic criteria evaluation." *European Journal of Operational Research*, vol. 142, no. 1, pp. 174-186, 2002.

[52] S. Bodjanova, "Median alpha-levels of a Fuzzy Number." *Fuzzy Sets and Systems*. vol. 157, no. 7, pp. 879-891, 2006.

[53] C. W. Tang and C. T. Wu, "Obtaining a picture of undergraduate education quality: a voice from inside the university." *Higher Education*, vol. 60, pp. 269–286, 2010.

[54] M. A. Bashir and N. A. Ali, "Impact of customer brand relationship through Facebook on the level of customer engagement," *Pakistan Business Review*, no. April 2016, pp. 159–178, 2016.

[55] J. Gutiérrez-Cillán, C. Camarero-Izquierdo, and R. San José-Cabezudo, "How brand post content contributes to user's Facebook brand-page engagement. The experiential route of active participation," *BRQ Business Research Quarterly*, vol. 20, no. 4, pp. 258–274, 2017.

[56] M. Mohd Jamil, S. Siraj, Z. Hussin, N. M. Noh and A. Arifin, *Pengenalan asas kaedah Fuzzy Delphi dalam penyelidikan rekabentuk pembangunan*. Minda Intelek, Bangi, 2014.

[57] N. Habibah, A. Ramlie, Z. Hussin and M. Mohd Jamil, "Aplikasi Teknik Fuzzy Delphi Terhadap Keperluan Aspek ' Riadhah Ruhiyyah ' Untuk Profesionalisme Perguruan Pendidikan Islam." *The Online Journal of Islamic Education*, 2014.

[58] T. Wiesel, K. Pauwels and J. Arts, "Practice prize paper—Marketing's profit impact: quantifying online and off-line funnel progression." *Marketing Science*, vol. 30, no. 4, pp. 604-611, 2011.

[59] B. Grosser, "What do metrics want? How quantification prescribes social interaction on Facebook", *Computational Culture*, 2014.

# Validating Antecedents of Customer Engagement in Social Networking Sites using Fuzzy Delphi Analysis

Noraniza Md Jani[1], Mohd Hafiz Zakaria[2], Zulisman Maksom[3]
Fakulti Teknologi Maklumat dan Komunikasi,
Universiti Teknikal Malaysia Melaka, Malaysia

Md. Shariff M. Haniff[4]
Faculty of Business and Management,
Universiti Teknologi MARA
(Alor Gajah Campus), Melaka, Malaysia

Ramlan Mustapha[5]
Institute of Teacher Education, Tengku Ampuan Afzan
Campus, Kuala Lipis,
Pahang, Malaysia

***Abstract*—The concept of online customer engagement is getting imperative in modern business due to the uncontrolled conversation via cyber-avenue. This study validates the antecedents of customer engagement conceptualized in Social Networking Sites (SNS) by benefitting the Fuzzy Delphi method. Through purposive sampling, a total of 12 experts from academics and practitioners have participated in the verification of items through 7-point linguistic scales of the questionnaire instrument. The findings show that invited experts have reached agreement on the elements shown within the framework through a 75% percent agreement for each construct. The analysis of this study has highlighted the implications of the relevant theories on the direction and the new dimensions of customer engagement concept especially in SNS to future researchers. Businesses are clearly able to gain stronger knowledge and information about their customer-related factors and their prospects at SNS.**

***Keywords*—Customer engagement; antecedents; fuzzy delphi; SNS; online community**

## I. INTRODUCTION

The rapid emergence of mobile devices and applications have provided options and power for customers in brand communities of Social Networking Sites (SNS) to participate in market conversations. Varieties of proliferated social data have generated User-Generated-Content (UGC) in which manifesting dialogues among experienced and new customers as well as information seekers who are looking for valuable product reviews and knowledge. SNS major worldwide currently are Facebook, Instagram, YouTube, Qzone and Weibo [1]. Facebook, as a leading site, has served as an effective platform for many top brands to set up a page with an important goal of connecting with customers and maintaining two-way communication [2]-[3].

Firms typically use SNS by creating corporate profiles through brand pages to place and reinforce their brand and disseminate information about products. This brand page will aggregate the brand community to perform activities, among which most often are experience sharing of brands and products [4], thoughts and opinions [5], a sense of satisfaction over brands and firms [6], [15] and sentiment expressions [7] circulated through the electronic Word-of-Mouth (eWOM) mechanism.

On different notes, businesses are losing more control over their customers as online communities comprising peers, new and experienced members are interacting and influencing each other when making decisions about the use of brands and products based on information that becomes a matter of conversation [8]-[10]. Therefore, a different approach is required to measure the SNS marketing investment. Firms must change the approachable ways to new dimensions that focus on retaining, sustaining, and nurturing their prospects and existing customers [11]-[12] as customer insights remain a top research priority especially among practitioners [13]. On top of that awareness, many firms prioritise customer engagement activities as their main agenda in SNS digital medium. As the firm's key focus has shifted to customers, SNS is crucial as an intermediate medium capable of reaching greater promotion and campaigning for larger audiences, facilitating and accelerating customer service, raising awareness to new customers and prospects and facilitating the sharing of ideas and information from marketers and followers in an online community [14]-[16]. The Facebook and Instagram sites that have an increasingly significant number of active users in each month [1], [17]-[18] have been able to engage in many prospects and help retain customers through community participation in brand sites.

Research in recent years has explored customer engagement not just in theoretical but in an empirical ways by scholars and practitioners alike. This emerging concept can be approached and conceptualized in different dimensions, definition and object. However, this study, by its own lens, is focus to embrace it as a behavioural manifestation [19]-[20], [23] as an attempt to aggregate multiple ways that might influence the firm [21]-[23]. Besides affective and cognitive dimensions, behaviours remain strong indicators of engagement, in addition they manifest in a number of ways on social media [20]. Furthermore, the plethora of emergence social media are giving potentials for more research to be conducted in order to explain the customer engagement behaviour [24]-[27].

Customer engagement in an online community has been defined by this study as "a customer's behavioural manifestations that go beyond transactions and have a brand or firm focus, allow firms to involve specific interactive experiences within the online community, to build long-term relationship, and create more meaningful and sustainable interactions between the brand and/or the brand community and the customer, or among customers, resulting from motivational drivers" [19], [21]-[22], [28]. The motivations that drive customers to be engaged may comprise attitudinal factors such as their own goals, satisfaction, and trust, social benefits such as brand-relationship and interaction, economic benefits such as reward, and other expecting benefits such as enhanced knowledge and reputation [19], [29].

Previous empirical studies on customer engagement behaviour on SNS are still scarce [30]. Only a few researches have been known to explore the relevant elements of online customer engagement at SNS community [31]-[34], moreover, their research extends to many dimensions and is not limited to behavioural factor. Meanwhile a considerable numbers of the related works were focused on more generic engagement in social media and web platform [35]-[37]. There are researchers concentrate only on the specific elements of engagement in digital spaces (see [38]-[39]). On the other hand, some studies suggest many elements and are presented in the theoretical approach, but their study does not provide any primary data (see [19], [21]). Thanks for a few research on providing the diverse mixed scales for user engagement in SNS, nevertheless they were revealing the validation evidence from the previous works and not empirically tested the new specific context (see [40]). Less scientific studies cover more holistic aspects, such as verifying multi-dimensional factors that induce customer engagement in the behavioural dimension on SNS. As there are limited works on harnessing mechanism to measure this perspective [37], this study investigates the validation of the proposed customer engagement framework via Fuzzy Delphi technique.

## II. Purpose of the Study

The present study was conducted to obtain a consensus of experts relating to the development and validity of the proposed diverse constructs in which is being called antecedents for the SNS customer engagement framework in Malaysian context. In the knowledge of this study, the Fuzzy Delphi technique works well as a research instrument to elucidate the agreement of experts on the examined questions [41]-[42] and to improve the understanding of an incomplete knowledge about a phenomenon [43]. Therefore, this effective tool is intentionally applied in the study to answer the research questions as follows:

*1) What is the level of agreement among experts on the conceptualized SNS customer engagement framework?*

*2) What is the level of agreement among experts towards the indicators for antecedents of customer engagement via the SNS platform?*

## III. Conceptual Framework

A considerable numbers of authors have highlighted customer engagement as a core role that has an array of antecedents, thus mediating the effects to the other consequences (see [19], [44]-[48]). Antecedents are lined up by decisive factors that stimulate customers to engage with brands and/or brand communities. Understanding the antecedents of customer engagement is important since it is likely to provide significant managerial guidance for firms regarding how to make their SNS efforts in brand/products offerings useful and influential for customers. Consequently, brands can take many benefits from their engagement activities to attract greater attention from the SNS community. For example, when satisfied customers share their positive experiences with products among Facebook contacts, it is likely that brand followers will increase dramatically, making friends or prospects affected or willing to engage with the brand via *like* or *follow* indicators on the board of SNS brand page.

Although the marketing communication and socialization offered in online engagement still cannot be replaced by an offline medium [30], it has been observed that online and offline world are closely linked in which business activities in the offline world will be eventually reflected in their online activities [49]. Either in the digital or offline environment, the determinants towards an engagement is very cross-related. In that case, this study validated the multi-dimensional operational structure of customer engagement as found in the existing literature in both online and offline marketing mechanism [28], [48].

This study is taking an inspiration from the past relevant and established research in engagement's domain to build up the conceptual framework. This one-layer conceptual framework consists of ten antecedents believed to have a very significant effect on the involvement of SNS online customers. By exploring conceptual and empirical research as being shown in Table I, the initiative has taken an approach to collectively examine a set of well-cited antecedents (*Trust, Commitment, Interaction, Brand Post, Reward,* and *Satisfaction*) that drives customer engagement, combined with the newly determined in which is a gap to previous researchers (*Perceived Relationship, Information Platform and Processes, Consumption Goal,* and *Reputation*). This way of integration will highlight the diverse set of multi-dimensional drivers of customer engagement (antecedents) with the aim of looking into the positive future anticipation from experts.

TABLE I. ANTECEDENTS OF CUSTOMER ENGAGEMENT

| Author(s) | Antecedents of customer engagement | | | | | | | | | | Context, research type |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Relationship* | *Interaction* | *Trust* | *Commitment* | *Reward* | *Brand Post* | *Information Platform & Processes* | *Satisfaction* | *Reputation* | *Consumption Goal* | |
| [19] | | | √ | √ | | | √ | √ | | √ | Offline, conceptual |
| [21] | | | √ | √ | | | | √ | | | Offline, conceptual |
| [38] | | √ | | | | | | | | | Online social platform, conceptual |
| [28] | | √ | | | | | | | | | Offline, conceptual |
| [50] | √ | √ | | √ | | | | √ | | | Social media, conceptual |
| [10] | | | | | √ | √ | | | | | Online brand community, conceptual |
| [30] | √ | √ | | | | √ | | | | | SNS, empirical |
| [71] | | | | | | | | √ | | | Online brand community, empirical |
| [32] | | | | | √ | | | | | | Online brand community, empirical |
| [20] | √ | | √ | | √ | √ | | √ | | | Online brand community, empirical |
| [46] | | | √ | √ | | | | √ | | | Social media, empirical |
| [39] | √ | | | | | | | | | | SNS, empirical |
| [33] | | | | | √ | √ | | | | | SNS, empirical |
| [3] | | √ | | | √ | √ | | | | | SNS, empirical |
| [34] | | √ | | | | √ | | | | | SNS, empirical |
| [48] | | | √ | | | | | √ | | | Mobile devices, empirical |
| [37] | | | | | | √ | | | | | Social media, empirical |

To a greater scholar extent, customer engagement is clearly advocated by the principle of relationship marketing theory in consumer behaviour research [50]-[51]. As the present study concerns online community platform, the definition of relationship marketing that has been employed by this study is closely linked to what [51] suggested that marketing is seen as a relationship, network, and interaction. Acting as a vehicle for the majority of businesses, past studies have proved social media, including SNS, as a key relationship marketing variable that leads to building and maintaining relational customers [52]. This principle theory are anchored with diverse theories of service-dominant logic, social exchange, and user gratification. Relationship marketing and its expansion theories were also embraced by many research work in different study context to cultivate meaningful relationships with customers and prospects, thus spotting the engagement [53].

Formed based on previous theories and works, this study postulates the conceptualization of customer engagement in SNS as illustrated in Fig. 1 below. The left box forms a group of ten antecedents that are argued as a determinant of customer motivation towards their engagement with the community and/or brand within the SNS pages. The following section of the methodology will describe the detailed sequence practiced to confirm these antecedents as the precise determinants of the customer engagement, as well as respond to the purpose of this study.

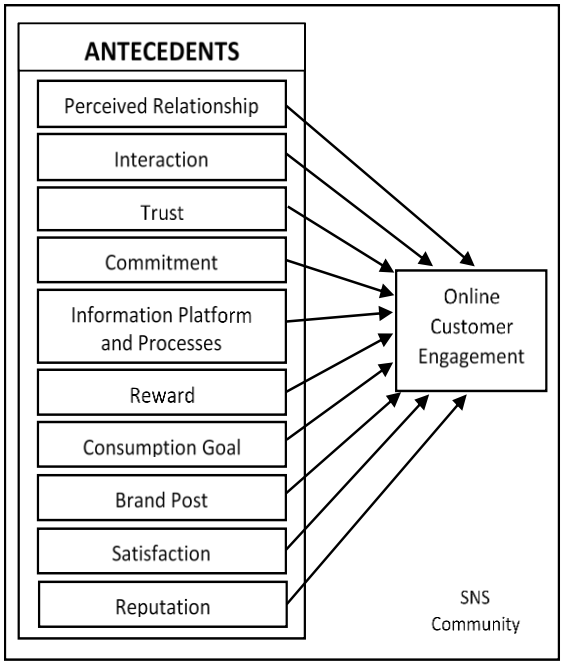



Fig. 1. Conceptual Framework of Customer Engagement in SNS

## IV. METHODOLOGY

This study has performed the Fuzzy Delphi technique to validate the conceptualized SNS customer engagement framework and the measurement scales throughout the whole research process. Successfully adopted in a wide variety of research variations including the IT/IS projects [43], it is a combination of traditional Delphi and Fuzzy set theory. This tool is a very helpful method when the level of research requires approval from a group of experts. Furthermore, Fuzzy Delphi technique is an attractive method to draw a collective decision on the blurred ideas as a result of the alignment of the opinions of the chosen experts [54].

While traditional Delphi uses infinite assessment rounds until experts' agreements are achieved, the new Fuzzy Delphi has the advantage of getting feedback in shorter time and lower costs, thereby increasing the number of survey results. Interestingly, experts can fully express their real reactions without misinterpretation to ensure completeness and consistency of opinion [56].

### A. Instrument Development

The first phase of the Fuzzy Delphi technique began with the development of questionnaire scripts used for experts to verify the research items. The questionnaire was comprised with two parts: 1) demographics, and 2) 53 items of 7-Lickert point scales in 10 tested constructs or antecedents. The 7-point scale was chosen because the evidence shows that the higher the scales in the questionnaire, the more accurate the data obtained [41]. The measurement items are developed based on literature and further followed by a round of interviews by practitioners or real SMEs to confirm the verification of captured items [43].

TABLE II. LIST OF CONSTRUCTS AND ADAPTED ITEM SOURCES

| Antecedents | Adapted Scales | Number of item scales |
|---|---|---|
| Perceived Relationship | [52] | 4 items |
| Interaction | [58]-[59] | 8 items |
| Trust | [57] | 8 items |
| Commitment | [60] | 6 items |
| Information Platform and Processes | [19] | 4 items |
| Reward | [61] | 3 items |
| Consumption Goal | [37] | 3 items |
| Brand Post Characteristics | [33], [64] | 4 items |
| Satisfaction | [48], [62] | 5 items |
| Reputation | [63] | 8 items |
| **Total** | | **53 items** |

In details, this has been achieved by three sequential steps. Initially, the SNS customer-conceptual framework of engagement was set up as illustrated in Figure 1. Second, the identified indicators are reviewed by practitioners through the individual close-interviews. Five SMEs were invited as participants in this stage with the task of assessing the reasonable aspect of engagement in the current practical context and then giving priority to each construct. Once they confirm agreement with the antecedents of customer engagement in SNS as shown, the final step is working with the determination of scale items for the engagement framework. Indicators for this purpose are taken from the comprehensive as well as relevant sources and adapted to be studied in relation to the brand and brand community engagement in SNS environment. Refer Table 2 for the list of adapted sources for each construct. These items are then collected and formatted into the form of a questionnaire to be disseminated to selected experts in the evaluation process.

### B. Data Collection and Analysis

The second phase of the Delphi Fuzzy technique is critical in which the item scale in the questionnaire provided will be verified by the appointed experts and finally collected and analyzed. With the aim of achieving good results, this phase is implemented through five processes as follows:

### C. Respondents and sampling studies

The purposive sampling is the best way in the experts' selection as the researcher wants to gain views and consensus on a matter [55]. Using this method, a total of 12 heterogeneous experts from different profiles and backgrounds were identified and invited to participate in the survey via a scheduled and face-to-face approach following the description procedures that may be required when issues arise from the items reviewed. This number of 12 respondents was taken from Gordon's suggestion [65] which states that heterogeneous sampling requires between 10 and 35 experts. They consist of two distinct groups: academician and practitioner. Comprising of 7 academicians and 5 practitioners representing SMEs, they have been identified to meet the criteria required as experts based on knowledge and experience with the issues being investigated, the ability and willingness to participate, the qualifications, the individual character, the comparative ability, the communication skills as well as the consistent and credible in their respective fields [54], [66], as shown in Table 3. To be more specific, these specialists are skilled in either one or more of these areas: IT, social media, online governance, marketing, and consumer behaviour.

TABLE III. CLASSIFICATION OF HETEROGENEOUS EXPERTS

| Experts | Institution/Sector | Number of participants |
|---|---|---|
| Academicians | UTeM (IT domain) | 2 |
| | UTeM (Business domain) | 1 |
| | UITM (IT domain) | 1 |
| | UITM (Business domain) | 1 |
| | UPM (IT domain) | 1 |
| | Polytechnic (IT domain) | 1 |
| Practitioners (SMEs) | Manufacturing | 2 |
| | Services | 3 |
| **Total** | | **12** |

On each item provided in the questionnaire of 7-Lickert scale assessment, feedback is obtained from each expert through the marking scales 1 to 7 to show the level of agreement. They are also provided by the particular spaces at each construct for any suggestion to upgrade the accuracy and relevancy of questions. The forms then collected by hand and the item scales were analysed by applying Fuzzy Delphi in the spreadsheet of Microsoft Excel.

TABLE IV. 7-LICKERT FUZZY SCALES OF AGREEMENT LEVEL

| Linguistic variables | Fuzzy scales | | |
|---|---|---|---|
| Strongly Disagree | 0.0 | 0.0 | 0.1 |
| Disagree | 0.0 | 0.1 | 0.3 |
| Somewhat Disagree | 0.1 | 0.3 | 0.5 |
| Neither agree or disagree | 0.3 | 0.5 | 0.7 |
| Somewhat agree | 0.5 | 0.7 | 0.9 |
| Agree | 0.7 | 0.9 | 1.0 |
| Strongly agree | 0.9 | 1.0 | 1.0 |

### D. The Determination of Linguistic Variables and Average Responses

After obtaining the response from the selected respondents, all linguistic variables in the form of 7-Lickert scales were transformed to triangular fuzzy numbers [70], as displayed in Table IV. Triangular Fuzzy Numbers represent the value of $m_1$ (minimum value), $m_2$ (simple value), and $m_3$ (maximum value) written as $(m_1, m_2, m_3)$. The average value of each expert's response will then be calculated with the following equation (1) as below:

$$m = \frac{\sum_1^n m_i}{n} \quad (1)$$

*E. The determination of threshold value "d"*

To identify the level of agreement between experts, the threshold value between two Fuzzy numbers ($m = m_1, m_2, m_3$) and ($n = n_1, n_2, n_3$) will be calculated using the given formula equation (2) :

$$d(\bar{m}, \bar{n}) = \sqrt{\frac{1}{3}[(m_1 - n_1)^2 + (m_2 - n_2)^2 + (m_3 - n_3)^2]} \quad (2)$$

According to [67]-[68], the "threshold" ("d") value must be <= 0.2 to reach an agreement of all experts.

*F. The Determination of Percentage of Group Consensus*

Afterwards, the overall value of the items in each construct are computed to ensure the percentage must exceed 75% to achieve the group agreement [67]. Otherwise the second round of this technique should be implemented. The formula is as below:

TABLE V. VALUE "D" OF ITEM, VALUE "D" OF CONSTRUCT, % OF EXPERT CONSENSUS, DEFUZZIFICATION AND ITEM RANKING FOR *PERCEIVED RELATIONSHIP* AND *INTERACTION*

| **Expert** | **PERCEIVED RELATIONSHIP** | | | | **INTERACTION** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Q1** | **Q2** | **Q3** | **Q4** | **Q5** | **Q6** | **Q7** | **Q8** | **Q9** | **Q10** | **Q11** | **Q12** |
| 1 | 0.09 | 0.19 | 0.16 | 0.06 | 0.23 | 0.11 | 0.07 | 0.04 | 0.05 | 0.26 | 0.04 | 0.01 |
| 2 | 0.09 | 0.19 | 0.03 | 0.06 | 0.03 | 0.11 | 0.07 | 0.04 | 0.05 | 0.03 | 0.04 | 0.01 |
| 3 | 0.23 | 0.19 | 0.23 | 0.06 | 0.16 | 0.11 | 0.07 | 0.13 | 0.05 | 0.03 | 0.04 | 0.01 |
| 4 | 0.09 | 0.07 | 0.23 | 0.49 | 0.16 | 0.24 | 0.07 | 0.04 | 0.11 | 0.03 | 0.04 | 0.01 |
| 5 | 0.09 | 0.07 | 0.16 | 0.06 | 0.03 | 0.44 | 0.08 | 0.13 | 0.11 | 0.14 | 0.12 | 0.01 |
| 6 | 0.23 | 0.32 | 0.16 | 0.19 | 0.16 | 0.11 | 0.08 | 0.13 | 0.11 | 0.26 | 0.12 | 0.01 |
| 7 | 0.46 | 0.37 | 0.03 | 0.20 | 0.23 | 0.15 | 0.07 | 0.56 | 0.28 | 0.03 | 0.04 | 0.25 |
| 8 | 0.09 | 0.07 | 0.16 | 0.19 | 0.03 | 0.44 | 0.08 | 0.13 | 0.11 | 0.14 | 0.12 | 0.01 |
| 9 | 0.09 | 0.19 | 0.03 | 0.06 | 0.03 | 0.15 | 0.07 | 0.04 | 0.28 | 0.03 | 0.28 | 0.01 |
| 10 | 0.16 | 0.07 | 0.03 | 0.20 | 0.23 | 0.24 | 0.08 | 0.04 | 0.11 | 0.14 | 0.04 | 0.15 |
| 11 | 0.46 | 0.37 | 0.23 | 0.19 | 0.16 | 0.24 | 0.08 | 0.13 | 0.11 | 0.14 | 0.12 | 0.15 |
| 12 | 0.09 | 0.07 | 0.03 | 0.06 | 0.03 | 0.10 | 0.07 | 0.13 | 0.05 | 0.03 | 0.04 | 0.01 |
| Value "d" of each item | 0.18 | 0.18 | 0.12 | 0.15 | 0.12 | 0.20 | 0.07 | 0.12 | 0.12 | 0.10 | 0.08 | 0.05 |
| Value "d" construct | **0.16** | | | | **0.11** | | | | | | | |
| Percentage (%) of expert consensus for each item | 67% | 75% | 75% | 92% | 75% | 58% | 100% | 92% | 83% | 83% | 92% | 92% |
| Percentage (%) of expert consensus for construct | **77%** | | | | **84%** | | | | | | | |
| Defuzzification | 0.83 | 0.77 | 0.87 | 0.82 | 0.85 | 0.78 | 0.91 | 0.87 | 0.88 | 0.86 | 0.88 | 0.86 |
| Item ranking in each construct | 2 | 4 | 1 | 3 | 7 | 8 | 1 | 4 | 2 | 5 | 3 | 6 |
| Item content | Satisfied relationship with [brand] | Ideal relationship with [brand] | **Valuable relationship with [brand]** | Quality relationship with [brand] | Access to specific knowledge about [brand] | Use of knowledge about firms and other customers | **Guidelines or support from other customers about [brand]** | Quick reaction from other customers | Receive advice and helps to solve problem | Time to support online activities about [brand] | Provides with experience on [brand]. | Provides informational and interpersonal exchanges |

*G. The Identification of Defuzzification and Ranking of Items*

According to [69], approximation of a fuzzy set by an $\alpha$-cut (alpha-cut) is the oldest and the simplest defuzzification method. The value of $\alpha$ = 0.5 is the median of interval [0, 1]. To calculate the defuzzification score, there are three formulas that can be used (choose either i, ii or iii):

| | |
|---|---|
| i | $A_{max} = \frac{m_1 + m_2 + m_3}{3}$ |
| ii | $A_{max} = \frac{m_1 + 2m_2 + m_3}{4}$ |
| iii | $A_{max} = \frac{m_1 + 4m_2 + m_3}{6}$ |

The resulting score numbers (defuzzification scores) must be in the range 0 to 1 [41] and the value of $\alpha$-cut must be exceeding 0.5 [69]-[70] to indicate the expert consensus. The present work can identify the priorities of the constructs and items according to the scores obtained.

## V. Result

The Fuzzy Delphi analysis on the present study has shown a satisfactory and good overall outcome. Following are the result of the "d" value obtained from each antecedent's item (see Table V, Table VI, Table VII, and Table VIII). The box marked with "d" values exceeds 0.2, indicating that there are experts who are at the midst of agree and disagree with the items submitted in the questionnaire. However, the most unchecked boxes conclude that the items are relevant and agreeable to the experts. The first two constructs, *Perceived Relationship* and *Interaction*, has an average value of "d" threshold of less than 0.2. Accordingly, both have reached the percentage of expert consensus of more than 75%, and the defuzzification scores greater than 0.5, making them acceptable as antecedents for the customer engagement concept studied.

According to expert opinion, not all adapted items from the literature are applicable in the context of the SNS community to be studied. For example, the calculation template was attempted to adjust towards two constructs, namely *Trust* and *Commitment* (refer Table VI). This study has removed 3 items from *Trust* due to the low % of "d" value, leaving only 5 as valid questions to measure this construct. That action was done to able the whole % item "d" $\leq$ 0.2 has achieved the agreement of 78%, making this construct successfully maintained. However, the removal of items from the *Commitment* due to the low% "d" value still does not support agreements on this construct to reach 75%. Therefore, it has to be removed from the framework.

TABLE VI. Value "D" of Item, Value "D" of Construct, % of Expert Consensus, Defuzzification and Item Ranking for *Trust*, *Commitment* and *Consumption Goal*

| **Expert** | **TRUST** | | | | | **COMMITMENT** | | | | **CONSUMPTION GOAL** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Q13** | **Q14** | **Q15** | **Q16** | **Q17** | | | | | **Q18** | **Q19** | **Q20** |
| 1 | 0.12 | 0.19 | 0.10 | 0.09 | 0.18 | 0.18 | 0.12 | 0.14 | 0.24 | 0.03 | 0.05 | 0.03 |
| 2 | 0.12 | 0.05 | 0.06 | 0.09 | 0.06 | 0.06 | 0.12 | 0.14 | 1.10 | 0.03 | 0.05 | 0.03 |
| 3 | 0.12 | 0.19 | 0.10 | 0.22 | 0.18 | 0.18 | 0.25 | 0.26 | 0.25 | 0.03 | 0.05 | 0.13 |
| 4 | 0.13 | 0.20 | 0.29 | 0.17 | 0.21 | 0.21 | 0.15 | 0.42 | 0.09 | 0.03 | 0.05 | 0.03 |
| 5 | 0.12 | 0.05 | 0.10 | 0.22 | 0.18 | 0.18 | 0.25 | 0.26 | 0.36 | 0.12 | 0.11 | 0.13 |
| 6 | 0.26 | 0.05 | 0.10 | 0.22 | 0.18 | 0.18 | 0.25 | 0.26 | 0.36 | 0.12 | 0.11 | 0.13 |
| 7 | 0.13 | 0.20 | 0.29 | 0.17 | 0.21 | 0.21 | 0.15 | 0.14 | 0.32 | 0.28 | 0.29 | 0.27 |
| 8 | 0.12 | 0.05 | 0.10 | 0.22 | 0.18 | 0.18 | 0.25 | 0.26 | 0.36 | 0.12 | 0.11 | 0.13 |
| 9 | 0.13 | 0.20 | 0.10 | 0.46 | 0.21 | 0.21 | 0.44 | 1.00 | 0.09 | 0.03 | 0.05 | 0.03 |
| 10 | 0.13 | 0.20 | 0.06 | 0.17 | 0.50 | 0.50 | 0.44 | 0.14 | 0.09 | 0.03 | 0.11 | 0.13 |
| 11 | 0.43 | 0.19 | 0.10 | 0.17 | 0.18 | 0.18 | 0.15 | 0.26 | 0.32 | 0.12 | 0.11 | 0.27 |
| 12 | 0.12 | 0.05 | 0.06 | 0.09 | 0.06 | 0.06 | 0.12 | 0.14 | 0.24 | 0.03 | 0.05 | 0.03 |
| Value "d" of each item | 0.16 | 0.14 | 0.12 | 0.19 | 0.19 | 0.22 | 0.28 | 0.32 | 0.28 | 0.08 | 0.09 | 0.11 |
| Value "d" construct | **0.20** | | | | | **0.28** | | | | **0.10** | | |
| Percentage (%) of expert consensus for each item | 83% | 100% | 83% | 58% | 67% | 45% | 36% | 27% | 55% | 92% | 92% | 83% |
| Percentage (%) of expert consensus for construct | **78%** | | | | | **41%** | | | | **89%** | | |
| Defuzzification | 0.81 | 0.82 | 0.89 | 0.82 | 0.83 | **NOT COMPUTED** | | | | 0.88 | 0.89 | 0.90 |
| Item ranking in each construct | 5 | 3 | 1 | 4 | 2 | | | | | 3 | 2 | 1 |
| Item content | Product look with [brand] | Meet expectations with [brand] | **Confidence by [brand]** | Reliability in addressing concern | Dependence on problem-solving | **ITEMS WERE REMOVED** | | | | Helpful resource on using products | Specific benefits within [brand]'s use | **Valuable information about [brand]** |

The next six constructs (*Consumption Goal, Reward, Brand Post, Information Platform and Processes, Satisfaction* and *Reputation*) displayed through Table VII and Table VIII have a generous overall outlook after being analysed. Each obtained average value "d" beyond 0.2 [67]-[68], the percentage of experts' agreement surpassed 75% [67], as well as the defuzzification that exceeds 0.5 [69]-[70]. These results confirm that all of these constructs are influential as a determinant of a customer's engagement through the SNS platform.

Given the findings, 9 constructs for antecedents of customer engagement are having consensus among experts, showing by the "d" <= 0.2 [67]-[68] and the overall percentage "d" <= 0.2 exceeding 75% [67]. However, one of the antecedents (*Commitment*) is analyzed as unsuccessfully beyond the agreement of the experts involved. According to the results of defuzzification in Tables V - VIII, the scores obtained for all items of antecedents not including *Commitment*, are within the range of 0 to 1 and the $\alpha$-cut values have exceeded 0.5 [41], [69]-[70]. From the aspect of item updating, Fuzzy Delphi analysis has resulting the reduction of 9 out of 53 items reviewed by the expert group, bringing the current amount of items to 44 (labelled by Q1 – Q44 in Tables V - VIII). This has resulted in a collective question that is really appropriate and useful to target the objective of the actual study on the next level.

TABLE VII. VALUE "D" OF ITEM, VALUE "D" OF CONSTRUCT, % OF EXPERT CONSENSUS, DEFUZZIFICATION AND ITEM RANKING FOR *REWARD, BRAND POST* AND *INFORMATION PLATFORM AND PROCESSES*

| **Expert** | **REWARD** | | | **BRAND POST** | | | | **INFO PLATFORM & PROCESSES** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Q21** | **Q22** | **Q23** | **Q24** | **Q25** | **Q26** | **Q27** | **Q28** | **Q29** | **Q30** | **Q31** |
| 1 | 0.14 | 0.04 | 0.05 | 0.04 | 0.03 | 0.04 | 0.04 | 0.06 | 0.16 | 0.06 | 0.06 |
| 2 | 0.05 | 0.04 | 0.05 | 0.04 | 0.03 | 0.04 | 0.04 | 0.06 | 0.11 | 0.06 | 0.06 |
| 3 | 0.14 | 0.14 | 0.16 | 0.04 | 0.03 | 0.04 | 0.04 | 0.10 | 0.23 | 0.10 | 0.10 |
| 4 | 0.25 | 0.04 | 0.23 | 0.56 | 0.23 | 0.54 | 0.54 | 0.30 | 0.45 | 0.30 | 0.06 |
| 5 | 0.14 | 0.14 | 0.16 | 0.13 | 0.16 | 0.15 | 0.15 | 0.10 | 0.23 | 0.10 | 0.10 |
| 6 | 0.14 | 0.14 | 0.16 | 0.13 | 0.16 | 0.15 | 0.15 | 0.10 | 0.23 | 0.10 | 0.10 |
| 7 | 0.25 | 0.26 | 0.23 | 0.04 | 0.23 | 0.24 | 0.24 | 0.06 | 0.16 | 0.06 | 0.06 |
| 8 | 0.14 | 0.14 | 0.16 | 0.13 | 0.16 | 0.15 | 0.15 | 0.10 | 0.23 | 0.10 | 0.10 |
| 9 | 0.54 | 0.55 | 0.53 | 0.04 | 0.03 | 0.04 | 0.04 | 0.06 | 0.45 | 0.06 | 0.06 |
| 10 | 0.14 | 0.14 | 0.16 | 0.13 | 0.23 | 0.15 | 0.15 | 0.10 | 0.16 | 0.10 | 0.10 |
| 11 | 0.14 | 0.14 | 0.16 | 0.13 | 0.16 | 0.15 | 0.15 | 0.10 | 0.23 | 0.10 | 0.10 |
| 12 | 0.05 | 0.04 | 0.05 | 0.04 | 0.03 | 0.04 | 0.04 | 0.06 | 0.11 | 0.06 | 0.30 |
| Value "d" of each item | 0.18 | 0.15 | 0.17 | 0.12 | 0.12 | 0.14 | 0.14 | 0.10 | 0.23 | 0.10 | 0.10 |
| Value "d" construct | **0.17** | | | **0.13** | | | | **0.13** | | | |
| Percentage (%) of expert consensus for each item | 75% | 83% | 75% | 92% | 75% | 83% | 83% | 92% | 42% | 92% | 92% |
| Percentage (%) of expert consensus for construct | **78%** | | | **83%** | | | | **79%** | | | |
| Defuzzification | 0.86 | 0.85 | 0.87 | 0.87 | 0.85 | 0.85 | 0.85 | 0.90 | 0.78 | 0.90 | 0.92 |
| Item ranking in each construct | 2 | 3 | 1 | 1 | 2 | 3 | 4 | 2 | 4 | 3 | 1 |
| Item content | Monetary incentives by [brand] | Functional incentives by [brand] | **Psychological appreciation benefits by [brand]** | **Vivid features in posted ads, campaigns and messages** | Sufficient content posted ads, campaigns and messages | Fun and excitement with posted ads, campaigns and messages | Comments influenced by posted ads, campaigns and messages | Information searching through any search engine | Notification through the newsletter and mailing list | Presence of [brand] in various SNS channels | **Converse with well-trained, knowledgeable and proactive marketers** |

TABLE VIII. VALUE "D" OF ITEM, VALUE "D" OF CONSTRUCT, % OF EXPERT CONSENSUS, DEFUZZIFICATION AND ITEM RANKING FOR *SATISFACTION* AND *REPUTATION*

| Expert | SATISFACTION | | | | | REPUTATION | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Q32 | Q33 | Q34 | Q35 | Q36 | Q37 | Q38 | Q39 | Q40 | Q41 | Q42 | Q43 | Q44 |
| 1 | 0.08 | 0.05 | 0.22 | 0.12 | 0.15 | 0.11 | 0.16 | 0.06 | 0.05 | 0.23 | 0.04 | 0.03 | 0.03 |
| 2 | 0.08 | 0.05 | 0.09 | 0.12 | 0.15 | 0.05 | 0.05 | 0.06 | 0.05 | 0.05 | 0.04 | 0.03 | 0.03 |
| 3 | 0.16 | 0.14 | 0.16 | 0.18 | 0.21 | 0.11 | 0.16 | 0.10 | 0.11 | 0.16 | 0.15 | 0.13 | 0.13 |
| 4 | 0.18 | 0.21 | 0.17 | 0.14 | 0.10 | 0.05 | 0.05 | 0.06 | 0.05 | 0.05 | 0.04 | 0.03 | 0.03 |
| 5 | 0.08 | 0.05 | 0.09 | 0.12 | 0.15 | 0.11 | 0.16 | 0.10 | 0.11 | 0.16 | 0.15 | 0.13 | 0.13 |
| 6 | 0.08 | 0.18 | 0.09 | 0.12 | 0.15 | 0.11 | 0.16 | 0.10 | 0.11 | 0.16 | 0.15 | 0.13 | 0.13 |
| 7 | 0.18 | 0.21 | 0.17 | 0.43 | 0.40 | 0.28 | 0.23 | 0.30 | 0.28 | 0.23 | 0.24 | 0.27 | 0.27 |
| 8 | 0.08 | 0.05 | 0.09 | 0.12 | 0.15 | 0.11 | 0.16 | 0.10 | 0.11 | 0.16 | 0.15 | 0.13 | 0.13 |
| 9 | 0.08 | 0.21 | 0.46 | 0.43 | 0.10 | 0.28 | 0.23 | 0.06 | 0.28 | 0.53 | 0.54 | 0.27 | 0.27 |
| 10 | 0.47 | 0.05 | 0.17 | 0.14 | 0.10 | 0.05 | 0.53 | 0.06 | 0.05 | 0.05 | 0.04 | 0.03 | 0.03 |
| 11 | 0.21 | 0.18 | 0.22 | 0.26 | 0.40 | 0.05 | 0.05 | 0.10 | 0.11 | 0.16 | 0.15 | 0.13 | 0.13 |
| 12 | 0.08 | 0.05 | 0.09 | 0.12 | 0.15 | 0.11 | 0.16 | 0.10 | 0.11 | 0.16 | 0.04 | 0.03 | 0.03 |
| Value "d" of each item | 0.15 | 0.12 | 0.17 | 0.19 | 0.19 | 0.12 | 0.17 | 0.10 | 0.12 | 0.17 | 0.14 | 0.11 | 0.11 |
| Value "d" construct | **0.16** | | | | | **0.13** | | | | | | | |
| Percentage (%) of expert consensus for each item | 83% | 75% | 75% | 75% | 75% | 83% | 75% | 92% | 83% | 83% | 83% | 83% | 83% |
| Percentage (%) of expert consensus for construct | **77%** | | | | | **83%** | | | | | | | |
| Defuzzification | 0.80 | 0.83 | 0.79 | 0.76 | 0.78 | 0.88 | 0.85 | 0.89 | 0.87 | 0.84 | 0.85 | 0.87 | 0.87 |
| Item ranking in each construct | 2 | 1 | 3 | 5 | 4 | 2 | 6 | 1 | 3 | 8 | 7 | 4 | 5 |
| Item content | [brand] purchases | Support decision for [brand] | **Satisfactory experience with [brand]** | Completely pleased with [brand] | [Brand] is turning out better than expectation | Overall satisfied with [brand] | Good feeling about [brand] | Admire and respect [brand] | **Innovative products and services by [brand]** | High quality products and services by [brand] | [Brand] is an environmentally responsible | [Brand] maintains high standards in treating people | [Brand] tends to outperform its competitors | [Brand] has strong prospects for future growth |

## VI. DISCUSSION AND CONCLUSION

It has been concluded that all the constructs (except *Commitment*) are maintained and certified as antecedents of online customer engagement at SNS based on the consensus of experts. Using Fuzzy Delphi analysis, this study has proven the importance of customer engagement emerging concept in a systematic and trusted approach. It demonstrates that antecedents obtained from literature are validated to be drivers of new as well as existing customers to engage with brand and/or brand communities in SNS environment. The findings of the study are in line with it's purpose to answer the questions pertaining to the agreement of experts on antecedents of customer engagement in popular SNS marketing channels. Therefore, the listed items are feasible and appropriate to be forwarded to the next study sample, consisting of active customers interacting with the SNS community for brand-related activities.

The defuzzification process is greatly used to filter the priority of items. In *Perceived Relationship*, valuable relationships with brand have the greatest contribution to enabling customer engagement in the SNS channel (refer to Table V). For *Interaction* construct, experts have evaluated guidelines and support from firms and customers when interacting with SNS is the most important to engage someone (refer to Table V). Whereas in *Trust*, customers will only engage to the brands that he/she has a high level of confidence and can provide the services for problem-solving as being shown in Table VI. This table also displays that customers involved with SNS activities is hoping to aim valuable information acquisition in the first place from the point of view of brand and product use (see *Consumption Goal*).

Gathering psychological appreciation such as membership and entertainment in SNS brand community conversations have become the item that gets the most consent from experts (referring to the *Reward* in Table VII). In the same table, through the *Brand Post* dimension, the existence of vivid features in the ads, campaigns, and messages displayed through brand sites (including animations, different colors, videos, and pictures) also highlighted priorities over the other three items. *Information Platform and Processes* dimension show the most numerous deals from experts are through customer dependency with trained, knowledgeable and proactive marketers on the SNS brand page. Meanwhile, Table

VIII shows the position of the customer's items in engaging with the brand within SNS through the antecedents of *Satisfaction* and *Reputation*. The result has confirmed that a satisfactory experience from existing customers dominates the expert agreement on *Satisfaction*. In parallel, factor *Reputation* concludes that innovative products and services by brands acquire the highest expert's choice.

The above findings clearly provide twice the implications. In view of the contribution of knowledge, research in online customer behaviour supports the limited, theoretical and empirical scientific materials, on the SNS platform. This finding can also be a space for future researchers who stick to the same theory (relationship marketing, service-dominant, social exchange and user gratification theories) to translate the conceptual elements in existing models into more useful practical studies. Fuzzy Delphi technique, as far as researcher knowledge, has not been used to verify the study material in the domain of customer engagement in SNS. Throughout the lens of this study, the Fuzzy Delphi method is proved to be the effective tool for validating the fuzzy items of multi-dimensional determinants for customer engagement, which are adapted from the literature with different context and setting. It also increases the statistical use of Fuzzy Delphi analysis among IT/IS research that has been proven extensively used before.

The study also offers practical implications that support important concepts to the business world such as customer engagement. In addition to the dominant elements in theories such as *Trust*, *Reputation*, and *Satisfaction*, stakeholders, especially business owners and marketers, need to take into account the other determinants dimensions that greatly affect digital subscribers to engage, such as the content and quality of postage advertisements published on the SNS interface, and the effectiveness of the delivery and information processing provided as an intermediary between their parties and customers. Firms accordingly can proactively cultivated the life cycle of marketing initiatives through the SNS channel where their brands and products need to be successfully maintained.

## ACKNOWLEDGEMENT

The author would like to thank Universiti Teknikal Malaysia Melaka (UTeM) for providing research facilities at the time of conduction of this study and for the financial incentive for publishing this paper.

## REFERENCES

[1] P. Kallas, "Top 15 Most Popular Social Networking Sites and Apps [August 2018]", Dreamgrow, 2018. [Online]. Available: https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/. [Accessed: Sep. 16, 2018].

[2] L. Dokyun, K. Hosanagar and H. Sasikumar Nair, "The effect of social media marketing content on consumer engagement: evidence from Facebook. Available at SSRN, 2290802, 2018.

[3] I. Jamid UI and Z. Rahman, "Islam, Jamid Ul, and Zillur Rahman. "Linking customer engagement to trust and word-of-mouth on Facebook brand communities: An Empirical Study", Journal of Internet Commerce, vol. 15, no. 1, pp. 40-58, 2018.

[4] A. Sahabi Y and A. H. Busalim, "Antecedents of eWOM in social commerces", International Conference on Research and Innovation in Information Systems, ICRIIS. 2018, 2018, pp. 1-6.

[5] S. Chu and Y. Kim, "Determinants of consumer engagement in Electronic word-of-mouth (eWOM) in Social Networking Sites", International Journal of Advertising, vol. 30, no. 1, pp. 47-75, 2011.

[6] T. Aichner and F. Jacob, "Measuring the degree of corporate social media use," International Journal of Market Research, vol. 57, no. 2, pp. 257–275, 2015.

[7] Z. Zhang, X. Li, and Y. Chen, "Deciphering word-of-mouth in social media," ACM Transactions on Management Information Systems, vol. 3, no. 1, pp. 1–23, Apr. 2012.

[8] V. Blazevic et al., "Beyond traditional word-of-mouth: an expanded model of customer-driven influence." Journal of Service Management, vol. 24, no. 3, pp. 294–313, 2013.

[9] T. Hennig-Thurau, C. F. Hofacker, and B. Bloching, "Marketing the pinball way: understanding how social media change the generation of value for consumers and companies," Journal of Interactive Marketing, vol. 27, no. 4, pp.237-241, 2013.

[10] J. Wirtz et al., "Managing brands and customer engagement in online brand communities," Journal of Service Management, vol. 24, no. 3, pp. 223–244, Jun. 2013.

[11] R. Hanna, A. Rohm, and V. L. Crittenden, "We're all connected: the power of the social media ecosystem." Business Horizons, vol. 54, no. 3, 265–73, 2011.

[12] S. D. Vivek, S. E. Beatty, and R. M. Morgan, "A generalized multidimensional scale for measuring customer engagement," Journal of Marketing Theory and Practice, vol. 22, no. 4, pp. 401–420, 2014.

[13] "2018-2020 Research Priorities : Cultivating the customer asset", Marketing Science Institute, 2018. [Online]. Available: http://www.msi.org/research/2018-2020-research-priorities/cultivating-the-customer-asset/. [Accessed: Nov. 07, 2017].

[14] C. Yu-Ting, H. Yu and H. Lu, "Persuasive messages, popularity cohesion, and message diffusion in social media marketing", Journal of Business Research, vol. 68, no. 4, pp. 777-782, 2015.

[15] O. M. Zain and M. B. Saidu, "The customers satisfaction on retailers' brand products: a study on selected areas in Klang Valley," Procedia-Economics and Finance, vol. 35, pp. 418–427, 2016.

[16] L. Carolyn A and T. Kim, "Predicting user response to sponsored advertising on social media via the technology acceptance model", Computers in human behavior, vol. 64, pp. 710-718, 2016.

[17] J. Titcomb, "Facebook now has 2 billion users, Mark Zuckerberg announces", Telegraph, 2017. [Online]. Available: https://www.telegraph.co.uk/technology/2017/06/27/facebook-now-has-2-billion-users-mark-zuckerberg-announces/. [Accessed: Oct. 15, 2017].

[18] "Number of monthly active Facebook users worldwide as of 2nd quarter 2018 (in millions)", Statista, 2018. [Online]. Available: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/. [Accessed: Aug. 12, 2018].

[19] J. Van Doorn et al., "Customer engagement behavior: theoretical foundations and research directions." Journal of Service Research, vol. 13, no. 3, pp. 253-266, 2010.

[20] L. Dessart, C. Veloutsou, and A. Morgan-Thomas, "Consumer engagement in online brand communities: a social media perspective," Journal of Product & Brand Management, vol. 24, no. 1, pp. 28–42, Mar. 2015.

[21] A. Ilic, R. J. Brodie, L. D. Hollebeek, and B. Juric, "Customer engagement : conceptual domain, fundamental propositions, and implications for research," Journal of Service Research, vol. 14, no. 3, pp. 252-271, 2011.

[22] P. C. Verhoef, W. J. Reinartz, and M. Krafft, "Customer engagement as a new perspective in customer management," Journal of Service Research, vol. 13, no. 3, pp. 247–252, 2010.

[23] E. Jaakkola and M. Alexander, "The role of customer engagement behavior in value co-creation : a service system perspective," Journal of Service Research, vol. 17, no. 3, pp. 247–261, 2014.

[24] V. Kumar, L. Aksoy, B. Donkers, R. Venkatesan, T. Wiesel, and S. Tillmanns, "Undervalued or Overvalued Customers: Capturing Total Customer Engagement Value," Journal of Service Research, vol. 13, no. 3, pp. 297-310, 2010.

[25] T. H. Bijmolt, M. Eisenbeiss, B. G. Hardie, A. Lemmens and P. Saffert, “Analytics for customer engagement.” Journal of Service Research, vol. 13, no. 3, pp. 341-356, 2010.

[26] A. Javornik and A. Mandelli, “Behavioral perspectives of customer engagement: an exploratory study of customer engagement with three Swiss FMCG brands,” Journal of Database Marketing & Customer Strategy Management, vol. 19, no. 4, pp. 300–310, 2012.

[27] W. Kunz, L. Aksoy, H. Hall, Y. Bart, K. Heinonen, and R. Marketing, “Customer engagement in a big data world,” Journal of Services Marketing, vol. 31, no. 2, pp. 1–39, 2017.

[28] S. D. Vivek, S. E. Beatty and R. M. Morgan, “Customer Engagement: Exploring customer relationships beyond purchase.” Journal of Marketing Theory and Practice, vol. 20, no. 2, pp. 122-146, 2012.

[29] S. Nambisan and R. A. Baron, “Virtual customer environments: testing a model of voluntary participation in value co-creation activities,” Journal of Product Innovation Management, vol. 26, no. 4, pp. 388–406, Jul. 2009.

[30] G. Greve, “ScienceDirect The moderating effect of customer engagement on the brand image – brand loyalty relationship,” Procedia - Social Behavior. Science, vol. 148, no. 148, pp. 203–210, 2014.

[31] M. R. Habibi, M. Laroche, and M.-O. Richard, “The roles of brand community and community engagement in building brand trust on social media,” Computers in Human Behavior, vol. 37, pp. 152-161, 2014.

[32] B. J. Baldus, C. Voorhees, and R. Calantone, “Online brand community engagement: Scale development and validation,” Journal of Business Research, vol. 68, no. 5, pp. 978–985, 2015.

[33] F. Kujur and S. Singh, “Engaging customers through online participation in social networking sites,” Asia Pacific Management Review, vol. 22, no. 1, pp. 16–24, Mar. 2017.

[34] J. Gutiérrez-Cillán, C. Camarero-Izquierdo, and R. San José-Cabezudo, “How brand post content contributes to user’s Facebook brand-page engagement. The experiential route of active participation,” BRQ Business Research Quarterly, vol. 20, no. 4, pp. 258–274, 2017.

[35] A. Mollen, H. Wilson, M. Oxon, and C. Dipcompsci, “Engagement, Telepresence and Interactivity in Online Consumer Experience: Reconciling Scholastic and Managerial Perspectives,” Journal of Business Research Special Issue Internet Customer Behavior, vol. 63, no. 910, pp. 919–925, 2010.

[36] D. E. Schultz and J. Peltier, “Social media’s slippery slope: challenges, opportunities and future research directions,” Journal of Research in Interactive Marketing, vol. 7, no. 2, pp. 86–99, May 2013.

[37] J. Carlson, M. Rahman, R. Voola, and N. De Vries, “Customer engagement behaviours in social media: capturing innovation opportunities,” Journal of Services Marketing, vol. 32, no. 1, pp. 83–94, Feb. 2018.

[38] C. M. K. Cheung, M. K. O. Lee, and X.-L. Jin, “Customer Engagement in an Online Social Platform: A Conceptual Model and Scale Development.” in Thirty Second International Conference on Information Systems, 2011.

[39] M. A. Bashir and N. A. Ali, “Impact of customer brand relationship through Facebook on the level of customer engagement,” Pakistan Business Review, no. April 2016, pp. 159–178, 2016.

[40] L. Sigerson and C. Cheng, "Scales for measuring user engagement with social network sites: A systematic review of psychometric properties", Computers in Human Behavior, vol. 83, no. 2018, pp. 87-105, 2018.

[41] M. Mohd Jamil, S. Siraj, Z. Hussin, N. M. Noh and A. Arifin, Pengenalan asas kaedah Fuzzy Delphi dalam penyelidikan rekabentuk pembangunan. Minda Intelek, Bangi, 2014.

[42] M. Mohd Jamil, S. Siraj, Y. Farazila, M. N. Nurulrabihah, Z. Hussin and S. Ahmad Arifin. " Aplikasi teknik Fuzzy Delphi terhadap keperluan elemen keusahawanan bagi pensyarah kejuruteraan politeknik Malaysia." International Journal of Business and Technopreneurship, vol. 5, no. 1, pp. 135-150, 2015.

[43] S. Gregory, J, F. T. Hartman and J. Krahn, "The Delphi method for graduate research.", Journal of Information Technology Education: Research, vol. 6, pp. 1-21, 2007.

[44] R. C. Gambetti, G. Graffigna, and S. Biraghi, “The Grounded Theory approach to consumer–brand engagement,” International Journal of Market Research, vol. 54, no. 5, pp. 659-687, 2012.

[45] L. D. Hollebeek, M. S. Glynn, and R. J. Brodie, “Consumer brand engagement in social media: conceptualization, scale development and validation,” Journal of Interactive Marketing, vol. 28, no. 2, pp. 149–165, May 2014.

[46] A. Abdullah and D. S. Siraj, “Brand equity through customer engagement in social media: a critical review,” IOSR Journal of Business and Management, vol. 18, no. 08, pp. 38–46, 2016.

[47] F. J. Martínez-López, R. Anaya-Sánchez, S. Molinillo, R. Aguilar-Illescas, and I. Esteban-Millat, “Consumer engagement in an online brand community,” Electronic Commerce Research and Applications, vol. 23, pp. 24–37, 2017.

[48] R. Thakur and S. P. Jain, “Customer engagement and online reviews,” Journal of Retailing and Consumer Services, 41, pp.48-59, 2017.

[49] H. Marsden, Guard your reputation online. Smartebookshop, Birmingham, 2013.

[50] C. M. Sashi, “Customer engagement , buyer - seller relationships , and social media.” Management decision, vol. 50, no. 2, pp. 253-272, 2012.

[51] E. Gummesson, “Making relationship marketing operational”, International Journal of Service Industry Management, vol. 5, no. 5, pp. 5-20, 1994.

[52] M. Clark and J. Melancon, “The Influence of social media investment on relational outcomes: A relationship marketing perspective”, International Journal of Marketing Studies, vol. 5, no. 4, pp. 132 – 140, 2013.

[53] K. Filo, D. Lock and A. Karg, “Sport and social media research: A review.” Sport Management Review, vol. 18, no. 2, pp. 166-181, 2015.

[54] R. Mustapha and G. Darusalam, Aplikasi kaedah Fuzzy Delphi dalam Sains Sosial. University of Malaya Press, 2017.

[55] R. Mustapha, Z. Hussin and S. Siraj, "Analisis faktor penyebab ketidakjujuran akademik dalam kalangan mahasiswa: aplikasi teknik Fuzzy Delphi.", Jurnal Kurikulum & Pengajaran Asia Pasifik, vol. 2, no. 2, pp. 1-17, 2017.

[56] N. Mat Noh, S. Abd Razak, N. Alias, S. Siraj, M. Mohd Jamil and Z. Hussin, "Usage of Facebook : the future impact of curriculum implementation on students in Malaysia", Procedia-Social and Behavioral Sciences, vol. 103, pp. 1261-1270, 2013.

[57] E. Delgado-Ballester, J. L. Munuera-Aleman and M. J. Yague-Guillen, “Development and validation of a brand trust scale,” International Journal of Market Research, vol. 45, no. 1, pp. 35–54, 2003.

[58] U. M. Dholakia, V. Blazevic, C. Wiertz, and R. Algesheimer, “Communal service delivery: how customers benefit from participation in firm-hosted virtual P3 communities,” Journal of Service Research, 2009.

[59] T. Hennig-Thurau, K. P. Gwinner, G. Walsh, and D. D. Gremler, “Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the Internet?,” Journal of Interactive Marketing, vol. 18, no. 1, pp. 38–52, Jan. 2004.

[60] A. Gustaffsson, M. D. Johnson, and I. Roos, “The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention,” Journal of Marketing, vol. 69, no. 4, pp. 210-218, 2006.

[61] H. Jang, L. Olfman, I. Ko, J. Koh and K. Kim, “The influence of online brand community characteristics on community commitment and brand loyalty.” International Journal of Electronic Commerce, vol. 12, no. 3, pp. 57-80, 2008.

[62] J. J. Cronin, M. K. Brady, and G. T. M. Hult, “Assessing the effects of quality, value, and customer satisfaction on consumer behavioral intentions in service environments,” Journal of Retailing, vol. 76, no. 2, pp. 193-218, 2000.

[63] C. J. Fombrun, N. A. Gardberg and J. M. Sever, “The Reputation QuotientSM: a multi-stakeholder measure of corporate reputation.” Journal of Brand Management. vol. 7, no. 4, pp. 241-255, 2000.

[64] L. De Vries, S. Gensler, and P. S. H. Lee, “Popularity of brand posts on brand fan pages : an investigation of the effects of social media marketing,” Journal of Interactive Marketing, vol. 26, pp. 83–91, 2012.

[65] T. J. Gordon, The Delphi Method. Future Research Methodology, 1994.

[66] M. Adler and E. Ziglio, Gazing into the oracle: the Delphi Method and its application to social policy and public health. London: Jessica Kingsley Publishers, 1996.

[67] P. L. Chang, C. W. Hsu and P. C. Chang, "Fuzzy Delphi Method for evaluating hydrogen production technologies." International Journal of Hydrogen Energy. pp. 14172 – 14179, 2011.

[68] C.H. Cheng and Y. Lin, "Evaluating the best main battle tank using Fuzzy Decision Theory with linguistic criteria evaluation." European Journal of Operational Research, vol. 142, no. 1, pp. 174-186, 2002.

[69] S. Bodjanova, "Median alpha-levels of a Fuzzy Number." Fuzzy Sets and Systems. vol. 157, no. 7, pp. 879-891, 2006.

[70] C. W. Tang and C. T. Wu, "Obtaining a picture of undergraduate education quality: a voice from inside the university." Higher Education, vol. 60, pp. 269–286, 2010

[71] Dovaliene, A. Masiulyte, and Z. Piligrimiene, "The relations between customer engagement, perceived value and satisfaction: the case of mobile applications," Procedia-Social and Behavioral Sciences, vol. 213, pp. 659–664, Dec. 2015.

# Effect of Fusion of Statistical and Texture Features on HSI based Leaf Images with both Dorsal and Ventral Sides

Poonam Saini[1], Arun Kumar[2]
Sir Padampat Singhania University
Udaipur, India

***Abstract*—The present work involves statistically analyzing and studying the overall classification accuracy results using Hue channel images of different plant species using their dorsal and ventral sides, and then subjecting them to the process of feature extraction using first order statistical features and texture based features. These extracted features have been subjected to the classification process using KNN and Random Forest algorithms. Further, this work studies the fusion of two different kinds of features extracted for dorsal and ventral plant leaf images and studying the effect of fusion on the overall classification accuracy results. This work also delves into the feature selection task using random forest algorithm and studies the effect of reduced dataset with unique features on the overall classification accuracy results. The most important outcome of this investigation is that the ventral leaf images can be a suitable alternative for plant species classification using digital images and further, the fusion of features does improve the classification accuracy results.**

***Keywords*—Dorsal; ventral; leaf classification; random forest; texture features; statistical features***

## I. Introduction

The human beings have been gifted with brain and eyes for proper discrimination of objects found in their daily lives. While discriminating the objects, the images formed in the eye seek important valuable features from the objects of concern and store those cues for future reference. The eyes often seek color, shape and texture features for identification process. The color is a very important discrimination parameter, as one always sees objects with color as almost first criteria for discrimination. As the picture is formed in the eyes, the brain interprets these visual cues, the color is one of the most important factors that is stored as valuable information for discrimination purpose. Secondly, the shape of the object being regular or irregular polygon, forms the other important parameter for discrimination by the human brain. Thirdly, the property of smoothness or roughness of the surface of the object, i.e. the sense of touch has its role in the discrimination of the objects by the human brain.

The plants have been studied for their fruits, seeds, roots, flowers and leaves etc. The scientific methods for the proper taxonomic classification of the plants and their species have been there with the biologists for quite a long time, but the role of image processing professionals by applying the latest state of the art machine learning techniques in discriminating the plants through their digital images is a recent one, and this has become possible with the improvement of computational efficiency of the computers and also the decrease in the prices of the hardware. There are millions of plants and their subspecies and not all of them have been taxonomically classified and this all accounts for the slow identification tools and techniques of the past and vastness of the flora on this earth. Therefore, in order to preserve the plants for future generation, one needs to understand them and taxonomically classify them for any future reference.

The research work carried out by [1] has prepared a mobile biodiversity informatics tools for identifying and mapping Indonesian medicinal plants. The system - called MedLeaf – has been developed as a prototype data resource for documenting, integrating, disseminating, and identifying of Indonesian medicinal plants. The results indicate that combination of leaves features outperform than by using single features with an accuracy value of 88.5%.

The researchers in [2] have implemented an image feature extraction method using morphological techniques and have worked on leaf diseased dataset containing leaf spot disease and leaf blight. The results have been obtained using morphological features.

The researchers in [3] have carried work on identification of whitefly on the plants using image processing technique. The segmentation results using triangle method has an accuracy of 75.36%.This suggests that triangle method can be used for the whitefly segmentation process in vegetable crops leaf images.

The leaf image texture features have been extracted by [4], through Gabor based techniques and then subjecting them to PSO-CFS based search method for identifying the best set of features from the complete feature set and then classifying them using four classification algorithms like KNN, J48, CART and RF. Another objective of [4] was to utilize the two faces available on the plant leaves (Dorsal and Ventral), instead of one (i.e. Dorsal) for classification of plants on the basis of digital leaf images and to analyze the effects on classification accuracy values for dorsal and ventral sides of leaf images. The accuracy achieved using this work has been 92.09%.

The researchers in [5] have worked on extraction of standard tobacco leaves based on HIS color space, wherein the H, S and I components have been quantified and extracted by

using color histogram, and then the average value corresponding to every color component was calculated and the grading for tobacco leaves was done on this basis.

The researchers in [6], have utilized eight varieties of crop images. In this work, the authors have worked with texture features obtained using GLCM method and color features using HSV have been deployed. The artificial neural network (ANN) has been used for classification of plants in this work. The highest value of classification accuracy achieved by this work is 84.375%.

The GLCM and FFT techniques for feature extraction from the color images have been used by [7] and have achieved an accuracy value of 86% for the GLCM based color feature set.

The researchers in [8] have used the concept of chromaticity moments for the images and has extracted the texture features for the images. The accuracy value achieved is as high as 90% using SVM algorithm for classification of diseased plants.

The work carried out by [9] considers the classification of a household plant used for its medicinal properties by using the shape and color features obtained from the leaves of the plant. This work has utilized the shape descriptors represented by Dyadic Wavelet transform and Zernike complex moments along with the HSV based color features and have obtained a classification accuracy value of 81.77%.

In the present work, the section II explains about the methodology adopted for carrying out the research work. It explains the role of HSI color models for extraction of color features. This section also explains the extraction of first order statistical features and texture features and the preparation of feature database and along with it, it also describes about the computation of variance values between various features using PCA. The section III describes about the feature selection methodology adopted to prepare the feature subsets. The section IV discusses the results obtained for various subsets of data created in the previous sections along with their statistical analysis and comparison of the present results with other works of similar nature which is followed by the conclusion drawn from this work in section V.

## II. Materials and Methods

The plant leaves have two faces viz.: dorsal and ventral, therefore there is a need for paying critical attention on both the sides of the leaves, as both have independent and unique set of features. The existing leaf image databases available on the internet contain the leaf images of dorsal sides, but to achieve the objective of this work, there is a need to create an independent leaf image database with dorsal as well as ventral sides of the leaves. Therefore, 25 dorsal side and 25 ventral side leaf images were clicked for each plant species and a database of ten plant species has been created with 250 dorsal and 250 ventral sides of leaf images and 500 images in totality. A sample of such colored leaf images with dorsal and ventral sides is shown in Fig. 1 by [10]. The 500 images captured were subjected to the process of background removal, size reduction to 256 X 256.

Fig. 1. A Sample of Collected Leaf Images with Dorsal and Ventral Sides.

The human eye considers the color as important characteristic feature for the recognition of the objects visible through naked eye in the images as well as for various image processing tasks like segmentation and detection etc. The color features are defined subject to a particular color space or model. There are a number of color spaces that have been used by different researchers, such as RGB, LUV, HIS. By using the different color spaces, the different color features can be extracted from images or the regions of interest in the images.

### A. Extraction of Color Features using HSI Color Model

There has been numerous works carried out on the principles of binary and gray image processing techniques. The image processing tasks carried out on the basis of color image processing have high computational dependency and the earlier works reduced the images into two dimensions color maps to reduce the time consumed in computing. Certainly by doing so, there is loss of information upon the conversion of color images to binary or gray levels. The growth of computing power, the storage capacity and the reduced costs of capture systems and printing images have made the image acquisition and further processing in color easier and reasonably computationally cheaper. The RGB color model has three color components (Red, Green or Blue) in varying proportions. To overcome the disadvantages of RGB model in various image processing tasks, HSI model fares better and is a simple substitute of human visual system. The HSI model contains three matrices of *mXn* dimensions containing values with respect to hue, saturation and intensity. Each pixel of the image prepared according to this model has hue and saturation values indicating the color information contained in it and intensity indicates the brightness.

Therefore, the HSI color space model is suitable for detection and analysis of the color characteristic properties. By using (1), (2) and (3), the five hundred RGB images captured for ten different plant species were subjected to the process of RGB to HSI conversion process. The HSI images prepared using above equations were converted into a stack of images and then the stack was split into three independent channels (H, S and I) respectively.

As discussed earlier, the color information is contained in the hue channel, therefore the hue channel images were further subjected to the process of feature extraction using two techniques namely the general first order statistical features and second order texture features using the co-occurrence

matrix. There were ten statistical features: Mean, Standard Deviation(StdDev), Max, XM, YM, Integration Density (InteDen), Median,Skewness (Skew), Kurtosis (Kurt), Area_Percent. that were extracted for this study using Fiji by [11] and the different values were stored in a CSV file.

Texture is another important property of images. It is generally believed that human visual systems use texture for recognition and interpretation. The texture is a measure of the intensity variation of a surface which quantifies properties such as smoothness and regularity. The texture, on its own does not have the capability of finding similar images, but it can be used to classify textured images from non-textured ones and then be combined with another visual attribute like color to make the retrieval more effective.

The researchers in [8] have proposed 14 texture features, of which 11 have been chosen and they are Contrast(Contr), Homogeneity, Angular Second Moment (ASM), IDM, Energy, Entropy, Variance, correlation, Inertia, Shade and Prominance. The GLCM matrix was prepared for $0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}$ with an offset of unity in ImageJ using the GLCM plugin for batch processing given by [11]. The GLCM matrix of the dorsal and ventral sides of the leaf image was prepared and combined together. The complete matrix contains 600 rows per degree with unit offset of data for the above fourteen attributes and in totality 2400 rows for all the four degrees and with unit offset values was prepared.

$$I = \frac{1}{3}\left(R + G + B\right) \tag{1}$$

$$S = 1 - \frac{3}{\left(R + G + B\right)}\left[\min\left(R, G, B\right)\right] \tag{2}$$

$$H = \cos^{-1}\left\{\frac{\frac{1}{2}\left[\left(R - G\right) + \left(R - B\right)\right]}{\left[\left(R - G\right)^2 + \left(R - B\right)\left(G - B\right)\right]^{1/2}}\right\} \tag{3}$$

### B. Preparation of Database with Different Feature Set

The first order statistical database has been prepared using the 10 first order features mentioned in section (IIA). Two independent features sets have been prepared for dorsal and ventral side images of leaves using hue channel based images. The ten features extracted are Mean(SD1 or SV1), StdDev(SD2 or SV2), Max(SD3 or SV3), XM(SD4 or SV4), YM(SD5 or SV5), IntDen(SD6 or SV6), Median(SD7 or SV7), Skew(SD8 or SV8), Kurt(SD9 or SV9), Area_Percent(SD10 or SV10). Here D in $SD_i$ indicates the dorsal and V in $SV_i$ indicates ventral leaf images.

The two datasets are called HDIM (Hue-Dorsal-Istorder-Measure) and HVIM (Hue-Ventral-Istorder-Measure) as represented through (4) and (5) respectively. Here $SD_1, SD_2, ..., SD_{10}$ indicate all the 10 different values of first order statistical features obtained for dorsal images and the $SV_1, SV_2, ..., SV_{10}$ indicate all the 10 different values of statistical first order features obtained for the ventral images.

$$HDIM = \left(SD_1, SD_2, SD_3, SD_4, SD_5, SD_6, SD_7, SD_8, SD_9, SD_{10}\right) \tag{4}$$

$$HVIM = \left(SV_1, SV_2, SV_3, SV_4, SV_5, SV_6, SV_7, SV_8, SV_9, SV_{10}\right) \tag{5}$$

By using the gray level co-occurrence feature extraction methodology as mentioned in section (IIA), the texture feature dataset has been prepared for the dorsal as well as the ventral leaf images for which the Hue channel has been extracted.

$$HDGM = \left(TD_1, TD_2, TD_3, TD_4, TD_5, TD_6, TD_7, TD_8, TD_9, TD_{10}, TD_{11}\right) \tag{6}$$

$$HVGM = \left(TV_1, TV_2, TV_3, TV_4, TV_5, TV_6, TV_7, TV_8, TV_9, TV_{10}, TV_{11}\right) \tag{7}$$

The two datasets are called HDGM(Hue-Dorsal-GLCM-Measure) and HVGM(Hue-Ventral-GLCM-Measure) as represented through (6) and (7) respectively. Here $TD_1, TD_2, ..., TD_{11}$ indicate all the 11 different values of texture features obtained for the dorsal images and the $TV_1, TV_2, ..., TV_{11}$ indicates all the 11 different values of texture features obtained for the ventral images.

There are two more datasets prepared by combining all the dorsal and ventral features together and named as HDGIM(Hue-Dorsal-GLCM-Istorder-Measure) and HVGIM(Hue-Ventral-GLCM-Istorder-Measure) and are represented through (8) and (9) respectively, where HDGIM has been obtained by combining the features obtained through (4) and (6), whereas HVGIM has been obtained by combining features of (5) and (7).

$$HDGIM = HDIM + HDGM \tag{8}$$

$$HVGIM = HVIM + HVGM \tag{9}$$

### C. Studying the Variance amongst the Various Features Extracted using PCA

The feature sets prepared in subsection B have been thoroughly studied for the correlation or variance amongst each other so that only those features could be selected which are least correlated and are unique in nature. By using unique features, the size of the dataset is reduced and the overall computation time is reduced as well for discrimination of the datasets using images.

The PCA component plots have been shown in the Fig. 2, 3, 4 and 5 for different datasets. These plots demonstrate the importance of unique features required for the classification process. The concept of PCA (Principal Component Analysis) algorithm as discussed by [11] and [12], is an unsupervised method, it has been in vogue for the dimensionality reduction process in almost all the literature concerning the classification of data.

This technique is used for extracting important variables (in form of components) from a large set of variables available in a data set. It extracts low dimensional set of features from a high dimensional data set with a sole motive to capture as much information as possible. By using a fewer variables, the process of visualization of data also becomes much more meaningful and the behavior of the data can be studied in a

better way. The concept of PCA is more useful when dealing with three or higher dimensional data variables. It is a technique to combine similar or correlated items or variables. The normalized data is subjected to the process of PCA and only those dimensions are selected which have the highest value of variance and this process churns out the highly correlated items from the variable set. The Fig. 2, 3, 4 and 5 show the plots with two PC's (principal components, PC1 and PC2) obtained for the various variables used. The principal component is a normalized linear combination of the original predictors in a data set.

**The first principal component (PC1)** is a linear combination of original predictor variables which captures the maximum variance in the data set. It determines the direction of highest variability in the data. The larger the variability captured in first component PC1, the larger is the information captured by that component and no other component can have variability higher than first principal component. The first principal component results in a line which is closest to the data i.e. it minimizes the sum of squared distance between a data point and the line. **The second principal component** (PC2)**,** is also a linear combination of original predictors which captures the remaining variance in the data set and is uncorrelated with PC1. In other words, the correlation between first and second component should is zero. If the two components are uncorrelated, their directions should be orthogonal. The Fig. 2 and 3 show the plot for principal component and their values. The Fig. 2 shows that the first principal component of HDIM dataset has a value of 42.08% and the values for other components are lower than this value. For HVIM dataset the Fig. 3 shows, the highest value of principal component is 52.79%. In Fig. 4 for HDGM dataset, it is 55.51% and that of HVGM, it is 60.16% as shown in Fig. 5.

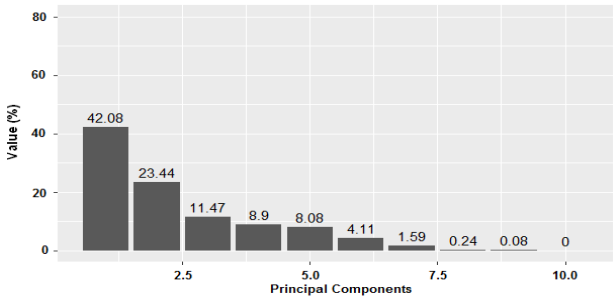



Fig. 2. PCA Component of HDIM.

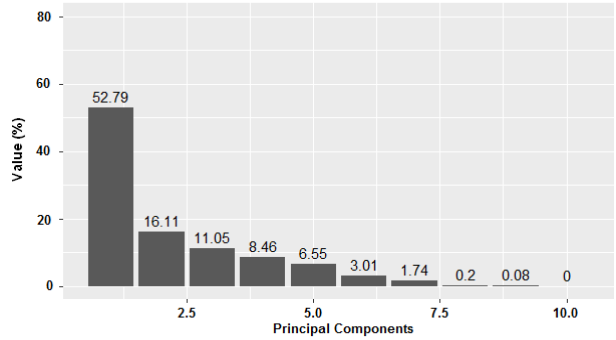



Fig. 3. PCA Component of HVIM.

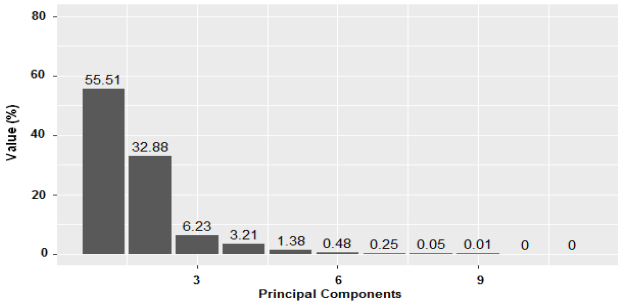



Fig. 4. PCA Component of HDGM.

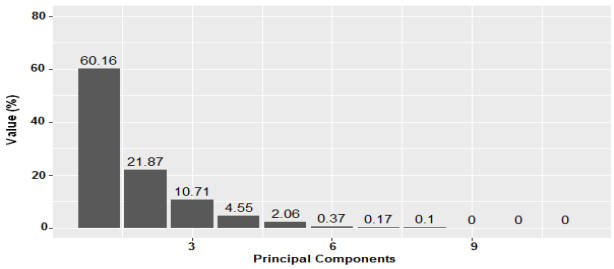



Fig. 5. PCA Component of HVGM.

## III. Adoption of Feature Selection Methodology for Feature Subset Selection

The concept of feature selection also known as variable selection, feature reduction, attribute selection or variable subset selection, is a widely used dimensionality reduction technique, which has been the focus of much research in machine learning and data mining and has found applications in text classification, web mining, and so on . It allows faster model building by reducing the number of features, and also helps removing irrelevant, redundant and noisy features. This begets simpler and more comprehensible classification models with better classification performance results.

Hence, selecting relevant attributes are a critical issue for competitive classifiers and for data reduction. In this present work, random forest technique has been used for feature subset selection which identifies unique features from large datasets.

### *A. Use of Random Forest as a Feature Selector for Feature Subset Selection*

Random Forest directly performs feature selection while a classification rule is built. The two commonly used variable importance measures in RF are Gini importance index and permutation importance index (PIM) .In this paper, two step approach has been used for feature selection. In first step, permutation importance index are used to rank the features and then in second step, Random Forest is used to select the best subset of features for classification. This reduced feature set is then subjected to the process of plant species classification using images of their leaves.

The high dimensional nature of many tasks in pattern recognition has created an urgent need for feature selection techniques. The goal of feature selection in this field is manifold, where the two most important are: to avoid over fitting and improve model performance, and to gain a deeper

insight into the underlying processes that generated the data [12]. The interpretability of machine learning models is treated as important as the prediction accuracy for most life science problems. Unlike most other classifiers, Random Forest directly performs feature selection while a classification rule is built by [13]. Permutation importance measure (PIM) given by [14] is arguably the most popular variable's importance measure used in RF.

### B. *Preparation of Different Feature Subsets using Random Forest Feature Selection Technique*

By suitably applying the RF, one can examine which variables are working the best or worst in each of the trees. In this study, CARET package developed by [13] has been used for finding the importance of the features which builds up a model. The decision trees help in determining the variable importance.

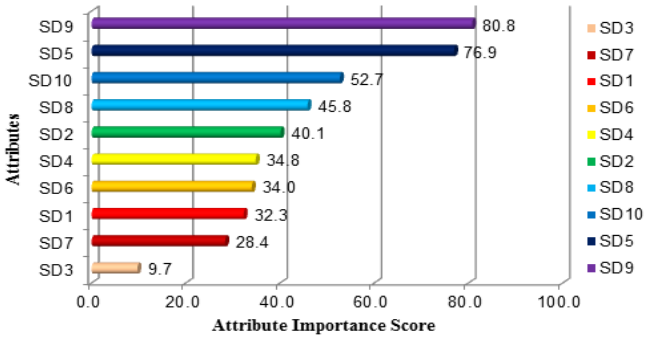



Fig. 6. Attribute Importance Plot for HDIM.

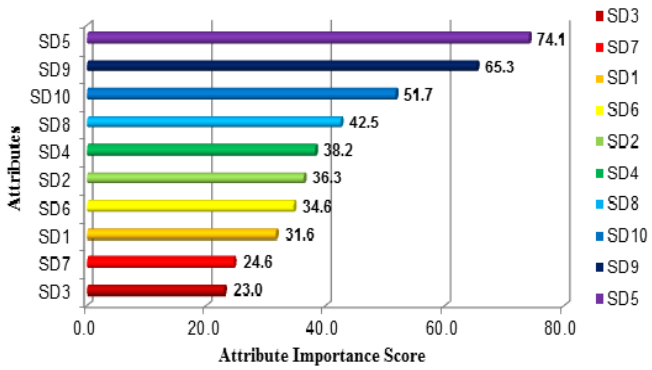



Fig. 7. Attribute Importance Plot for HVIM.

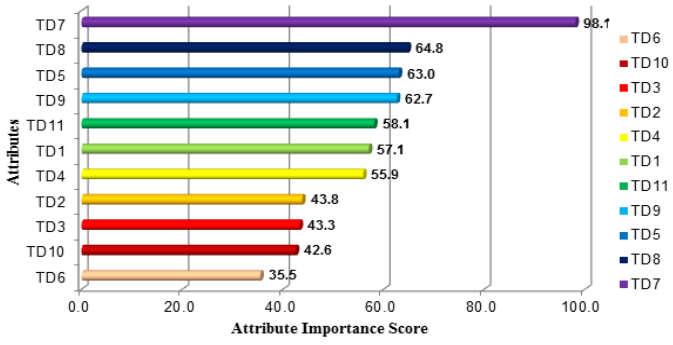



Fig. 8. Attribute Importance Plot for HDGM.

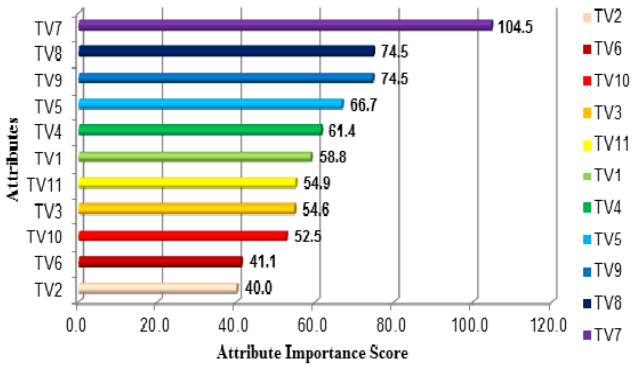



Fig. 9. Attribute Importance Plot for HVGM.

The cross validation technique has been utilized for identifying the error rate and this helped in identifying the fitness of the individual features. The Fig. 6, 7, 8, 9, 10 and 11 clearly depict the attribute importance value for each variable which helps in choosing the unique variables and forming the feature subsets to be utilized for the classification process and represent HDIM, HVIM, HDGM, HVGM, HDGIM and HVGIM feature datasets respectively and arranged in ascending order of variable importance. In the present work; seven unique features have been selected from each of the feature datasets mentioned in Fig. 6, 7, 8, 9, 10 and 11 for further classification process. These features have been selected on the basis of feature score.

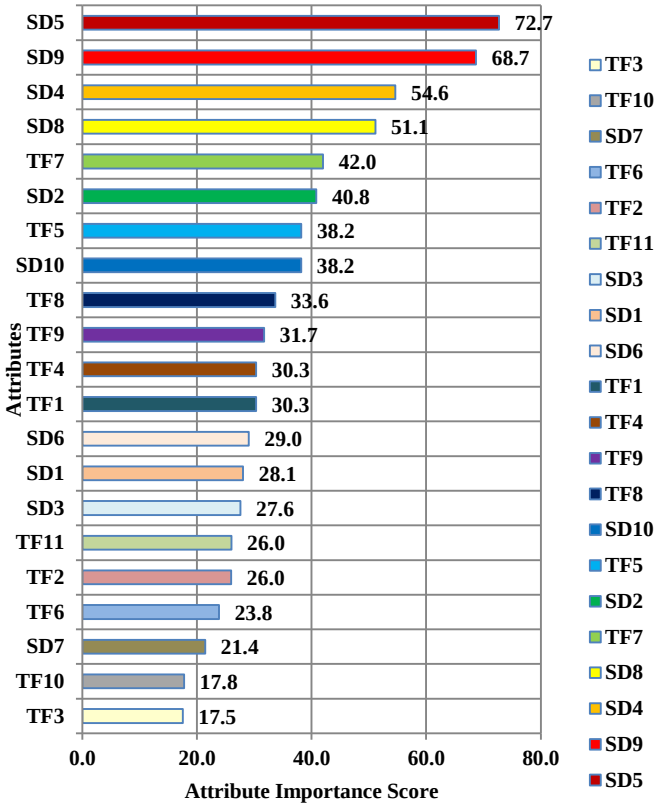



Fig. 10. Attribute Importance Plot for HDGIM.

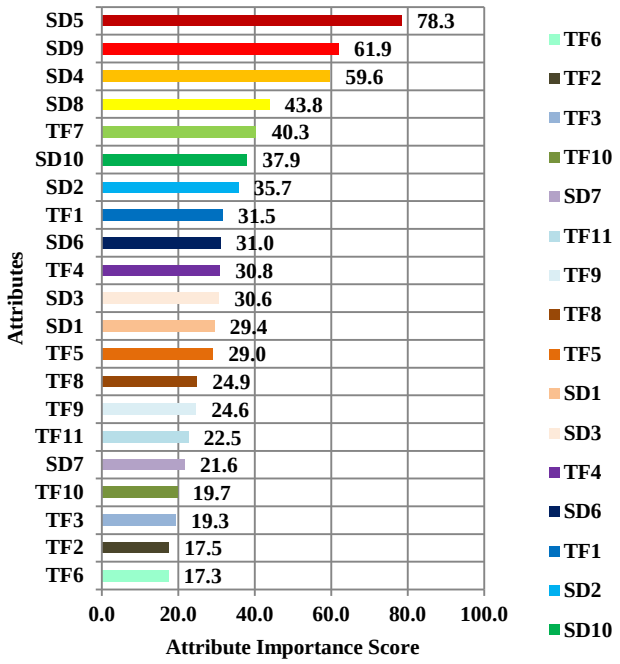



Fig. 11. Attribute Importance Plot for HVGIM.

In the present work, the statistics based features, texture based features have been computed for the dorsal and ventral leaf images of the different plant species.

By using the random forest feature selection method, six more datasets have been prepared and mentioned through (10), (11), (12), (13), (14), (15) and each of the subset datasets contain only seven most important features which are unique in nature and thereby reducing the datasets considerably.

$$HDIM\_S = \left(SD_2, SD_4, SD_5, SD_6, SD_8, SD_9, SD_{10}\right) \quad (10)$$

$$HVIM\_S = \left(SV_2, SV_4, SV_5, SV_6, SV_8, SV_9, SV_{10}\right) \quad (11)$$

$$HDGM\_S = \left(TD_1, TD_4, TD_5, TD_7, TD_8, TD_9, TD_{10}\right) \quad (12)$$

$$HVGM\_S = \left(TV_1, TV_4, TV_5, TV_7, TV_8, TV_9, TV_{11}\right) \quad (13)$$

$$HDGIM\_S = HDIM\_S + HDGM\_S \quad (14)$$

$$HVGIM\_S = HVIM\_S + HVGM\_S \quad (15)$$

HDIM_S(Hue-Dorsal-Istorder-Measure_Subset), HVIM_S (Hue-Ventral-Istorder-Measure_subset), HDGM_S(Hue-Dorsal-GLCM-Measure_Subset), HVGM_S(Hue-Ventral-GLCM-Measure_Subset), HDGIM_S(Hue-Dorsal-GLCM-Istorder-Measure_Subset), HVGIM_S(Hue-Ventral-GLCM-Istorder-Measure_Subset) are four feature subsets prepared with unique features which have the highest scores for attribute importance as shown through Fig. 6, 7, 8, 9, 10, 11 and through (10), (11), (12), (13), (14) and (15). HDGIM_S and HVGIM_S are the two datasets prepared by using the combined features.

## IV. Results

The classification accuracy results have been computed for all the twelve different kinds of datasets prepared in the present work and has been compared with [4, 15, 16] and have been shown through Fig. 15 and 16.The predictive accuracy values have been calculated using KNN and Random Forest algorithms. The corresponding Kappa accuracy results have been calculated as well and shown through Fig. 12 and 13. The result Fig. 12 shows the Kappa results calculated for the complete dataset for HDGIM and HVGIM which have fared better with values of 98.61% and 99.45% respectively and these results substantiate the point that the fusion of features improves the predictive accuracy results. The Fig. 13 shows the kappa values for the feature subsets prepared and HDGIM_S and HVGIM_S datasets have fared better as compared to other datasets with values 98.91% and 99.51% respectively. Further it can be substantiated that the ventral leaf image dataset has proved to be better performer.

By using seven features in dorsal as well as ventral feature subset of datasets, ten features in HDIM and HVIM, eleven features in case HDGM and HVGM and twenty one features in case of HDGIM and HVGIM; a difference of maximum percentage accuracy results achieved by using Random forest algorithm, through the creation of different datasets and has been shown in Fig. 14. By using subset of features, there is a rise in percentage accuracy values, but exception in the case of HVIM_S and HVGM_S. Therefore, it has been observed that there is a positive effect on predictive accuracy results, by using a small subset of feature datasets as compared to datasets with more number of features.

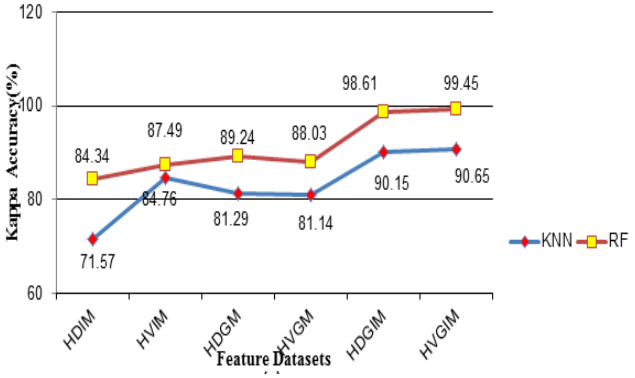



Fig. 12. Predictive Kappa Accuracy Results for Complete Feature Set.

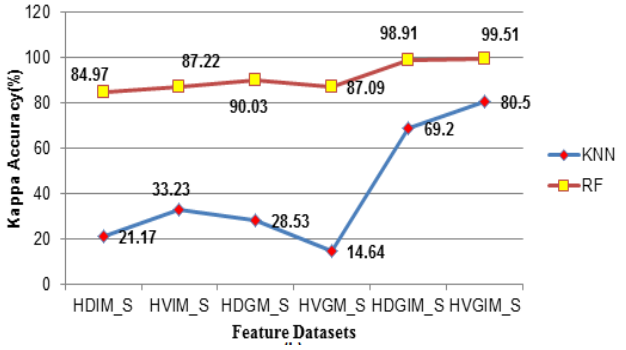



Fig. 13. Predictive Kappa Accuracy Results for Feature Subset.

On observing Fig. 15, on fusing the dorsal datasets (HDIM and HDGM), the rise in percentage predictive accuracy is approximately 10.63% and in the case of fusion of ventral datasets (HVIM and HVGM), the rise in predictive accuracy is approximately 10.52%. Therefore, in the result Fig. 15, HDGIM (98.75%) and HVGIM (99.56%) have shown the maximum value of predictive accuracy results using Random forest based classification algorithm amongst all the complete datasets created. This shows that the fusion of features have comparative effect on the overall classification accuracy results.

On observing Fig. 16, on fusing the subset features obtained for dorsal datasets (HDIM_S and HDGM_S), the rise in percentage predictive accuracy is approximately 10.26% and in the case of fusion of subset features obtained for ventral datasets (HVIM_S and HVGM_S), the rise in predictive accuracy is approximately 11.10%. The result Fig. 16 is for the subsets of features and HDGIM_S (99.02%) and HVGIM_S (99.56%) have shown better results and these results are even better than the combined feature subsets as well.

The Fig. 15 and 16 also depicts the comparison of results with [4, 15, 16]. Complete datasets prepared in the present work have been compared with [4, 15, 16] as shown in Fig. 15 and it has been observed that the HDGIM and HVGIM fare better over all the datasets. On the other hand, the results obtained in the present work with HDGIM_S and HVGIM_S have fared better than mentioned by [4] as shown in Fig. 16.

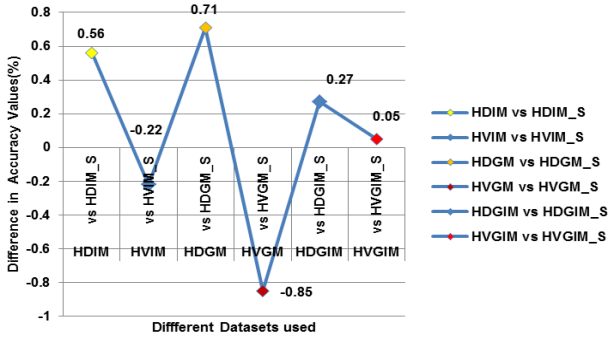



Fig. 14. Difference in Maximum Percentage Accuracy Values of Different Datasets.

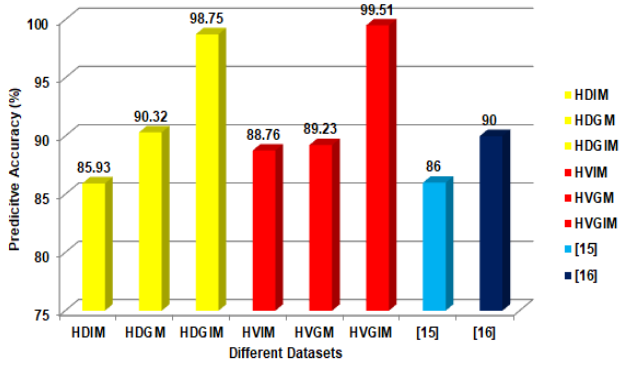



Fig. 15. Comparison of the Complete Dataset Results with the Work of [15, 16].

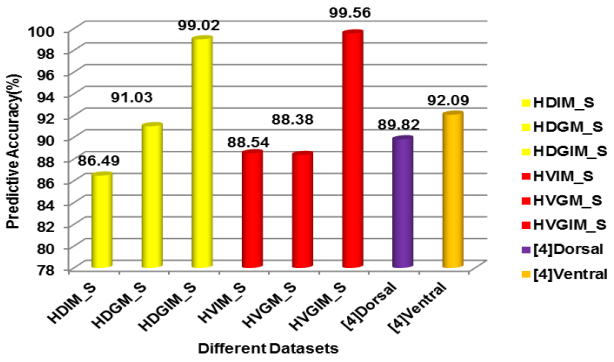



Fig. 16. Comparison of the Feature Subset Results with the Work of [4] having Dorsal and Ventral Sides of Leaf Images.

## V. Conclusions

The present work portrays the efficacy of ventral leaf image dataset based classification of plant species over the dorsal image datasets, which are in vogue. The concept of color as used in the present work, especially the hue channel which carries the color information, is based on human perception of classification of objects into classes based on visual perception of color in human minds and the concept of texture features based on feeling of touch and their subsequent fusion to obtain a combined set of features, has resulted in a new understanding of the fact that the combination of features can be utilized for improved classification accuracy results in the case of plant species classification using digital images. The application of PCA on results in understanding various features thereby helping in overall reduction of the dataset.

The random forest based feature selection technique for minimizing the datasets has provided improved classification accuracy results.

## References

[1] Herdiyeni, Y., Ginanjar, A. R., Anggoro, M. R. L., Douady, S., & Zuhud, E. A. (2015, November). MedLeaf: Mobile biodiversity informatics tool for mapping and identifying Indonesian medicinal plants. *Proceedings of 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, 54-59. IEEE, Fukuoka, Japan.

[2] Manik, F. Y., Herdiyeni, Y., & Herliyana, E. N. (2016). Leaf morphological feature extraction of digital image anthocephalus cadamba. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *14*(2), 630-637.

[3] Hutomo, D. T., Herdiyeni, Y., MKom, M., & Rauf, A. (2013). Whitefly (Bemisia tabaci) Calculation based on image processing using triangle method. *Proceedings of International Conference on Computer, Control, Informatics and its Applications*, 165-168, Indonesia.

[4] Kumar, A., Patidar, V., Khazanchi, D., & Saini, P. (2016). Optimizing feature selection using particle swarm optimization and utilizing ventral sides of leaves for plant leaf classification. *Procedia Computer Science*, *89*, 324-332.

[5] Ji, J. T., Deng, M. L., He, Z. T., Jia, S. T., Du, X. W., He, Y. K., & Liu, J. J. (2014). Extraction of tobacco leaves color features based on HSI color space. *Applied Mechanics and Materials*, Trans Tech Publications, 651, 2424-2429.

[6] Blotta, E., Bouchet, A., Ballarin, V., & Pastore, J. (2011). Enhancement of medical images in HSI color space. *Journal of Physics: Conference Series*, *332*(1), 012041. IOP Publishing.

[7] Gonzalez and Woods (1992). The HSI color model, derivations of HSI to RGB and RGB to HSI conversion equations. In: Digital Image Processing, (1st ed.). Addision Wesley.

[8] Haralick, R. M., and Shanmugam, K. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, *3*(6), 610-621.

[9] Arai, K., Abdullah, I. N. and Okumura H.(2013) Image Identification Based on Shape and Color Descriptors and Its Application to Ornamental Leaf. *Interbational Journal of Image, Graphics and Signal Processing*, 10, 1-8.

[10] Kumar, A., Patidar, V., Khazanchi, D., Purohit, G. & Saini, P. (2016) An approach to improve the classification accuracy of leaf images with dorsal and ventral sides by adding directionality features with statistical feature sets. In R. Choudhary, J. Mandal, N. Auluck, H. Nagarajaram (Eds), *Advanced Computing and Communication Technologies. Advances in Intelligent Systems and Computing* (pp. 89-97), 452. Springer, Singapore

[11] Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T. & Tinevez, J. Y. (2012). Fiji: an open-source platform for biological-image analysis. *Nature methods*, *9*(7), 676-682.

[12] Charpentier, A. Variable importance plot and variable selection (2015, *June, 17).* Retrieved from https://feakonometrics.hypothesis.org/19835

[13] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Retrieved from http://www.R-project.org.

[14] Latte, M. V., Shidnal, S., Anami, B., & Kuligod, V. (2015). A combined color and texture features based methodology for recognition of crop field image. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, *8*(2), 287-302.

[15] Ghazali, K. H., Mustafa, M. M., & Hussain, A. (2007). Color image processing of weed classification: a comparison of two feature extraction technique. *Proceedings of the International Conference on Electrical Engineering and Informatics*, 607-610.

[16] Tian, Y., Niu, Y., & Li, T. (2008, October). A study on the method of image processing and feature extraction for cucumber diseased. *Proceedings of International Conference on Computer and Computing Technologies in Agriculture*, Springer, Boston, MA. 967-972.

# Design of an Error Output Feedback Digital Delta Sigma Modulator with In–Stage Dithering for Spur–Free Output Spectrum

Sohail Imran Saeed[1]
Department of Electrical Engineering
Iqra National University
Peshawar, Pakistan

Khalid Mahmood[2]
Department of Electrical Technology
University of Technology
Nowshera, Pakistan

Mehr e Munir[3]
Department of Electrical Engineering
Iqra National University
Peshawar, Pakistan

***Abstract*—Digital Delta Sigma Modulator (DDSM) is responsible for generation of spurious tones at the output of fractional n frequency synthesizer due their inherent periodicity. This results in an impure output spectrum of frequency synthesizer when they are used to generate the fractional numbers in the divider of Phase Locked Loop (PLL) based frequency synthesizer. This paper presents the design of Error – Output feedback modulator based third order Multi – stage noise Shaping (MASH) structure with lesser hardware and effective error compensation network to break the underlying periodicity of DDSM. The DDSM is also analyzed by using non-shaped, shaped and self – dithering mechanism to achieve a pure output spectrum and reduced quantization noise.**

***Keywords—Digital delta sigma modulator; fractional N – frequency synthesizer; phase locked loop; error feedback modulator; spur; dither; MASH; HK – MASH***

## I. INTRODUCTION

Phase Locked Loop based fractional N- frequency synthesizer is pillar of modern wireless communication system due to its wide range of application. They are also a favorite choice due their high frequency resolution and fast settling time. But the performance of the synthesizer is limited by the presence of spurious tones that result in an impure spectrum at its output.

The divider in the feedback path of fractional N – Frequency synthesizer as shown in Fig. 1 is based on Digital Delta Sigma Modulator (DDSM) to produce the fractional part of the divide value. The DDSMs are used to oversample and re-quantize the high resolution discrete – time input in order to produce an output with lower resolution. [1] The DDSM is a finite state machine that produces the periodic output when it is subjected to an input which is either periodic or constant input. As a result, the quantization noise in a DDSM is also periodic in nature. The output of DDSM suffers the unwanted tones due to its inherent periodicity which affect its performance. [2], [3] These tones are referred to as spurs. The position of spur in the output spectrum depends on the length of cycles produced by the DDSM. Short DDSM cycles result in strong spurs. Two classes of techniques, stochastic and deterministic are used break the length of cycles. Stochastic techniques refer to addition of random dither signal at the input while deterministic techniques deal with the structural changes in the modulator.

The recent researches have focused on developing the deterministic methods to increase the cycle length of DDSM. The extended cycle length reduces the quantization noise power in spurious [4], [5]. The techniques using error masking for reduced hardware [6] and lower power consumption using lower order DDSMs have also been proposed [7].

On the other hand, the dithering techniques are also focus of the recent researches to eliminate the spurious tones and achieve the spectral purity at the output of DDSM. [8], [9], [10].

This work inspired from idea in [10] and [11]. Authors have presented the Hybrid Key (HK) EFM structure in [11] with long cycle lengths to reduce the power of quantization noise tones. The DDSM and MASH structure is modified and efficient dithering schemes are implemented to reduce quantization noise and achieve the spectral purity.

The paper is organized as follow. The review of conventional DDSM is provided in Sec. II. In Sec. III we discuss the and HK – EFM and MASH structure based on it. Proposed Model and the simulation results with observations are discussed in Sec. IV and Sec. V simultaneously. The paper is summarized and conclusion is presented in Sec. VI.

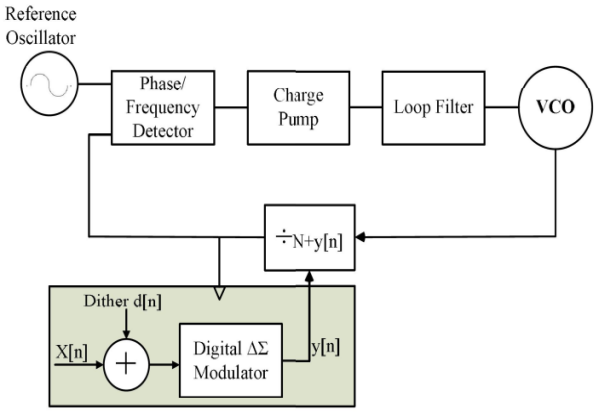



Fig. 1. PLL based Fraction N - Frequency Synthesizer.

## II. CONVENTIONAL DDSM

The first order Error Feedback Modulator (EFM1) in Fig. 2 is a basic building block of MASH DDSM. It receives a digital input of N bits which passes through quantizer *Q(.)* with the step size of $M = 2^N$. The output is '1' when the quantizer overflows and the it is '0' when it does not overflow.

Mathematically,

$$y[n] = \begin{cases} 0, & b[n] < M \\ 1, & b[n] \geq M \end{cases} \quad (1)$$

The discrete output of EFM is given by

$$y[n] = \frac{1}{M} x[n] + \left( eq[n] + eq[n-1] \right) \quad (2)$$

In Z-domain the output is represented as

$$Y(z) = \frac{1}{M} X(z) + (1 - z^{-1}) E(z) \quad (3)$$

Where $\frac{1}{M}$ is Signal Transfer Function (STF) and $(1 - z^{-1})$ is the Noise Transfer Function (NTF)

A MASH DDSM is formed by cascading the EFM1s and a noise cancellation network. Each EFM block has the quantization error of previous stage at its input while the carry out of each EFM1 block is fed to noise cancellation network to compensate the error at the output. A 3rd order MASH DDSM is presented in Fig. 3.

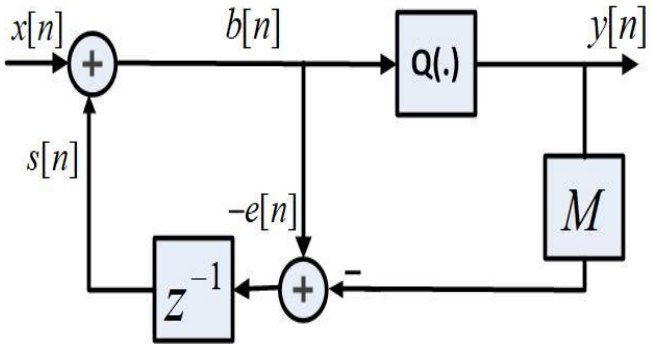



Fig. 2. First Order Error Feedback Modulator.

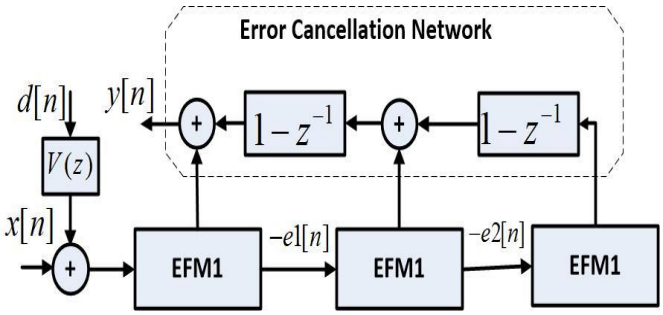



Fig. 3. MASH 1-1-1 DDSM based on EFM1.

The output of a 3rd order composed of cascaded EFM DDSM in Z – domain is as follows

$$Y(z) = \frac{1}{M} X(z) + (1 - z^{-1})^3 E(z) \quad (4)$$

Where NTF $(1 - z^{-1})^3$ is used as noise cancellation network for a better and cleaner output.

## III. HYBRID KEY – EFM

Hybrid Key Error Feedback Modulator (HK – EFM) based MASH is proved to have long cycles which are vital to reduce the number of unwanted tones at its output. The HK – MASH is based on HK – EFM1 which is different from conventional EFM in a way that it has an additional output feedback path is denoted by $az^{-1}$ [11], [12].

The output of a first order HK EFM1 is expressed as

$$Y(z) = \frac{1}{1 - \alpha z^{-1}} X(z) + \frac{(1 - z^{-1})}{1 - \alpha z^{-1}} E(z) \quad (5)$$

Where $\alpha$ is the total number of carry bits from all the stages of a MASH and is given by

$$\alpha = \frac{a}{M} \quad (6)$$

and,

$$STF = \frac{1}{1 - \alpha z^{-1}} \text{ and } NTF = \frac{(1 - z^{-1})}{1 - \alpha z^{-1}} \quad (7)$$

The $l^{th}$ order HK – EFM based MASH structure developed in [11] is shown in Fig. 4. It is noteworthy that authors have considered the normalization factor and have used $(1 - z^{-1})$ in the noise cancellation network. The error of one stage is fed to the next stage as in standard MASH. The output of $l^{th}$ order HK – EFM based MASH is given as

$$Y(z) = \frac{1}{2^N} \frac{1}{1 - \alpha z^{-1}} X(z) + \frac{1}{2^N} \frac{(1 - z^{-1})^l}{1 - \alpha z^{-1}} E(z) \quad (8)$$

with,

$$STF = \frac{1}{1 - \alpha z^{-1}} \text{ and } NTF = \frac{(1 - z^{-1})^l}{1 - \alpha z^{-1}} \quad (9)$$

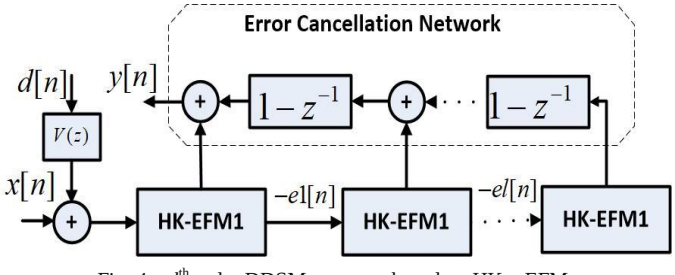



Fig. 4. $l^{th}$ order DDSM structure based on HK – EFM.

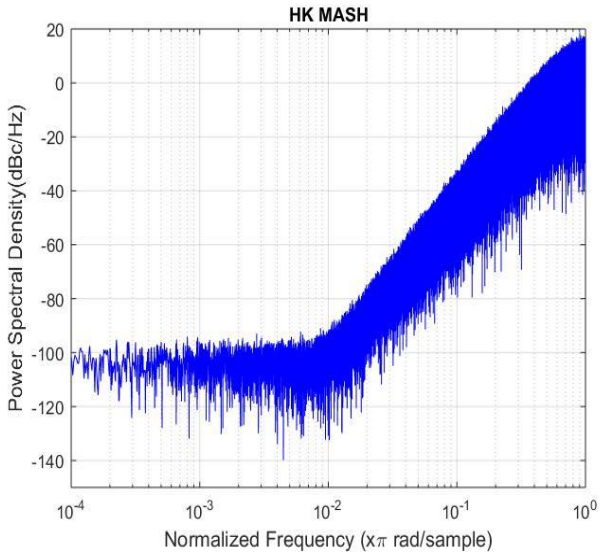



Fig. 5. PSD of MASH 1-1-1 based on EFM1 with Unshaped Dither at the Input of First Stage.

Simulated Power Spectral Density (PSD) of the system with a 15 bit quantizer setting and zero-order (or unshaped) dither applied to first stage of system is shown in Fig. 5.

## IV. Proposed Error Output Feedback Modulator and MASH with Modified Error Cancellation Network

The modifications to first order HK – EFM1 and MASH structure are proposed in this work. The proposed modulator after the modification is referred to as Error Output Feedback Modulator (EOFM1) in the remainder of this paper. The proposed changes to existing HK - EFM1 and the MASH based on it are:

*1)* The output feedback in EOFM is without any scaling factor, instead, the actual output is fed back to the modulator. This will reduce the extra hardware required to implement the scaling factor.

*2)* The noise cancellation network in EOFM based MASH consists of $NTF = \frac{(1-z^{-1})}{M-z^{-1}}$ whereas, in previous researches, the denominator is ignored in the noise cancellation network. Possibly, because of the normalization effect in both the STF and NTF. However, in case of HKEFM, the STF is not only the scaled version of the input but is also passed through a system with transfer function $\frac{1}{1-\alpha z^{-1}}$.

*3)* Finally, the techniques for efficient dithering proposed by Gonzalez *et al.* in [13] along with other in–stage dithering techniques have been implemented on the proposed design to analyze the performance of proposed.

Fig. 6 shows the proposed EOFM where the output of internal signals of the modulator are as follow.

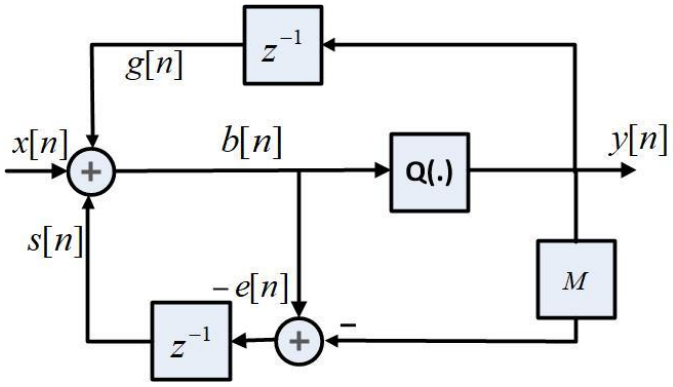



Fig. 6. Proposed Error Output Feedback Modulator with Reduced Components.

The quantization noise $e_q[n]$ is added to the output when the input passes through an N bit quantizer with $M = 2^N$ quantization levels. Increasing the number of quantization bits can reduce quantization noise but it is not possible to eliminated entirely. The error signal $e[n]$ is calculated by subtracting the actual signal from the quantized signal and is fed back to the input after addition of delay. The delayed output $g[n]$ is also added with the input. Feedback signals are:

$$e[n] = -Me_q[n] \tag{10}$$

$$g[n] = y[n-1] \tag{11}$$

$$s[n] = -Me_q[n-1] \tag{12}$$

The output of the system is then

$$y[n] = \frac{1}{M}\left(x[n] + y[n-1]\right) + e_q[n] - e_q[n-1] \tag{13}$$

And in Z – domain as

$$Y(z) = \frac{1}{2^N}\frac{1}{M-z^{-1}}X(z) + \frac{1}{2^N}\frac{M(1-z^{-1})}{M-z^{-1}}E(z) \tag{14}$$

with the following signal and Noise transfer functions

$$STF = \frac{1}{2^N}\frac{1}{M-z^{-1}} \quad \text{and} \quad NTF = \frac{(1-z^{-1})}{M-z^{-1}} \tag{15}$$

A 3rd order MASH DDSM is designed using EOFM. The noise cancellation network of equation no. 15 and shown in Fig. 7 is used to cancel the effect of noise from the output of each cascaded stage of the MASH structure. The error of each stage is passed to the input of next stage as in the standard MASH DDSM

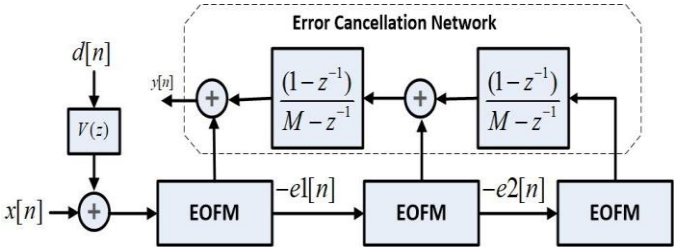



Fig. 7. EOFM based MASH DDSM with Modified Error Cancellation Network.

The output of 3rd order MASH based on EOFM is given as

$$Y(z)=\frac{1}{2^N}\frac{1}{M-z^{-1}}X(z)+\frac{(1-z^{-1})^3}{M-z^{-1}}E(z) \quad (16)$$

## V. Simulation Resutls

MATLAB and Simulink have been used to simulate and analyze the performance of the proposed model. The detailed block diagram used to simulate the system is presented in Fig. 8. The system is implemented using a 15 bit quantizer with $2^N$ quantization levels.

A comparison 3rd order EOFM MASH with HK MASH and standard MASH is provided in Fig. 9. PSD of output is plotted using a 15 bit quantizer and zero – order dither applied to first stage of all the systems. Other configuration for all the system are kept similar for the coherent analysis. It has been noticed that EOFM MASH does not show the performance degradation in comparison to the HK – MASH structure. However, the reduction of hardware due to decrease in each stage of the MASH is an advantage.

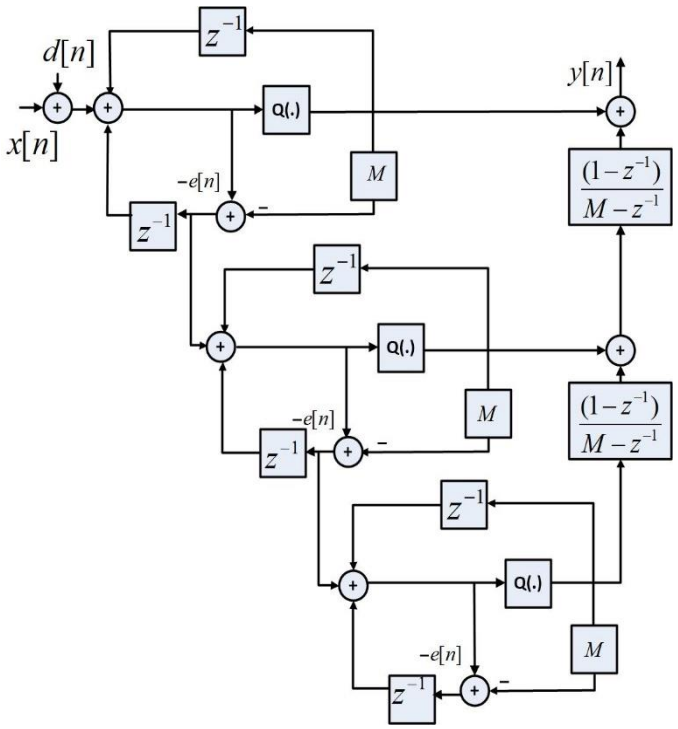



Fig. 8. Detailed Block Diagram of EOFM based MASH DDSM with Proposed Error Cancellation Network.

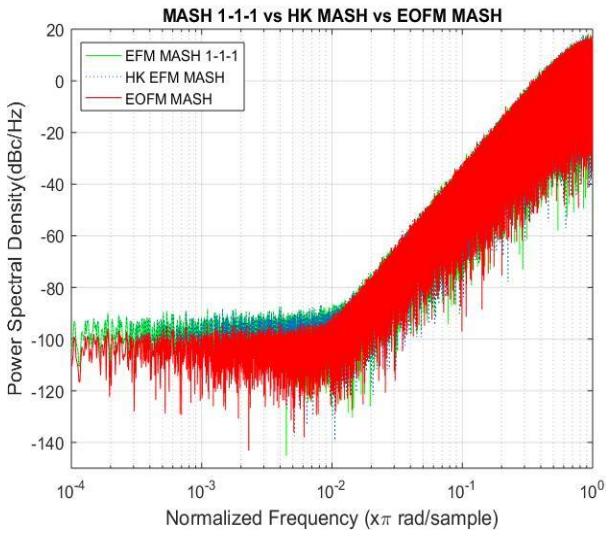



Fig. 9. Comparison of Output of PSDs EFM based MASH and HK MASH with Proposed EOFM MASH Under the Similar Conditions for all Models.

Next we have investigated the performance of our proposed DDSM by applying the efficient dithering technique that was proposed for EFM1 based MASH 1-1-1 structure where the dither is injected at the input of second and third stage together to obtain the spur free output spectrum along with the reduction of noise floor at low frequencies. We have modified the model for EOFM MASH as in Fig. 10

PSD of the model is presented in Fig. 11 which shows a reduction in noise floor in low frequencies region from –100 dBc to –170 dBc.

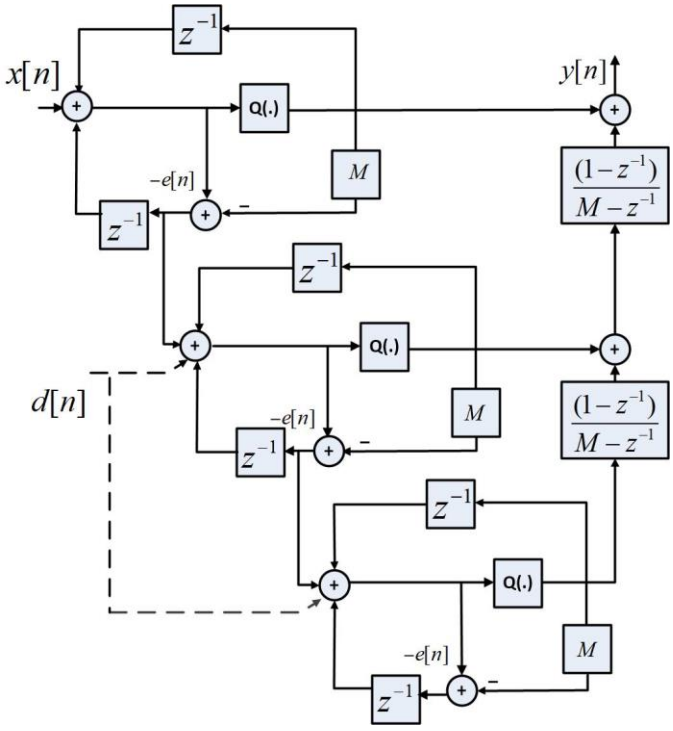



Fig. 10. Detailed Block Diagram of Proposed DDSM along with the Application of Efficient Dither Model of Gonzalez *et. al.* [13].

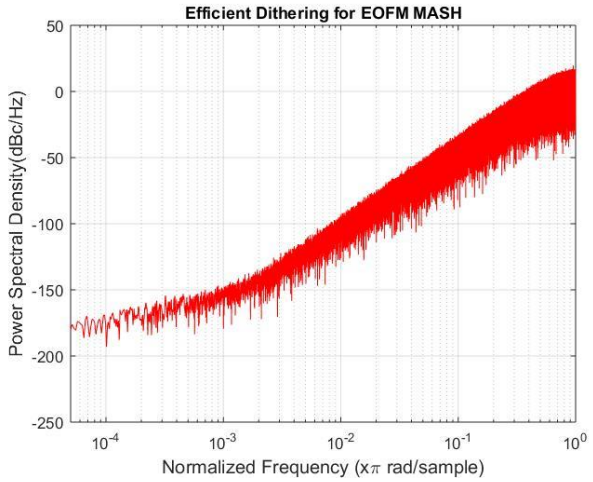



Fig. 11. PSD of Proposed 3rd Order MASH after Application of Efficient Dithering Strategy.

It is seen in the Fig. 11 that the noise floor has significantly fallen in low frequencies by applying the efficient dithering strategy vis a vis system where only the unshaped dither is applied to it and pure spectrum is also achieved. But it is further found that the obtained spectrum is similar to that when the dither is applied to the second stage only i.e. first order shaped dither. The implementation of this strategy costs the hardware overhead of applying the dither to two stages of MASH without any significant improvement in comparison to application of first order dither. Therefore, this strategy does not seem to work in case of EOFM MASH. The comparison is provided in Fig. 12.

In another investigation, the second order shaped dither is applied to proposed EOFM MASH i.e. dither is passed through second order high-pass filter $V(z) = (1 - z^{-1})^2$. The system is simulated with previous configuration settings and second order shaped dither. The PSD in Fig. 13 shows further reduction in noise floor to –250 dBc in low frequency zone with a spur free spectrujm.

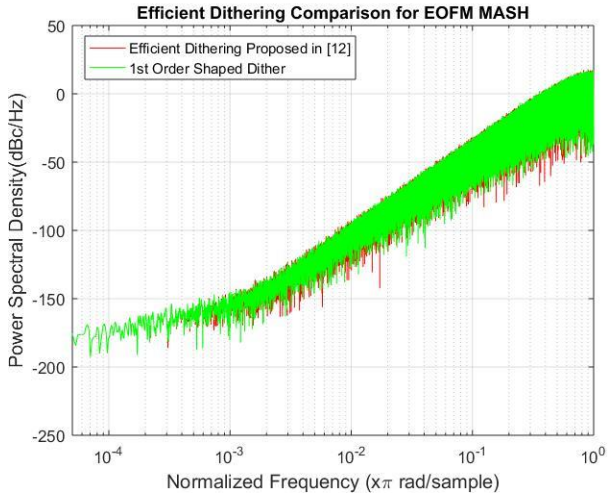



Fig. 12. PSD of Proposed 3rd Order MASH Showing the Comparison Application Of Efficient Dithering Strategy with First Order Shaped Dithering Strategy.

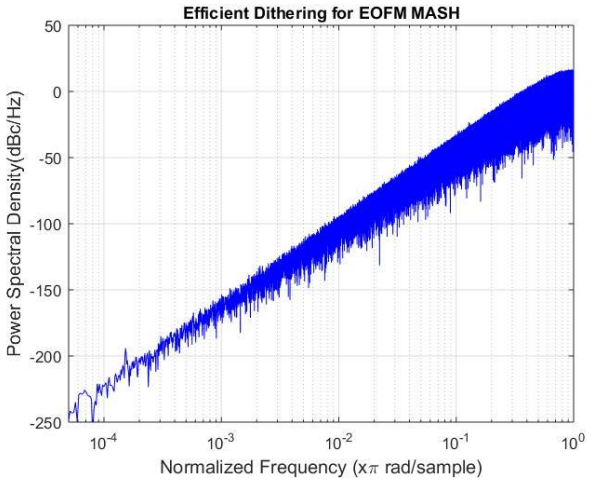



Fig. 13. PSD of Proposed 3rd Order MASH after Application of Third Order Shaped Dithering.

The comparison of the investigated dithering mechanisms is provided in Fig. 14. The graph on black color shows the output PSD when unshaped, zeroth ordered dither is applied to the proposed model. The output PSD when first order shaped dither is applied to proposed model is shown in green color while output PSD of efficient dithering strategy of Gonzalez *et. al.* is plotted in red. It is seen that both these strategies have same output performance but applying the first ordered dither can help to achieve the same performance with lesser hardware.

Finally, the graph in blue color shows the output PSD when second order shaped dither is applied. It is seen that the noise floor in lower frequencies has further fallen –250dBc in comparison to first order shaped dither where noise floor falls till –170 dBc and to unshaped dither where the noise floor stands at –170 dBc. This shows that applying the second order shaped dithering provides the best results in terms of lesser quantization noise in low frequency range and also retains the clean and spur free spectrum at the output of proposed EOFM based 3rd order MASH DDSM.

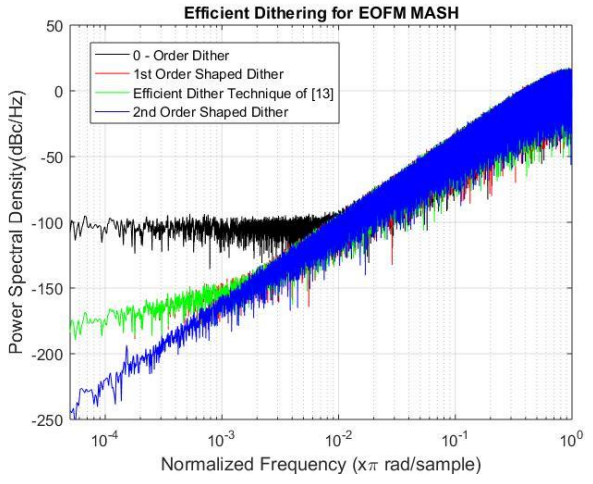



Fig. 14. Comparison of Output of PSDs of Proposed 3rd Order MASH with Different Techniques of Dither Applied to it.

## VI. Conclusion

Digital delta sigma modulators are notorious for their spurious output due to periodic quantization noise generated due to its periodicity. The HK – MASH are helpful to reduce this quantization noise because of their long cycle length. In this paper we have presented an EOFM DDSM that has long cycle length of HK MASH with smaller hardware cost without degradation in performance.

The efficient in-stage dithering model as proposed by Gonzalez *et al.* to inject the dither at multiple points has been investigated for the proposed MASH along with the single point in–stage dither injection schemes. It has been concluded that second order shaped dithering provides better noise reduction and a pure output spectrum for EOFM based MASH in comparison to the other dithering models.

## References

[1] K. Hosseini, M. P. Kennedy, "Minimizing Spurious Tones in Digital Delta-Sigma Modulators" Springer Science & Business Media. 2011

[2] V. S. Sadeghi, S. I. Saeed, S. Calnan, M. P. Kennedy, H. M. Naimi, and M. Vesterbacka, "Simulation and experimental investigation of a nonlinear mechanism for spur generation in a fractional-N frequency synthesizer" In Irish Signals and Systems Conference (ISSC 2012), 2012

[3] S. I. Saeed, "Investigation of Mechanisms for Spur Generation in Fractional-N Frequency Synthesizers" Electronic Press, Linköping University. 2012

[4] B. Fitzgibbon, M. P. Kennedy and F. Maloberti, "A novel implementation of dithered digital delta-sigma modulators via bus-splitting," 2011 IEEE International Symposium of Circuits and Systems (ISCAS), Rio de Janeiro, 2011, pp. 1363-1366. doi: 10.1109/ISCAS.2011.5937825

[5] Y. Donnelly, H. Mo, and M.P. Kennedy, "High-Speed Nested Cascaded MASH Digital Delta-Sigma Modulator-Based Divider Controller" In IEEE International Symposium onCircuits and Systems, 2018.

[6] C. Y. Yao, and C. C. Hsieh, "Hardware simplification to the delta path in a MASH 111 delta–sigma modulator" IEEE Transactions on Circuits and Systems II: Express Briefs, 56(4), pp.270-274, 2009

[7] R. LAAJIMI, and M. MASMOUDI, "Design of A high performance low-power consumption discrete time Second order Sigma-Delta modulator used for Analog to Digital Converter". International Journal of Advanced Computer Science and Applications (IJACSA), vol. 3(11), 2012.

[8] Mo, H. and Kennedy, M.P., 2017. Masked dithering of MASH Digital delta-sigma modulators with constant inputs using multiple linear feedback shift registers. *IEEE Transactions on Circuits and Systems I: Regular Papers*, *64*(6), pp.1390-1399.

[9] Z. Xu, J. G. Lee, "Self-dithered digital delta-sigma modulators for fractional-N PLL. *IEICE transactions on electronics*, *94*(6), 2011 pp.1065-1068.

[10] H. Mo and M. P. Kennedy, "A high-throughput spur-free hybrid nested bus-splitting/HK-MASH digital delta-sigma modulator," 2013 European Conference on Circuit Theory and Design (ECCTD), Dresden, 2013, pp. 1-4.

[11] B. Fitzgibbon and M. P. Kennedy, "Calculation of the cycle length in a HK-MASH DDSM with multilevel quantizers," Proceedings of 2010 IEEE International Symposium on Circuits and Systems, Paris, 2010, pp. 245-248.

[12] A. Telli and I. Kale, "The practical limits of MASH Delta-Sigma Modulators designed to maintain very long controllable sequence lengths for structured tone mitigation," 2009 IEEE 10th Annual Wireless and Microwave Technology Conference, Clearwater, FL, 2009, pp. 1-5.

[13] V. R. Gonzalez-Diaz, M. A. Garcia-Andrade, G. E. Flores-Verdad and F. Maloberti, "Efficient Dithering in MASH Sigma-Delta Modulators for Fractional Frequency Synthesizers," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 57, no. 9, pp. 2394-2403, Sept. 2010.

# Applied Artificial Intelligence in 3D-game (HYSTERIA) using UNREAL ENGINE4

MuhammadMuzammul[1], Muhammad Awais[2], Muhammad Umer ghani[3]
Department of software engineering
Government College University
Faisalabad, Pakistan

Muhammad Imran Manzoor[4]
Department of Computer Science & IT
University of Lahore
PakPattan Campus

Muhammad Kashif [5], Muhammad Yahya Saeed[6]
Department of software engineering
Government College University
Faisalabad, Pakistan

***Abstract*—Game development industry spreading it roots at wider level. With the advancements in gaming technologies industries adopted latest trends for developing modern games. Artificial intelligence (AI) with programming provided countless support for latest technology adoption in game industry. This paper aims to highlight some major points of our research "Creation of third person shooter game in unreal engine 4". We discussed how we can use one of the most powerful current generation game engines in an attempt to create our own game "Hysteria". Endeavoring used to replicate the process of the major game production cycle .It is used by modern gaming industries. We attempted it to create an action adventure shooting game by creating its own original storyline. The game Hysteriaisplayedfromathirdperson perspective in which the player must go through multipleenvironmentsfightinghordesof enemies and try to reach the end of level. Depending on the difficulty level that the player sets, there will be the number of enemies and their fighting intensity. The game has been developed but running at initial stages; further enhancement will be required to give it a much professional impression so that in near future it could be successfully commercialized.**

***Keywords—Applied AI; UNREAL ENGIN 4; technology adoption***

## I. INTRODUCTION

The video game industry is one of the most commercially successful extensions of the computer science industry. With over billion video game enthusiasts around the world, the video game industry produces annual revenue in billions of dollar each year. Witheachgamingtitles ranging from console titles of Xbox one and Play Station4 to the free android games, grossing over millions of dollars, video game industryare at epitome of its commercializationat is still evolving year by year.Keeping in mind, our project aim to produce our own game title which will be available for commercialization in the near future. Consideringthe countrieswhich considered as key players in the video game industry such as China,JapanandAmerica,theyallhavemanynumbersof video developmentstudios andpublishingfirms who follow a certaindevelopment process of video games. Whilethe process may be somewhatsimilar to the typical software development process, it is different considering aspects of the different categories of video game. Since our game title falls into the category of thirdperson action shooter, aspects such as characterdesign, environment design and artificial design are to be considered.With studyingthe development process of the video game industry,a core game engine is required which provides a basic skeleton structure to the game body, with arrangement of different assets sequentially, it provides a linear game play experience in an exe.They all have large number of video development studios and publishing firms which follow a certain development process of video games. While the process may be somewhat similar to the typical software development process, it is different considering aspects of the different categories of video game. Since our game title falls into the category of third person action shooter, aspects such as character design, environment design and artificial design are to be considered.

With studyingthe development process of the video game industry,a core game engine is required which provides a basic skeleton structure to the game body,with arrangement of different assetssequentially,it provides a linear game play experience in an exe file of the development mode. The overall system design objective is to lay down the basic structure using the most powerful game engine of theindustryunrealengine 4 anduse variousothersoftware to createa sequentiallyincrementingAAA video game title.The game will be available to play on PC, PS4 and Xbox one, and thus will be a current generation game title running on advance recent consoles.The actual game play will be in third person shooting covering structure and with its original storyline; it will be available for commercialization upon further development. The scope of this project is to design and develop a game title using Epic games most recent incarnation of its unreal engine game engines series, unreal engine 4. The game title will be a third person action adventure game. Thegraphicswill be developedusing unrealengine 4 andZbrushwhich are bothprofessional industrytools andthusthegoal is to provide the game with high resolutiongraphics,capable of runningfrom 1080 p to 360 p screens. The game play does not require active internet connection to run and will be a single player storyline game play.

The software requirement for the development of this software is: I. UNREAL ENGINE 4 ii. BLENDER iii. ZBRUSH IV.AUTODESK MAYA v. GIMP VI. MIXAMO .The Current game system will be available as an exe installation to run on PC and the PC is required to possess a

minimum requirement of:4GB Ram. 15GB hard disk space. A Direct X 11 Graphic card (Nvidia or Geforce) Keyboard and Mouse or Game Joystick. The game will also be able to run on different resolutions according to the capabilities of the graphic card from 360 p to 1080 p. There will be a user interface for arranging the setting of the game visual and game play and it will be up to the player to decide. Also there will be setting for the difficulty level for the game that will dynamically affect the game play.And the player can save and load their game progress and exit at will, a fully functional user interface is provided. We used Linear Sequential Model sometime it also called classical Life cycle or water flow model.Sequential model suggests a systematic sequential approach to software development that begins at the system level and progresses through the following.

Through research of the industry market, the action adventure genre is one of the highest demanding and grossing among the gamers.So the action adventure genre was chosen and the basic structure such as the assets and the components were decided and designed as a prototype. With the working title of Project Hysteria the development process begin.Unreal Engine 4 was the core component used as all the assets were arranged in it in a sequential order and an alignment of various assets and process were decided.After implementing all the component, the project was ready for testing. Our first objective in the testing phase will be the connectivity between the components on which the game is executing. The game went through various beta testing phase in which students and adults alike played the game and the first response was positive.Second testing phase was in collages and schools were students played the initial version and were curious to know when the first commercial version will be released for further experience of the game.

The project under creation is to provide a gaming revolution in Pakistan Fig1. Videogame industry is one of the highest grossing sectors in the technologicalfield [1].Withthedevelopmentofa game project, the aim is to motivate the aspiring gamedevelopersworkingonminorIndiegame projects, to take a major step and work on a major gametitleandthismayleadtointroductionof several game development teams and development studios [2]. We use one of the most powerful game enginesUnrealEngine4,whichisthecurrent version of the Unreal Game Engine series by Epic Games. With Unreal engine we use employ several powerful soft ware's such Autodesk Maya, Zbrush and Gimp.

Thegameisathird-personshootervideogame which goes on par with the graphically advanced games being released on current generation of the gaming consoles (Play Station 4 and Xbox One). Withtheprojectunderincrementaldevelopment and facing major changes and being continuously updated, upon release, it is certain to create the identity of Pakistan Game Industry in the international entertainment market andcompete with other current generation advance game titles.

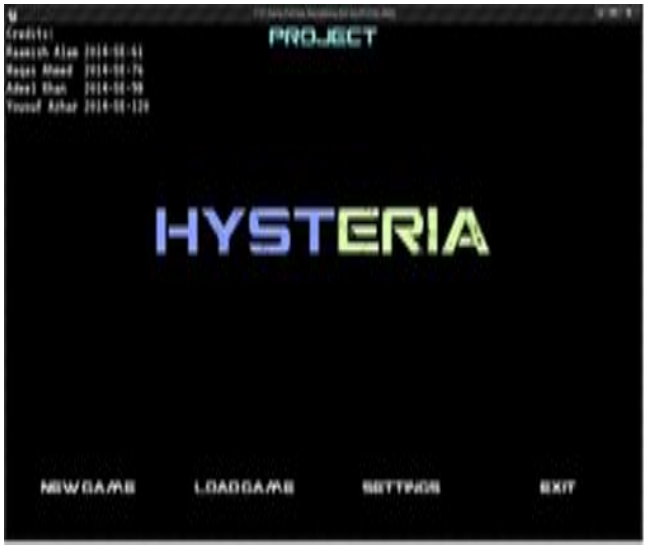



Fig. 1. The Main Title Screen of Hysteria Displaying All the Typical basic Options of a Title Screen.

## II. HYSTERIA STORYLINE

There is now almost a tradition for a game of any genre to have an interesting and catchy storyline in order to captivate the attention of consumers. We also took the liberty of designing our original storyline.

Considering various factors such as survival, action, drama,adventureandheavysciencefiction,our story is set in a fictional universe in which human beings have reached out and conquer other planets and their reach have extended to outer space. The main character of our story Commander Anderson fig2. is a Galactic Patrol Federation commando whose space ship crashed in an unfamiliar planet of which the environment heavily resembles that of earth but much more exotic. This planet is HYST-51891 an unidentified planet is occupied by hostile and savage human beings whose origins are unknown. They arrived suddenly and drove the previous occupants to extinction.

Fig. 2. Commander Anderson, The Main Character Donning the Galactic Patrol Federation Armor and Wielding a Laser Assault Rifle.

Fig. 3. The Game Play Screenshot.

On the planet Commander Anderson is attempting to fix his space ship or get any sort of help when he is ambushed by the hostile native of the planet fig3. These human beings are ruthless to intruders and wield the same advance armors and weaponry used by the Galactic Patrol Federation and are known as the Demons. He must know battle these creatures and escape the planet as the planets climatic changes are eventually drawing its catastrophic destruction. As he traverses through many environments from exotic green ruins to the deep jungle, from the frozen frost wastelands to the fire landsofHYST-51891alsoknownasHysteria's Hell , he must defeat all the enemies that cross his path and find a way to escape the impending doomsday on the planet.

The story is science fiction themed and manages to bestow upon the player a sense of survival as the number and fighting intensity of adversaries will grow as the player passes through each level.

## III. Softwares used in Developing Hysteria

Hysteria is a science fiction action adventure game played from a third person perspective in which the player must shoot the enemies and avoid enemy fire by using the cover objects. We used industry standard tools that are used by the modern game development studios to create professional commercial games. Some of them are listed below:

*1) Unreal engine 4*: There are a number of professional tools that we used to bring our vision to life, the most notable and the trump being the unreal engine 4 game engine which acts as a compiler which arranges each of the individual objects created sequentially and in proper order fig 4. It is also used to give the color, textures and contrast of the environments and characters. While the assets of the game may be developed in other software unreal engine 4 is used to modify these assets in order to make them look realistic and beautiful [6]. It also provides the basic graphic options of scaling, translation and rotation which are heavily used for modifying and arranging the environment together.Scalingandtranslationis used to extend the platforms created using blenders whilerotationisusedtorotateandarrangethe scaled environment in perfect order.


Fig. 4. Hysteria is Powered by Unreal Engine 4.

It is also used for giving the player and the enemies' characters skeletons. This is accomplished through the in game rendering. Unreal engine 4 has blueprint technology implement into it which can be used as a sort of visual coding [3]. It is important as it has been used to create the main artificial intelligence between the player characters.

*2) Blender*: Blender is an open-source 3D computer graphic software that is used for creating 3d printed models used in our game [4] [11]. Many of our assets are that are used in game are created using blender. For e.g. the guns that are wielded by Commander Anderson and the enemies was modelled in blender and then exported to unreal engine 4 in which it was given propertextureandfinally implementedin on the joint point of the player and enemies skeleton to make an appearance that they are actually wielding the gun.

The blender is only used for modelling an object as itscolor;contrastandtexturearegiveninany graphicsoftwarethatitisimportedto,suchas Autodesk Maya, 3DsMax or in our case, the unreal engine 4 fig 5.

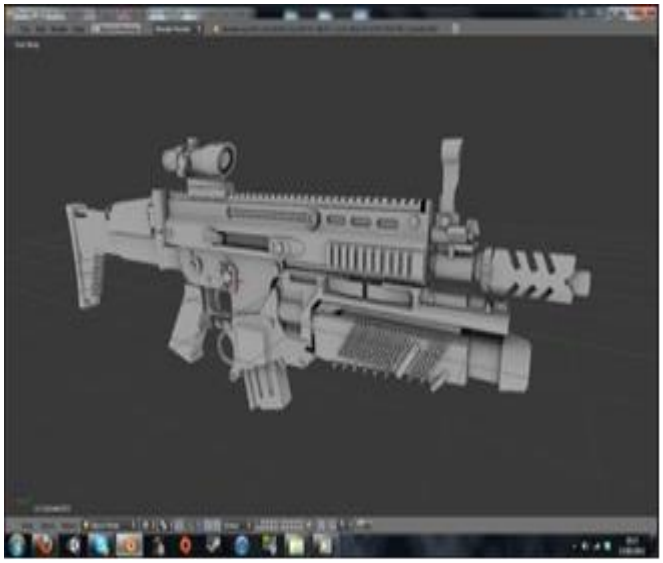


Fig. 5. The Assault Rifle Is Modelled using Blender and then Exported to Unreal Engine 4.

*3)* Zbrush: Zbrush is used in the animation and gaming industry tocreatehighresolutionmodes(ableto reach40+million polygons) [5] . In our project, although it is used for a small role, it plays a critical small role as some of the contrast for enemies and player character and the environments (mountains, rocky terrai8n and statues) were modified using Zbrush.It is an advanced tool that could have been used to create even more detailed characters for our gamesbuttheideawasomittedwelackthe necessary time to modelthe charactersin such a complicated tool.

*4)* Autodesk maya: Unreal engine 4 consists of an A.R.Ts (Animation and Rigging Tools) which is used for animating and rigging in game characters using Autodesk Maya [7]. This plug in of Autodesk Maya possess a default character rig model which is used to create animation (character moving and running) fig 6. These animationscan be exported to unreal engine 4 to test in game play and can be replaced by our default model.

This is optional as the Unreal Engine 4 also has the in game rendering system but the downside is that it is of inferior quality than the A.R.Ts. That why A.R.Tsis usedby mostanimatorsas it not only eases up the animation process but have far more versatile options while animating characters.

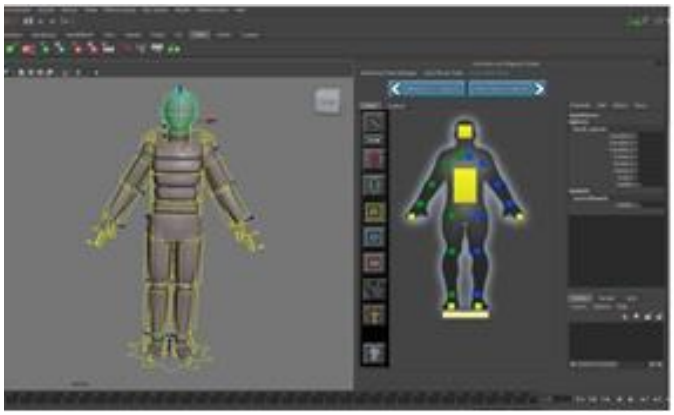


Fig. 6. The Unreal Engine 4 A.R.Ts Plug in Running on Autodesk Maya.

## IV. Development of Hysteria

*1)* Artificial intelligence: The artificial intelligence is the most crucial part of our game. It is how the player and enemy react to each other [8]. The enemies have been programmed with AI to react in certain situations such as:

- To attack the player when in their line of sight.
- To petrol an area under which they are stationed.
- To become aware of the player character when hit by player fire and the player is not in their line of sight.
- To keep firing at the player until the player character runs out of health/ life bar.
- To cease fire once the player dies.
- To take cover when under player fire.
- To use assault rifle when fighting player.
- To use shotgun when fighting player.
- To use sniper rifle while taking cover using cover object when fighting player.
- To lay down dead when their health / life bar runs out (Death animation).

The following AI situations have been programmed using the blueprint technology of the unreal engine 4. It is its visual scripting system.

As shown in Fig 7, the blueprint programming consists of various nodes that are arranged and scripted to make the characters react to certain situations. The nodes as shown in the above diagram are connected together to make an enemy do certain things. It is basically programming their reaction and their default behavior, for e.g. in An Enemy blueprint, there will be a node for enemy to be in stationary position and that exact node extends to patrolling a certain area and that node will extend to move to one waypoint and then move to the other waypoints andthenback,bygoingbackand forwardto the waypoints,a patrolof the areais beingcompleted,andiftheplayercharacteris enemy pre-programmed sight, there is a node that will engage the enemy character with the player in gun fight and as their health bar runs out, there will be a node of death animation for the enemy.

The player character will have similar nodes but most of their actions will be consisted in the node which gives player the overall control of the player character.

The nodes have to be pre-programmed by the developer and when certain actions and reactions are programmed in to the node, there is virtual artificial intelligence within game [10].

The Fig 8. also gives a clearer example as there is node arrangement in the player character session as there is a node arrangement for updating the camera view of the game as the player progress forward using the L key. Notice how the nodes are all connected together so that when the player makes the player character move, the camera will be programmed to automatically follows.

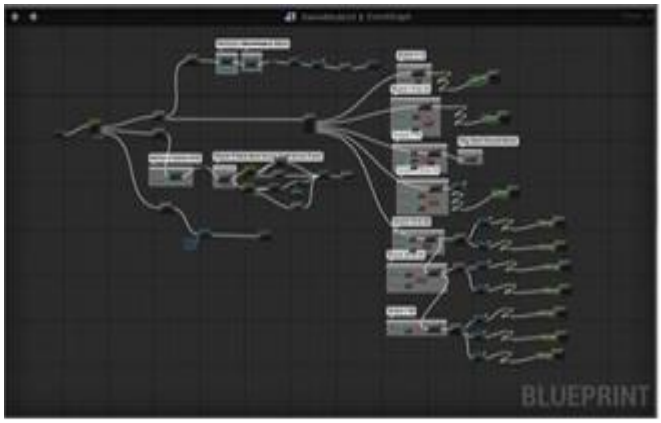



Fig. 7. The Blueprint Technology is used to Orchestrate the in Game Artificial Intelligence.


Fig. 8. Nodes of Camera, Character and Key Toggle.

Hence blueprint technology is one of the most important features of the unreal engine 4 as it is directly responsible for planning down and developing the in game functionalities and sequencing [9].

*2)* Environment creation: The environment in game was developed using unreal engine 4 and blender. Blender was first platforms on which the player and enemy character move fig 9. It was used to create objects that were later on programmed to avoid enemy fire and hence became covers.

The platforms are exported into the unreal engine 4 andmanipulated byusingtheingame functionalities of scaling (the platforms were extended in order to create the floor and ground), translation (use to transit the scale floor and grounds to proper place) for creating mountains and rotation was used to rotate the objects into proper order fig 10. Besides the usual movement of these manipulated objectsinto properplace,unrealenginewasalso usedtogivethosecontras,texturesandoverall realistic feelings.

Blender is also used heavily in creating the hallway levels of the game as the platform were heavily modified and arranged together (using scaling, translation and rotation) with blender to give it an appearance of the hallway, only after arranging the hallway together was it exported into the unreal engine to give it contrast and textures.

Fig. 9. Environment Manipulation of Scaling To Create Extended Floor from Platforms.

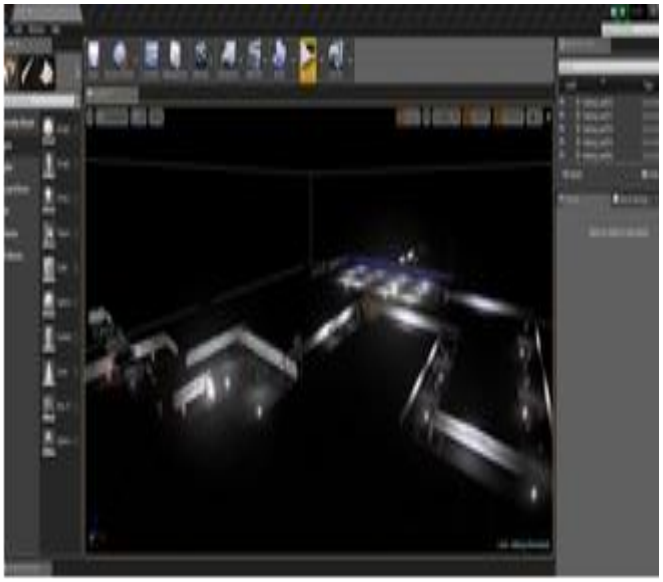


Fig. 10. The Hallway Arranged through Many Scaled, Translated and Rotated Platforms in Unreal Engine 4 with Contrast.

The unreal engine 4 is the most recent version of the unreal game engine by Epic Games. Due to it developed for making games on next generation consoles(8th generationgames)itssuperior graphics is credited for the realistic look of the environmentas it is used to developgame using high end graphic cards fig 11. Only graphic cards with a direct x 11 is appropriate for developing a game in unreal engine 4 or playing a game title developed in unreal engine 4.

Some of the environments consist objects such as trees and statues that are too difficult and complicated for contrasting and texturing in unreal engine 4, this is why software like Zbrush can be appropriate for such job fig 12.

Fig. 11. The Hallway after giving the Frozen Wastelands Contrast, Notice the Superior Graphics.

Fig. 12. The in Development Screenshot, the Trees are Textured in Zbrush.

*3)* Character creation: The characters and enemies all follow same character skeleton render and are contrasted and textured using unreal engine 4 and Zbrush.

They are programmed to wield both one assault rifle and one shotgun or one sniper rifle.

Their artificial intelligence programmed with unreal engine 4 blueprints allows them to move, attack each other, and take cover when under enemy fire (either player with player character or enemy with enemy AI) fig 13.

## V. Conclusion

With the release of one of the first major studio console game title HYSTERIA from Pakistan, an era of gaming revolution is to be inspired. Many of the indie-game developers will be inspired to create creative major game titles in advanced game engines such as Unreal Engine and Cry Engine and thedevelopmentofPakistanigamestudiosown game engine will become in-evitable and this may leadtomajorgamingrevenuetobegenerated within Pakistan. More advance games with artificial intelligence implementation can also be developed for business prospective.

## VI. Future Work

In future more advance version of (HYSTERIA) game expected with latest graphics tools such as Unity 3D.This game introduced new trend of game development using Unreal engine 4 in game industry of Pakistan. Artificial intelligence can also provide new ways in field of game revolution.

Fig. 13. Some of the Enemy Characters in Development View Window of the Game.

## REFERENCES

[1] Zackariasson and t. L. Wilson (eds). "The video game industry: formation, present state, and future", published book, June 8, 2012

[2] The observation that "indie games" as a movement within game development are a contemporary phenomenon has been made before (for example, see Lipkin, 2013, p. 9)

[3] Brian Tayler, "Introduction to Unreal Engine 4, 2014.

[4] BLENDER, "Open Source 3D creation. Free to use for any purpose, forever", (https://www.blender.org/), Retrieved 15 Aug. 2018

[5] Zbrush Manual, www.Pixelogic.com, Retrieved May 20, 2008.

[6] Unreal Engine Mechanics, www.epicgames.com.Retrieved July 10,2018

[7] Todd Palmer, "Mastering AutoDesk Maya", Book published by Autodesk official press (2016)

[8] Arc Sosangyo. "Artificial Intelligence: The Most Important Technology for Our Future". Updated on May 24, 2018,

[9] Kevin Smith and Andersonbadk, "Blue print technology in Unreal Engine, 2015.

[10] Peter L. Newton and Jie Feng,. "Unreal Engine 4 AI Programming Essentials" June 6, 2016

[11] Justin Plowman., " 3D Game Design with Unreal Engine 4 and Blender", June 29, 2016

# Comprehensive Classification Model for Diagnosing Multiple Disease Condition from Chest X-Ray

Savitha S K[1]
Research Scholar,
Department of CSE
Visvesvaraya Technological University
Bangalore, India

N.C. Naveen[2]
Professor
Department of CS & E, JSSATE
Bangalore, India

***Abstract*—Classification plays a significant role in the diagnosis of any form of radiological images in the healthcare sector. After reviewing existing classification approaches carried out over chest radiographs, it was explored that existing techniques are highly restricted to perform binary classification that is not comprehensive for assisting in an effective diagnosis process of chest disease condition. This paper presents a novel approach to classifying chest x-rays on the basis of the practical disease condition. Harnessing the potential features of content-based image retrieval, the proposed system introduces a novel concept of attribute map that not only performs comprehensive classification but also makes the complete computational model extremely lightweight. The study outcome proved to offer better accuracy with the proposed non-iterative process in contrast to existing classifier design.**

***Keywords*—Chest x-ray; classification; supervised learning; radiographs; accuracy***

## I. INTRODUCTION

There has been a significant improvement in the area of medical image processing with an evolution of various devices and forms of medical images [1][2]. Essential body parts that are found to be investigated most frequently in the majority of the hospital using medical image processing is the chest region. At present, there is a various form of advanced radiological techniques used for assessing the condition of the chest, e.g., Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Positron Emission Tomography (PET), etc. [3]. Medical assessment using these processes are quite less frequent as they are used either to assess the advance critical condition of the chest. In any cases of diagnosis, a chest x-ray is the most frequently adopted diagnosis method [4] recommended by physicians. Various research papers have discussed challenges associated with diagnosing medical images, but very few are found to be associated with chest x-ray [5][6]. Over the period of time, certain researchers have dedicated themselves to discuss the existing challenges as well as problems associated with image processing techniques over chest radiographs [7][8][9]; however, there is no effective disclosure about effectivity in the approaches presented by researchers. The automated diagnosis processes (on the basis of research) are categorized into two problems, i.e., identification process and classifying the disease. There are various significant techniques for assisting in detection operation for chest radiographs [10]-[15]. The detection process is mainly dependent on segmentation techniques [16] and feature-based aspects [17] while the classification techniques are more into involving machine learning approaches [18]-[20]. However, there are various associated problems of existing classification techniques that are found unspoken in any research papers. The *first* problem is associated with the classification objective itself. Almost all the researchers have performed binary classification problems that are quite essential but only in preliminary level. For example, if the queried image is found to be malignant in binary classification process than it should further show the stage of cancer or exhibit more comprehensive information. Disease conditions of lungs are quite lethal, and it requires precise diagnosis, whereas all the existing researchers have only presented the solution to the just preliminary stage. The *second* problem associated with any of the existing approaches is related to the cardinality of the outcomes. 99% of the research work uses a trained dataset to be compared with the test image with the aid of features-based method. The outcome is the only one trained image to show the similarity match with the test image. This reduces the scope of inferring the outcome of classification and this lead the researchers to highly depend on the flow of the techniques used for database building as well as for classification. Adoption of such a process is evaluated only with respect to the empirical approach adopted, or the feature-based approached utilized and possibly it does so by overlooking the actual clinical context of it. For example, a chest x-ray of pleural effusion, as well as a chest x-ray for any pulmonary edema within the lungs, is visibly the same. There are good chances that if existing classification approaches are applied that it could not be able to identify the difference between them. It should be noted that pleural effusion, as well as pulmonary edema, are two different disease condition of the chest which has to be diagnosed discretely.

Hence, the proposed paper discusses a novel technique of classification that is capable of addressing the problem discussed above. The organization of the present manuscript is as follow: Section II presents a discussion of existing research approaches towards classification, Section III highlights the identified research problems addressed in this paper, Section IV briefs about research methodology undertaken, Section V discusses algorithms implemented for accomplishing the goal of this paper. Finally, result analysis is discussed in Section VI, and summary of contribution is briefed in Section VII in the form of a conclusion.

## II. Related Work

There are various literatures contributing towards unique techniques of classification in medical images. This section updates more research work as an extension to our prior review work [21].

Classification performance can be optimized if a closer connection is set with the identification technique. One such attempt was found to be seen in the work of Katre and Thakare [22]. However, an in-depth study shows that the study is more focused on implementing *the watershed algorithm* for segmentation as well as it addresses the basic noise removal process. The overall claim of detection and classification of chest x-ray is not seen in this work, and hence the work just offers elementary theoretical guidelines. The work of Mukherjee et al. [23] has also claimed combinely addressing detection as well as classification problems in lung images. The authors have mechanized a unique bilateral filtering approach for eliminating the noise followed by segmentation and thresholding process. Finally, a binary classification technique was applied over the extracted features (e.g., eccentricity, circularity, and aspect ratio). The study offers good detection implementation; however, the binary classification process is not a novel or unique approach. Apart from this, the applicability of approach towards multiple disease condition of the chest is very unlikely. Research towards identification problems has been carried out by Bar et al. [24] where *convolution neural network* has been used along with binary supervised classification technique, which is quite a common approach adopted by frequent researchers. Unfortunately, binary classification is only meant for critical disease condition of the chest and doesn't relate to other potential disease condition.

Early research work has also focused on using fuzzy logic for the purpose of classification of the lung nodules. Roy et al. [25] have used contour-based model for accomplishing segmentation. Hence, the study was more inclined towards detection and less toward classification. Similar research flow was also adopted by Zhang et al. [26]. Zhang et al. [27] have presented a classification process that uses rank for classifying. Researchers have also presented a mechanism for unique binary classification process that is meant for assists in further classification process but not directly for disease diagnosis (Xue et al.[28]). A simplified *rule-based method* was reported to assist in classification as reported in work of Taher et al. [29]. It was also seen that the usage of the Bayesian classifier as reported in the work of Dhaware and Pise [30] is found to improve identification of lung cancer. *Fuzzy-c means* clustering was applied for improving the detection as well as for addressing directionality reduction. However, the study lacks extensive evaluation to prove its robustness against different disease condition of lungs. One of the early initiatives for classification process has been initiated by Kulkarni and Panditrao [31] where the authors have signified the essential contribution of medical image processing for diagnosis of lung cancer and importance of classification. The neural network is another frequently used approach that offers the simple classification approach along with feature extraction. However, the outcomes are not subjected to comparative evaluation for assessing the outcomes practicality. A block-based approach has been used by Ahmad et al. [32] where the classification has been carried out by *the Bayes classifier*. The target of classification was mainly to identify the pleural effusion in chest x-ray to show 100% accuracy with Gabor filter. However, the study can only classify any case that comes under the pleural effusion condition of the chest x-ray. Learning approach plays a significant role in the classification of chest x-rays. A study conducted by Dong et al. [33] have shown that it is feasible to train the multi-class model for assisting in further classification using a *convolution neural network*. The study has carried out analysis on the dataset of different disease condition using standard accuracy parameters; however, the technique adopted was quite iterative and less challenging datasets. Similar usage of convolution neural network was also proven to assists in classifying lung nodule as claimed by Liu et al. [34]. The study outcome shows the comprehensive benefits of performing classification operation; however, the classification rate is slightly diminished compared to another existing system.

Investigation towards classifying lung nodule was also reported in the work of Mao and Deng [35] where a *pattern-based approach* was used for performing representation of effective features. The study uses a linear classifier, e.g., supports vector machine with the impressive rate of classification outcomes. However, there is no evidence if it can be used for classifying other disease condition of a chest. Usage of deep learning algorithm can also improve the representation performance of features. Literature also claimed of using an encoding technique of local binary pattern for classification. [36] The work of Mao et al. [37] has used deep learning approach to perform classification of the lung nodule. The outcome shows better performance in contrast to other conventional classifiers. Work carried out by Kumar et al. [38] has presented the discussion of usage of *deep features* for assisting in solving classification problems of chest radiographs. Harnessing in the potential of deep learning, the author discussed that it is an effective classifier for binary classification of lung cancer stage. Unfortunately, deep learning is the quite advanced version of the machine learning approach using a neural network, and it is capable of doing more than binary classification, which was not found to be attempted. Deep learning and convolution neural network were also reported to improve the classification performance. This fact was discussed by Li et al. [39] where *region-of-interest* was used for performing the classification process. However, this complexity of analysis by the computational model was not checked owing to the usage of the region of interest based approach with almost information defined for critical classification. Wang et al. [40] have discussed the impact of database towards benchmarking problems in the classification of chest x-rays. Apart from this, there are also reported usages of Principal Component Analysis (PCA) for solving classification problems in chest x-ray [41]-[45]. The next section briefs the research problems associated with the existing research techniques of classification.

## III. Research Problem

From the prior section, it can be seen that there are various existing techniques of classification in medical images especially considering chest x-ray. However, are various

significant problems have been found unaddressed. This section briefs about such forms of unaddressed research problems as follows:

- *The gap between detection and classification problems*: It has been observed that detection, as well as classification, has always been treated as separate problems in an existing system. The integrated study of connection of detection and classification with respect to accuracy has been less emphasized.
- *Existing Classification Uncertainty*: There are various research works towards classification where the outcomes show significant accuracy. However, it should be known that such accuracy is claimed on the basis of the single instance of a match with the database or trained images. The classifications are numerically proven correct, but there was no explicit discussion of accuracy with respect to disease-based classification.
- *Unprepared Dataset*: The existing publically used dataset doesn't have any well-structured categorization of the chest x-ray on the basis of any disease. However, there are disease-specific dataset too [46]-[49].
- *Untapped Potential of Features*: Features are one of the significant information set to represent the complete dataset. Various empirical attributes within the features have been less explored for hence granularity of the classification process has never been testified in this regards.

Therefore, the proposed study targets to address the entire above problem in order to evolve up with a novel and robust classification process of the chest x-ray.

## IV. Research Methodology

The proposed study is a continuation of our prior research work [50]. By adopting an analytical research methodology, the proposed study presents a novel classification process of chest x-ray very differently than existing approaches. In order to address the *first* research problem, the proposed study introduces a novel attribute map-based method for assisting in the classification process. This mechanism is used in both detection and classification problems. The *second* research problem is addressed by applying content-based image retrieval technique where classified images have multiple options to be double checked for its accurate prediction. The *third* research problem is addressed by synthetically developing a well-structured dataset of trained images. The training is carried out using attribute maps and a supervised learning technique. The *fourth* problem is addressed by completely developing the model using attribute and not features. The difference between an attribute and features is that features are a subset of information for attribute and hence working with attribute offer more granularities in system design as well as assist in assessing the classification performance. Various empirical variables associated with the image-related pixel and numerical information forms an attribute. Fig.1 highlights that complete model consists of i) training phase and ii) classification phase. The training phase consists of applying enhanced supervised learning algorithm over the disease-specific dataset. The classification phase consists of applying a series of algorithms in order to obtain classified images. For better study accuracy, the proposed model contributes to offering better inference in its outcome by incorporating ranks with the outcome of the similar set of images and thereby redefining the conventional process of classification of a chest x-ray. The next section illustrates about the algorithm implemented for the proposed system.

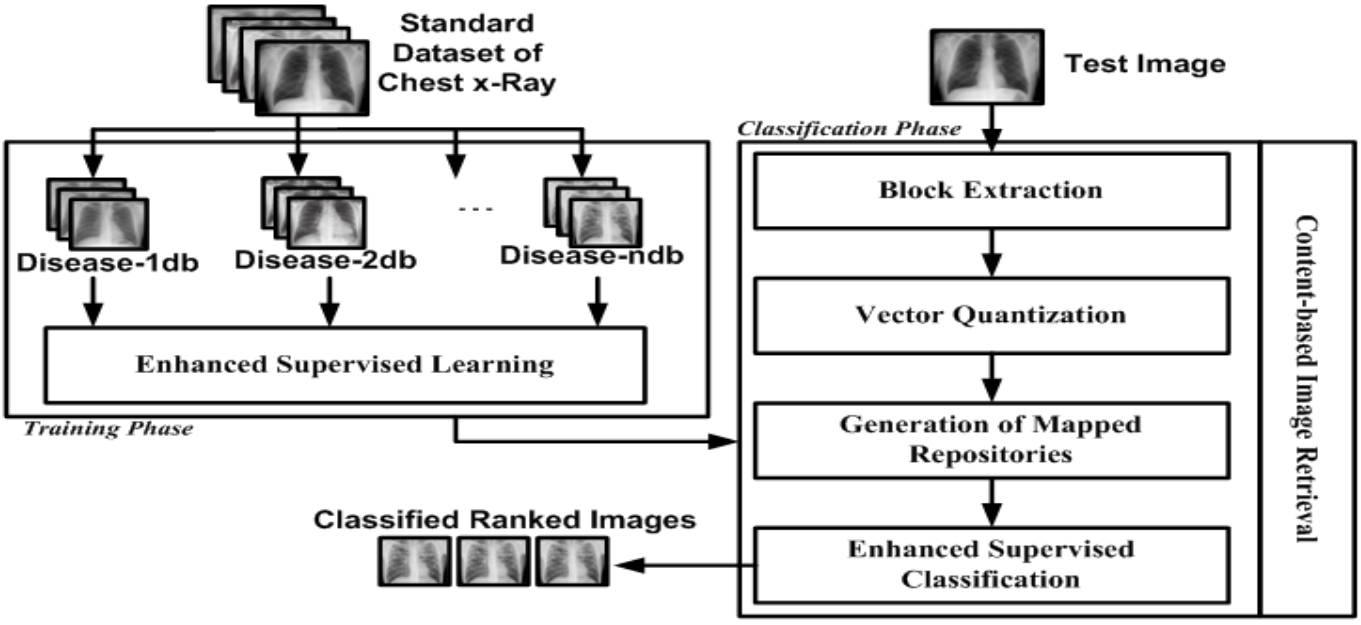



Fig. 1. Schematic Architecture of Proposed System.

## V. Algorithm Implementation

This section discusses the series of algorithms that have been constructed in order to perform classification of the chest radiographs. There are various inclusions of enhanced machine learning approach as well as content-based image retrieval approach implemented in order to perform classification of the chest x-ray. The discussions of the implemented algorithms are as follows:

### A. Enhanced Supervised Learning

This is the first algorithm that is responsible for generating a trained dataset of a chest x-ray. The complete training is carried out using enhanced supervised learning approach (as Shown in Fig.2). From the application viewpoint, this algorithm can be considered to be configured on various terminals of the healthcare facility that takes the feed of different forms of chest x-rays. All the input images are not directly stored in physical/cloud storage units, but they are trained, categories, and the arranged in a highly structured manner.

According to the above scenario, all the input chest x-rays are forwarded to the terminals by all respective radiologists. However, they are all mixed and will be needed to be well categories. .The initial categorization is carried out manually as only these images will be later treated as a reference image. Hence, it is considered that such images are error-free and involved all forms of information that are required to be used by the physician to perform an effective classification process during the query. It is also considered that each category represents a specific form of cancer or certain abnormalities of the chest x-ray. The prime target is that while performing a query, it is essential that the queried image be subjected to classification with a higher degree of similarity with a maximum number of images, unlike conventional approaches of disease classification. Therefore, the process of generating trained data is essentially important to be effectively maintained. The algorithmic steps of the supervised learning are as followed:

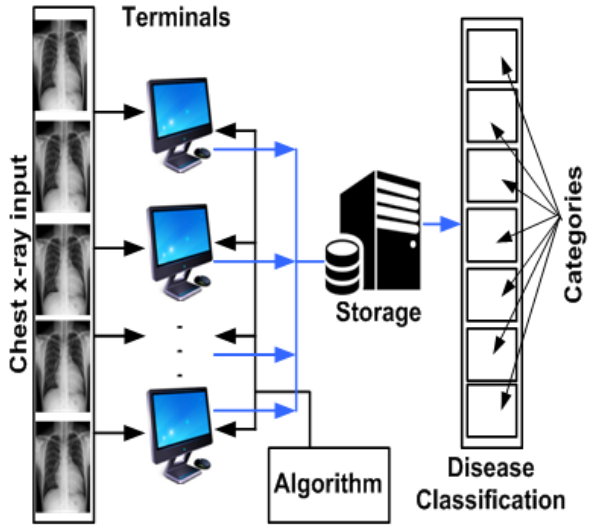



Fig. 2. Considered Scenario of Algorithm Implementation.

**Algorithm for Enhanced Supervised Learning**

**Input**: $I$, $d$, $n$

**Output**: $t_d$

**Start**

1. $I=\{d(I_{n1}), d(I_{n2}), \ldots, d(I_{nn})\}$
2. **For** i=1: $n$
3. $a_{map} \rightarrow f_1(d(I_i))$
4. $c_{class} \rightarrow f_2(I, a_{map})$
5. $c_{mat} \rightarrow f_3(c_{class}, I)$
6. **End**
7. $t_d=\{a_{map}, c_{class}, c_{mat}\}$

**End**

The illustration of the lines of the above algorithms is as follow: The algorithm takes the input $I$ (chest x-ray), $d$ (dataset), and $n$ (number of the dataset) that after processing yields an outcome of $t_d$ (trained data). The algorithm doesn't take a single image as input, but it takes the complete folder with category mentioned for the pre-defined classified disease as the input. This design principle also offers the advantage of investigating by considering various types of dataset $d$ towards the classification process. The complete dataset of an image I is formed by considering each individual dataset $d$ with different types, formats, and the number of images within it. Consider that dataset $d(I_{n1})$ consists of $n_1$ number of images, dataset $d(I_{n2})$ consists of $n_2$ number of images, and so on (Line-1). Considering all the $n$ number of images of chest x-ray to be trained (Line-2), the first processing carried out by the proposed system is to extract an attribute map $a_{map}$ (Line-3). This extraction process is carried out by an explicit function $f_1(x)$ that returns a matrix of all the visual-based attributes. As it is a matrix-based operation, therefore, the input $d(I_i)$ is required to be in an array form of the chest $x$-ray image. Owing to the matrix-based form of an input data set, it is feasible to extract the potential attributes from all the digitized chest x-ray irrespective of any shape and size. The algorithm specifies the location of an image ($I_{loc}$) by defining the aggregated images of chest x-rays. The study represents $I_{loc}$ in the form of the string that represents an array of input images. The algorithm does this operation recursively for all the input chest x-ray images. It then generates a set of the map against all the significant visual attributes of an input image. The process of execution of $f_1(x)$ is carried out in three different steps, i.e., identification of attributes, describing the attributes, and generation of the final matrix of reference attribute. The proposed study considers high-level extraction of attributes in its first stage followed by abstraction of all the images by individual blocks as a step to describe the attributes. The mechanism of representing attribute is all about the mechanism of representing this block as the numerical vector. The final process leads to a generation of referential attribute matrix that maps the information about the similar form of blocks. The next step in algorithm implementation is to performing training over the targeted classifier in order to obtain better results in the trained dataset. For this purpose, a different function $f_2(x)$ is constructed (Line-

4) that takes the input of the input images as well as generated attribute map (Line-4). The prominent potential of this function $f_2(x)$ is that it constructs a specific category for the different set of chest x-ray images by harnessing the potential of machine learning and statistical approach. All the computed elements of I will now be represented by a specific set of categories. Further, this process is optimized using a supervised learning approach of Support Vector Machine in order to perform the more enhanced form of classification and further results in category class matrix $c_{class}$ (Line-4). The final step of this algorithm is to carry out an assessment using third explicit function $f_3(x)$ considering input arguments of category class $c_{class}$ and input images *I* (Line-5). This also leads to a generation of an error matrix that can be used for further analysis. All the outcomes, i.e., attribute map $a_{map}$, category classifier $c_{class}$, and category matrix $c_{mat}$ are stored in the form of trained data $t_d$ (Line-7). This completes the algorithm operation, and now the system is ready to perform classification with the aid of a test image.

### *B. Block Extraction*

This algorithm is responsible for generating an image block from the input of the test image, where each block is the direct representation of a localized form of the chest x-ray. The blocks also represent some of the essential information pertaining to image maps.

According to the Fig.3, the outcome of the block extraction process is basically an attribute matrix, which contributes towards the classification process. Each block represents the overall test image in multiple form and dimension and is obtained by specifying a particular size of cells in the form of a number of rows *nr* and columns *nc*. A statistical operation is further applied in this algorithm to ensure that the outcome is least affected by any form of variability factors (e.g., illumination, brightness, contrast, etc.). The steps of the algorithm are discussed as follows:

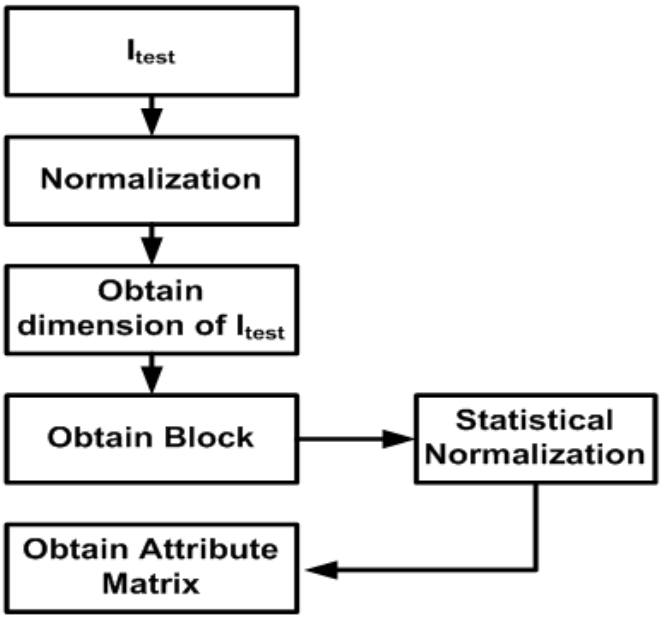



Fig. 3. Proposed Block Extraction.

**Algorithm for block extraction**

**Input**: I, b

**Output**: B

**Start**

1. init b
2. I=g(I)
3. **For** i=1:b:$nr_b$
4. **For** j=1:b:$nc_b$
5. B→I(i:$i_b$, j:$j_b$)
6. B=apply *norm*(B)
7. A→[A, B]
8. **End**
9. **End**

**End**

The algorithm takes the input of *I* (queried image) and *b* (block size) in order to generate the blocks. After initializing with the number of block size *b* (Line-1), the algorithm performs slight fine-tuning of an image using a function $g(x)$ in order to offer better precise during the classification process. The fine-tuning is carried out by resizing the image followed by increasing the image to double precision (Line-2). The size of the test image is obtained to generate a matrix of dimension *nr* x *nc*; however, for blocking operation, this size will be required to be changed. Hence, a modified size of all rows $nr_b$ and columns $nc_b$ is obtained as nr-(*b*-1) and nc-(*b*-1) respectively. The abstract representation of this is $nr_b$ and $nr_c$ (Line-3 and 4) is used in order to construct a new block B (Line-5). Although the steps till now assist in constructing a new block, its outcome will be affected if the original test image is subjected to a different form of events that results in significant variance to the constructed blocks. This could also reduce the identification as well as classification performance. This problem is mitigated by applying a standard deviation to the obtained block B followed by further normalization where B is subtracted from the mean value of the blocks. The resultant is further divided by the obtained value of standard deviation that leads to the generation of final block B (Line-6). These extracted blocks also act as the specific attribute of high level for assisting in the classification process.

### *C. Vector Quantization*

After the blocks have been extracted in the prior algorithm, now the system is nearly ready to perform classification. The Fig.4 represents the flow of vector quantization. However, the information obtained from the attribute matrix is so massive that it requires performing the certain form of clustering operation. The proposed system applies vector quantization in order to carry out clustering operation. It was seen in the prior algorithm implementation step that vector is transformed into blocks after the process of generation of attribute map that finally results in forming a map repository (refer next algorithm). The proposed system applies vector quantization as a process of clustering that segregates all the vectors into

different groups on the basis of their attributes. The center of cluster obtained from this assists in forming the elements of mapped repositories.

According to the above flow of the vector quantization technique in clustering, the proposed system takes the input in the form of cluster definition in terms of a number followed by computing the center of cluster and computing spatial distance of all objects to the center of clusters. Using the criteria of the reduced distance scale, the system performs clustering. This results in obtaining a faster clustering operation. The steps of the algorithm are discussed as follow:

**Algorithm for Vector Quantization**
**Input**: A (attribute)
**Output**: $C_{id}$ (center of cluster)
**Start**
1. $C_{id}$→*vect_quan* (A, M)
2. Function *vect_quan*
3. $(s_1, s_2, ...., s_m)$←arb$((x_1, x_2, ....x_N)M)$
4. **For** m←1 to M
5. **Do** $\mu_m$←$s_m$
6. **While** the stopping criterion has not been met
7. **Do** for m←1 to M
8. **Do** $\omega_m$←{}
9. **For** n←1 to N
10. **Do** j←$\arg_{min}||\mu j-xn|$
11. $\omega_j$←$\omega_j \cup \{x_n\}$//reassignment of vectors
12. **For** m←1 to M
13. **do** $\mu_m = \frac{1}{|\omega_m|}\sum_{x\in\omega_m} x$
14. return $\{\mu_1, ....\mu_m\}$
15. **End**
16. **End**
17. **End**
**End**

Owing to the generation of a massive set of information, the Euclidean distance among the data points will actually increase and will be highly uncertain. Therefore, the prime motive of the presented vector quantization technique is to reduce the mean of the squared distance of all the data points with respect to the center point of the cluster. The algorithm computes the center of cluster $\mu$ of all images corresponding to $\omega$ cluster. The computation of the center of the cluster is carried out considering a function called as vector quantization (Line-1). The input to the algorithm is an attribute matrix *A* and limit the size of *M* for clustering. The algorithm formulates seeds by capturing different data points of order *M* and selects them randomly (Line-3). For all the specified number of clusters *M* (Line-4), the algorithm assigns respective seeds to the center of clusters $\mu$ (Line-5). In case the criteria of terminating the iteration have not been met than the algorithm attempts for minimization of the Euclidean space (Line-9-11) by performing reassignment of the vectors as well as re-computation of the center-of-clusters (Line-13). The outcome of the algorithm is finally a center of cluster $C_{id}$ that will be used as an input to the next algorithm for generating mapped repositories.

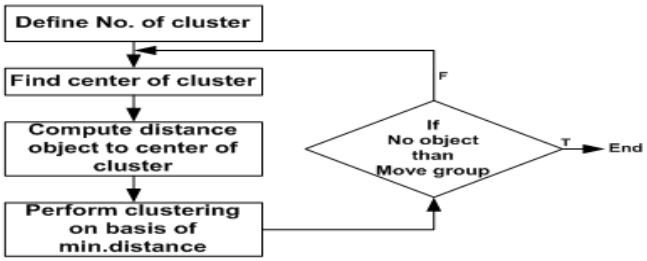



Fig. 4. Vector Quantization.

*D. Generating Mapped Repositories*

This algorithm is responsible for generating mapped repositories, which is a collection of all significant attributes from the images. The operations are two-sided. The first side of the operation is associated by identifying significant attributes followed by storing all this mapped information in one matrix called as map repository. This repository generates a perfect representation of an image. This repository will also be used by the test image in order to perform classification of the disease categories corresponding to the unique dataset of chest x-ray (refer Fig.5).

**Algorithm for Generating Mapped Repositories**

**Input**: $C_{id}$

**Output**: MR

**Start**

1. MR→mat (axb)
2. **For** i=1:p: M
3. **For** j=1:p:M-1
4. $a_{vec}$→$C_{id}(q)$
5. q=q+1
6. $a_{vec}$→$\delta(a_{vec})$
7. MR←$a_{vec}$
8. **End**
9. **End**

**End**

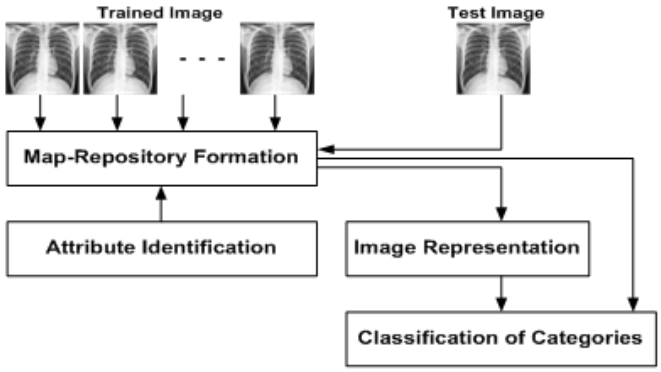



Fig. 5. Process of Map Repositories.

The algorithm begins by developing a matrix of size a *x* band it takes the center of cluster $C_{id}$ as an input (Line-1). An effective center of cluster $C_{id}$ is considered by following empirical expression,

$C_{id} \rightarrow C_{id}\text{-}arg_{min}(C_{id})$ (1)

The above empirical expression (1) represents the displacement of the scope of $C_{id}$ from 0 to a positive value. The complete process of generation of the map repositories *MR* is carried out for *M* number of clusters. For all the maximum value of the clusters (Line-2 and 3), the attribute vector is obtained for the center of clusters corresponding to its specific index value (index *q*) (Line-4). The index is incremented (Line-5) further followed by reshaping $\delta$ of the attribute vector (Line-6). Finally, the obtained result of attribute vector $a_{vec}$ is stored back to a matrix MR (Line-7).

### *E. Enhanced Supervised Classification*

The prime purpose of this algorithm is to perform disease classification for the test image of the chest x-ray. The operation is meant for forecasting the relevance of the input image of a chest x-ray (test image) with the most relevant trained dataset.

As the proposed concept is designed on the basis of content-based image retrieval process, therefore, the outcome of the classification will be more comprehensive, unlike the existing system. The proposed system uses a support vector machine for this purpose in order to perform classification operation. One of the significant enhancements of supervised technique applied in this algorithm is that it offers a significant control of iteration towards performing forecasting for a given set of data. The forecasting process carried out for the proposed system is done only once for one image, and that makes its non-recursive learning algorithm, unlike the existing applications of using a supervised technique that is highly recursive in order. Another significant contribution of the proposed algorithm is that it offers a ranking mechanism to exhibit the classified images as an outcome. The complete process of enhanced supervised classification is exhibited in Fig.6.

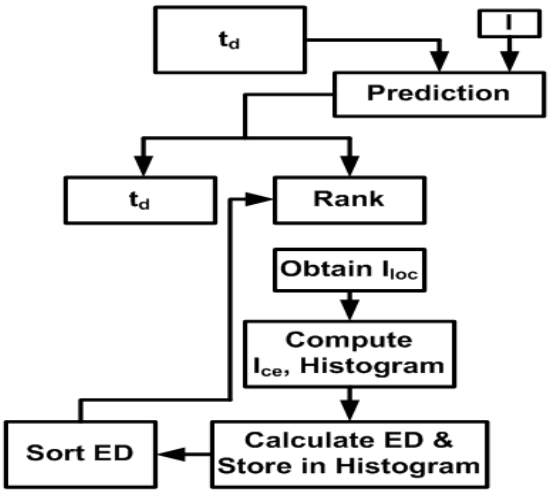



Fig. 6. Process of Enhanced Supervised Classification.

Fig.6 showcases the complete process of applying the predictive-based approach in order to obtain highly precisive classification process. A similar illustration is given in the form of algorithmic steps as following:

**Algorithm for Enhanced Supervised Classification**

**Input**: $t_d$, I

**Output**: $I_{type}$, $I_{ranked}$

**Start**

1. load $t_d$
2. $[l_q, \text{rank}] \leftarrow \rho(c_{class}, I)$
3. $I_{type} \leftarrow c_{class}(l_q)$
4. $svm_{op} \leftarrow l_q$
5. flag $I_{type}$ as an identified category of trained data
6. $I_{loc} \rightarrow f(svm_{op})$
7. $[H_o \text{ bins}] \rightarrow \tau(I_1)$
8. **For** i=1:size($I_{loc}$)
9. $I_t \rightarrow [I=(i)]$
10. Compute $E_{dist}(H, H_o)$
11. $ED_{vec} \rightarrow [ED]$
12. $I_{ranked} \rightarrow$ show ranked images (sort($ED_{vec}$))
13. **End**

**End**

The algorithm takes the input of $t_d$ (trained data) and I (queried image) that after processing leads to the generation of $I_{type}$ (Classified image) and $I_{ranked}$ (sorted images). This algorithm has dependency over the trained data $t_d$ (Line-1) that will be used mainly for checking the relevancy between the queried image and trained images. A function $\rho(x)$ is constructed to carry out a predictive operation using input arguments of category class $c_{class}$ and test image I (Line-2). The process results in label $I_q$ and rank (Line-2) where the labels stored back in an explicit matrix $svm_{op}$ (Line-4). Upon this step of implementation, the algorithm starts exhibiting the nearest and similar dataset that corresponds to the queried image. The next process involved in the algorithm is to obtain the information related to the location of an image $I_{loc}$ (Line-6). A new function $\tau$ is applied over resized image $I_1$ (Line-7) in order to obtain histograms $H_o$ as well as a bin. The algorithm also obtains another form of histogram *H* that is applied over the trained dataset. An interesting point to observe is that this step of histogram calculation is not carried out in database construction stage (first algorithm). Therefore, for all the sizes of image locations (Line-9), the algorithm applies Euclidean distance formula considering histogram from queried image Ho and that from trained image *H*. Finally, correlation is obtained between the queried image $I_1$ and trained image $I_t$ and the value is stored back in matrix $E_{dist}$ (Line-10). After obtained vectors of distance (Line-11), the sorting is carried out using Euclidean distance $ED_{vec}$ to complement the ranking process. The accomplishment of this final step of algorithm results in the

display of relevant images that are ranked based on their relevancy level with the test image.

## VI. Result Analysis

As the proposed system targets to offer a classification of the chest x-ray; therefore, the performance parameters selected in this process has to be closely related to the accuracy of the classification process. However, unlike any existing classification process where there is the only a single outcome of the classified image, the outcome of the proposed system will result in exhibiting multiple relevant images from the dataset. The outcome also comes with an identification of a relevant dataset that represents multiple disease categories. This section discusses the result analysis with respect to the assessment and implementation strategy.

### *A. Database Considered*

All the images for assessing the algorithm implementation is carried out considering the Japanese Society Radiological Technology (JSRT) dataset which hosts various forms and modalities of chest x-rays [51]. The average size of all the grey-scaled chest x-rays is approximately 8192 kilobytes, and it bears 12 bit of color. However, one of the biggest hurdles in using this database is that it has combined collection of all images together whereas the proposed system demands a highly structured categorized database as per the disease condition. Therefore, after obtaining the complete database, it was subjected to further manual categorization as per critical disease condition of the lungs. The proposed system considers 7 disease conditions of the lungs e.g. i) Chronic obstructive pulmonary disease [52], ii) Emphysema [53], iii) Bronchitis [53], iv) Cystic Fibrosis [51], v) Pleural Effusion [51], vi) Tuberculosis [54], and vii) Adenocarcinoma [55]. All the above mentioned disease conditions are most frequently encountered and the present database has possession of chest x-ray corresponding to this disease condition.

### *B. Visual Outcomes*

Implemented on MATLAB, the visual outcomes of the proposed study was feasible to be exhibited properly. Fig.7 highlights the input image that is subjected to blocking operation (Algorithm-B) to obtain the blocked image as shown in Fig.8. The blocks are then subjected to a consecutive algorithm for obtaining the attribute space (Fig.9). Finally, the center of the cluster is obtained (Fig.10) while the map repository is constructed (Fig.11).

Fig. 7. Input Image.

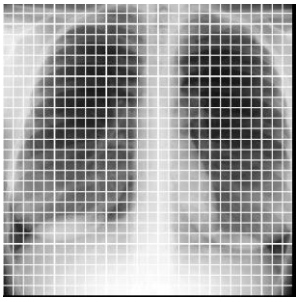


Fig. 8. Image with Blocks.

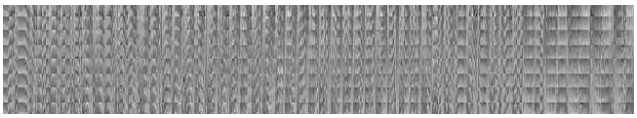


Fig. 9. Extracted Attribute Space.

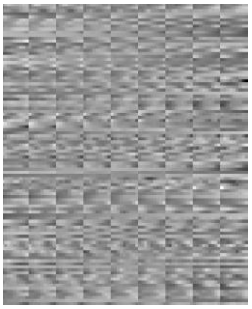


Fig. 10. Center of Cluster.

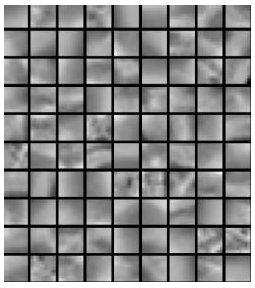
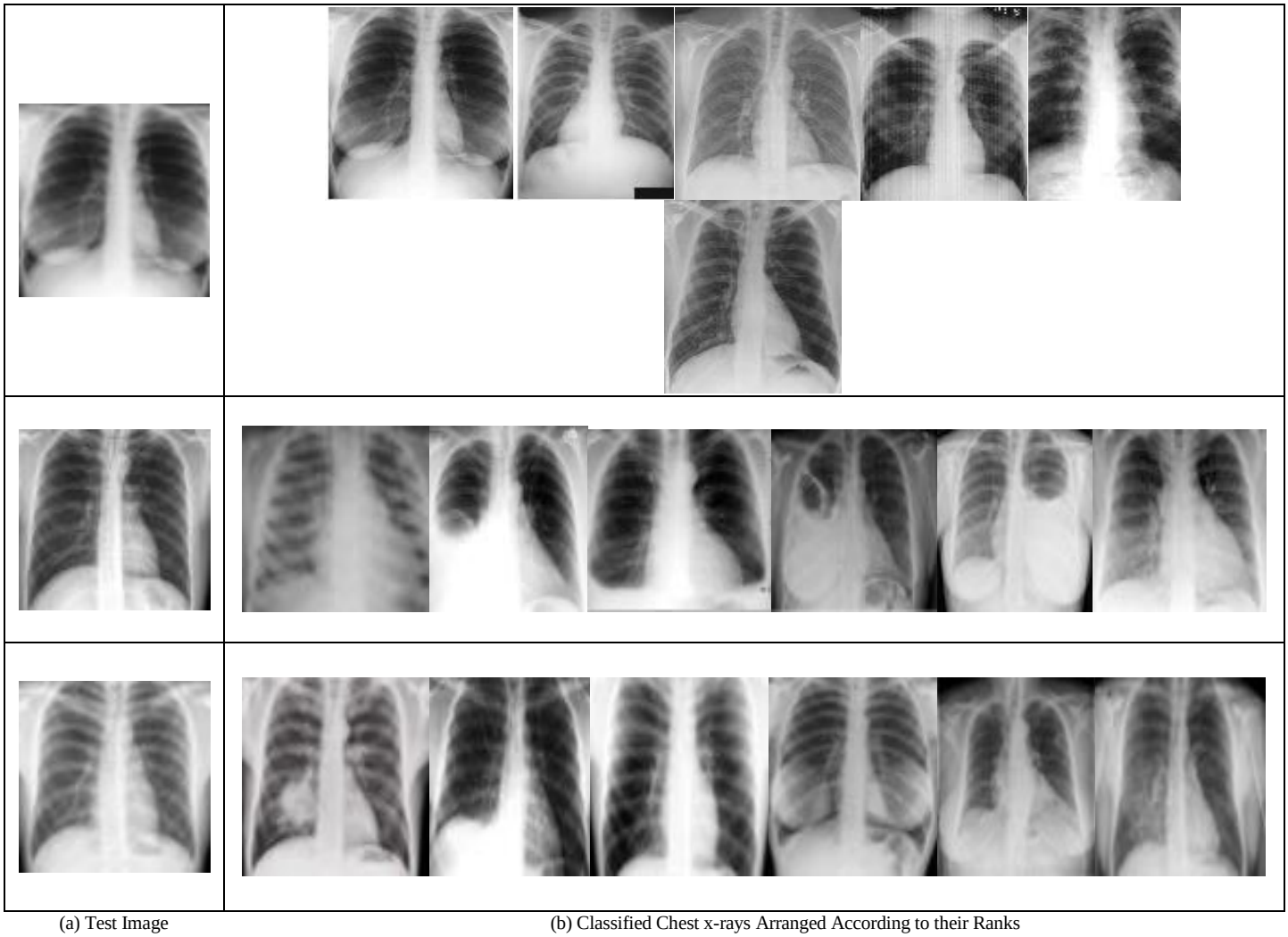


Fig. 11. Map Repository.

For final assessment of proposed classification using content-based image retrieval technique, the proposed system subjects multiple forms of test images to find the visual match of the most relevant images from the disease-specific dataset. One interesting part to observe is that the proposed system encapsulates complete information about the prominent attributes within a map repository, thereby making the complete process more lightweight when the classification is performed. The significant contribution are mainly two fold viz. i) the massive information of the large number of images in the dataset can be now stored in the form of map repository, which is comparatively very much smaller in size in contrast to total size of all images considered for training, Although, the map-repository shown in Fig.12 belongs to a single test image, but the same process can also be applied over the trained attributes of the images. This process makes the classification quite faster and highly percussive. Fig.13 highlights some of the sample outcomes of the classification process.



(a) Test Image (b) Classified Chest x-rays Arranged According to their Ranks

Fig. 12. Outcomes of Classification.

Exhibition of all visual outcomes is out of the scope of this paper. The novelty/contribution of this visual outcomes are again multi-fold viz. i) existing algorithms discussed in literature mainly classify one medical image (test image) with multiple trained images, but ultimately only one trained images are considered as the most matched outcome. Although such outcomes of the existing system may be technically correct from an image processing viewpoint, they could be clinically wrong. This can be illustrated in following-usually, there are good possibilities that chest x-ray of a subject suffering from two different diseases bears similar visual appearance, but clinically they are wrong. The root cause of this problem is the existing database system which is not arranged or organized with respect to multiple disease condition of a chest x-ray. Hence, the proposed system performs the classification of chest x-rays corresponding to some critical disease condition of the chest and performs the visual analysis on that basis. The outcome shown in Fig.13 only highlights that irrespective of any form of test images, the proposed system could successfully perform an effective classification of respective disease, ii) another significant benefit of this visual outcome is that a physician is offered with multiple referential images from the trained dataset that matches with the queried image. This phenomenon allows the physician to correlate the disease condition with more helping referential / clinical files associated with the extracted ranked classified images. Hence chances of true and reliable diagnosis of a disease condition are highly made possible using content-based image retrieval system implied on the classification technique. Another significant advantage is iii) proposed system offers highly precisive display of the classified image irrespective of different forms, shape, orientation, and format. Hence, the proposed system offers true test-bed to perform diagnosis of critical chest disease.

### *C. Numerical Outcomes*

The numerical assessment of outcomes obtained from the proposed system has been assessed with sensitivity and precision as core performance parameters. Although, there are many numbers of images within the JSRT dataset, the discussion of the outcomes is carried out using a small number of images in order to practically understand the numerical parameters, e.g., attributes are represented in Table.1. The assessment has been carried out considering 12 images each in a different dataset.

TABLE I. NUMERICAL INFORMATION OBTAINED FROM A DATASET

| A dataset with disease Name | Number of images | Number of extracted attributes |
|---|---|---|
| COPD | 12 | 858336 |
| Emphysema | 12 | 228108 |
| Bronchitis | 12 | 434380 |
| Cystic Fibrosis | 12 | 1072636 |
| Pleural Effusion | 12 | 617056 |
| Tuberculosis | 12 | 342112 |
| Adenocarcinoma | 12 | 273416 |

A simple table like this one above with only 12 images shows that it could have a massive number of attributes. This is important as the proposed system performs enhancement of the supervised learning approach using an attribute map. The analysis has exhibited that there are approximately 80% of the potential attributes that can be utilized with each image of the chest x-ray. The analysis also performs balancing of the attributes across all the chest x-rays in order to enhance the performance of clustering using vector quantization. For an effective analysis of the proposed system, the outcomes have been compared with another frequently used classifier, i.e., neural network and principal component analysis.

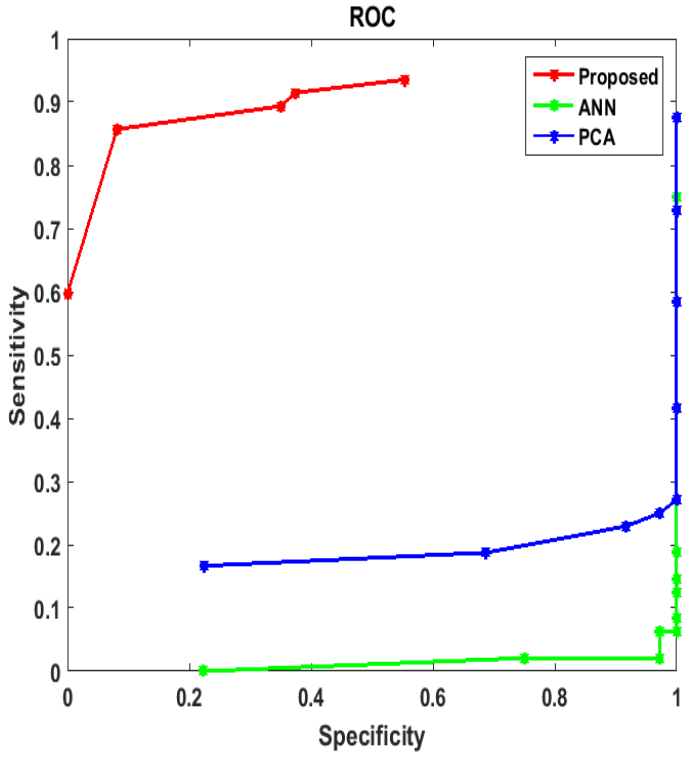



Fig. 13. Comparative Analysis of Sensitivity Vs. Specificity.

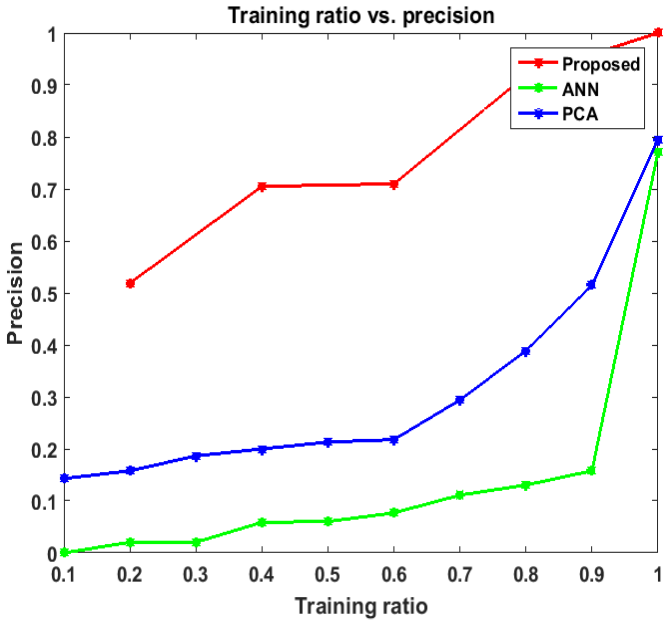



Fig. 14. Comparative Analysis of Precision Vs. Training Ratio.

A closer look at both the outcomes of Fig.13 and Fig.14 shows that the proposed system offers good accuracy performance as compared to existing classified design using machine learning (ANN) and statistical approach (PCA). The prime reasons behind this are multifold-the filtration process of ANN to obtain the elite outcome is highly accurate, but on the process of training, it cancels out the prior outcomes that it believes will adversely affect the faster convergence performance. Apart from this, the training process of ANN is quite iterative that could offer significant computational complexity while performing training. Similarly, usage of PCA is more emphasized on exploring orthogonal projections associated with the images of chest x-rays. It also looks for images with a higher degree of variance values. The good part of using PCA is to understand the correlation of the data within the same diseases as the data points are linearly correlated. However, a problem starts to surface when the PCA algorithm needs to read another dataset that corresponds to the different disease. Hence, while building / training dataset, PCA just filters out the data point with less variance value thereby leading to dropping of some good cases of similarity with the test data. Hence, it doesn't perform better than the proposed system. However, it performs better than ANN only for the reason that it is less iterative and it retains the potential amount of information associated with attributes for each dataset, which is not carried out by the neural network. The prime reason behind the outcomes of the proposed system is the mechanism of constructing attribute map that offers an efficient collection of all potential attributes specific to each disease for which reason classification can be carried out effectively. Analysis of computational complexity has been carried out with respect to algorithm processing time. Following are the outcomes.

The analysis has been carried out on the core i3 processor with all the trained images in a dataset where the algorithm processing time has been programmatically executed. A closer

look at the outcome (Table 2) shows that the processing time of the proposed system is slightly lower than a neural network based approach and slightly higher than principal component analysis. Although, with an increase of more images will definitely increase the processing time, but they are found to be within the tolerable limits. However, memory and processor is widely available in existing healthcare industry but what is required is a device with the capability of precise diagnosis of disease condition that can be effectively done by the proposed system in contrast to the existing system.

TABLE II. COMPARATIVE ANALYSIS OF ALGORITHM PROCESSING TIME

| Algorithms | Training Time | Classification time |
|---|---|---|
| Neural Network | 6.29388 sec | 0.14 sec |
| Principal Component Analysis | 7.89224 sec | 0.07 sec |
| Proposed | 9.29331 sec | 0.11 sec |

## VII. CONCLUSION

This paper has presented a unique approach to classifying the disease condition of chest x-ray using non-conventional and out of mainstream research methodologies. Review of existing approaches shows that binary classifier design assists in foretelling the preliminary stage of disease condition of the lung, but it is not sufficient for diagnosis. For an effective diagnosis, it is essential to extract more clinical information from the chest x-ray. Therefore, this paper presents one such solution to address this problem and introduce a new avenue of classification algorithms concerning chest x-rays. The significant contributions of the proposed study are as follows viz. i) unlike existing classifiers, the proposed classifier can perform comprehensive classification to offer maximum information associated with the test image. When the test image is found to be matched with multiple numbers of images from the trained dataset, more information from the trained images can be used to relate the actual disease condition. ii) one significant contribution of the proposed technique is that it doesn't use any of the conventional classifier found in literature e.g. neural network, deep learning, convolution neural network, fuzzy logic, etc. Unlike these existing algorithms, the proposed algorithm offers non-iterative operations, and hence they are likely to be more practically compatible with existing mobile healthcare devices. iii) The novel introduction to attribute map and map repository offers a highly compact mechanism of converging the information associated with attributes, which is a very good alternative of dimensional reduction concept (e.g., PCA used in existing times) with better accuracy performance. Hence, the applicability of the proposed study is quite more and comprehensive as it addresses some of the practical requirement of classification of a chest x-ray. The limitation of the paper is that it only deals with the classification time while other parameters like accuracy, PSNR and MSE were not considered.

The paper can be considered in future for other medical imaging like brain MRI, Mammogram classification for disease analysis. Further, the paper can be implemented as base paper to analyze other performance parameters like accuracy, PSNR, MSE, etc.

## REFERENCES

[1] R. Suganya, S. Rajaram, A. Sheik Abdullah, Big Data in Medical Image Processing, CRC Press, 2018

[2] Nilanjan Dey, Amira Ashour, Fuquian Shi, Valentina E. Balas, Soft Computing Based Medical Image Analysis, Academic Press, Technology and Engineering, 2018

[3] Terry Des Jardins, George G. Burton, Clinical Manifestations & Assessment of Respiratory Disease - E-Book, Elsevier Health Sciences, 2015

[4] Kaplan, Kaplan Medical, USMLE Step 2 CK Lecture Notes 2018: Internal Medicine, Simon, and Schuster, 2017

[5] U. E. Peter, U. Parampalli, C. I. Uchechi and O. Nnaemeka, "Challenges and prospects of blind spread spectrum medical image watermarking," *2017 IEEE 3rd International Conference on Electro-Technology for National Development (NIGERCON)*, Owerri, 2017, pp. 10-18.

[6] C. Cavaro-Menard, L. Zhang and P. Le Callet, "Diagnostic quality assessment of medical images: Challenges and trends," 2010 2nd European Workshop on Visual Information Processing (EUVIP), Paris, 2010, pp. 277-284.

[7] C. Cavaro-Menard, L. Zhang and P. Le Callet, "Diagnostic quality assessment of medical images: Challenges and trends," 2010 2nd European Workshop on Visual Information Processing (EUVIP), Paris, 2010, pp. 277-284.

[8] G. Niranjana and M. Ponnavaikko, "A Review on Image Processing Methods in Detecting Lung Cancer Using CT Images," 2017 International Conference on Technical Advancements in Computers and Communications (ICTACC), Melmaurvathur, 2017, pp. 18-25.

[9] K. Devaki and V. MuraliBhaskaran, "Study of computed tomography images of the lungs: A survey," 2011 International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, Tamil Nadu, 2011, pp. 837-842.

[10] K. Devaki and V. MuraliBhaskaran, "Study of computed tomography images of the lungs: A survey," 2011 International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, Tamil Nadu, 2011, pp. 837-842.

[11] R. Agarwal, A. Shankhadhar and R. K. Sagar, "Detection of Lung Cancer Using Content Based Medical Image Retrieval," 2015 Fifth International Conference on Advanced Computing & Communication Technologies, Haryana, 2015, pp. 48-52.

[12] X. Li, L. Shen and S. Luo, "A Solitary Feature-Based Lung Nodule Detection Approach for Chest X-Ray Radiographs," in IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 2, pp. 516-524, March 2018.

[13] H. Chung, H. Ko, S. J. Jeon, K. H. Yoon and J. Lee, "Automatic Lung Segmentation With Juxta-Pleural Nodule Identification Using Active Contour Model and Bayesian Approach," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 6, pp. 1-13, 2018.

[14] R. Sarkar and S. T. Acton, "SDL: Saliency-Based Dictionary Learning Framework for Image Similarity," in IEEE Transactions on Image Processing, vol. 27, no. 2, pp. 749-763, Feb. 2018.

[15] A. Zamani, S. A. Rezaeieh and A. M. Abbosh, "Lung cancer detection using frequency-domain microwave imaging," in Electronics Letters, vol. 51, no. 10, pp. 740-741, 5 14 2015.

[16] Awais Mansoor, Ulas Bagci, Brent Foster,"Segmentation and Image Analysis of Abnormal Lungs at CT: Current Approaches, Challenges, and Future Trends", PMC-Radiographics. July-August 2015; 35(4): 1056–1076.

[17] X. Li, L. Shen and S. Luo, "A Solitary Feature-Based Lung Nodule Detection Approach for Chest X-Ray Radiographs," in IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 2, pp. 516-524, March 2018.

[18] QingZeng Song, Lei Zhao, * XingKe Luo, and XueChen Dou, "Using Deep Learning for Classification of Lung Nodules on Computed Tomography Images", PMC-J Healthc Eng. 2017; 2017: 8314740.

[19] Yu.Gordienko, Yu.Kochura, O.Alienin, "Dimensionality Reduction in Deep Learning for Chest X-Ray Analysis of Lung Cancer", arXiv.org, Cornell University Library, 2018

[20] R. Hooda, S. Sofat, S. Kaur, A. Mittal and F. Meriaudeau, "Deep-learning: A potential method for tuberculosis detection using chest radiography," 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuching, 2017, pp. 497-502.

[21] Savitha S.K. and N C Naveen. Article: Study for Assessing the Advancement of Imaging Techniques in Chest Radiographic Images. Communications on Applied Electronics 4(5):22-34, February 2016. Published by Foundation of Computer Science (FCS), NY, USA.

[22] P. R. Katre and A. Thakare, "Detection of lung cancer stages using image processing and data classification techniques," *2017 2nd International Conference for Convergence in Technology (I2CT)*, Mumbai, 2017, pp. 402-404.

[23] J. Mukherjee, A. Chakrabarti, S. H. Shaikh and M. Kar, "Automatic Detection and Classification of Solitary Pulmonary Nodules from Lung CT Images," *2014 Fourth International Conference of Emerging Applications of Information Technology*, Kolkata, 2014, pp. 294-299.

[24] Y. Bar, I. Diamant, L. Wolf, H. Greenspan, "Deep learning with non-medical training used for chest pathology identification", Research Gate, 2015

[25] T. S. Roy, N. Sirohi and A. Patle, "Classification of lung image and nodule detection using fuzzy inference system," *International Conference on Computing, Communication & Automation*, Noida, 2015, pp. 1204-1207.

[26] F. Zhang, Y. Song, W. Cai, Y. Zhou, S. Shan and D. Feng, "Context Curves for Classification of Lung Nodule Images," 2013 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Hobart, TAS, 2013, pp. 1-7.

[27] F. Zhang et al., "A ranking-based lung nodule image classification method using unlabeled image knowledge," 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), Beijing, 2014, pp. 1356-1359.

[28] Z. Xue *et al*., "Chest X-ray Image View Classification," *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, Sao Carlos, 2015, pp. 66-71.

[29] F. Taher, N. Werghi and H. Al-Ahmad, "Rule based classification of sputum images for early lung cancer detection," *2015 IEEE International Conference on Electronics, Circuits, and Systems (ICECS)*, Cairo, 2015, pp. 29-32.

[30] B. U. Dhaware and A. C. Pise, "Lung cancer detection using Bayasein classifier and FCM segmentation," 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune, 2016, pp. 170-174.

[31] A. Kulkarni and A. Panditrao, "Classification of lung cancer stages on CT scan images using image processing," *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, Ramanathapuram, 2014, pp. 1384-1388.

[32] W. S. H. M. Wan Ahmad, W. M. D. Wan Zaki, M. F. Ahmad Fauzi and W. H. Tan, "Classification of Infection and Fluid Regions in Chest X-Ray Images," *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Gold Coast, QLD, 2016, pp. 1-5.

[33] Y. Dong, Y. Pan, J. Zhang and W. Xu, "Learning to Read Chest X-Ray Images from 16000+ Examples Using CNN," *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, Philadelphia, PA, 2017, pp. 51-57.

[34] X. Liu, F. Hou, H. Qin, A. Hao, "Multi-view multi-scale CNNs for lung nodule type classification from CT images", Elsevier, Vol. 77, pp. 262-275, 2018

[35] K. Mao and Z. Deng, "Research Article Lung Nodule Image Classification Based on Local Difference Pattern and Combined Classifier", Hindawi Publishing Corporation, pp. 7, 2016

[36] W. Wong, A. A. Abu-Shareha, M. F. Pasha and R. Mandava, "Enhanced local binary pattern for chest X-ray classification," *2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)*, Shimla, 2013, pp. 271-275.

[37] K. Mao, R. Tang, X. Wang, W. Zhang, and H. Wu, "Research Article Feature Representation Using Deep Autoencoder for Lung Nodule Image Classification", Hindawi, pp. 11, 2018

[38] D. Kumar, A. Wong and D. A. Clausi, "Lung Nodule Classification Using Deep Features in CT Images," *2015 12th Conference on Computer and Robot Vision*, Halifax, NS, 2015, pp. 133-138.

[39] W. Li, P. Cao, D. Zhao, and J. Wang, "Research Article Pulmonary Nodule Classification with Deep Convolutional Neural Networks on Computed Tomography Images", Hindawi Publishing Corporation, pp. 7, 2016

[40] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri and R. M. Summers, "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 3462-3471.

[41] R. N. R, A. Susanto, I. Soesanti and Maesadji, "Thoracic X-ray features extraction using thresholding-based ROI template and PCA-based features selection for lung TB classification purposes," 2013 3rd International Conference on Instrumentation, Communications, Information Technology and Biomedical Engineering (ICICI-BME), Bandung, 2013, pp. 65-69.

[42] S. Widodo, R. N. Rohmah and B. Handaga, "Classification of lung nodules and arteries in computed tomography scan image using principle component analysis," 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, 2017, pp. 153-158.

[43] S. Widodo, R. N. Rohmah and B. Handaga, "Classification of lung nodules and arteries in computed tomography scan image using principle component analysis," 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, 2017, pp. 153-158.

[44] X. Li, S. Luo, Q. Hu, J. Li and D. Wang, "Rib suppression in chest radiographs for lung nodule enhancement," 2015 IEEE International Conference on Information and Automation, Lijiang, 2015, pp. 50-55.

[45] J. Bai, X. Huang, S. Liu, Q. Song and R. Bhagalia, "Learning orientation invariant contextual features for nodule detection in lung CT scans," 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), New York, NY, 2015, pp. 1135-1138.

[46] "ELCAP Public Lung Image Database", http://www.via.cornell.edu/lungdb.html, Retrived on 24-September-2018

[47] "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning", https://stanfordmlgroup.github.io/projects/chexnet/, Retrived on 24-September-2018

[48] "gregwchase/nih-chest-xray", https://github.com/gregwchase/nih-chest-xray, Retrived on 24-September-2018

[49] "COPD Machine Learning Datasets", http://bigr.nl/research/projects/copd, Retrived on 24-September-2018

[50] Savitha S. K and N. C. Naveen, "Algorithm for pre-processing chest-x-ray using multi-level enhancement operation," 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, 2016, pp. 2182-2186

[51] "Digital Image Database", Digital Image Database, Retrived on 24-September-2018

[52] "Chronic obstructive pulmonary disease (COPD) Dataset", https://healthcaregateway.co.uk/chronic-obstructive-pulmonary-disease-copd-dataset/, Retrived on 24-September-2018

[53] "Random Sample of NIH Chest X-ray Dataset", https://www.kaggle.com/nih-chest-xrays/sample, Retrived on 24-September-2018

[54] "Tuberculosis Chest X-Ray Image Data Sets", https://ceb.nlm.nih.gov/repositories/tuberculosis-chest-x-ray-image-data-sets/, Retrived on 24-September-2018

[55] "Cancer Data Access System", https://biometry.nci.nih.gov/cdas/datasets/plco/21/, Retrived on 24-September-2018

# Analysis of End-to-End Packet Delay for Internet of Things in Wireless Communications

Imane Maslouhi[1]
Dept: Electronics, Informatics and Telecommunications
ENSAO, Mohammed I University,
Oujda, Morocco

El Miloud Ar –reyouchi[2]
Dept. Telecommunications and Computer science
Abdelmalek Essaadi University and SNRT, Tetouan, Morocco

Kamal Ghoumid[3], Kaoutar Baibai[4]
Dept. Electronics, Informatics and Telecommunications, ENSAO,
Mohammed I University
Oujda, Morocco

***Abstract*—Accurate and efficient estimators for End to End delay (E2EPD) plays a significant and critical role in Quality of Service (QoS) provisioning in Internet of Things (IoT) wireless communications. The purpose of this paper, on one hand, is to propose a novel real-time evaluation metrics, on the other hand, addresses the effects of varying packet payload (PP) size. These two objectives rely on the analysis of E2EPD for QoS provisioning in multi-hop wireless IoT networks through multiple hops count from source to destination. The results of this study show the critical effect of PP size, hops count and interface speed on the improving E2EPD use of applications requiring real-time IoT communications.**

***Keywords—End to end delay; internet of things; multi hop; wireless communication***

## I. INTRODUCTION

End-to-End delay is the time taken by a packet to travel from source to destination [1] , [2] ,[3]. It is an important design and performance characteristic of IoT wireless communications networks. It is especially important for delay-sensitive applications and for which need transmitting packet data with average delay constraints [4]. E2ED is a common term in IP network monitoring and differs from Round-Trip Time (RTT) [5].

The large turnout of real-time communication to the IoT gives much importance to improve E2ED. Reducing delay metric is exposed in different contexts such as access delay in [6].Delay improvement for the remote management of renewable energy using a random NC is also evaluated according to [7]. The evaluations of average E2ED and jitter in wireless tele-ultrasonography medical systems has been carried in [8].

E2E delay depends on number of hops in the path, congestion on the network and it is affected by various parameters as interface speed in intermediate nodes [9].

In IoT, an excessive E2E delay can significantly affect throughput. Higher delays could result in rejecting the packets by routers due to breaching the limit of Time to Live (TTL), then Internet Control Message Protocol (ICMP) packets are sent to the source and hence results in re-transmissions. E2E delay is also infected directly by the retransmission timeout (RTO) as [10]. IoT can also strongly secure the intelligent networks platforms which is studied in [11].

The E2EPD is especially important for delay-sensitive applications. Packets are delivered to destination nodes with delays, which may vary from packet to packet (one measure is jitter). E2EPD distribution calculation method it is an analytical model to calculate E2E delays in packet networks according to [12]. Recently, simulation results demonstrate the accuracy and effectiveness of analytical E2EPD modeling for achieving delay aware as in [13].

TCP (Transmission Control Protocol) is more widely used protocol on the Internet because of their errors correction. UDP (User Datagram Protocol) is another more frequently used protocol on the Internet. However, UDP is never used to transmit valuable data such as database information, webpages, etc. UDP is commonly used for streaming audio and video. Therefore, UDP is characterized by high-speed data communication. In IoT, protocol specification allows interoperability among things with different communication standards as CTP (Communication Things Protocol) according to [14]. Other study gives an overview of some technical details that pertain to the IoT enabling technologies applications as [15].

In the Remote Management field popularity of IoT is increasing day by day in the area of remote monitoring system as in [16]. The remote monitoring systems include, remote satellite monitoring, DVB stream management, data acquisition in remote areas, energy grid monitoring etc., In IoT network management, the real-world objects communicate with each other using source-destination which source and destination can be Supervisory Control and Data Acquisition (SCADA) and remote terminal unit (RTU) respectively.

The efficiency of applying the delay distribution from a single node and using convolution to find the E2E delay is given in [17]. The E2ED distribution in a linear network is derived for homogeneous networks as in [18].

This paper gives a simplistic overview of the role that can play the payload length (based on some basic parameters) for improving the E2E delay in IoT network performance.

The rest of this paper is structured as follows. Section II discusses the general conception of payload transmission.

Section III gives a general description of end-to-end packet delay in wireless multi-hop network, and its mathematical model is described . The experimental and simulation setup are given in Section IV. Section V lists results and discussion .In Section V, the authors provide the conclusions.

## II. General Concept

### A. Transmission Average Message Size

The basic structure of a packet varies between protocols, a typical packet includes two sections a header and payload. In this paper, the authors focus on transmitting the average message which vary in length from o to 1500 bytes, and their headers vary in complexity from five to 50 bytes. The E2ED in IoT network can be strongly dependent of the message size.

Data transmission over an E2E (source to destination) communication channel is being carried within a packet does not integrate the overhead data. These real data are referred as the payload.

For a communications layer that requires some of the overhead data to do its operation, the payload is sometimes designed to include the part of the overhead data. However, in this operational network,the payload is the bits that get delivered to the source (SCADA) at the destination( RTUn).

The most important factors that directly influence E2E delay are the arrival rate, the service rate, the number (count) of hops in a path, Ethernet interface speed and the baud rate [kbps] for the serial interface (COM).

When data is sent over network and Internet connections, each IoT nodes sent incorporate both header information and the real information being sent. The header contains various things depends on the used protocol , it can detects the source and destination of the packet, while the real information is assigned to the payload. Header information is applicated singly in the transmission process, it stripped off from the packet when it has just arrived at its destination. Therefore, the payload is the only data collected by the destination IoT node. In this paper the transmission message size is compared and analyzed.

### B. Data Transmission Speed

The results considered different value of the payload which is the data itself it needs to transfer (usually the user message size without any headers (IP, TCP, UDP,).

When crossing more than one device in IoT network, interface speed has always played a primordial role for improving E2E delay. It is necessary to configure the interface speed of IoT devices. In this paper it is referred to as the baud rate [bps] for the serial (COM) interface or an Ethernet interface speed.

In typical serial interface communication systems, the available bit rate values are: 2400 kbps, 4800 kbps, 9600 kbps, 19200 kbps, 38400 kbps and 115200kbps.

When using TCP instead of UDP lowers the total network capacity due to the higher TCP overhead (ETH – UDP/IP and serial options are equal. The performance evaluation of E2E delay of Randomized TCP is presented in [19].

UDP sends the packets which contains just simple things in the header as source IP/PORT and destination . TCP, on the other end contains some interesting information , namely the sequence number of the packet( to guarantee ordered delivery), a lot of flags (to guarantee the packet actually received in it's destination) and checksum of the data (to ensure it didn't get corrupted) and received correctly in its destination.

The protocols TCP and UDP are used for transmitting bits of data over the Internet. They can build on top of the IP protocol.

The packet sent directly via TCP or UDP are processed similarly, as they're forwarded from a source intermediary IoT nodes and to the destination

## III. A Theoretical Study of End-to-End Packet Delay

The E2E delay is typically measured in multiples or fractions of seconds, in that only path in the one direction from source to destination is measured . It is very interesting to specify how long it takes for a user data size without any headers to travel across the network from source to destination.

### A. End to End Delay

In telecommunication and data networks, the end-to-end packet delay $D$ usually consists of following four elements [20]:

Transmission delay $D^{(t)}$, Radio propagation delay $D^{(r)}$, Signal processing delay $D^{(s)}$ and Queueing delay $D^{(q)}$.

Their mathematical relation can be simply expressed as

$$D = D^{(t)} + D^{(r)} + D^{(s)} + D^{(q)} \quad (1)$$

The E2E delay equation above describes the packet delay at a single IoT node along its path from source to destination.

Assuming that the radio transmission delay $D^{(t)}$ and signal processing $D^{(s)}$ delay are small enough to neglect, (1) becomes

$$D = D^{(t)} + D^{(q)} \quad (2)$$

In the IoT connections, if $n$ is the number of IoT nodes in the connected path from source to destination, the number of links is $n+1$, the end to end delay is

$$D_n = (n+1)(D_n^{(t)} + D_n^{(q)}) \quad (3)$$

Where $n$ are the nodes can be considered for measuring the end-to-end delay.

### B. Network Model and Performance

Consider the problem of analyzing the E2E delay over the paths from a node source $S$ to a node destination $T$ in a complete IoT network. The following Fig.1 shows an Internet of Things Network model.

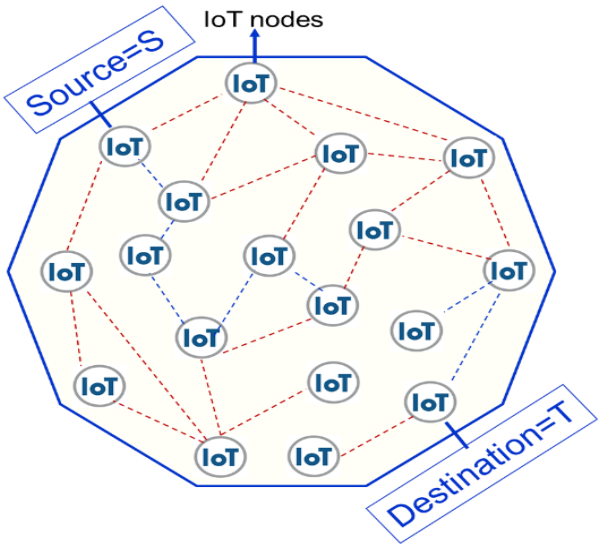



Fig. 1. Internet of Things Network model.

Assume that the signal processing delay $D_n^{(s)}$ , in the IoT node, and Radio propagation delay $D_n^{(r)}$ are equal to 0.

Suppose Source $S$ begins to transmit the first packet at $D_{(t=0)} = D_0$ delay over a path with $n$ IoT nodes. This packet is received by node $T$ at $D_0 + D_n^{(t)}$ , where $D_n^{(t)}$ is the transmission delay of a packet.

The packet is re-transmitted to the destination S across $n$ IoT nodes, while he next packet is transmitted on a same path.

With $n$ IoT nodes, the first packet $P_1$ reaches destination node $T$ at:

$$(n+1)D_n(t) \tag{4}$$

If $p$ is the number of packets, the last packet sent to:

$$(p-1)D_n(t) \tag{5}$$

The last packet arrives at (which corresponds to the end of the transfer):

$$(p-1)D_n + (n+1)D_n \tag{6}$$

Either again,

$$D_n(t)(p+N) \tag{7}$$

By posing $D_n(t) = L / pR$ , where $p$ and $R$ are the number of packets and data rate respectively, the crossing delay of the network $D_n$ is obtained as :

$$D_n = (L / pR)(p+N) \tag{8}$$

Or

$$D_n = (L / R)(1+N / p) \tag{9}$$

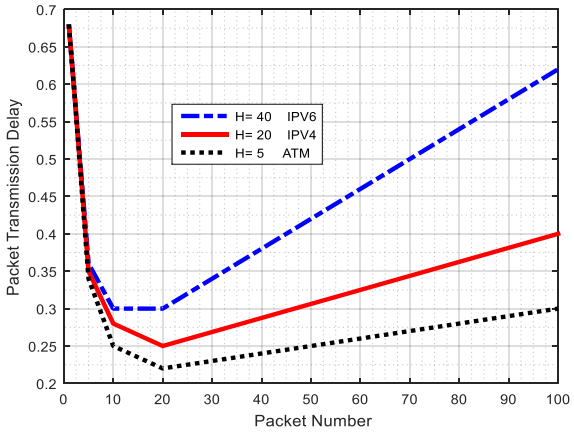



Fig. 2. Evolution of Packet Transmission Delay vs Packet Number.

There however, this formula does not consider the protocol data (H), Which should be added to each packet, hence:

$$T_p = (\frac{L+pH}{R})(1+\frac{N}{p}) \tag{10}$$

The curves given in Fig.1graphically illustrates the theoretical result with the following hypothesis: L = 1500 bytes, N = 5

The packet transmission delay increases significantly versus of packet number. The values are expressed depending on the message switching ( $p=1$ and $n \geq 0$ ) that is a network switching technique in which data is routed in its entirety from the source $S$ to the destination $T$ .

*If* $p=1$ , $n \succ 0$ *Message Switching*

*if* $p=1$ , $n=0$ *Circuit Switching*

Fig.2 compares the performance according to a header: ATM (5 bytes), IPV4 (20 bytes), and IPV6 (40 bytes).

The transmission delay, in the network, is even lower than is small. This leads to search for routes that minimize the number of nodes crossed (routing algorithms) and to increase the network (increase the probability of finding a more direct route).

It should be noted that the influence of the service header size is not negligible.

## IV. Experimental and Simulation Setup

In this practical application, the topology showed in Fig. 1 is well-respected. This paper studies the E2ED of an IoT wireless network, the system is configured as a single source $S$ node sends packets to single destination $T$ node across several IoT nodes.

The IoT Network system consists of two parts, related hardware and management software. The system hardware is divided into on source node, wireless transmission IoT nodes and destination node. The software adopts a centralized control management model, providing users basic information management for real-time monitoring,

TABLE I. SIMULATION PARAMETERS

| *Critical parameters* | *Value* |
|---|---|
| Packet Payload [bytes] | 0-1500 |
| Processing time[ms] | 20 |
| Interface speed [kbps] | 2400,38400,115200 |
| Hops | 1-9 |
| ACK | off |

The SunSet E20c is a device used to measure E2E delay which provides a full transmission testing according to [21]. It can also verify Datacom circuits by monitoring the received information, control leads, and physical layer results.

The simulation results were developed using Matlab software.

Data transfer rate and interface speed of each IoT nodes are same as IoT nodes wireless communication module. If more IoT nodes are used in E2E delay path it will increase and it performance can improved by reducing the packets size as in [22].

Table I gives some simulation parameters used for the analysis of E2E listing conditions selected.

## V. RESULTS AND DISCUSSION

This section describes the experiments and simulations results. The presented results illustrate how E2E delay (between *S* and *T* ) varies depending on the PP for several hops and UMS (100 Bytes and 1500 Bytes) of wireless edges in wireless communication for IoT, then compare the simulation result between Ethernet TCP/IP and UDP/IP as interface speed using fixed values of PP of multi-hop wireless IoT networks.

Fig.3 reveals the E2E delay measurements as a function of hops count per static path for different sizes of data without any headers. It also shows the multi hop transmission from a source to destination through IoT nodes.

TABLE II. END TO END DELAY RESULT OF MULTI-HOP PATHS

| *PP bytes* | *Hops* | | | |
|---|---|---|---|---|
| | 1 | 3 | 6 | 9 |
| 100 | 66 | 178.5 | 347.5 | 516 |
| 800 | 106.5 | 299.5 | 589 | 879 |
| 1500 | 147 | 420 | 831 | 1241 |

For tree, six and nine hops transmission, the average E2ED is around 420 ms, 831ms, and 1241 ms for the maximum payload offered (PP=1500 bytes) respectively, while 178 ms ,347 ms, and 516 ms respectively for PP=100 bytes.

It is observed that the respective E2ED were linearly increased with increasing hops count or packet payload.

Table II below summarizes some of the simulation results.

Adding one more hop in transmission path increases the E2E delay by 53.878%, 52.463% and 51.852% for 100 bytes, 800 bytes and 1500 bytes respectively. Consequently, each of those hops introduces some delay according to a payload size.

Fig.4 illustrates the results of indirect transmissions through IoT nodes over 9 hops for different packet payload size and give details of how E2E delay varies in terms of different interface speed (2400kbps up to 115200 kbps) of nine hops.

Results indicated in Fig.4 interface speed has a direct and significant effects on E2ED.

The PP=0 bytes in Fig .4 means that data size is zero because the packet is only acknowledging data; it is not transmitting any data. Packets with an ACK flag and 0 size can be TCP keep alive packets. There are other circumstances in which a system will send TCP packets with zero length.

The Table 3 below represents some simulations results of E2E delay using the baud rate [kbps] for the serial (COM) interface (three different payload sizes for a path with 9 hops).

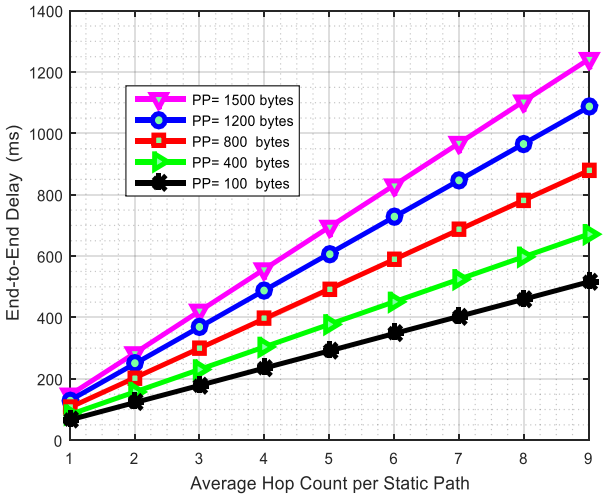



Fig. 3. End to End Delay vs Hop Count Per Static Path.

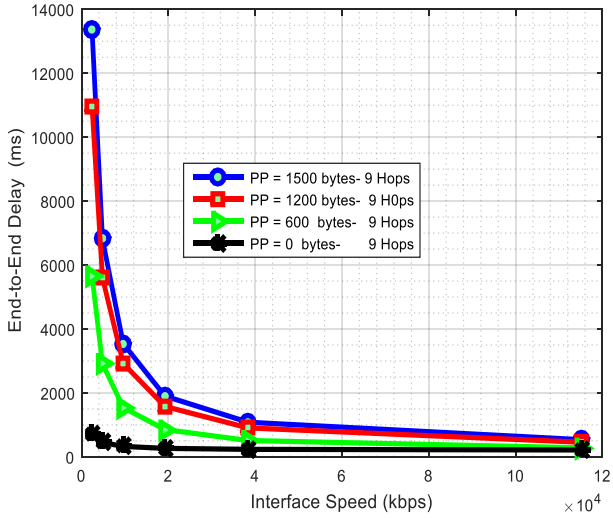



Fig. 4. End to End Delay vs Interface Speed.

TABLE III. E2E DELAY RESULTS OF CRITICAL INTERFACE SPEED

| *PP bytes* | *Interface speed* | | |
|---|---|---|---|
| | 2400 | 38400 | 115200 |
| 0 | 347 | 117 | 107 |
| 600 | 5658 | 511 | 282 |
| 1500 | 13370 | 1082 | 535 |

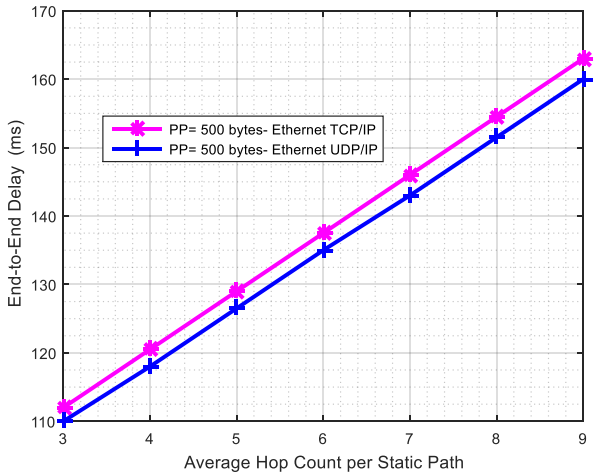



Fig. 5. End to End Delay vs Hop Count Per Static Path using ETH TCP/IP and ETH UDP/IP.

From this table it can be seen that the more the interface speed increase, the more there is improvement of E2E delay.

Fig. 5 shows experimental results of the comparison between Ethernet TCP/IP and UDP/IP as interface speed results for PP=500 bytes.

Fig.5 illustrates that the UDP is speedily than TCP. The reason is because there is no form of flow control or error correction or its absent acknowledge packet (ACK) that allows a continuous packet flow, instead of TCP that acknowledges a determined packets.

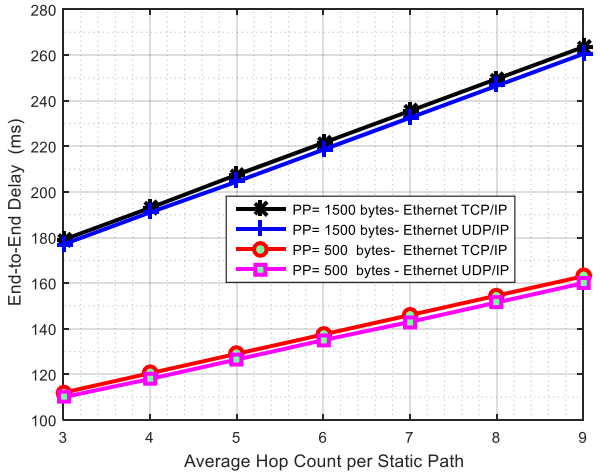



Fig. 6. Comparison End to End Delay vs Hop Count Per Static Path (PP=500 Bytes and 1500bytes).

When using UDP, packets are just sent to the recipient. The sender continues transmitting the next packets (without waiting that the recipient received of the previous packet) If the IoT destination misses a few UDP packets, they are lost. The sender will not resend them. Losing all this overhead means the IoT nodes can communicate more speedily.

Fig. 6 shows simulation result of the comparison between Ethernet TCP/IP and UDP/IP as interface speed using two values of PP (500 bytes and 1500 bytes).

Table 4 lists the comparison of the average delay from simulation and estimation results using Ethernet interface speed.

It can also be further concluded that the values developed performs well under different conditions.

TCP/IP is a suite of protocols used by IoT nodes to communicate over the Internet. UDP/IP is used by applications to deliver a speedily stream of information by doing away with error-checking.

In this paper, the analysis of E2E packet delay for internet of things in wireless communications was developed and illustrated by both experimental and simulations results. This analysis, compared to the analysis results, gives more a simplistic and quickly overview for improving the E2E delay in IoT network performance as in [17].

Consider the links from node 1 to node 10, keeping the same assumptions of [12]. For 2 Mbit/s links and constant packet lengths of 400 bytes, the comparison between the E2EPD distribution calculation method given by [12] and the proposed method clearly shows the efficiency and accuracy of the results of E2EPD obtained by the proposed method as in Fig.7.

TABLE IV. COMPARISON OF E2E DELAY USING TCP/IP AND UDP/IP

| *PP (bytes)* | *4 Hops* | | *8 Hops* | |
|---|---|---|---|---|
| | TCP/IP | UDP/IP | TCP/IP | UDP/IP |
| 500 | 120 | 118 | 154 | 151 |
| 1500 | 193 | 191 | 249 | 246 |

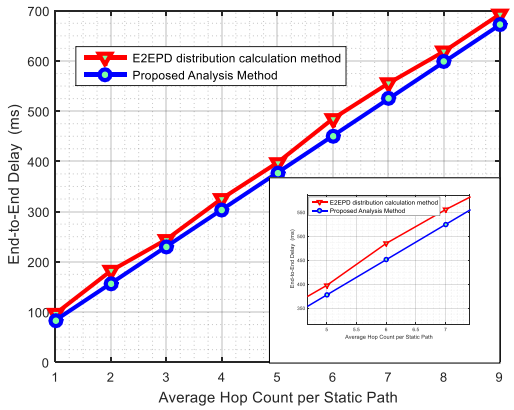



Fig. 7. Comparison between the E2EPD Distribution Calculation Method and the Proposed Method.(PP=400 bytes ).

It was also observed (in Fig.7 with zoomed curves) that the respective E2EPD distribution calculation method does not increase linearly with increasing hops count or packet payload while the E2EPD of the proposed method increase in straight line. The simple and the efficient proposed method for analyzing E2ED gives a very accurate and robust results.

## VI. Conclusions

The authors have performed both an experimental and theoretical analysis of the End to End delay which is influenced by critical parameter. The simulation, measurement and estimation results were shown the impacts of payload size, hops count and interface speed on E2E delay performance. The results showed that this estimator provides good estimates of payload packets, End to end delay, and jitter gave a key insight into the QoS provisioning for multi-hop wireless networks. Ethernet UDP/IP is exploited when a speed is requested and error correction is not needful . The setting parameters discussed in this paper gives a rapid and easy idea of the E2E delay improvement in the IoT network. The E2EPD for narrowband Internet of Things in Wireless Communications nodes should be analyzed and minimized finding new techniques and methods which will be the future works.

## Acknowledgment

This work was partially supported by the SNRT (laboratory of regional service) Morocco.

## References

[1] S.-N. Yeung and J. Lehoczky, “End-to-end delay analysis for real-time networks,” in Proc. IEEE RTSS, London, U.K., Dec. 2001, pp.299–309

[2] El.M..AR-REYOUCHII, K. Ghoumid, K. Ameziane, and O. El Mrabet, “The Improvement of End to End Delays in Network Management System Using Network Coding,” International Journal of Computer Networks & Communications (IJCNC),vol.5,n 6 (Nov. 2013), :65–84. DOI: 10.5121/ijcnc.2013.5604.

[3] Nasim Abbas; Fengqi Yu ,”Performance analysis of end-to-end delay and reliability for linear wireless multimedia sensor networks ,” IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC), 3-5 Oct. 2017 2017

[4] R. A. Berry and R. Gallager, “Communication over fading channels with delay constraints,” IEEE Trans. Inform. Theory, vol. 48, May 2002, pp. 1135–1148.

[5] El. .Ar-Reyouchi, K. Ghoumid, K.Ameziane, O. El Mrabet,“Performance Analysis of Round Trip Time in Narrowband RF Networks For Remote Wireless Communications,” International Journal of Computer Science & Information Technology (IJCSIT),vol.5,n 5,October 2013,pp.1–20. DOI: 10.5121/ijcsit.2013.5501

[6] L.Zhao ; W. Sun ; Y. Shi ; J. Liu,” Optimal Placement of Cloudlets for Access Delay Minimization in SDN-Based Internet of Things Networks”, IEEE Internet of Things Journal,vol.5,n..2(2018), 1334 - 1344 DOI: 10.1109/JIOT.2018.2811808

[7] El .Ar-reyouchi,M.Hammouti,I. Maslouhi, K.Ghoumid “The internet of things: Network delay improvement using network coding,” ACM ,ICC ,Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing, Cambridge, United Kingdom, ,March 22 - 23, 2017 . DOI: 10.1145/3018896.3018902

[8] Y. Chen, N. Philip, R. Istepanian, and I. Darwazeh, “End-to-end delay distributions in wireless tele-ultrasonography medical systems,” in Global Telecommunications Conference (GLOBECOM 2013), 2013 IEEE, 2013.

[9] El. Miloud Ar Reyouchi, K.Ghoumid, K. Ameziane, O. El Mrabet, S.Mekaoui,” Performance Analysis of Round Trip Delay Time in Practical Wireless Network for Telemanagement,” World Academy of Science, Engineering and Technology International Journal of Electronics and Communication Engineering,vol7, n.11( 2013). 1413–1419. scholar.waset.org/1999.5/17295.

[10] E.G. Davis, A. Calveras and I. Demirkol, "Improving packet delivery performance of publish/subscribe protocols in wireless sensor networks,” Sensors, 2013,vol. 13, no. 1, pp. 648-680.

[11] Maslouhi I., Ghoumid K., Zaidouni J., Benzekri W., Ar-reyouchi E.M., Lichioui A. (2016) ,”Network Higher Security of Intelligent Networks Platforms in Telecommunications Areas,”. In: Proceedings of the Mediterranean Conference on Information & Communication Technologies 2015. Lecture Notes in Electrical Engineering, vol 380. Springer.

[12] O. Osterbo, “Models for calculating end-to-end delay in packet networks,” ITC-18, Berlin, Germany, August 31-September 5, 2003, pp. 1231-1240.

[13] Qiang Ye,; Weihua Zhuang , Xu Li , Jaya Rao, “End-to-End Delay Modeling for Embedded VNF Chains in 5G Core Networks,” IEEE Internet of Things Journal,2018 ,DOI: 10.1109/JIOT.2018.2853708.

[14] A.Asensio,A. Marco, R. Blasco and R. Casas, "Protocol and Architecture to Bring Things into Internet of Things,” International Journal of Distributed Sensor Networks, 2014.

[15] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, “Internet of things: A survey on enabling technologies, protocols, and applications,” IEEE Communications Surveys Tutorials, vol. 17, no. 4, pp. 2347–2376, 2015.

[16] El Miloud Ar Reyouchi, “Optimisation des performances des réseaux de communications sans fil : Performances des réseaux de communications sans fil pour la télégestion des stations de la Télédiffusion TV/FM ,” Paf, ,2017, ISBN 978-3-8416-4286-8.

[17] R. Stewart,“End-to-end Delay analysis for Small/Medium Scale IP networks,” PhD Thesis, University of London 2003.

[18] M. Xi and M. Haenggi, “Towards an end-to-end delay analysis of wireless multihop networks,” Ad Hoc Netw., vol. 7, no. 5, pp. 849–861, Jul. 2009

[19] M.Shorfuzzaman, M.Masud and Md. M. Rahman,"Characterizing End-to-End Delay Performance of Randomized TCP Using an Analytical Model,” International Journal of Advanced Computer Science and Applications(IJACSA), 2016. Vol.7,n.3, 2016.

[20] D. Bertsekas and R. Gallager, Data Networks. Prentice-Hall, 1987.

[21] Y.Chatei,M.Hammouti, .E.M. Ar Reyouchi, and K. Ghoumid,“Downlink and Uplink Message Size Impact on Round Trip Time Metric in Multi-Hop Wireless Mesh Networks,”vol.8, 2017/03/01-DOi - 10.14569/IJACSA.2017.080332.

[22] G. Hasegawa, T. Tomioka, K. Tada, and M. Murata, “Simulation studies on router buffer sizing for short-lived and pacing TCP flows,,”Computer Communications, vol. 31, no. 16, pp. 3789 – 3798, 2008.

# The Generation of a Stable Walking Trajectory of a Biped Robot based on the COG based-Gait Pattern and ZMP Constraint

Arbia Ayari[1], Jilani Knani[2]
LA.R.A, in Automatic Control, National Engineers School of Tunis
BP 37, Le Belvédère, 1002.
ENIT, University of Tunis El Manar
Tunis, Tunisia.

***Abstract*—The research works contained in this paper are focused on the generation of a stable walking pattern of a biped robot and the study of its dynamic equilibrium while controlling the two following criteria; the centre of gravity COG and the zero-moment point ZMP. The stability was controlled where the biped have to avoid collision with obstacle. The kinematic constraints were also taken into consideration during the walking of the biped robot. In fact, the generation of the walking patterns is composed of several stages. First, we used the Kajita method for the generation of the COG trajectory, based on the linear inverted pendulum LIPM during the simple support phase SSP and linear pendulum model LPM during double support phase DSP. After that, we used two $4^{th}$spline function to generate the swing foot trajectory during the SSP and we used exact formulate for the foot trajectory during DSP. Finally, Newton's algorithm was performed (at the level of the inverse geometric model), in order to calculate the different joints according to the desired trajectories of the hip and the feet. Ground reaction forces were also determined from the dynamic model to satisfy the kinematic constraints on both feet of the biped. The generation of walking is done for two different speeds. To study the biped balance, ZMP generation algorithm was performed during the different walking phases and the results obtained for the two cases were compared.**

***Keywords*—Biped robot; COG; ZMP; stability; LIPM; LPM; walking gait**

## I. INTRODUCTION

Research in the field of humanoid robotics has generally focused on how to reproduce the stable human walking as faithfully as possible ([1], [2]). Walking a biped robot at a certain speed and under certain constraints (roughness and inclination of the ground, carried additional weight, avoid collision with obstacle and performing some tasks) while keeping the stability of the biped is a challenge for researchers. On the work of [3] a novel method of Zero-Moment-Point (ZMP) compensation is proposed to improve the stability of locomotion of a biped, which is subjected to disturbances(carried an additional weight while walking and it walked up a 10◦ slope and walked down a 3◦ slope).The effectiveness of the method is verified on a humanoid robot, MANUS-I.

On the work of [4], a motion pattern generator of humanoid robots that walks on a flat plane, steps and a rough terrain was proposed. If the biped robot walks on a terrain other than a flat plane and/or if it exist a contact between the hand of the robot and its environment, some approximations should be introduced to judge the contact stability in association with the ZMP. In this paper, other conditions have been added to guarantee the stability of the robot.

Compared to other types of walking robots, bipedal walking is a much less stable activity than say four-legged walking, as multilegged robots have more footholds for support. Bipedal walking allows instead greater maneuverability, especially in smaller areas [3].

During the displacement of the biped robot, there are two types of stability characterizing the way of walking the biped; static stability and dynamic stability. The static stability is characterized by the fact that the walking robot always keeps its COG projected vertically inside the support polygon. Unlike, at the dynamic stability, the robot's COG leaves the support polygon for some periods of time and the robot is in imbalance during the step, when raising one of its legs [5]. The problem of imbalance of the biped is posed generally at the level of the SSP. However, such imbalance periods of time must be short and the ZMP constraint must be controlled in order to avoid a possible fall of the robot. Previous research works has found a relationship between COG and ZMP to study the stability of walking robots. Several references have worked on the generation of stable walking of a two-legged walking robot based on the generation of the trajectory of the COG and controlling the dynamic constraints ZMP ([6], [7]).

The generation of the walking cycle for walking robots can be classified according to three main approaches: Model-based gait, Biological mechanisms-based gait and Natural dynamics-based gait [8].Our work is interested in the first method and more precisely on the approach of Centre of gravity based gait.

The structure of the paper is as follows: Section 2 describe the biped gait cycle. Section 3 represents static and dynamic constraints. Section 4 includes the different steps of generating

the robot's walking cycle. Section 5 describes the inverse dynamic modelling. Simulation results of stability control have been shown in Section 6.The conclusion is given in Section 7.

## II. Biped Robot Gait Cycle

The reproduction of a walking cycle for a walking robot is based on the generation of a step sequence. Each step is composed of two main successive phases SSP and DSP and an intermediate phase of impact. A race cycle, meanwhile, is a succession of SSP, flight, and impact phase. Depending on the duration of a step, SSP occupies the longest duration of time relative to the DSP (about 80% for the SSP versus 20% for the DSP).

Thanks to the complexity of the reproduction of a human walking cycle, most researchers simplify the gait cycle and studies have shown four essential configurations to generate a complete walking cycle: the first pattern contains one SSP and one DSP, the second one contain a consecutive phase of SSP, the third one contain one SSP and two sub-phases of DSP and the last pattern composed of two sub-phases of SSP and two sub-phases of DSP. Previous research has shown the importance of integrating a dual support phase into the biped cycle. The work of [9] compared the results obtained on the speed of the center of gravity and the trajectory of the ZMP for a walk with DSP and without it. He showed that for a cycle that contains only SSP, the discontinuity in COG velocity when the biped switches their feet and it becomes more unstable when it faces impulsive forces. The presence of the DSP ensures the continuity of the COG speed and improves the stability of the biped during the switching phase of two feet.

## III. Static and Dynamic Constraint

### A. Unilateral Contact Constraint

During walking, the robot must maintain contact with the ground through the support points (one foot in single support phase and two feet in double support phase). Indeed, these points can neither penetrate on the ground nor leave it, and this can be expressed in the form of holonomic constraints as follows:

$$C_h(\mathrm{x}) = 0 \tag{1}$$

With $C_h(\mathrm{x})$ represent the vector of the position constraints on the support foot and,

$$C_h(\mathrm{x}) = \begin{pmatrix} x_{cs} \\ z_{cs} \\ q_0 \end{pmatrix} \tag{2}$$

With $x_{cs}$ and $z_{cs}$ represent the direct geometric model MGD of the support foot ankle and $q_0$ represent the support foot joint angle.

$$x_{cs} = x_H + l_2 * \sin q_2 + l_1 * \sin q_1 \tag{3}$$

$$z_{cs} = x_H - l_2 * \cos q_2 - l_1 * \cos q_1 \tag{4}$$

With $x_H$ represent the position of the hip, $q_1$ and $q_2$ represent the two joint angles of the support leg and $l_1$ and $l_2$ represent respectively the length of the thigh and the tibia.

The first derivation of equation (1) allows us to determine the speed constraint represented by the following equation:

$$J_1(\mathrm{x}) * \dot{X} = 0 \tag{5}$$

With J(x) represent the Jacobian matrix and $\dot{X}$ represent the first derivation of the state vector.

$$\dot{X} = (\, \dot{q}_0 \quad \dot{q}_1 \quad \dot{q}_2 \quad \dot{q}_3 \quad \dot{q}_4 \quad \dot{q}_5 \quad \dot{q}_6 \quad \dot{x}_H \quad \dot{z}_H \,)^{\mathrm{T}} \tag{6}$$

$$J_1 = \begin{pmatrix} 0 & l_1 * \cos q_1 & l_2 * \cos q_2 & 0 & 0 & 0 & 1 & 0 \\ 0 & l_1 * \sin q_1 & l_2 * \sin q_2 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{7}$$

The same calculation procedures were applied to determine the kinematic constraint on the swing foot during the DSP. The expression of the Jacobian matrix is the following:

$$J_2 = \begin{pmatrix} 0 & 0 & 0 & l_2\,\mathrm{C}(q_3) & l_1\,\mathrm{C}(q_4) & h_p\,\mathrm{C}(q_5) + (l_p - L_p)\,\mathrm{S}(q_5) & 0 & 1 & 0 \\ 0 & 0 & 0 & l_2\,\mathrm{S}(q_3) & l_1\,\mathrm{S}(q_4) & h_p\,\mathrm{S}(q_5) + (L_p - l_p)\,\mathrm{C}(q_5) & 0 & 0 & 1 \end{pmatrix} \tag{8}$$

Where $q_5$ represent the swing foot joint angle, $q_3$ and $q_4$ represent the two joint angle of the swing leg, $h_p$ and $L_p$ represent respectively the high and the length of each foot.

**Note**: the two capital letters C and S respectively denote the two operators cosinus and sinus. We used these designations to reduce the size of the matrix.

### B. The Non Slip Constraint

If we consider the robot's leg and the ground as being two solid bodies in contact then according to the principle of Amontons-Coulomb, There is no slip between these two bodies if the modulus of the tangential component is increased by a term proportional to the modulus of the normal component. This can be expressed by:

$$|\lambda_T| \leq \mu_0 |\lambda_N| \tag{9}$$

With $\lambda_T$ and $\lambda_N$ denote the Lagrange multipliers ([10], [11]) for ground contact forces (normal and tangential, respectively). $\mu_0$ is called friction coefficient of adhesion. It depends essentially on the nature of the materials in contact.

### C. ZMP Constraint

The ZMP is defined as a point on the ground where the resultant ground reaction produces a zero moment along the x (front-back) and z (transverse) axes [12].

The ZMP is a dynamic balance criterion of bipedal robots and has been widely used in controlling the balance of biped robots: the main task of the control is to maintain the ZMP inside the support polygon to prevent the foot to switch. It is a defined criterion for checking the assumptions of contact with the ground, and thus managing the balance. It has many advantages: easy to understand and implement in a real robot, and taking into account the dynamics [13].

The main objective of our work is to generate a balanced walking of the biped robot following a well-defined trajectory under the constraint of avoiding an obstacle. It is for this reason that the trajectory of the ZMP must not leave the polygon of sustenance to guarantee the stability of our robot.

According to the dimensions of the feet and the length of the step realized by the biped, during the running cycle, the imposed conditions on the trajectory of the ZMP are the following:

$-L_p \leq X_{zmp} \leq P$ for SSP1

$0 \leq X_{zmp} \leq P$ for DSP

$L_p \leq X_{zmp} \leq P$ for SSP2 (10)

With: $L_p$ is the length of the foot, P is the length of the step and $X_{zmp}$ is the real position of the ZMP.

$$X_{zmp} = \frac{\sum_{j=1}^{k} m_j*(\ddot{Z}_{gj}+g)*X_{gj} - \sum_{j=1}^{k} m_j*\ddot{X}_{gj}*Z_{gj} - \sum_{j=1}^{k} I_j*\ddot{q}_j}{\sum_{j=1}^{k} m_j*(\ddot{Z}_{gj}+g)} \quad (11)$$

With k is the articulation number of the biped, $m_j$ is the weight of each segment, $X_{gj}$, $Z_{gj}$, $\ddot{X}_{gj}$ and $\ddot{Z}_{gj}$, designed respectively the position and the acceleration of the COG of each segment accordingly of the x and z axes, $I_j$ is the moment inertia of each segment, and $\ddot{q}_j$ is the acceleration of each joint.

The calculation of the ZMP is considered only according to the sagittal plane [14].

## IV. Generation of the Walking Trajectory

### A. Trajectory Planning using LIPM and LPM

We are chosen to work with the walking pattern that contain one SSP and one DSP and the generation of the COG using the two methods LIPM and LPM consecutively during the SSP and DSP to generate the COG trajectory. For stable and smooth walking, double support phase is important. According to the biped locomotive, both legs are considered as an inverted pendulum during the SSP phase and a pendulum during the DSP phase. Several references have worked with this method to generate the trajectory of COG ([15], [16], [9]). Figure 1 shows a walking pattern that contains a DSP.

To simplify the computing of the COG trajectory, the height of the COG $Z_{COG}$ is assumed fixed throughout the walking trajectory and the position of the COG coincides with the position of the hip.

The two following equations represent the trajectory and the speeds of the COG during the SSP phase:

$$X_{COG,s} = C_{1s}*e^{w_s*t} + C_{2s}*e^{-w_s*t} \quad (12)$$

$$\dot{X}_{COG,s} = w_s * C_{1s}*e^{w_s*t} - w_s * C_{2s}*e^{-w_s*t} \quad (13)$$

Where: $C_{1s}$ and $C_{2s}$ are two constants that have been determined from the initial and boundary conditions of the position and velocity of the COG and $w_s$ represent the characteristic frequency defined as follow:

$$w_s = \sqrt{\frac{g}{Z_{COG}}} \quad (14)$$

The two following equations represent the trajectory and the speeds of the COG during the DSP:

$$X_{COG,d} = C_{1d}*\cos(w_d * t) + C_{2d}*\sin(-w_d * t) \quad (15)$$

$$\dot{X}_{COG,d} = -w_d*C_{1d}*\sin(w_d * t) - w_d* C_{2d}*\cos(-w_d * t) \quad (16)$$

$$w_d = \sqrt{\frac{g*(\frac{1}{a_d}-1)}{Z_{COG}}} \quad (17)$$

We simulate the motion of the COG through the walking pattern using SIMULINK/MATLAB. The figure 2 shows the evolution of the cog trajectory and its speed as a function of time. The simulation is done for two SSPs and one DSP. As the figure shows, the integration of the DSP into the trajectory gait ensures the continuity of the COG velocity between the two phases. The velocity curve alternately follows deceleration during the SSP against acceleration during the DSP. Indeed, the robot speed decreases and the swing foot strike the ground smoothly which avoids the increase of the impulsive forces resulting from the shock between the robot's foot and the ground. The presence of the DSP improves the stability of the robot and the switching of two feet will be smoother.

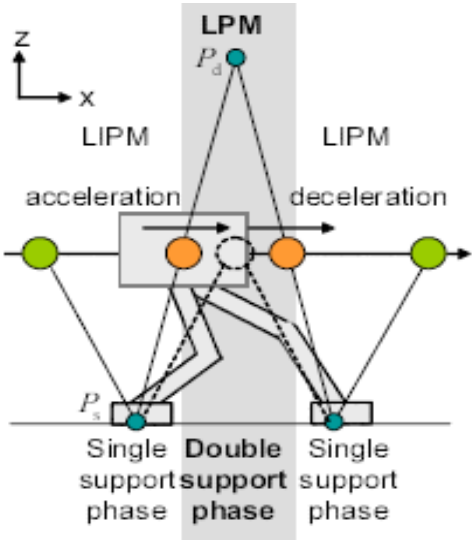



Fig. 1. Walking Cycle that Contains DSP [9].

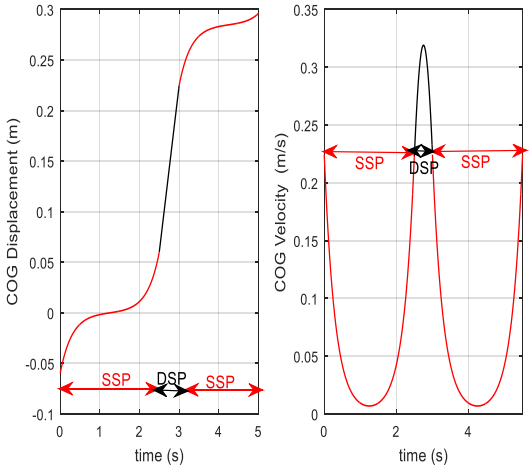



Fig. 2. The Position and the Velocity of COG.

### B. The Motion of the Feet

*1) Swing foot trajectory during SSP*: The phase of the SSP is defined by a single contact between the support foot and the ground. At the beginning of this phase the swing foot is in contact with the ground on the tip of the toe making an angle with the horizontal axis. During SSP, the oscillating foot moves forward while advancing one step. The support foot is in full contact with the ground and it does not leave it. So what interests us during this phase is the movement of the swing foot oscillating in the air while avoiding an obstacle. We used two piecewise 4$^{th}$spline functions to model this movement. The first function used to reproduce the displacement along the x-axis and the other along the z-axis.

$$F_x = \sum_{j=0}^{4} K_{xj}(t - t_0)^j \qquad t_0 \le t \le t_1 \tag{18}$$

$$F_z = \sum_{j=0}^{4} K_{zj}(t - t_1)^j \qquad t_1 \le t \le t_{ss} \tag{19}$$

Where F represents the function that model the x or z coordinates of the heel, ankle, or toe of the oscillating foot, $t_0$ , $t_1$ and $t_{ss}$ represent respectively the initial time of SSP, the time at which the biped have to avoid obstacle and the end time of SSP. $K_{xj}$ and $K_{zj}$ are constants determined by initial and boundary conditions.

The front foot is always flat to the ground during the SSP, so it is simple to determine coordinates of the heel, ankle and toe.

Figure 3 shows the motion of the swing foot during two successive SSP with switching of the two feet.

The three colors represent the positions of the heel, ankle and toe of the swing foot during the SSP and along the x-axis and the z-axis.

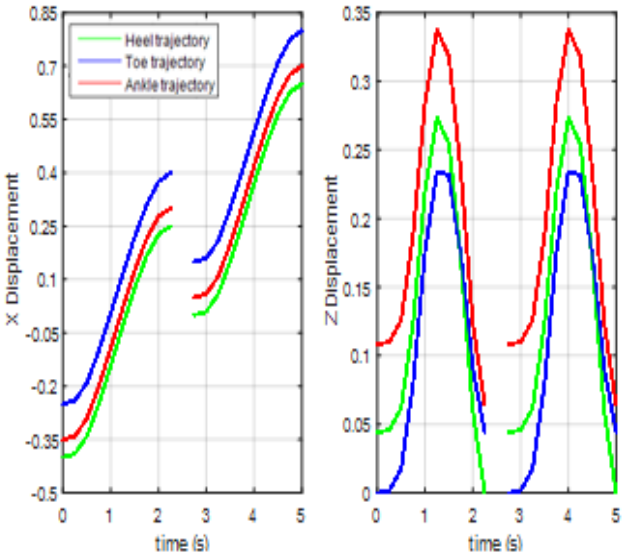



Fig. 3. The Trajectory of the Ankle, toe and heel of the Swing Foot during 2 Successive SSP Long the x-axis and z-axis.

The displacement of the swing foot along the x-axis shows the switching of two feet between the two phases. In fact at the end of the first step, the support foot is placed behind the swing foot. This is why the support foot coordinates at the beginning of the second step are lower than the one at the end of the first step (at the following step the support and the swing foot switch their positions) then it follows the same trajectory and it advances along the x-axis. The positions values of each point increase in function of time. The coordinates of the foot along the z-axis follows the same trajectory for the two steps. It modeling the motion of the swing foot during the SSP. The biped raises the swing foot and balances it forward to advance one step. The three curves increase as a function of time at the beginning of the SSP then it decrease and the value of the toe height becomes zeros at the end of the SSP (blue curve). At this moment the swing foot strikes the ground with the toe point and the DSP starts.

*2) Feet trajectory during DSP*: The phase of DSP is the defined by the contact of the both feet with the ground. At the beginning of this phase the swing foot is in contact with the ground on the tip of the heel making an angle with the horizontal axis. The front foot is flat to the ground. During this phase the biped raise the heel of the front foot without leave the ground until it make an angle with the horizontal axis. The swing foot descends until it becomes flat to the ground. We determined the trajectory of two feet during the DSP phase using exact equations.

*3) Feet orientation*: The orientations of the two feet are obtained from the simple polynomials of second degree which serves to generate the values of two angles $q_0$ and $q_5$ as a function of time.

Figure 4 shows the values of the two angles as a function of time during the two steps (two SSPs and two DSPs). $q_0$ represents the angle of the swing foot and $q_5$ represents the angle of the support foot. the figure 5 shows that during SSP, the angle of the support foot q5 is zero (returns to the fact that the support foot is in full contact during SSP), and at the end of the double support phase the two angles are zero because the swing foot becomes in full contact with the ground and the supporting foot makes a rotation on the toe stop and starts to leave the ground. The two legs switch their positions and the second step starts.

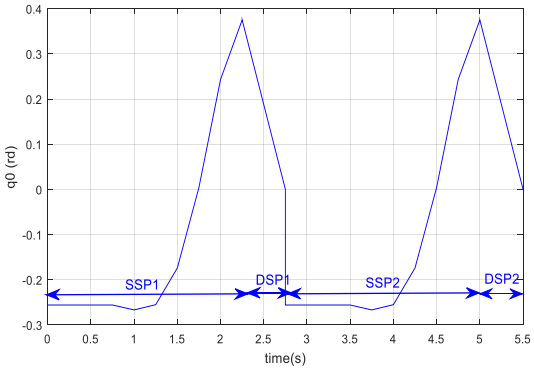



Fig. 4. The Swing Foot Angle as a Function of Time.

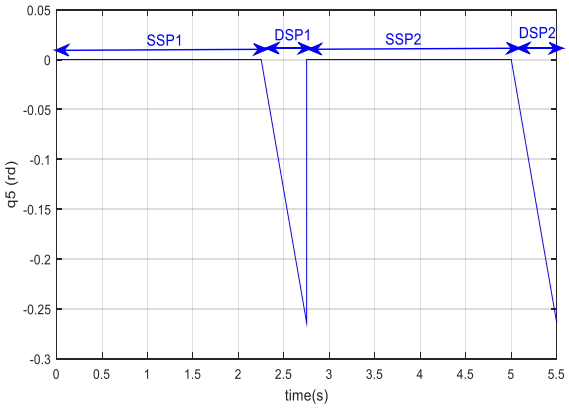



Fig. 5. The Support Foot Angle as a Function of Time.

### C. Inverse Geometric Model using Newton's Algorithm

The resolution of the inverse geometric model MGI is used to find the joint coordinates of the two legs from the Cartesian coordinates of the hip and two feet. The position of the hip and the positions of each point of the foot (heel, ankle and toe) were determined in the preceding paragraphs.

We used Newton's algorithm to numerically solve the inverse geometric model. Newton's algorithm is an efficient algorithm used to find numerically a precise approximation of a zero (or root) of a real function of a real variable and used to solve non-linear systems, it consists in the specification of a calculation schema, in the form of a series of elementary operations obeying a determined sequence. For our case, to solve the MGI, Newton's algorithm can be presented briefly as follows: During the gait cycle and at each iteration, it receives at the input the positions of the hip and the ankle, then it calculates the length of the virtual leg L and it solves the system of non linear equations to find the right values of the two joint positions $q_1$ and $q_2$for the front leg and $q_3$and $q_4$ for the swing leg. The values of the joint coordinates are initialized at the beginning of the algorithm and the convergence of the algorithm is conditioned by a margin of error and by avoiding the positions of singularity.

Newton's algorithm is simulated on MATLAB and the simulation results give the following curves.

Figure 6 shows the evolution of the four angles of the two legs as a function of the time during the two steps. The values of the four angles were calculated by the Newton's Algorithm and it corresponds to the biped configurations (positions of the hip and the feet calculated previously).

The speeds of the four joints angles were calculated from the derivation of direct geometric model and they were subsequently used later for calculating the ZMPs throughout the biped walking cycle. The four curves are represented by figure 7.

Figure 8 shows the stick diagram of the walking pattern that contains one SSP and one DSP with the rotation of the swing foot at the toe stop. We have used direct geometric model to simulate the motion of the biped robot. The biped walks two steps with a speed of 0.23m/s.

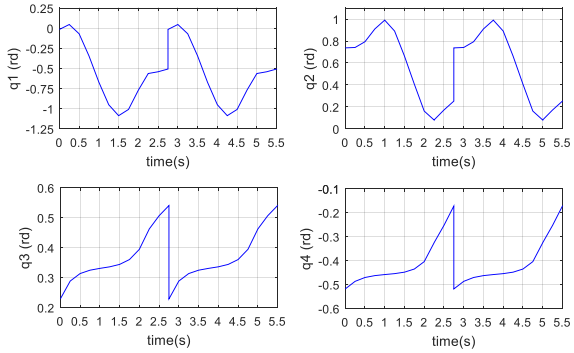



Fig. 6. The four Angles of the Left and Right Legs Calculated by the IGM.

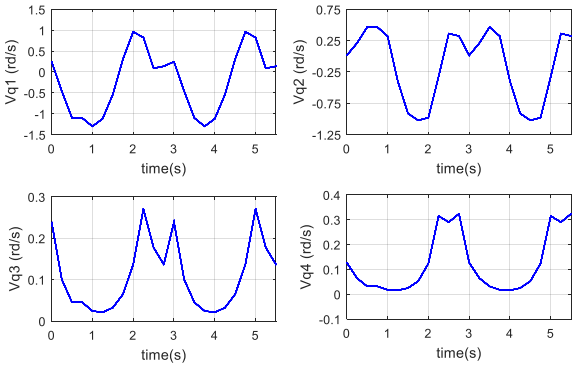



Fig. 7. The Speed of the Four Angles.

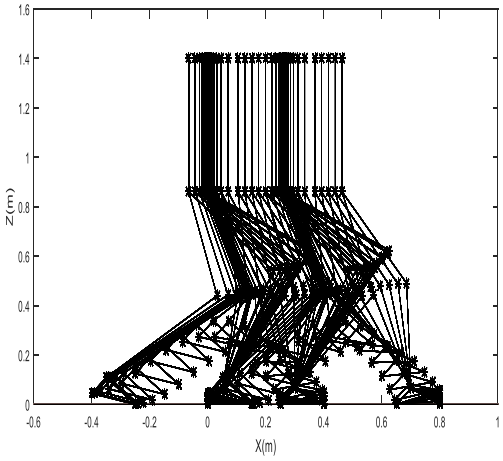



Fig. 8. The Displacement of the Biped Robot for Two Steps.

## V. Dynamic Modelling

There are several formalisms for calculating the inverse dynamic model of a robot whose Euler-Lagrange formalism, the principle of virtual works and the Newton-Euler method. The choice of the Euler-Lagrange formalism simplifies the number of the inverse dynamic model computing in symbolic.

When the effort on the terminal organ is zero, Lagrange's equations are as follows:

$$F = \frac{d}{dt}\frac{\partial L}{\partial \dot{X}_i} - \frac{\partial L}{\partial X_i} \quad (20)$$

Where L = E – U is the Lagrangian of the system, E represent the total kinetic energy of the system and U is the total potential energy of the system.

The resolution of the inverse dynamic model allows to determine the torques and forces (tangential and normal force) applied on both feet of the biped during the walking cycle according to positions, speeds and articular accelerations ($q, \dot{q}$ and $\ddot{q}$).

With the method adopted and the Lagrange equation, the dynamic model can be written in the following form:

$$M(X)\ddot{X} + N(X,\dot{X}) + G(X) = F \quad (21)$$

Where $M(X) \in \mathbb{R}^{9X9}$ is the inertia matrix, is defined as a positive symmetric matrix, $N(X,\dot{X})$ represents the vector of the centrifugal and Coriolis effects of dimension (9X1), $G(X)$ is the vector of the effects of gravity of dimension (9X1).

If we consider that both feet in full contact with the ground, we must consider the forces generated on both feet of the biped by the ground, the complete dynamic model becomes:

$$M(X)\ddot{X} + N(X,\dot{X}) + G(X) = A\Gamma(q, \dot{q}, \ddot{q}) + J_1{}^T\lambda_1 + J_2{}^T\lambda_2 \quad (22)$$

Where A represents the actuation matrix of dimension (9X6), $J_1$ and $J_2$ represent the two Jacobian of contact on both feet, $\lambda_1$ and $\lambda_2$ are the ground interaction forces on each foot of the robot.

### A. Resolution of the Inverse Dynamic Model in Simple Support

The SSP is defined by one foot in contact with the ground and the other foot in motion therefore free. The fixed foot arrives in flat impact and the swing foot takes off instantly, while swing foot in motion, the robot is in SSP.

During the SSP, the force on the swing foot is zero. The equation of the simplified dynamic model becomes.

$$M(X)\ddot{X} + N(X,\dot{X}) + G(X) = A\Gamma(q, \dot{q}, \ddot{q}) + J_1{}^T\lambda_1 \quad (23)$$

The knowledge of the trajectory of the hip and q(t) makes it possible to determine entirely the vectors $\dot{X}$ and $\ddot{X}$.

For the case of the SSP, the number of equations being equal to the number of unknowns, the resolution is directly possible and is put in the following form:

$$\begin{bmatrix}\Gamma(q, \dot{q}, \ddot{q}) \\ \lambda_1\end{bmatrix} = \left[A\ J_1{}^T\right]^{-1}\left[M(X)\ddot{X} + N(X,\dot{X}) + G(X)\right] \quad (24)$$

### B. Resolution of the Inverse Dynamic Model in Double Support

During the DSP the two feet of the robots are in flat contact with the ground. The dynamic modelling of the robot must therefore take into account the efforts of the ground on both feet. The dynamic model is described by the equation 22.

In the DSP, the robot is in an over-actuated configuration. Indeed, 9 coordinates are needed to describe the configuration of the robot in a general case and the robot has 6 actuators. However, in the DSP, the conditions of contact of the two feet on the ground add 4 constraints. The degree of over-actuation is therefore 1.

There are two solutions to solve the dynamic model during the dual support phase:

*1)* Parameterization of the forces exerted on the swing foot by mathematical functions [17].

*2)* The internal resolution of these forces by an optimal control resolution[18].

We adopted the first method to solve the dynamic model, the forces exerted on one of the feet requires expressing in an explicit form of time. We have chosen to fix the expression of the forces exerted on the foot 1 by a polynomial of order 2, $\lambda_1 = [\lambda_{1x}\lambda_{1z}]^T$.

$$\lambda_{1x} = \sum_{i=0}^{2} R_{1xi}t^i \quad (25)$$

$$\lambda_{2x} = \sum_{i=0}^{2} R_{2xi}t^i \quad (26)$$

Subsequently, the resolution of the system is expressed by this equation:

$$\begin{bmatrix}\Gamma(q, \dot{q}, \ddot{q}) \\ \lambda_2\end{bmatrix} = [B]^{-1}\left[M(X)\ddot{X} + N(X,\dot{X}) + G(X) - J_1{}^T\lambda_1\right] \quad (27)$$

Where:

$$B = [A\ J_1{}^T] \quad (28)$$

B is an invertible square matrix of size (9X9).

### C. Resolution of the Impact Model

The impact between two rigid bodies can be translated as a collision, which is modeled by an algebraic equation at the moment of impact. In the case of biped walking the impact phenomenon takes place at the end of the single support phase. It is characterized, under the assumption of rigid bodies, by a collision between the swing foot and the walking surface.

The joint velocities undergo sudden and instantaneous changes against the generalized coordinates after impact are the same as those before the impact.

After the consideration of certain hypotheses, the equation linking the ground forces to the robots' feet and its articular velocity before and after the impact is as follows:

$$M(q)*(\dot{q}^+ - \dot{q}^-) = F_{ext} \quad (29)$$

Where $\dot{q}^-$ (Respectively $\dot{q}^+$ ) represents the velocities before (respectively after) the impact.

$$F_{ext} = J_2{}^T(q)*\lambda \quad (30)$$

Where $J_2$ is the Jacobian matrix of constraints.

$$\lambda = \begin{pmatrix}\lambda_N \\ \lambda_T\end{pmatrix} \quad (31)$$

The system resolution is expressed by this equation:

$$\lambda = [J_2\ M^{-1}J_2{}^T] \quad (32)$$

## VI. Simulation Results

The computing of the ZMP values during the walking cycle of the biped robot is done according to the equation 11.

Figure 9 shows the curve of the ZMP before the regulation of the biped walking stability. However, Figure 10 reperesent the curve of the ZMP after the regulation of the biped walking stability. In fact, we have implemented an algorithm that calculates at each itération the value of the ZMP and it checks the stability conditions of the biped robot. According to the static and dynamic constraints and avoiding the singularity positions, the robot changes the value of the double support period and the running speed to bring the ZMP back inside the retention polygon and thus adjust the stability.

The polygon of sustenance is modelized by the curves in red. The biped walk is stable if the ZMP curve calculated during the walk cycle remains within the retention polygon. To ensure stability of the biped, we have to control the value of the ZMP and keep it within the bounds.

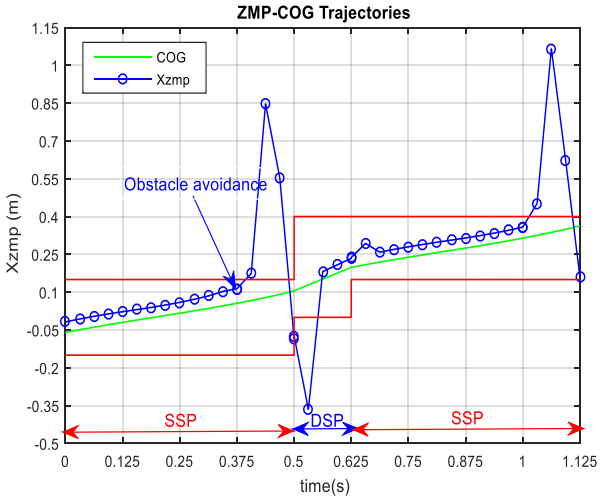



Fig. 9. The ZMP and COG Trajectories for Unstable Walking.

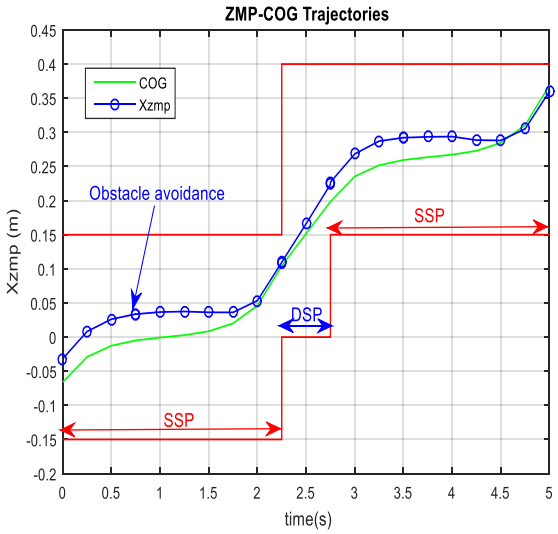



Fig. 10. The ZMP and COG Trajectories for Stable Walking.

The robot is generally in conditions of instability, especially during the SSP phase (only one foot in contact with the ground) while increasing its speed of movement, also the presence of an obstacle to avoid represents a disturbance for the stability of the biped. As the figure 9 shows, the ZMP leaves sustaining polygon when the biped starts to avoid collision with obstacle and it starts to rise the swing foot. This figure proves that the problem of biped imbalance is mainly located (in the conditions of a flat and non-ruthless ground) at the level of the SSP.

Figure 10 shows the effectiveness of our algorithm. The ZMP is still inside the polygon of sustenance. The robot keeps its balance throughout the cycle of walking and even the presence of an obstacle to avoid.

## VII. Conclusion

In this paper, an algorithm for controlling the biped walking stability has been implemented. The simulation results of the ZMP have shown the effectiveness of this algorithm. A constraint (the presence of an obstacle to be avoided) is added for the disturbance of the balance of the biped. Indeed, the limits of the sustenance polygon change with the movement of the biped robot through the gait walking and the values of the ZMP always remain inside the convex surface which constitutes the sustenance polygon.

The ZMP is a criterion used for the control of the dynamic stability and presents one of the most used methods for the generation of a stable trajectory for a biped robot because the control of balance has a common advantage with the other methods used for trajectory generation. The disadvantage of this method is that the ZMP is not always applicable. Indeed, its use is restricted to flat ground. As a result, it is not suitable for climbing up / down stairs, or on uneven ground. In addition, the ZMP cannot handle the balance during the flight phases during the race movement where there is no contact between the robot and the ground. Finally, the multi-contact between the arms and the environment is not taken into account in the ZMP. In a series of papers, the authors and colleagues have developed new feedback control strategies [19].

As a perspective of our work, we will study other scenarios and add other constraints on the walking biped robot as the climbing up / down of a stairs, walk on snow ground, carry an additional weight or integrate a phase of flight. Depending on the type of constraint added, we will associate with the ZMP another criterion for the management of the balance. An appropriate control method will be implemented in a closed loop to control and correct the stability of the biped robot according to the different scenarios.

## References

[1] H. Suzuki, J. H. Lee, and S. Okamoto, “Development of Semi-Passive Biped Walking RobotEmbedded with CPG-based Locomotion Control,” 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), June 28 - July 1, 2017 at Maison Glad Jeju, Jeju, Korea.

[2] I. W. Park, J. Y. Kim, J. Lee and J. H. Oh, “Online Free Walking Trajectory Generation for Biped Humanoid Robot KHR-3 (HUBO) ,’’Proceedings of the 2006 IEEE International Conference on Robotics and Automation, Orlando, Florida - May 2006.

[3] P. Vadakkepat,D. Goswami and M. H. Chia, “Disturbance rejection by online ZMP compensation,” Robotica volume 26, pp. 9–17, June 25, 2007, Cambridge University Press.

[4] H. Hirukawa and al. , “A Pattern Generator of Humanoid Robots Walking on a Rough Terrain,” IEEE International Conference on Robotics and Automation, Roma, Italy, 10-14 April 2007.

[5] A. Chemori, “ Quelques contributions à la commande non linéaire des robots marcheurs bipèdes sous actionnés,” Doctoral thesis, June 2005.

[6] N. D.K. Nguyen, B. L.Chu, V. T. A. Nguyen, V. H. Nguyen and T. T. Nguyen, “Optimal Control for Stable Walking Gait of a Biped Robot,” IEEE 14th InternationalConference on Ubiquitous Robots and Ambient Intelligence (URAI), June 28 - July 1, 2017 at Maison Glad Jeju, Jeju, Korea.

[7] J. P. Ferreira, M. Crisóstomo and A. P. Coimbra, “ZMP trajectory reference for the sagittal plane control of a biped robotbased on a human CoPandgait,’’ The 2009 IEEE/RSJ International Conference onIntelligentRobotsandSystems, October 11-15, 2009 St. Louis, USA.

[8] H. F. N. Al-Shuka and B. J. Corves, “On the Walking Pattern Generators of Biped Robot,” Journal of Automation and Control Engineering, vol. 1, No. 2, June 2013.

[9] M. Shibuya, T. Suzuki and K. Ohnishi, , “Trajectory Planning of Biped Robot Using Linear Pendulum Mode for Double Support Phase,” IEEE Trans, pp, 4094-4099, 2006.

[10] L. Sciavicco and B. Siciliano. “Modeling and control of robot manipulators,’’ McGraw Hill, New York, 1996.

[11] M. Spong and M. Vidyasagar. “Robot Dynamics and Control,’’ John Wiley & Sons, NewYork, 1989.

[12] Y. Wada, Y. Kanekoa, E. Nakano, R. Osu, and M. Kawato,“Quantitativeexaminationsfor multi joint arm trajectory planning using a robust calculation algorithmof the minimum commanded torque change trajectory,’’ Neural Networks, 14 :381–393, 2001.

[13] S. Omran, “ Generation de trajectoires de marche de robots humanoides etablies sur les modeles pendulaires bio-inspires de l’humain ,” Doctoral thesis, HAL Id: tel-01538626, http://tel.archives-ouvertes.fr /tel-01538626, Nantes University, June 2013.

[14] I. W. Park and J. Y. Kim, “Fourier Series-Based Walking Pattern Generation for a Biped Humanoid Robot,” IEEE-RAS International Conference on Humanoid Robots, Nashville, TN, USA, Decembre 6-8, 2010, pp. 461-467.

[15] E. Dong, D. Wang, C. Chen and J. Tong, “Realization of biped robot gait planning based on NAOrobot development platform”, Proceedings of 2016IEEEInternational Conference on Mechatronics and Automation, August 7 - 10, Harbin, China, pp. 1073-1077.

[16] S. Kajita and K. Tan, “Study of Dynamic Biped Locomotion on Rugged Terrain - Derivation and Application of the Linear Inverted Pendulum Mode-,” Proceedings of the 1991 IEEE Intemational Conference on Robotics and Automation, Sacramento, California - April 1991, pp. 1405-1411.

[17] D. Tlalolini, C. Chevallereau, and Y. Aoustin, “Comparison of different gaits with rotation of the feet for a planar biped,”Robotics and Autonomous Systems, 57(4): 371-383, 2009.

[18] T. Saidouni and G. Bessonnet, “Generating globally optimised saggital gait cycles of biped robot,” Robotica, 21(2): 199-210, 2003.

[19] C. Chevallereau, E.R. Westervelt, and J. W. Grizzle, “Asomptotically Stable Running for a Five-Link, Four–Actuator, Planar Bipedal Robot, ”4 march 2005 .

# Enhancing the Secured Software Framework using Vulnerability Patterns and Flow Diagrams

Nor Hafeizah Hassan[1], Nazrulazhar Bahaman[2], Burairah Hussin[3], Shahrin Sahib[4]
Faculty of Information Communication and Technology
University Teknikal Malaysia Melaka
Malaysia

***Abstract*—This article describes the process of simplifying the software security classification. The inputs of this process include a reference model from previous researcher and existing Common Vulnerabilities and Exposure (CVE) database. An interesting aim is to find out how we can make the secured software framework implementable in practice. In order to answer this question, some inquiries were set out regarding reference model and meta-process for classification to be a workable measurement system. The outputs of the process are the results discussion of experimental result and expert's validation. The experimental result use the existing CVE database which serves as an analysis when a) the framework is applied on three mix datasets, and b) when the framework is applied on two focus datasets. The first explains the result when the framework is applied on the CVE data randomly which consist mix of vendors and the latter is applied on the CVE data randomly but on selective vendors. The metric used in this assessment are precision and recall rate. The result shows there is a strong indicator that the framework can produce acceptable output accuracy. Apart from that, several experts' views were discussed to show the correctness and eliminate the ambiguity of classification rules and to prove the whole framework process.**

***Keywords*—Software secured framework; security classification; software security; common vulnerabilities and exposures***

## I. Introduction

In software application, it is observed that there are negative consequences when security is compromised. Security can be compromised when there is lack of understanding of the in hand situation. Various terms used for security and it's family, huge numbers of models and framework to refer to, had created confusions to the software practitioner to classify vulnerability that is accurate, consistence and correct.

It is observed that there is a challenge in forming a vulnerability classification scheme due to type of data used. For example, some vulnerability database like Common Vulnerabilities Exposures or CVE is very much using natural language structure but without proper English grammar as given in its web page of Common Vulnerabilities and Exposures: The Standard for Information Security Vulnerability Names. One way to extract the information is by using semantic analysis [1]. However, in security domain, some terms are used differently. For instance, the meaning of buffer overflow is to overwrite the adjacent memory by overrun buffer and is not simply means that buffer is more than full.

Therefore, it is learned that the terms must be specified with related to predefined rules of information security. Another challenge was to formally translate the domain terms into a schema that can be translated to a workable engine to extract the vulnerability given a historical database as debated in [2]. Therefore, this study is to focus on this scenario.

The current vulnerability classifications suffered from multiple dimensions of classifiers. They are either too specific or too complex [3] and [4]. Or they were only for dedicated cases. This leads to disability in performing a detection or protection from newer attack of vulnerability. The understanding of the taxonomy which are also various, requires a formal classification that can be used for generic cases regardless of applications, mobiles, networks or other devices [5]. This study focuses on the research and development of the design, formalization and translation of the vulnerability classification pattern through a framework using common vulnerabilities and exposures data pattern. It is achieved through the usage of syntax and semantic formal representation that not only accurate to produce a simplified set of vulnerabilities patterns but also consistently can be use within other incident cases. The final aim of this study is to measure the accuracy and correctness of the vulnerability classification procedures of algorithm, which already indicates the focal view and depth in security domain.

## II. Related Work

The early work on vulnerability classification focused on the knowledge of fault identification. It tried to solve the difficulty of identifying the origin of faults within a software [6],[7].

### A. The Origin of Fault

The initial works claimed there is a specific place a fault exist or known as origin. However, the newer research claimed that the place alone cannot be considered as the origin of fault but shall encompass the time it occurred, as debated in [8],[9]. The later work eventually pointed out that the development phase is frequently used as the time the faults are introduced.

### B. Fault, Failures, Attacks and Vulnerbailities

The knowledge of this fault identification leads to the terminology emergent of faults, failures, errors and attacks. Shortly, a differentiation of the definition is given in [8] as following. An error is a human action that produces an incorrect result, a fault is an incorrect step, process or data definition in a computer program and a failure is the inability

of a system to perform its required functions within specified performance requirements. As system becomes complex, [10] defined fault as an imperfection that able to result as failure in software.

However, their study focuses on the taxonomy of security terms and did not critically informing enough how the terms can be used to reduce numbers of vulnerabilities. Eventually, another finding is the study of relationship between the terms which perhaps gave new meanings to reduce vulnerability. Few researchers addressed fault as a concrete manifestation of an error within the software. One error may cause several faults, and various errors may cause identical faults which if encountered can cause system failure as deduced by [11] and [12]. However, [9] uniquely elaborates the relationship between vulnerability and error. They defined that vulnerability is an instance of an error, and error can be in specification or development or software configuration. We also learned that [13] in their work introduced the vulnerability term as a characterization of vulnerable states which is distinguish from any non-vulnerable states as in Fig. 1.

Fig.1 shows the relationship of the terms error, fault, failures and vulnerabilities. Perhaps, because of this consensus, later, in Common Vulnerabilities and Exposures: The Standard for Information Security Vulnerability Names, another definition of vulnerability is given as a state in a computing system that either:

- Allows an attacker to execute commands as another user
- Allows an attacker to access data that is contrary to the specified access restrictions for that data
- Allows an attacker to pose another entity
- Allows an attacker to conduct a denial of service

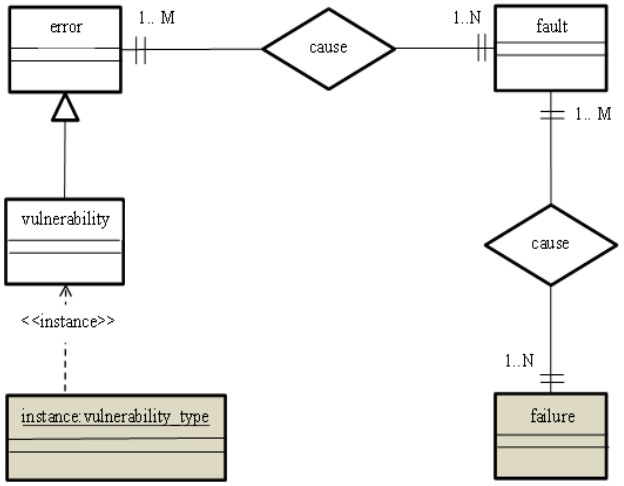



Fig. 1. Relationship of the terms-Error, Fault, Failures and Vulnerabilities

This definition shows, when connecting with definition from Fig. 1, that vulnerabilities are the negative consequences of faults that bring about from one or more errors. And they are always related to external user or attacker who manipulates the vulnerabilities to get access. As a result, a compromised vulnerability can cause a (or various) system failures.

### *C. Vulnerbailities Classification: The issues*

Consequently, the term vulnerability in vulnerability classification reflects an inclusive meaning in related to the reason of any software failures. But, still is the issue of how to classify these vulnerabilities given the fact that vulnerability, so that a developer could know which failure they are related to.

Very often, the common style of classification pro-motes the vulnerabilities to be placed in more than one class, due to the hierarchical style of classification. This technique is called data multiplication. Multiplication of data is the major drawback. For example, if vulnerability x is located into class A and class B, it will return a confusion for later analysis especially for treatment. This will cause an incorrect result. The method were later improved by [14] as discussed in [15, but still not much difference. On top of that, [13] argued the classification of flaws in PA project, RISOS project and in [9] which they meet neither the uniqueness nor the well-defined decision procedure requirement. They suggested that one can view a vulnerability as a containing class and attacks as elements of that class.

In 2005, two researchers from CMU/SEI published their work of new way of classification through technical report in [16]. In order to avoid multiplication issues in mentioned earlier, which using hierarchical style of classification, they suggested the attribute-pair values through object-roles definition of vulnerability. However, the method suffered the level of abstraction and viewpoint. Another researcher, [17] classified the vulnerabilities through characteristic trees, still is a hierarchical style of classification which suffer the tradeoff between different operating system used. In the nutshell, the gap is the vulnerability classification is still an issue where, it could be either too specific to certain system or too complex which, in addition, introduce the ambiguity.

All classification involves the goals and the perspectives. The goals and perspectives demonstrate of how the classification is being seen and carried out. A study on classifications goals and perspectives were done and showed that a trend of classification from as basic as error to as complex as data mining had been carried out. However, the diversity of classification patterns still can be refined as some of them had produced complex patterns with huge number of classes.

### *D. Vulnerability Classification: The generic Process*

The analysis during these study also reveal that each classification scheme by the researchers consist of few generic processes namely as identification, analysis, confirmation and elimination of flaw. The summary of the processes completed by the researchers are as in Fig. 1.

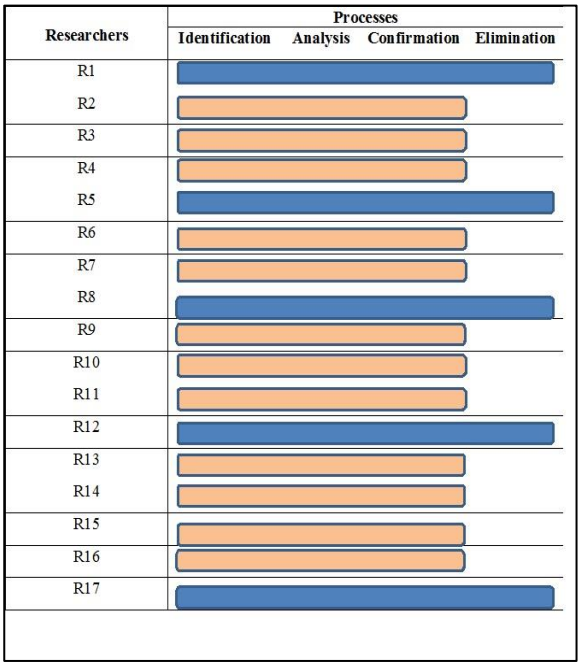



Fig. 2. Vulnerability classification based on process

Fig. 2 shows four generic processes in determining and classifying vulnerabilities: identification, analysis, confirmation and elimination of flaw. The identification process covers from non-unique identification of system resource to specific identification. The analysis process involves the labeling, add-on features and dissemination of identified flaws. The confirmation process is putting the flaw into the appropriate class. And the elimination process informs the reader that the respective perspective also able to assist in reducing the identified flaws. The reasons of having these four generic processes, according to [4], [18] and [19] are to: establish, act upon and maintain relationship about the threat in related with their life cycle.

The development team can gain better awareness of the larger threat viewpoint when this information is shared in advance of an incident, hence reducing the occurrence of next possible bugs and vulnerabilities. And from the analysis, not all past works contained every process. But, here, from the table showed that at least three initial processes must exist in any vulnerability classification scheme - identification, analysis and confirmation. Using these generic processes, many researchers develop their own models or frameworks to reduce the numbers of bugs and vulnerabilities as in [4], [20] and [21]. However, the definition of what kind of framework that should work with vulnerability classification is remained a dispute.

### *E. Secured Software Framework*

A framework is defined as the conceptual structure that dependent to each other to complete specific purposes. The structure could be an artifact of input, output, process, function, or boundaries [22] and [23]. Following the analysis in [13], the definition of framework is given as a sequence of decision procedure which when apply, will classify a state to exactly as one tuple. A decision procedure refers to the application of the function to a specific vulnerable state. Discriminating properties, as embodied in the decision functions, determine classification.

The specific purposes of such a framework are to provide a historical record of the vulnerabilities in a form that software developers can use: a) to anticipate flaws in their systems; b) to describe the vulnerabilities in a form useful for detection; c) to show common characteristics in related flaws for prevention and elimination; and d) to enable a security monitor to detect exploitation (or attempted exploitation) of the flaws.

According to [8], a framework is not simply a neutral structure for categorizing elements. Taking from here, a vulnerability classification is an element in a absolute framework, in which, this framework shall serve as the blueprint working process to do the vulnerability classification despite any aim it is meant for to be delivered. But still, in a context to have secured software.

Hence, the framework to describe the vulnerability classification working process is named as a secured software framework.

A comparison of security framework had been made by [24] which analyzed eleven frameworks from year range 1996 to 2004. The comparison significantly shows the needs for a systematic approach to embed security concerns into software process as early stage. Another work was conducted by [25] to summarize the security dimensions, such as cause, impact, and location, encountered in security frameworks.

However, there is still lack of research done of how to integrate secured process operation in software development process [26]. A comparison work on this secured software framework was done in [27]. The summary is given as in Table 1.

The researchers review had revealed that those frameworks start their security consideration as early as requirement stage as depicted in Table 1.

TABLE I. THE STAGES FOCUSED IN SOFTWARE SECURITY PROCESS FRAMEWORK

| Software security testing framework | Stage | |
|---|---|---|
| | Requirement | Design |
| KAOS (Security Extension) | ✓ | ✓ |
| MDS (Model Driven Security) | Not stated | ✓ |
| i* framework | ✓ | Not stated |
| ST (Secure Tropos) | ✓ | Not stated |
| SQUARE | ✓ | ✓ |
| SREP | ✓ | ✓ |
| SRE | ✓ | ✓ |
| TM (Threat Modeling) | ✓ | ✓ |

The summary demonstrates that in the mentioned security frameworks, security is a concern as early as during the requirement stage as in [28] and [29]. They highlighted the needs for a standardized methodological approach that taking into account security elements from the earliest stages of development till the completion.

### F. Measurement

Even though [30] highlight the quality attributes of CIA-triad (confidentiality, integrity and availability) in security studies, this research also emphasized measurement for the algorithm used. The successfulness of the framework is dependent how good the algorithm can extract the required patterns. The extraction-based procedures are measured using:

- precision rate
- recall rate

The precision measures the accuracy of the algorithm and the recall measures the effectiveness of it. However, the correctness of incident case-pattern matching process is measures using expert opinion approach.

## III. METHOD

This study focuses on two parts. The first part is to uncover current process of vulnerability classification by identifying the meta-process model and indicate its framework. The second part is to enhance the framework by analyzing the vulnerability classification patterns and confirming ways in accepting affirmation

In identifying meta-process model, task of this phase is to investigate the existing meta-process models of vulnerability by analyzing and synthesizing the core elements with the help of parsing technique in the existing common vulnerability and exposures database to determine whether or not the model satisfies with the objectives.

In analyzing the vulnerability classification patterns, the possible patterns exist in the database by using key aspects in a threat model and the metaprocess model. The list of patterns in threat model are synthesize with the current categorized of threat in the database. A formal classification method through an algorithm is proposed and used to deduce the patterns which will sustain the vulnerability classification framework to accept the subsequent process of affirmation.

And in accepting affirmation is to execute the formal algorithm by applying it within the classification framework and get effectually valid through experimental and expert views.

## IV. THE FRAMEWORK

A successful meta-process model highly depends on proper apprehension of their functionalities, contexts and architectures. To achieve this purpose, a reference model technique is used.

### A. Identifying Meta-Process Reference Model.

This is an input to the process of generalizing the meta-process model. In this activity, a qualitative study was used to collect, categorize and select the related reference model. The output from this procedure is the formulation of vulnerability classification meta-process. This meta-process is used as the blue-print for the framework. This study is inspired from the work of Linde [31]. In the work, a four steps of classification meta-process was introduced which consist four stages: knowledge of system control structure, the generation of an inventory of suspected flaw, confirmation of the hypothesis and making generalizations of specific flaw instance. The former study was aimed to focus on software attack through generic operating system or OS flaws [6], [7] and [13]. As the study of classification matured, the needs to enhance the methodology arise, for example to support the purpose of doing the classification varies among researches [31]. Therefore, in existing secured software framework, there are at least three stages involved in vulnerability cycle, however four is preferred as the fourth support the mechanism for elimination. The four steps are adapted as the meta-process for vulnerability classification which consists of i) identification, ii) analysis, iii) the confirmation and iv) elimination. On top of that, instead of using class name or perspectives for the countenance, we assign the label of characterization to represent the class showing the significant of a characterization embed in respective class. The label of characterization shall be named after considering the attributes that carry specific value describing the label as discussed in [4]. By answering the attributes with relevant value will help the researcher to present the characterization into a single tier or more. The following subsection elaborates the steps in detail.

### B. Analyzing the Process, Activities And Output

The identification process involved identifying the objective of the classification. Then, the researcher refers the previous taxonomy of classification to understand the subsequent philosophy. Within this process, the context (scenario) of each security flaw or attack is analyzed and any available repository is examined to find any match. As it is almost impossible to analyzed the information system flaws as a whole as pointed out by [32], a researcher will sub-task the system by addressing the flaws or attack from the encapsulate platform which either come as an operating system or application based security flaw (or attack). In addition, the nature of the platform (operating system or program) which also referred as target system (or victim), is explored and essential elements are captured.

The output of this process is to confirm any sensitive data that activate an interest to an attack (security objects) and to distinguish the protection mechanism that protects the respective data (control objects) [31].

The analysis process involves labeling and dissemination. These security flaw or attack contexts are further given certain labeling, often to map them with the captured elements of the target system as well as to meet the aim of the research perspective. These labeling contained common shared traits also known as classifiers or attributes of the flaws, which in many cases are very difficult to be precise. To initiate a labeling, a set of questions for the content shall be imposed. [6],[13] and [32]. In the dissemination process, the identified classifiers or attributes that gathered from defined scenario in the activation process are examined, and assigned a formal value. Using this value, the description of the security flaw is

transfer into hierarchical form such as list (flat or non-flat list) and tree (commonly in directed acrylic graph). Very often the hierarchy is going top-bottom or generic (inclusive) to specific [6],[32],[9],[13] and [10]. The selected hierarchical orientation must consider the complexity of security flaw or attack which may act as an amalgamation (the blended form of flaw/attack) and decelerate the process. As the process iteratively matured, the countenances shall update the classifier to suit with the analysis which also referred as categories refinement. The whole process of defining the countenance is also known as specific taxonomy (class syntax) and a good start for a (high-level) class syntax for OS flaws was introduced by [6].

The confirmation of instances process is the placement of any noticeable flaw or attack into the hierarchical form and expected to form a directory of security flaws or attack to assist in next production. A proper directory will indicate the flaw labeling; thus, facilitate the protection and testing process effectively. In common, the dissemination is done in bottom up approach or from a specific flaw towards achieving its generic class.

The elimination process is the detection, elimination and forecasting of faults in the next system development. It covers at least five aspects namely as fault detection, fault prevention, fault tolerance, fault removal and fault forecasting. The process is detailed out in terms of activity and output as simplified as in Table 2.

TABLE II. THE PROCESS, ACTIVITY AND OUTPUT OF A VULNERABILITY CLASSIFICATION FROM CRITICAL REVIEW

| **Process** | **Activity** | **Output** |
|---|---|---|
| Identification | Define objective, target system and identify data | Classification objectives, target system specification and a selection of data |
| Analysis | Select method, construct hierarchy using syntax or semantic and analyze impact | Method selected, hierarchy (list,tree or graph), list of impact |
| Confirmation | Refinement of the hierarchy | Refined hierarchy |
| Elimination | Detecting and eliminating the new flaw found | a. Fault detection,<br>b. Fault prevention,<br>c. Fault tolerance,<br>d. Fault removal<br>e. Fault forecasting |

### C. Conceptualization of Vulnerability Classification Meta-process

Next, we analyze the current approach on vulnerability classification. Each work is determined for their unique processes in conducting the classification; implicitly or explic itly stated in the research work. As a result, it is obviously seen that in every works, the researchers begin with identifying their purposes to align with the results (as identification process). Next, based on the purpose, they analyze the data or any vulnerability sources to determine the appropriate classes (as analysis process). Following to that, the data is assigned into the respective classes, iteratively (as dissemination process) In order to ensure that the assignment is correct, a benchmarking with secondary data such as technical reports or literature review is conducted (as confirmation process). Some classification proposed the solution to the vulnerability while others don't. In elimination process, the mechanism to solve the problem is identified. For example, a fault can be detected and remove. Or a frequent occurrence of fault may lead to fault forecasting and thus precede to fault preventing. The formulation can be presented as in Fig 3.

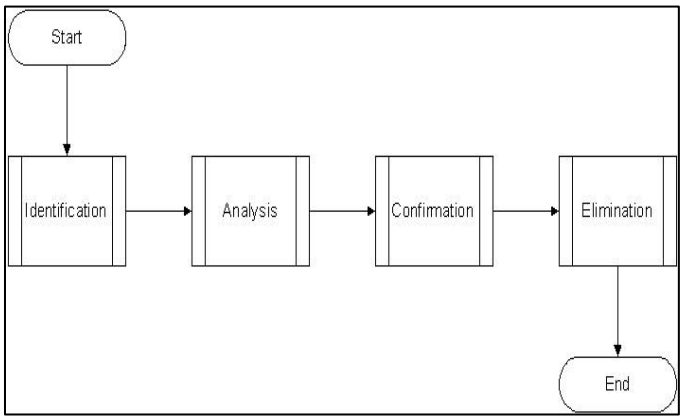



Fig. 3. The Formulation of Vulnerability Classification Meta-process

Fig 3 shows four generic processes involves as a vulnerability classification meta-process - identification, analysis, confirmation and elimination. Next, the details in each of it are explained in the next subsection.

### D. The Vulnerability Classification Framework (VulClaF)

This section summarized the study from Table 2: The process, activity and output of a vulnerability classification with Fig 3: The Formulation of Vulnerability Classification Meta-process. We discovered the high-level overall framework for secured software as in Fig 3. In order to detail out each processes, each process can be re-iterate for respective stages-analysis and deployment as in Fig 4.

For example in identification process, involves the activities of defining objective, target user, system and data. And for deployment or output stage, the detail of them is describes as in Fig 4.

However, this framework which was used unconsciously used during classification was suffered from issue of data multiplication that eliminate the uniqueness of a class. Hence, slow down the remedy process.

This issue is mainly due (as in Fig 4) in the Analysis process at activity Produce vulnerability tree. A vulnerability tree has high potential to create redundancy if the classification rules were not formally defined in activity Produce syntax-semantic schema. An enhancement to overcome this is proposed.

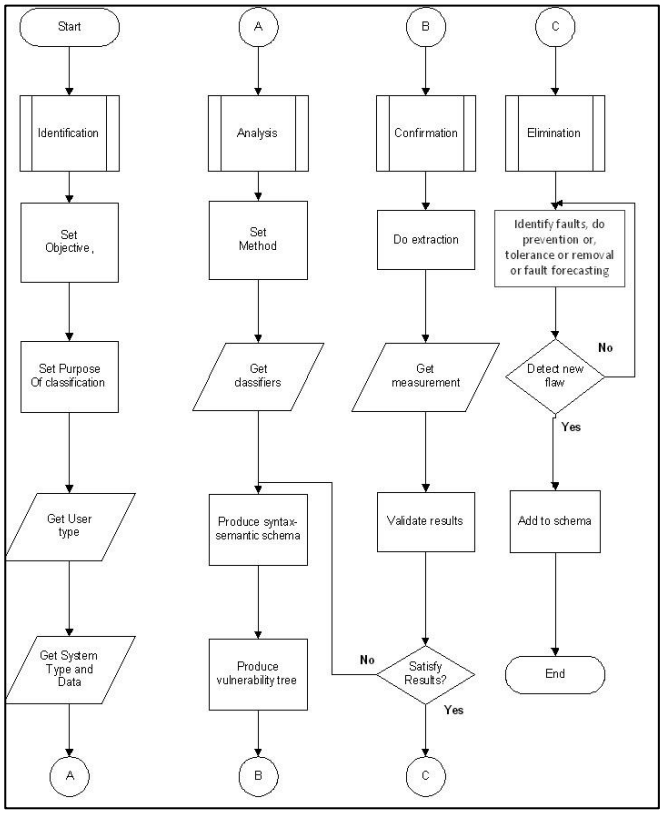



Fig. 4. The Vulnerability Classification Framework (VulClaF)

## V. The Enhanced Vulnerbaility Classifcation Framework

The proposed enhanced framework is based on these phases: a) attaining the vulnerability classification patterns and b) accepting affirmation.

### A. Attaining Vulnerability Classifcation Pattern

The classification pattern is attained from preprocessed raw data, with the help of the context-free grammar as input. Through this procedure, vulnerability classification patterns are identified after matching process with the domain specific wordlist. However, this procedure involves understanding of existing vulnerability classification trends. Analyzing Existing Vulnerability Classification Trend. This is another input to the process that involved identifying current trend in vulnerability classification.

The output from this procedure is to determine the label of characterization and establish their relationship.

- The Label of Characterization. A numbers of existing vulnerability classification were studied, scoped and analyzed for their similarities elements. It is found out that many terms were used to characterize the attributes or state of threats such as origin, time, OS etc. With that, in this study a generic term to address them is used as: label of characterization and served as partial classifiers because when on their own, they are not sufficient yet to do a good classification. The connection between them must be discovered. The output is to get a set of container with different motives of their existence.
- Discovering Relationship. The sequence of the labels with their cause and effect were later realized as a relationship notation. They were pre-analyzed on a sample of data and tested. The output is to define the cause-effect of each container from label of characterizations. Next, a set of context- free grammar then was established for the purpose to avoid ambiguity in experimental results.
- Context-free Grammar. This is an input to the process of generating the vulnerability classification algorithm. The output consist domain specific schema and domain specific marker.

### B. Label of Characterizations

The purposes of classification from past works clearly stated their aims are to do the classification for the purpose of assisting in testing and maintenance as in [4], [33] and [34]. It means that the purposes shall represent the level of abstraction. Level of abstraction implies the attempt how extensive the classification is [34]. The user perspective reflects the aimed users that the classification is made for, like for the usage of software developer or software designer.

The label of characterization depicts the elements or features considered that results as a class in the classification. In [34], the outcomes of classification consist of software development issues, location of flaws in the system and impact of flaws on the system. Each class comprises of second tier of subclasses. In [33], it is classified into seven groups: input validation and representation, API abuse, security features, time and state errors, code quality, encapsulation, and environment.

The research in [4] shows close similarities with the aim of this research which proposes the combination of : i) cause, ii) location, iii) attack vector and iv) impact as the essential label of characterizations for secured process, maintenance and assessment. The cause describes the reason for the existence of the vulnerabilities; the location determines the attacker community that defines the risk level and in turn determines the mitigation strategies. Attack vector defines the attack mechanism used by the attackers and impact describes the degradation of the system performance after an exploit takes place as in [4]. Therefore, this model of [4] is appropriated as the reference model for vulnerability classification within software developer perspective that aim of assisting security process, maintenance and assessment. And based on this model, we scoped the current study of other researchers as in Table 3.

TABLE III. Identifying the Vulnerability Classifiers From Past Works

| Purpose | User Perspective | Label of Characterization |
|---|---|---|
| To assist security process, maintenance and assessment | Software developer | Cause<br>Location<br>Attack vector<br>Impact |

Based on Table 5, we represent the significance of the purposes. Any decision to specify the label of characterization is established based on the purpose of the classification and the target users. The purpose responds to the issue of level of abstraction and the user's perspective counteracts the issues of point of view. However, the connections sequences between the labels remain unclear and the purpose scope is extensive try to cover process, maintenance and assessment – three dimensions of works. Based on the issues addressed, a new label of characterization is proposed to determine the relationship between each class with more specific scope of purpose.

The important aims within secured software are to confirm the presence of security attributes in the software and to increase the delivered reliability of the software to the user as in [35], [36] and [37]. As the vulnerability population is vast, it is impossible to examine each of them individually [37]. However, the secured software focuses on vulnerability exploit that demands the understanding of different states occurred as well as their relationship between each other when compromised. These states describe the behavior of an event that having certain attribute at any given condition. The existing studies describe how they perceive the existence (occurrence) of vulnerability in various ways such as domain, origin, operational and software development life cycle phases. Nevertheless, in software, the occurrence is about writing at least an executable statement of code. Then, regardless of how they are perceived, an occurrence of vulnerability consists of at least four states: i) creation, ii) discovered, iii) exploited and iv) resolved of an executable statement of code. Creation is when the code is created with a specific reason or cause at a definite place or location and still in an idle state. Discovered is when an attacker realizes the particular code as a vulnerability point in the location or application, potential to be compromised and this is a visible state. Exploited is when the attacker uses an attacking mechanism that is appropriate with the cause and location to compromise the vulnerability point and this is the utilize state. Resolved is when a targeted impact of the attack affects the system and an action is required to solve the problem, including appropriate test and this denotes the state of fix. These four states determine the characterization needed during the vulnerability classification model for a secured software framework. In any exploitation of vulnerability, it is essential for an attacker to identify the states of possible exploitable code using an attack mechanism and produce the targeted impact they have expected.

The states are the life cycle events of vulnerability and were represented as a waterfall sequence. However, the relationships between the events are currently not well explained. For example, whether there exists any kind of source and target events, or if there exists bidirectional relationship between them. In order to identify the type of relationship that exists, the study of the vulnerability class must be able to map with the respective states.

These states are further perceived as label of characterization introduced in the Table 4 by accepting the labeling of cause as source root (to refer to the vulnerable code statement), source location as application (to refer to the application container of where the code statement reside), target vector as target (to refer to the effect of attack mechanism use on the vulnerable code), target impact as impact (to refer to the expected result of the exploitation or damage), and this research introduce the notion an additional of attacker (to refer to the user who initiate the attack). As the purpose is to determine the relationship between each label, the component relationship is added as in Table 4.

TABLE IV. THE NEW LABEL OF CHARACTERIZATION AND RELATIONSHIP FOR SECURED SOFTWARE FRAMEWORK

| Objective | User Perspective | Labels and relationship of characteristics |
|---|---|---|
| To assist secured software framework | Software developer | Label of characterization<br>• Source Root<br>• Source Location<br>• Target Vector<br>• Target Impact |
| | | Relationship between label |

Based on Table 4, the pattern for CVE is produced using three steps: labeling, categorizing and producing schema grammar from CVE incidents cases. The first process is labeling the domain specific marker (DSMarker) for each classifier. The markers are used as indicator to pick and group the words into related classifier. The markers are used as input to the second process. The second process is to categorize the words into domain specific wordlist (DSWlist). The purpose of this process is to extract the similarities (domain-based) between the words and form the wordlist. The output of this process is used as an input for the next process, producing the domain specific schema (DSSchema) grammar. The study from the data shows the incidents have obvious marker that indicates which phrase shall belong to a classifier. Fig. 5 shows the DSMarker.

The top level is the primary marker. The second level is the secondary marker, which can be divided into two: the left side of the primary marker and the right side of the primary marker. The third level is the subsequent markers, those that immediately serve the phrase after into the specific classifier denotes as the lowest most level.

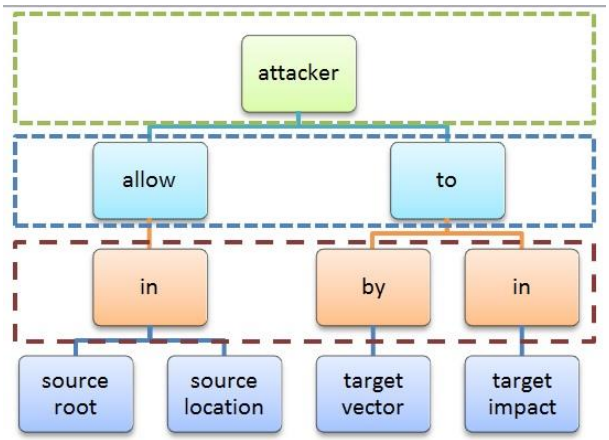



Fig. 5. Showing the generic hierarchical structure of the classifiers

The preliminary study also showed that there are four classifiers: source root, source location, target vector and target impact. Source root is the reason or the antecedent that imply to the impact of vulnerability. The source root can be (but not limited to) buffer overflow, configuration, connection, credential, injection or settings. These conditions are originated from the coding and implementation phases. They are a result of lacking concern of early security imposed in the requirement and design phases either lack of tools used or lack of knowledge about them. Source location is the emplacement or entity where the vulnerability antecedent exists. The place can be (but not limited to) an argument, a string, path, registry, password, host or script. These are places where certain vulnerability may be originated and developed. Interestingly, this analysis found that only few terms share two different classes. The 'mail' and 'directory' could be in class source location and target impact. Therefore, to differentiate them, the keyword 'in' before 'mail' or 'directory' is recognized as subsequent marker to determine them as source location.

Target vector is the magnitude and directions causing damage to occur on target. The magnitude and directions represent any doing that can cause damage such as (but not limited to) change, compromise, check, execute, modify, insert, read or obtain. These actions are showed by the usage of verbs in the pattern. The word magnitude is used as it carries certain weight to cause the damage and the word direction is used to show the aiming point of damage. Target impact is the entity that been affected by the target vector. The effected entity can be (but not limited to) data, database, file, directory, an account, a program or a system. The effect is the outcome or consequences that a target becoming upon a vector is carried out. In order to ease the reference to the classifiers, in the rest of this discussion, each of them are label as: SourceRoot, SourceLocation, TargetVector and TargetImpact. The examples of each classifiers class are as in Table 5.

TABLE V. EXAMPLE OF ITEMS FOR CLASSIFIERS

| SourceRoot | SourceLocation | TargetVector | TargetImpact |
|---|---|---|---|
| format | mail | compromise | data |
| configuration | authentication | authenticate | certificate |
| connection | cache | bypass | client |
| credential | function | cause | code |
| crossite | host | change | command |

Table 5 shows that there will be a tendency for word redundancy or similar meanings, e.g code and command in TargetImpact. This issue is taking care by constructing the similarity list discussed in the next subsection.

### C. Domain Specific Wordlist (DSWordlist)

Domain specific wordlist (DSWordlist) is the phase where the words are groups not only based on the grammar usage of singular or plural, or tenses, but also based of their semantic domain understanding, for example, an email and a mail , both should be considered as in one group of word. First, using one-to-one matching, the definitions of the words are listed as shown in the sample in Table 6.

TABLE VI. WORDS, DEFINITION AND CLASSIFIERS

| Words | Definition | Classifier |
|---|---|---|
| setting | change qualities of how the application works. | SourceRoot |
| configuration | is often where an application is customized for user or group | SourceRoot |
| function | a function returns a value | SourceLocation |
| parameter | are the strings/arguments used to pass value to functions or programs | SourceLocation |
| hijack | to stop and steal | TargetVector |
| spoof | to imitate | TargetVector |
| database | place where records stores in it | TargetImpact |
| script | written code for run-time environment | TargetImpact |

Table 7 shows the words, their meanings and the classifiers they belong to. However, if one search for the incident is using function and another is using program, there is chance of ambiguity to locate the patterns, as these two words may refer to the same object.

Therefore, the second process took place. Second process, the domain specific word list (DSWordlist) is defined as the list of words that are referring to the domain semantic meanings in their specific classifier. An example of (DSWordlist) is given in Table 7.

TABLE VII. DOMAIN SPECIFIC WORD LIST WITH SEMANTIC

| Wordlist | Semantic Meanings | Classifier |
|---|---|---|
| connection<br>request<br>session | activity between two machines | SourceRoot |
| function<br>program<br>system | workable code to perform instruction | SourceLocation |
| denial<br>block<br>terminate<br>crash | ability to deny | TargetVector |
| account<br>system<br>password | user belonging identity | TargetImpact |

Table 7 shows the semantic meaning for the words. The first column is the words from the data set. The second column is the meaning and the third column is the classifier they belong to. In this example, the function and program has been defined as in similar group of object. The wordlist and their classifiers are the input to the next phase: domain specific schema grammar (DSSchema).

### D. Domain Specific Schema Grammar (DSSchema)

Domain specific schema grammar is the general description how the classifiers worked and used to execute the classification. In this study, the schema is represents using Backus- Naur Form (BNF) notation. The words in the sentence are referred as field. The target of this classification is to define the pattern. Each pattern is characterized by the four classifiers. The classifiers are SourceRoot, SourceLocation, TargetVector and TargetImpact. In this study, the process begins with inquiry of the incidents existed in the file. Each incidents consists of incidents name ( the CVE number) and patterns (the sentence),

<incidents> = <incidentsname> + <patterns>

Upon this phase, the inquiry will focus on patterns as target. At this point, the DSMarker (the primary marker) is use to identify between the source and target.

<patterns> = <source> |< primarymarker> | <target>
<target> = <targetvector> | <subsequentmarker> |
<targetimpact>

At this point, the inquiry will focus on the first classifier, the sourceroot, to investigate the domain specific wordlist (DSSWordlist) within this classifier. Follow by the next classifiers and their DSSWordlist. The whole process is given as,

<patterns> = <sourceroot>| <subsequentmarker>|
< sourcelocation>|<primarymarker>|
<targetvector>|<subsequentmarker>|
< targetimpact>

This process produced the generative grammar and formed a new notation refer as vulnerability flow diagram as in [38] and later used in the new framework and refer as vulnerability classification pattern or VulClaP.

### *E. Accepting Affirmation*

Accepting affirmation phase is one of most important steps during the entire process, whose task is to perform the validation for the generic vulnerability classification framework.

Vulnerability Classification Pattern Algorithm. The purpose of the algorithm is to support the generic vulnerability classification framework. It should be validated to check its accuracy.

Vulnerability Classification Framework. This framework served as the final product, therefore an affirmation to accept it is important and execute in this phase.

Experimental Result. This result come from the quantitative approach to show the accuracy by using the precision and recall rate. Five random datasets were used and each consist 500 records.

Validation Result. This result comes from the qualitative approach to show the acceptance of the experts by using a questionnaire. The experimental result also was supported with feedback from the experts.

## VI. DATA COLLECTION

This study intends to observe the vulnerability pattern from a reported threat or attack incidents. These incidents are collected and monitored by few organizations such as NVD, CVSS, and CWE (Common Vulnerabilities and Exposures: The Standard for Information Security Vulnerability Names). Used as the case study, these reports were assessed and debated by the experts in the field.

In order to study the pattern of the reported incident, a sample is needed based on certain stratum. The target population is aim within the first five years of the reports. In this case, the random sampling is used. The size shall determine the generality of the results and the ability to detect true effects. The data collection must also adhere to the conditions as in Table 8.

TABLE VIII. DATA CHARACTERISTICS

| Characteristics | Description |
|---|---|
| Reported incidents | The dataset must be reported of original incidents |
| Endorsed by authorized organization | The data must be reported and verified by an authorized security organization |
| Open to public | The data is open for public as reference |
| Sample size must be more than 500 data | The quantity of sample in a data must be able to represent all possible patterns |

Table 8 shows that the data to be used in this study must be reported incidents, endorsed by authorized organization which having their own of third party security auditor, and should be open to public as it indicates a collective and ability to produce a predictive trend [39]. In market, the vulnerability database can be from local authorities or international authorities. The local authorities are varies and comes from both profit-based organisation and nonprofit-based organisation. However, the samples available are limited. Thus, this study turns to databases maintained by international authorities. There are two major open vulnerability databases that have international authorities:

- Open Source Vulnerability Database (OSVDB)
- National Vulnerability Database (NVD)

The OSVDB is an open to public, independent vulnerability database which founded in 2002. However, recently, in April 2016 this database had been closed as announced in an article by Jon Gold at www.networkworld.com. Hence, this research used the reported vulnerability database of Common Vulnerability Exposure or also known as CVE, as uses by numbers or researchers including [40],[41],[3] and [42]. Even though, there are some inconsistency issues in CVE as mentioned in their website, it is treat as minor and has been overcome during the implementation process [42]. In addition, the database is being used in security research such as [43] and [44] due to its ability to produce trends in threat or vulnerability. Nevertheless, the database also comprised a number of security organizations who found, reported and confirmed the cases such as, SANS and they claimed that despite the issues, the CVE has been used as a de facto standard in security industry [45].

### *A. Threat Model versus CVE*

According to the current CVE website, the data introduce 13 categories of vulnerabilities. They are: Denial of Service (DoS), Code Execution, Overflow, Memory Corruption, Sql Injection, XSS, Directory Traversal, Http Response Splitting, Bypass something,Gain Information, Gain Privileges, CSRF and File Inclusion. An example of listing from year 1999 to 2008 is given in Fig. 6.

Vulnerabilities By Type

| Year | # of Vulnerabilities | DoS | Code Execution | Overflow | Memory Corruption | Sql Injection | XSS | Directory Traversal | Http Response Splitting | Bypass something | Gain Information | Gain Privileges | CSRF | File Inclusion |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1999 | 894 | 177 | 112 | 172 | | | 2 | 7 | | 25 | 16 | 103 | | |
| 2000 | 1020 | 257 | 208 | 206 | | 2 | 4 | 20 | | 48 | 19 | 139 | | |
| 2001 | 1677 | 403 | 403 | 297 | | 7 | 34 | 123 | | 83 | 36 | 220 | | 2 |
| 2002 | 2156 | 498 | 553 | 435 | 2 | 41 | 200 | 103 | | 127 | 74 | 199 | 2 | 14 |
| 2003 | 1527 | 381 | 477 | 371 | 2 | 49 | 129 | 60 | 1 | 62 | 69 | 144 | | 16 |
| 2004 | 2451 | 580 | 614 | 410 | 3 | 148 | 291 | 110 | 12 | 145 | 96 | 134 | 5 | 38 |
| 2005 | 4935 | 838 | 1627 | 657 | 21 | 604 | 786 | 202 | 15 | 289 | 261 | 221 | 11 | 100 |
| 2006 | 6610 | 893 | 2719 | 663 | 91 | 967 | 1302 | 322 | 8 | 267 | 271 | 184 | 18 | 849 |
| 2007 | 6520 | 1101 | 2601 | 953 | 95 | 706 | 884 | 339 | 14 | 267 | 323 | 242 | 69 | 700 |
| 2008 | 5632 | 894 | 2310 | 699 | 128 | 1101 | 807 | 363 | 7 | 288 | 270 | 188 | 83 | 170 |

Fig. 6. CVE Vulnerability Type - An Example from 1999-2008 (source: CVE website)

Fig. 6 shows the vulnerability type from MITRE CVE website. However, it was admitted in their FAQ page that this preliminary classification and categorization of vulnerabilities were too rough to be used to identify and categorize the functionality offered. Therefore, a mapping to a threat model is needed to fine grain the categories. As mentioned in section 2.5 at Table 2 and comparison in [27], threat model are varies, and the STRIDE model is light but comprehensive and less likely to be extended but convinced enough to be used in major model threat cases - a reasonable fact to choose this model as mentioned in [46] and [47]. The mapping from this category to STRIDE model is given in Table 9.

TABLE IX. MAPPING CVE TYPE TO STRIDE MODEL

| STRIDE | CVE Threat Type |
|---|---|
| Spoofing | Http Response Splitting, Sql Injection, XSS |
| Tampering | Memory Corruption, File Inclusion |
| Repudiation | Bypass something, CSRF |
| Information Disclosure | Gain Information, Directory Traversal |
| Denial of Service | DoS, Overflow |
| Elevation of privilege | Gain privilege, Code Execution |

Table 9 shows the threat type as defined in CVE website which made it maps to the threat model of STRIDE. This mapping is used as threat type total when random selection of data was made to allow VulClaP execute on it. Meaning, for any randomly selected CVE record, it will be checked on what threat type that they roughly been categorized in before compare with the execution result for that record.

## VII. RESULT

This section explains the results discussion of experimental result and validation result.

### A. Experimental Output

The analysis was conducted on three datasets: DS1, DS2 and DS3. Each dataset contains 500 data. Data set DS1 consist the data from year 1991-2004. Data set DS2 consist data from year 2005-2008 and dataset DS3 contains data of 2009-2016. All data were randomly selected based on a simple random function. The CVE Threat Type for all of them were checked by referring the CVE number. Next, they are put into the STRIDE class. The details for each dataset are given in Table 10.

TABLE X. DATA SET FROM CVE – DS1, DS2 AND DS3

| From CVE Threat Type to STRIDE | DS1 | DS2 | DS3 |
|---|---|---|---|
| Spoofing | 10 | 10 | 8 |
| Tampering | 58 | 145 | 238 |
| Repudiation | 105 | 43 | 19 |
| Information Disclosure | 38 | 74 | 35 |
| DDoS | 96 | 118 | 98 |
| EoP | 123 | 89 | 85 |
| Not classified | 70 | 21 | 17 |
| Total | 500 | 500 | 500 |

Table 10 listed the numbers of patterns that been successfully gained from CVE database after mapping their category with STRIDE. Those that not detected, is categorized as Not classified. For example, in CVE-2002-1932: Microsoft Windows XP and Windows 2000, when configured to send administrative alerts and the 'Do not overwrite events (clear log manually)'option is set, does not notify the administrator when the log reaches its maximum size, which allows local users and remote attackers to avoid detection. This incident was not able to be classified by the algorithm therefore considered as Not classified. The algorithm had successfully detected the words that associate with attackers, analyzed them but however, could not mapped them to associate with TargetImpact category.

The rest of this subsection presents the execution of the vulnerability pattern algorithm into the CVE datasets. The execution was conducted on two modes: a) the mix-based datasets and b) the vendor-based datasets. The mix-based datasets refer to the random selection of data with different range of years. The vendor-based datasets refer to the difference of problems reported by either hardware or software vendors. For the purpose of this paper, we present the discussion on the mix-data set as the space constraints. The analysis was conducted on three datasets: DS1, DS2 and DS3. Each dataset contains 500 data. Data set DS1 consist the data from year 1991-2004. Data set DS2 consist data from year 2005-2008 and dataset DS3 contains data of 2009-2016.

The details for each dataset are given in Table 11. Table 11 listed the numbers of patterns that been successfully gained from each dataset by using the vulnerability classification for patterns (VulClaP) algorithm – patterns are classified into six classes with another one class that is Not classified. Those that not detected, is categorized as Not classified. For example, in CVE-2002-1932: Microsoft Windows XP and Windows 2000, when configured to send administrative alerts and the 'Do not

overwrite events (clear log manually)'option is set, does not notify the administrator when the log reaches its maximum size, which allows local users and remote attackers to avoid detection. This incident was not able to be classified by the algorithm therefore considered as Not classified. The algorithm had successfully detected the words that associate with attackers, analyzed them but however, could not mapped the avoid and detection to associate with TargetImpact category. The accuracy of the model is measured using two metrics: precision and recall rate.

Precision is the degree of confidence that the returned patterns are accurate when the VulClaP algorithm is applied on the data set. Recall is the degree of the ability to return the patterns when the VulClaP algorithm is applied on the data set. As given in the definition, between the two rates, the precision rate suggests a better understanding of accuracy in the model. Therefore, in later discussion the accuracy rate is used interchangeably with the precision rate. For the purpose of this paper, the discussion will use the first data set only, DS1. The data has been analyzed and the summary is given in Table 11.

Table 11 shows the precision and recall rate for DS1 data set. This table informs about how good or bad the prediction could be made from the 500 samples of incidents that range from year 1999 to 2004. The left column is the number of actual pattern classes from the data. The middle column is the number of true selected pattern that gained when analyzes using the VulClap. The aim of the analysis is to predict the value of the pattern based on several input of classifiers, which, in this research are the four classifiers SourceRoot, SourceLocation, TargetVector and TargetImpact. And the right most columns are the total counts of the predicted pattern and their precision rate, or accuracy.

TABLE XI. THE PRECISION AND RECALL RATE FOR DS1 DATA SET

| **Actual Pattern** | **True Selected Pattern** | | | | | | | **Total** | **P Rate** |
|---|---|---|---|---|---|---|---|---|---|
| | **S** | **T** | **R** | **I** | **D** | **E** | **NC** | | |
| S | 8 | 0 | 0 | 0 | 0 | 0 | 2 | 10 | 0.80 |
| T | 1 | 53 | 1 | 1 | 1 | 0 | 1 | 58 | 0.91 |
| R | 0 | 2 | 100 | 0 | 0 | 3 | 0 | 105 | 0.95 |
| I | 0 | 1 | 1 | 34 | 0 | 2 | 0 | 38 | 0.89 |
| D | 0 | 2 | 0 | 0 | 93 | 0 | 1 | 96 | 0.97 |
| E | 0 | 1 | 3 | 2 | 0 | 116 | 1 | 123 | 0.94 |
| NC | 1 | 1 | 0 | 0 | 0 | 1 | 67 | 70 | 0.96 |
| **Total** | 10 | 60 | 105 | 37 | 94 | 122 | 72 | 500 | 0.92 |
| **R Rate** | 0.80 | 0.88 | 0.95 | **0.92** | 0.99 | 0.95 | 0.93 | | 0.92 |

The last rows are the total counts of the recalled patterns and their recall rates. The last cell diagonally at bottom right denotes the sample size from either the sum total from row of precision or recall .The bold values are the average rate for the precision and recall rates. In this data set, the average rate for precision and recall are both 0.92. The shaded areas are the count of true predicted pattern in proportion of the total recall or precision. Or, reading top-down, the analysis also shows that VulClaP algorithm had recalled 105 samples to be in class Repudiation, denotes by the last total row. And from that total, 100 samples are truly recalled as the Repudiation class, and another five samples have been falsely recalled as EOP (3 samples), Tampering (1 sample) and Information Disclosure (1 sample), make the recall rate as 0.92.

### *B. Expert Validation*

The validation was also done through expert opinion to verify the correctness of the classification categories. The experts are the people who is working in academic or security industry with more than five years of experience, and may have additional related professional certificate on top of their knowledge and job experience. Five experts were selected to argue on the vulnerability categories and three experts validate on the frameworks pragmatic. The results justified that issues such as ambiguity of words, redundancy in category, the influenced of certain conjunction such as 'and' , 'or', the generality of certain terms regardless of the application version, the used of newer words are still able to be classified its respected class. In addition, the slight changes of CVE sentence structure and the usage of uncommon word were also highlighted in the expert questionnaire and resolved as less significant.

## VIII. CONCLUSION

After the proposed work on the formal syntax and semantic by using the BNF and the relationship notation, two activities under analysis process are enhanced. The enhanced framework is given in Fig. 7.

The red activities show the enhanced part. The issue of high potential in redundancy is resolved by using a formal BNF to represent the syntax and semantic schema. And the relationship of vulnerability classifier is representing using the vulnerability flow diagram.

In particular, this study contributes to the software industry by assisting to formulate a secured software process framework with vulnerability classification algorithm and vulnerability flow diagram - an area that has been underestimating by researches due to the difficulty to generalize the outcomes. The significant of this secured framework that makes it different from others is the ability of the vulnerability flow diagram to visualize the precedent and antecedent of an exploit using the simplified vulnerability pattern.

The main challenging task in this study is to formulate the domain specific wordlist and domain specific schema for the classification which is proposed to be further look into in future work.

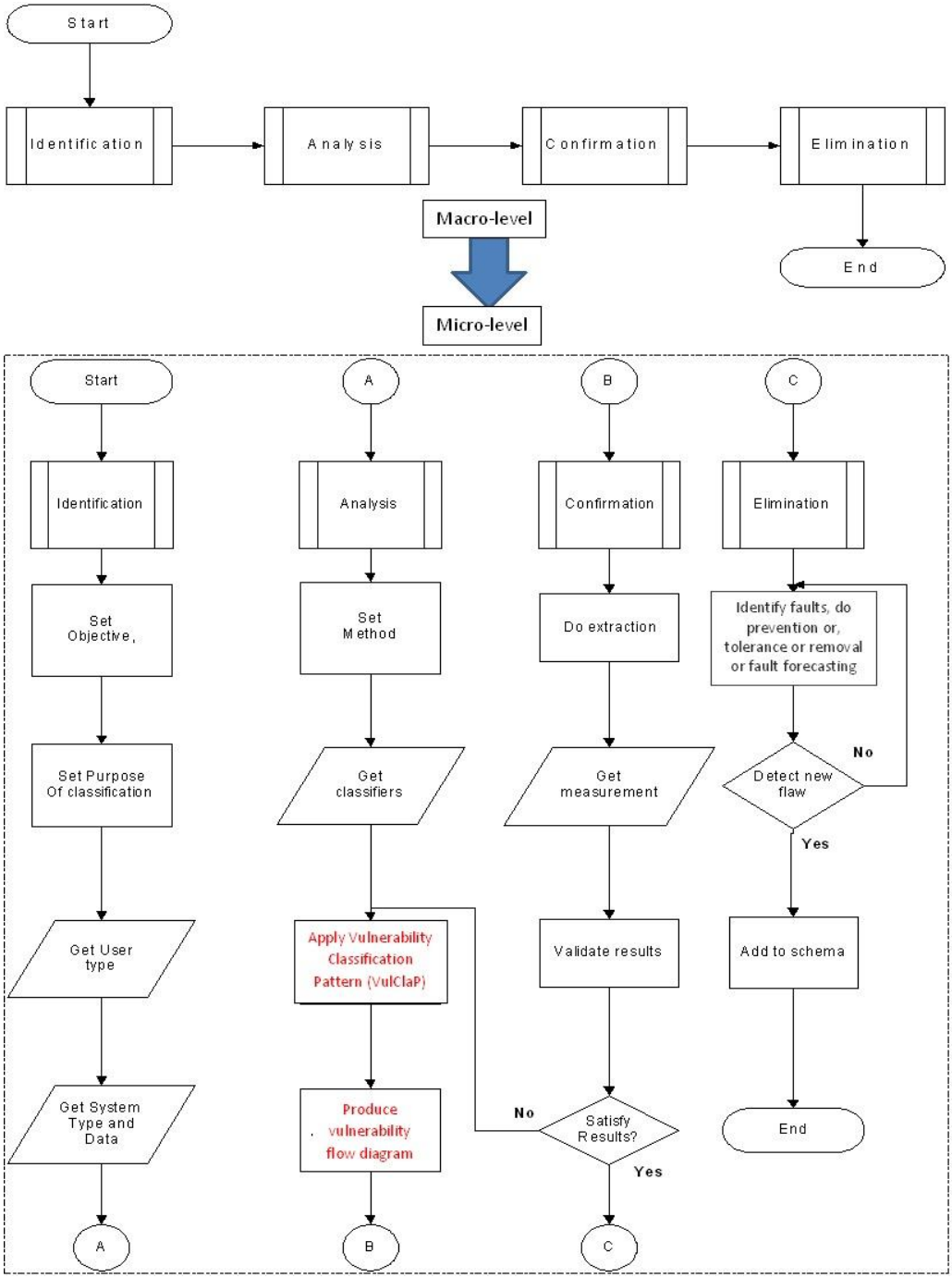



Fig. 7. The Enhanced Vulnerability Classification Framework (E-VulCaF)

## ACKNOWLEDGMENT

This project is supported by grant of PJP/2018/FTMK(4B)/S01631 from Center For Research and Innovation Management (CRIM), UTeM.

## REFERENCES

[1] O. Rebolloa, D. Melladob, E. Fernandez-Medinac, and H. Mouratidis, "Empirical Evaluation Of A Cloud Computing Information Security Governance", Information and Software Technology, vol. 58(3): pp. 44–57, 2015.

[2] R. Shaikh and M. Sasikumar, "Data classification for achieving security in cloud computing", Procedia Computer Science International Conference on Advanced Computing Technologies and Applications (ICACTA), vol.45, pp. 493 – 498, 2015.

[3] J. Ruohonen, S. Hyrynsalmi, S. Rauti, and V. Leppänen, "Mining Social Networks Of Open Source Cve Coordination", Proceedings of the 27th International Workshop on Software Measurement and 12th International Conference on Software Process and Product Measurement, IWSM Mensura '17, ACM, New York, NY, USA, pp. 176–188, 2017.

[4] A. Tripathi and U. Singh, "Taxonomic Analysis Of Classification Schemes In Vulnerability Databases", Computer Sciences and Convergence Information Technology (ICCIT), 6th International Conference on, pp. 686–691, 2011.

[5] E.W. Burger, M. D. Goodman, P. Kampanakis, and K.A. Zhu, "Taxonomy Model For Cyber Threat Intelligence Information Exchange Technologies", Proceedings of the ACM Workshop on Information Sharing, vol.38, pp. 51–60, 2014.

[6] R. Abbott, J. Chin, J. Donnelley, W. Konigsford, S. Takubo and D. Webb, Security Analysis and Enhancement of Computer Operating Systems, NBSIR pp.76-1041, 1976.

[7] R. Bisbey, and D. Hollingworth. "Protection Analysis: Final Report", Technical Report ISI/SR-78-13, Technical report, Information Sciences Institute, University of Southern California, 1978.

[8] E. L.Carl, R.B. Alan, P.M. John, and S.C. William, "A Taxonomy of Computer Program Security Flaws", ACM Computer Survey vol.26, pp. 211–254, 1994.

[9] Aslam, T., Krsul, I. and Spafford, E., "Use of A Taxonomy of Security Faults", 19th National Information Systems Security Conference, 1996.

[10] J. C. Munson, A.P. Nikora, and J.S Sherif, "Software Faults: A Quantifiable Definition", Advances in Engineering Software vol.35, pp.327–333, 2006.

[11] G. S. Walia, and J.C. Carver, "A Systematic Literature Review to Identify and Classify Software Requirement Errors", Information and Software Technology vol.51 pp. 1087–1109, 2009.

[12] J. Luo, K. Lo, and H.Qu, "A software vulnerability rating approach based on the vulnerability database", pp 1–9, 2014.

[13] M. Bishop, "Classifying Vulnerabilities", 1999.

[14] C. Weissman, "System Security Analysis Certification Methodology and Results", Technical Report SP-3728, Technical report, System Development Corporation, 1973.

[15] D.L. Lough, Thesis: A Taxonomy of Computer Attacks with Application to Wireless Network, Technical report, Virginia Polythechnic Institute, 2001.

[16] R.C. Seacord, and A.D. Householder, "A Structured Approach to Classifying Security Vulnerabilities", CMU/SEI-2005-TN-003, Technical report, Carnegie Mellon, Software Engineering Institute, 2005.

[17] S.Eagle, S.Whalen, D. Howard, and M. Bishop, "Tree approach to vulnerability classification", Technical Report CSE-2006-10, Technical report, Department of Computer Science, UC Davis, 2006.

[18] B. Biggio, , G. Fumera, and F. Roli, "Security Evaluation Of Pattern Classifiers Under Attack", IEEE Trans. Knowl. Data Eng. vol.36, pp. 984–996, 2014.

[19] A. Bollin, "Crossing the borderline- from formal to semi formal specifications", IFIP of Software Engineering Techniques: Design For Quality, Vol. 227, pp. 73–84, 2006.

[20] L. Lowis, and R. Accorsi, "Vulnerability Analysis In Soa Based Business Processes", IEEE Transaction on Services Computing, vol.4, pp.230–242, 2011.

[21] D. Last, "Using Historical Software Vulnerability Data To Forecast Future Vulnerabilities", IEEExplore, 2015.

[22] C. Haley, R. Laney, J. Moffett, J. and B.Nuseibeh, "Security Requirements Engineering:A Framework For Representation And Analysis", IEEE Transactions on Software Engineering, vol.34, pp.133–153, 2008.

[23] Lamsweerde, A. v., Brohez, S., Landtsheer, R. D. and Janssens, D., 2003. From System Goals to Intruder Anti-Goals: Attack Generation and Resolution for Security Requirements Engineering, Requirements for High Assurance Systems (RHAS'03), pp. 49–56.

[24] R. Villarroel, E. Fernandez-Medina, and M. Piattini, "Secure Information SystemsDevelopment - a Survey and Comparison", Computers & Security ,vol.24, pp.308–321, 2005.

[25] V.M. Igure and D.R. Williams, "Taxonomies of Attacks and Vulnerabilities in Computer Systems", IEEE Communication Surveys & Tutorials, vol. 10, pp 14, 2008.

[26] J. Maatta, J. Harkonen, T.Jokinen, M. Mottonen, P. Belt, M. Muhos, and H. Haapasalo, "Managing Testing Activities in Telecommunications: A Case Study", J. Eng. Technol.Manage. vol. 26: pp.24, 2009.

[27] N.H.Hassan, S.R. Selamat , S.Sahib and B.Hussin (2011) "Towards Incorporation of Software Security Testing Framework in Software Development". In: Mohamad Zain J., Wan Mohd W.M.., El-Qawasmeh E. (eds) Software Engineering and Computer Systems. ICSECS 2011. Communications in Computer and Information Science, Springer, Berlin, Heidelberg, vol 179, pp.16-30, 2011.

[28] A.Schipper, H.Fuhrmann, and R.V. Hanxleden, "Visual comparison of graphical models", vol.10, pp.335–340, 2009.

[29] F.A. Braz, E.B. Fernandez, E. B. and M.VanHilst, "Eliciting Security Requirements through Misuse Activities", 19th International Conference on Database and Expert Systems Application, pp. 328–333, 2008.

[30] M. Gupta, J. Walp, and R. Sharman, R., "Threats, Countermeasures and Advances in Applied Information Security", IGI Global, 2012.

[31] R.R. Linde, "Operating System Penetration", National Computer Conference, System Development Corporation, pp.8, 1975.

[32] S. Hansman, and R. Hunt, "A Taxonomy Of Network And Computer Attacks", Computers and Security vol.24, pp. 31 – 43, 2005.

[33] T. Katrina, C. Brian, and M. Gary, "Seven pernicious kingdoms: a taxonomy of software security errors", IEEE Security and Privacy, vol. 3, pp.81–84, 2005.

[34] K. Jiwnani, and M. Zelkowitz, "Maintaining Software with a Security Perspective", Proceedings of International Conference on Software Maintenance, pp. 194–203, 2002.

[35] B.Potter, and G. McGraw, "Software Security Testing", Security & Privacy, IEEE vol.2,pp. 81–85, 2004.

[36] L.Desmet, P. Verbaeten, W. Joosen, and F. Piessens, "Provable Protection against Web Application Vulnerabilities Related to Session Data Dependencies", IEEE Transactions on Software Engineering, vol.34, pp. 50–64, 2008.

[37] A.M. Memon, " A comprehensive framework for testing graphical user interfaces",2001.

[38] N.H.Hassan, S.R. Selamat and S.Sahib, "Establishing the Relationship in Vulnerability Classification for a Secure Software Testing", International Conference on Advances in Intelligent Systems in Bioinformatics, Chem-Informatics, Business Intelligence, Social Media and Cybernetics (IntelSys), 2014.

[39] K. Alzhrani, E.M. Ruddy, T.E. Boulty, and C.E. Chow, "Automated Big Text Security Classification", IEEE Intelligence and Security Informatics, pp. 1–6, 2016.

[40] Z. Moghbel, and N. Modiri, "A Framework For Identifying Software Vulnerabilities Within Sdlc Phases", International Journal of Computer Science and Information Security, vol.9, pp. 203, 2011.

[41] U.K. Singh, and C. Joshi, "Network security risk level estimation tool for information security measure", IEEE 7th Power India International Conference (PIICON), pp. 1–6, 2016.

[42] D.Papp, Z. Ma, and L. Buttyan, "Embedded Systems Security: Threats, Vulnerabilities, And Attack Taxonomy", 13th Annual Conference on Privacy, Security and Trust, Institute of Electrical and Electronics Engineers Inc., pp. 145–152, 2015.

[43] S.Neuhaus and T.Zimmermann, "Security Trend Analysis With CVE Topic Models", IEEE 21st International Symposium on Software Reliability Engineering (ISSRE), pp. 111–120, 2010.

[44] A.Barua, S.W. Thomas, and A.E. Hassan, "What are developers talking about? An analysis of topics and trends in stack overflow", Empirical Software Engineering, vol.19, pp. 619–654, 2014.

[45] J.Connolly, M. Davidson, M. Richard, and C. Skorupka, "The Trusted Automated Exchange Of Indicator Information", MITRE Corporation, pp. 1–20, 2012.

[46] R. Khan, K. McLaughlin, D. Laverty, D. and S. Sezer, "STRIDE-based Threat Modeling for Cyber-Physical Systems", IEEE, 2018.

[47] M. Ostkamp, C. Kray, and G. Bauer, "Towards A Privacy Threat Model For Public Displays", Proceedings of the 7th ACM SIGCHI Symposium on Engineering Interactive Computing Systems, EICS ACM, pp. 286–291, 2015.

# Intrusion Detection System with Correlation Engine and Vulnerability Assessment

D.W.Y.O.Waidyarathna[1], W.V.A.C.Nayantha[2], W.M.T.C.Wijesinghe[3], Kavinga Yapa Abeywardena[4]
Department of Information System Engineering
Sri Lanka Institute of Information Technology
Kandy Road, Malabe, Sri Lanka

***Abstract*—The proposed Intrusion Detection System (IDS) which is implemented with modern technologies to address certain prevailing problems in existing intrusion detection systems' is capable of giving an advanced output to the security analyst. Even though the network of an organization has been secured internally as well as externally the intruders find ways to penetrate the network. With the system that is proposed activities of those intruders can be identified with a higher probability even if managed to bypass security controls of the network. The goal of this project is to give a reliable output to the system users where all the alerts are more accurate and correlated using HIDS alerts and NIDS alerts which is similar to the modern SIEM concept. The system will perform as a centralized IDS by getting inputs from both HIDS and NIDS which gives data regarding the activities of hosts and network traffic. With those implementations, the system is capable of monitoring host activities, monitoring network traffic with existing tools and give a correlated output which is more accurate, advanced and reliable prioritizing the possible attacks by using machine learning techniques and rule-based correlation techniques. With all these capabilities final product is a fully automated Intrusion Detection System which gives correlated alerts as outputs with a less rate of false positives compared to the existing systems.**

***Keywords*—Intrusion detection system (IDS); intrusion detection message exchange format (IDMEF); network intrusion detection system (NIDS); host intrusion detection system (HIDS); security information and event management (siem); correlation; machine learning***

## I. Introduction

With ever growing technological solutions, computer systems and computer networks play a major role in the world. Today's enterprises, businesses, and organizations have mainly automated their old manual systems with the computerized solutions. As a result, while doing the processing, a lot of data is generated and consequently stored within these systems. Almost all devices of an organization are network aware and multiple devices are connected to the outside world via technologies such as virtual private networks and internet allowing outsiders to connect into the internal network of the organization. All these external users and processes, connected devices, networks and systems, security has become a major requirement in the field of information technology.

With these interconnected systems, there is a persistent risk for the organization as the probability of an attacker penetrating the network or sending a malicious payload to a system is comparatively high. Physical security is the first step of securing a network, however for a skilled intruder finding a small weak point to enter the internal network is not a challenge. Due to this concern physical security is merely not enough, hence internal security mechanisms are constantly applied and reviewed for network and system protection. Access Control Systems, Security Incident & Event Management Systems, Intrusion Prevention Systems, Intrusion Detection System and many more concepts are out there which are deployed against the intrusions done by various types of intruders. In this research, authors have focused on addressing prevailing drawbacks that exist in current Intrusion Detection Systems.

An intrusion detection system (IDS) is a device or software application that monitors a network or systems for malicious activity or policy violations. Any malicious activity or violation of a policy is typically reported either to an administrator or collected centrally using a Security Information and Event Management System (SIEM) [1]. Intrusion Detection Systems are mainly can be categorized as Host based Intrusion Detection Systems and Network Intrusion Detection Systems. A host-based intrusion detection system (HIDS) is an intrusion detection system that is capable of monitoring and analyzing the internals of a computing system such as user, operating system and application activities [2]. A network intrusion detection system (NIDS) monitors traffic on a network looking for suspicious activity, which could be an attack or unauthorized activity [3].

Another way of categorizing is according to the detection method, one of the most common way is signature-based detection which looks for known patterns of malicious network packets and malicious activities, referred to as signatures [4]. The other option available can be described as the anomaly-based detection which detects any abnormal behavior varies from the normal/legitimate traffic and operations [5]. Considering the above-mentioned detection methods, there are commonly studied problems especially with the signature-based models. For an example, most of the times systems tend to give false negatives as there is no way of identifying newly constructed attacks due to limited detecting capability with the known type of attacks/patterns. These newly constructed attacks are commonly referred to as zero day exploits. There is no way of detecting new attacks with signature based models [4]. The behavioral based models always try to generate alerts by identifying malicious traffic and operations even though it might belong to legitimate

events and network traffic. As a result, it creates a higher false positive rate [5]. Therefore, presently there is no perfect solution in the information security industry with respect to intrusion detection.

Even though security engineers cannot fully rely on the existing intrusion detection systems, the number of new types of attacks and the amount of attacks developed in the world hasn't been decreasing. Instead, zero day attack rate has increased considerably during the past few years. In order to address this problem, the security domain requires an advanced intrusion detection system that all types of organizations can afford without spending over the top. This solution should be capable of detecting intrusions with a higher rate of accuracy while maintaining a lower rate of false positives. Modern developments in the field of computing should be utilized to achieve this task.

## II. Research Gap and Research Problem

Traditional Intrusion Detection Systems are mostly built in a way that they identify attacks by using 'signatures' and 'anomalies'. Even though some researchers have published papers and ideas regarding applying machine learning techniques for existing Intrusion Detection Systems, any of those systems are not performing well in a way that it can stand against to the current cyberwar [6]. When we analyzed the existing systems we could discover the main problems that prevailing the existing IDSs. Those limitations can be listed as,

- A high rate of false positives and false negatives
- Separate NIDS and Separate HIDS
- No correlation of NIDS and HIDS alerts, using both rule-based and machine learning techniques
- Lack of integrated vulnerability management
- Small-scale businesses can't afford existing IDS systems due to the high cost of implementation

With these limitations, it has been difficult to achieve the core security components through the Intrusion Detection Systems. When these systems give a higher rate of false positives, security analysts cannot depend on those results. They have to manually do the process that was done by the IDSs by diving deeper into the system and analyzing raw logs [8]. It consumes a lot of time as well as it makes the intrusion detections system useless when the task is redone manually by a person.

The Intrusion Detection Systems always act separately as Host Intrusion Detection Systems or Network Intrusion Detection Systems [6]. However, with modern threats and their undiscoverable quality, if we can merge these two systems into one then the output shall be more effective. This can be defined as the main objective of this proposed solution. The proposed system attempts to correlate the alerts received from both HIDS and NIDS which are received in a common format, processed in an optimized environment and presented in a meaningful manner.

As far as the IDS functionality is concerned it is important to understand why and how these attacks are managing to bypass the existing solutions. Usually for each attack there exist major known vulnerability of a system i.e. the real cause of an attack is some critical vulnerability of the systems. However, most of the current Intrusion Detection Systems does not address the vulnerability management area [7]. Hence those IDSs are not equipped with any integrated vulnerability management/assessment tool. This also should be identified as a major concern with existing solutions.

## III. Methodology

The goal of this research was to develop a system which is mainly capable of detecting intruder activities with a higher rate of accurate alerts by minimizing the number of false positives using both rule-based and machine learning techniques for detecting the intrusion activities.

### A. Detailed Flow

Brief workflow of the proposed system has been depicted below in Fig. 1, and it emphasizes the overall architecture and placement of the solution in an existing network.

### B. Standardization of Alerts

*1) IDMEF*: The Intrusion Detection Massage Exchange Format known as IDMEF is focusing on defining data formats and exchange procedures for information sharing, which is a crucial factor for Security Incident and Event Management Systems. To define a standard representation of alerts, IDMEF is using XML based data models and this identical representation enables interoperability among different devices or systems [9]. The data from different devices are also allowed to be stored accurately by a standardized log format. The main purpose of the research is also focused on improving security by combining Snort NIDS System with OSSEC HIDS System. Many of past researches have used the IDMEF, as a protocol for exchanging intrusion detection massages which are being standardized by the IETF [13]. Intrusion Detection Message Exchange Format is basically an object-oriented depiction of generated alert data by Intrusion Detection Systems. Two types of implementation for IDMEF was proposed by Intrusion Detection Working Group (IDWG) [14]. One method is implemented using the Structure of Management Information (SMI) and the other is using XML.

### C. Optimization of NIDS

*1) Snort as NIDS*: Snort is the de facto standard for network-based intrusion detection. This network intrusion detection system is an open source and rule-driven language, which fuses the benefits of protocol, signature, and anomaly-based inspection methods. As this method is rule-based, it is generally a misuse detection system but apart from that, it has some anomaly recognition capabilities. It supports logging events to either log files or a database.

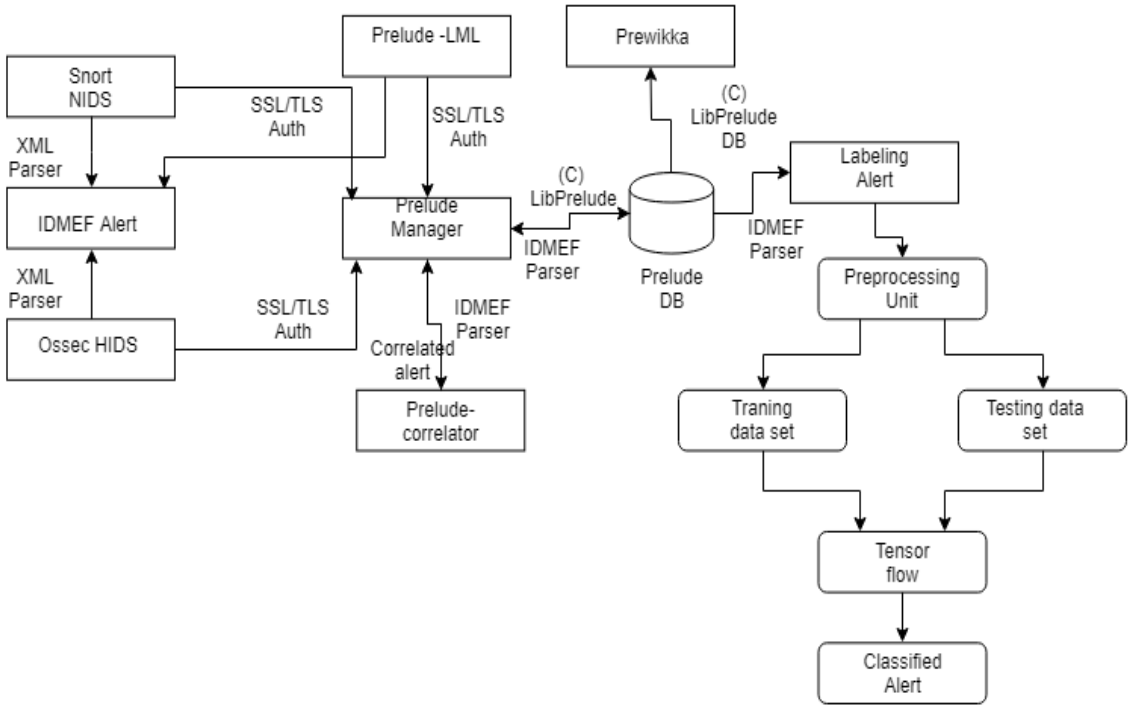



Fig. 1. Overall Diagram of the IDS.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<idmef:IDMEF_Message xmlns:idmef="http://iana.org/idmef version =1.0">
 <idmef:Alert messageid="1">
  <idmef:Analyzer analyzerid="527">
   <idmef:Node category="8" />
  </idmef:Analyzer>
 </idmef:Alert>
 <idmef:CreateTime ntpstamp="06/16-11:22:28.515219" />
 <idmef:Source>
  <idmef:Node>
   <idmef:Address category="ipv4-addr">
    <idmef:address>0.0.0.0</idmef:address>
   </idmef:Address>
  </idmef:Node>
  <idmef:Service>
   <idmef:priority> 2</idmef:priority>
   <idmef:protocol>IGMP</idmef:protocol>
  </idmef:Service>
 </idmef:Source>
 <idmef:Target>
  <idmef:Node>
   <idmef:Address category="ipv4-addr">
    <idmef:address>224.0.0.22</idmef:address>
   </idmef:Address>
  </idmef:Node>
  <idmef:Service>
   <idmef:protocol>IGMP</idmef:protocol>
  </idmef:Service>
 </idmef:Target>
 <idmef:Classification text=" Potentially Bad Traffic" />
</idmef:IDMEF_Message>
```

Fig. 2. Snort Rules.

NIDS is frequently placed between the edge firewall and a back-esnd firewall that protects the internal network from the publicly accessible network in between, called the DMZ or perimeter network or screened subnet [2].

Snort is configured to operate on NIDS mode. Whereas in Network IDS mode, Snort executes actual analysis to determine malicious traffic, based on that alerts are generated. To conduct testing DARPA 1998 datasets were downloaded from MIT Lincoln Labs website. Furthermore, the dataset comprises of replicated network traffic embedded with marked attacks. Snort was configured in the Network Intrusion Detection System to use this dataset. Example IDMEF messages obtained from a Snort alert file is shown below in Fig. 2.

### *D. Optimization of HIDS*

*1) OSSec as HIDS:* OSSEC is a host-based intrusion detection system. It has a powerful log analysis, registry monitoring, and integrity checking and rootkit detection engine. It is a signature-based IDS, which detects intrusions based on rules [12]. The following Fig. 3 depicts the OSSEC architecture.

The system is using OSSEC agents to collect logs in all the hosts and pass into the OSSEC server for the analysis process. In the server side, received logs are decoded in two phases which are pre-decoding and decoding as in the Fig. 3 depicted above [15]. Pre-decoding phase is used to extract static information like time, hostname, log message and date from received events. Non-static information like event ID, source

and destination IP addresses are extracted in the decoding phase [11]. After that decoded data are passed to analyzing phase. Analyzing is performed through rule matching, based on predefined signatures this extracted information are matched. If malicious patterns are detected, it will be stored these data in the database in IDMEF format for correlation and alerting purposes.

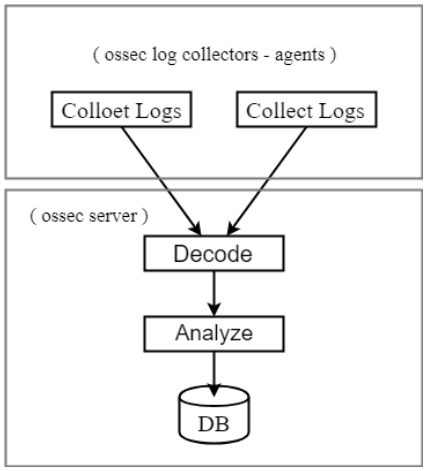



Fig. 3. OSSec Architecture.

### E. Alert Correlation of HIDS and NIDS

Deployment of one sensor and then consequent alert generation to the network administrator is the most basic setup of the NIDS. The bigger network setups frequently need to use more than one sensor to cover the entire network. Presuming that, all of these sensors used in the above mentioned basic setup, this will result in a large number of alerts being directed to the administrator. The most common solution for this problem is to deploy a database in a central server to store all the generated messages sent by sensors. Then the central server can send status reports to the administrator. This is an exhaustive process which causes excessive information generation. This is where the concept of alert correlation fits in. By correlating the alerts coming from different sensors, information can be merged together to reduce the volume of information. It can also help in detecting attacks which are going to be missed otherwise. When distributed attacks are needed to identify several nodes of different subnets, sensors distributed across the network can be used.

Prelude is a Security Incident and Event Management (SIEM) system which allows us to achieve correlation capabilities easily. It has a rule-based correlation engine known as Prelude Correlator. Prelude also has a relatively user-friendly interface called Prewikka to present the analyzed data. Hence Prelude is quite useful to group events, identify unique alerts and to identify which of the flagged events have been caught by connected sensors or Prelude itself. The Correlator of Prelude has limited correlation abilities since it has only a limited number of rulesets. The researchers intended to detect various types of advanced attacks using this approach. Authors intended to extend the correlation functionalities of Prelude Correlator by writing a set of custom rules. The correlation process of Prelude SIEM is displayed below in Fig. 4.

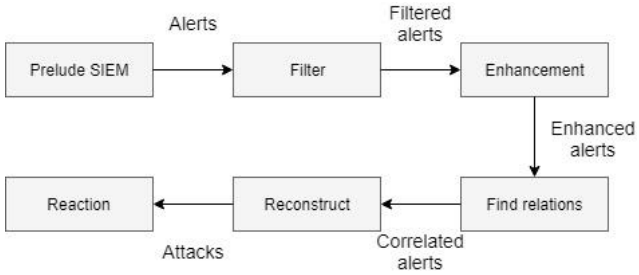



Fig. 4. Correlation Path.

In the proposed system vulnerability assessment tool will be integrated into the main system. This will be implemented via an API. Vulnerability Scanner will be called as a service when the system admin wants to run a vulnerability assessment. The vulnerabilities will be listed down according to the Common Vulnerability Scoring System (CVSS) [7]. Then the user will be notified which vulnerabilities should be addressed first and patches should be applied.

## IV. Implementation

The important contributions of this work were to the preliminary execution of mandatory decoders, the integration of Prelude components, detection modules of Snort and OSSEC sensors, Prelude configurations for Snort and OSSEC analyzers, building classification algorithm, and the monitoring events within the overall IDS framework.

When malicious packets reach the perimeter, packets are captured and immediately sent to the Snort IDS preprocessor for inspection. By detecting the attack, the alerts produce from Snort IDS reach Prelude instantly, provided the particular sensor is already TLS authenticated.

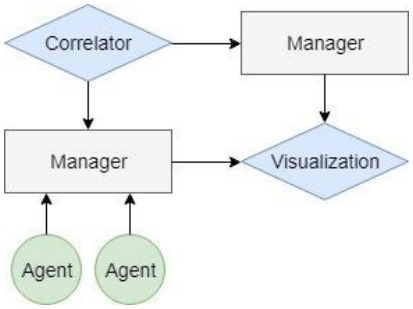



Fig. 5. Prelude Architecture.

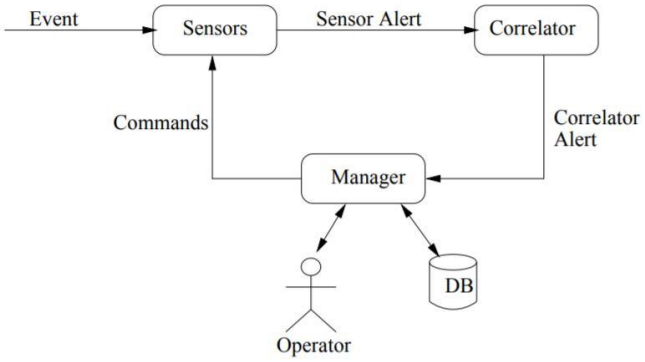



Fig. 6. Prelude Detailed View.

Prelude offers a great flexibility by combining a broad range of security tools under one powerful monitoring system. By correlating received alerts from other monitoring equipment such as Snort, OSSEC etc., it is possible to reduce the false positive alerts generated. Below figures, Fig. 5 & Fig. 6, illustrate the overall architecture of the Prelude SIEM.

All events have been normalized into the Intrusion Detection Message Exchange Format (IDMEF) by Prelude. Events from different devices are allowed to be stored in a structured format by normalization and more importantly, it allows all events collected, to be stored in the same database in the same format. It also makes the stored events well organized in order to maintain all the processed data. It is important to note that it does not need multiple storage devices to achieve this [3].

By reducing overheads and increasing efficiency Prelude helps to reduce security cost as well. Prelude SIEM records the events and filters them to eliminate non-threatening alerts, as well as to see whether if threats are connected via correlation.

Through Prelude web interface Prewikka, real-time event monitoring can be obtained. Manual reviewing of logs can easily result in a missed attack in the past. The real-time event analyzing allows, decrease in response time to the possible incidents. In an organization, it also allows seeing up to date activities easily on their entire network. Most recent events are showing automatically when the web page is set to refresh.

Libprelude gives an API (Application Programming Interface) which allows communicating with the Prelude subsystems for third-party software [10]. If any disturbance takes place between any of the components in the system, libprelude also makes sure that re-transmission of data is performed. Any device acting as a manager or a sensor, the libprelude package is required. As well, converting the logs into Preludes binary IDMEF format also entail. It ensures management servers and sensors use secure transmissions such as Secure Socket Layer/Transport Layer Security (SSL/TLS) to communicate.

For storing IDMEF alerts in the database, libpreludedb is the library that supplies an abstraction layer, and this library makes easier management of the database [10]. It allows the user to access the database without depending on the log format by hiding the inner workings. The hosting machine should require installing libpreludedb, in order to use the Prewikka web interface.

To analyze various different types of logs, Prelude requires Prelude LML (Log Management Library) component. The Prelude LML log analyzer determines whether activity within the logs is malicious by using a set of rules, and it is comparable to the way Snort uses the rules file to analyze packets. The rules files of Prelude LML attempt to match data within the log files instead of network packets.

The component which allows the correlation of events between various Prelude Management Servers is Prelude Correlator. Prelude Correlator is a python rule-based correlation engine and has the ability to connect and fetch alerts from a remote Prelude Manager Server. The users are allowed writing correlation rules using the Python programming language by Prelude Correlator. A correlation message is generated, once the streams of events match a correlation rule.

The web-based GUI (Graphical User Interface) for Prelude is Prewikka interface [10]. When a user logged into Prewikka, the Alerts tab act as the default page. Under the classification column, event summary is listed, by clicking it, a user can view more information about a particular event. This link will display all of the events that match a particular event description, source, target, and sensor. A user can view the actual event detail by clicking on the Threats tab. The Agents tab gives a detailed analysis of the agents which has reported to the Prelude Manager. Then the sensor is currently offline or online will be displayed on the system. The heartbeat analysis and heartbeat listing will appear when clicking on the Heartbeats tab and also it shows a list of recent heartbeats received by Prelude Manager Server. Heartbeat can be defined as a simple message which indicates that the agent is reporting, running and sending messages, and also it ensures whether the device is properly working, although it is not generating alerts.

### *A. Alert Classification using Machine Learning*

The main purpose of the Intrusion Detection System is to differentiate between normal events and attacks as discussed earlier. The common situation about the generation of high false alarms is caused by the most of Intrusion Detection Systems. The research proved that the IDS System is more efficient when it holds a fewer number of false negatives and false positives. The use of machine learning techniques is a one way to deal with this problem, and the machine learning can be used to differentiate between attacks and false alarms [1].

In the proposed system, the format dissimilarity of alerts come from different sensors overcome by using the IDMEF (Intrusion Detection Message Exchange Format) format. Analyzed IDMEF alerts obtained from prelude API, classified into false alarms and true attacks using machine learning techniques [9]. Thereafter alerts from various Intrusion Detection Systems is gathered and the process as follows,

- Obtain collected alerts from common IDMEF format
- Labeling of alerts
- Constructing the dataset
- Modeling of the supervised machine learning algorithm classifier
- Classification of alerts into false alarm or true attacks using the below mentioned machine learning technique

Algorithm for labeling alerts was executed in python, and the labeled alert file is used for categorization, which was attempted using a supervised machine learning algorithm K-Nearest Neighbors (K-NN) classifier. The proposed solution was implemented using the Tensorflow framework to build the algorithm classifier. According to the Fig. 7 below, the KNN classifier has the best accuracy rate [16]. Hence the classification process has adopted the KNN classifier.

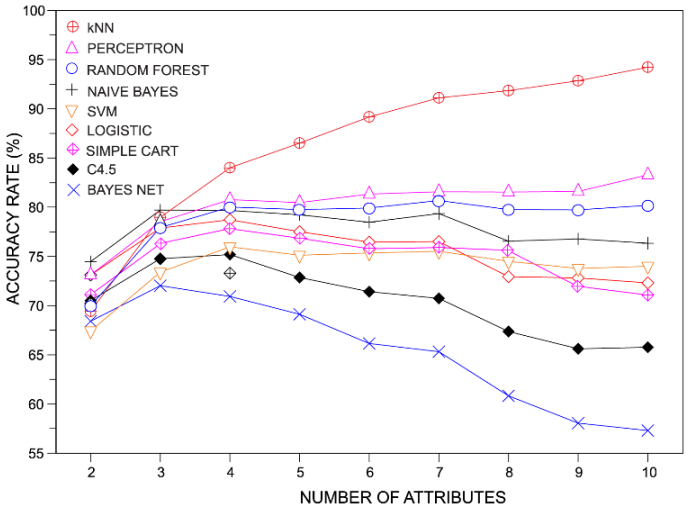



Fig. 7. Different Classifiers Accuracy Rate Comparison Graph.

## V. Results & Discussion

Traditional IDSs available today has its own relative weaknesses and strengths. While one solution may be strong at host-based intrusion detection, the other solution may be strong at network-based intrusion detection. The organizations are highly concerned about their network and system performance; hence they use multiple IDSs from various vendors as they do not wish to take a chance with security. Different IDSs generate alert events in different formats, as well as use different protocols. If the outputs alerts are not integrated properly, false positive rates may increase hence interrupting the legitimate performance of a system or a network. False alarms caused by the large volume of IDSs is intolerable to the administrators as it delays the smooth functioning of an organization. It is necessary to decrease the excessive of false alarms to reduce the operational cost and excel in the reliability of a security system. Hence, this research was conducted intending to advance a procedure to obtain alerts from different sensors and standardizes them into IDMEF.

Rule-based architecture and machine learning techniques were used to compare security events. These methods analyze alerts generated from various sensors, which are normalized and combined into meta-alerts, then it used to classify true alerts or false alarms.

## VI. Future Works & Conclusion

The research intended to introduce an advanced machine learning and rule-based, HIDS and NIDS correlated intrusion detection system. The system gives an optimized and reliable output which creates a fewer false positive rate compared to the past researches and existing IDS solutions. Further research can be conducted in developing an advanced intrusion detection system using the proposed approach. There are various open source IDS tools which can further be integrated with the proposed architecture to compare findings to fin the best possible combination. The overall objective is to achieve a more successful result in order to persevere against the modern types of attacks, which cannot be discovered by the traditional standalone Intrusion Detection Systems.

## Acknowledgment

We would like to acknowledge the Sri Lanka Institute of Information Technology for providing us the necessary lab facilities and the test beds to conduct our research activities. We would also like to acknowledge everyone who continuously supported during the period of our research.

## References

[1] A B. Athira, V. Pathari, “Standardisation and Classification of Alerts Generated by Intrusion Detection Systems”, IJCI, International Journal on Cybernetics & Informatics, Vol 5 Issue 2, 2016.

[2] Johansson Daniel, Andersson Par, “Intrusion Detection Systems with Correlation Capabilities”

[3] Yasm Curt, “Prelude as a Hybrid IDS Framework”, March, 2009

[4] Kumar Vinod, Sangwan Prakash Om, “Signature Based Intrusion Detection System Using SNORT”, IJCAIT, International Journal of Computer Applications & Information Technology, Vol. I, Issue III, November 2012.

[5] Singh Deepak Kumar, Gupta Jitendra Kumar, “An approach for Anomaly based Intrusion detection System using SNORT“, IJSER, International Journal of Scientific & Engineering Research, Volume 4, Issue 9, September 2013.

[6] S, Vijayarani, and Maria Sylviaa S. “Intrusion Detection System – A Study”, IJSPTM, International Journal of Security, Privacy and Trust Management ,Vol 4, Issue 1, pp. 31–44, February 2015.

[7] Yang Guangming, Chen Dongming, Xu Jian, Zhu Zhiliang, “Research of Intrusion Detection System Based on Vulnerability Scanner”, ICACC, Advanced Computer Control, March 2010.

[8] Chakraborty Nilotpal, “Intrusion Detection System and Intrusion Prevention System: A Comparative Study”, IJCBR, International Journal of Computing and Business Research , Volume 4 Issue 2, May 2013.

[9] Kothari Pravin, “Intrusion Detection Interoperability and Standardization”, February, 2002.

[10] TIMOFTE Jack, “Intrusion Detection using Open Source Tools”, Revista Informatica Economică nr.2(46), pp. 75-79, 2008.

[11] A T. Oğuz, R. Maraş, “Host-Based Intrusion Detection Systems Ossec Open Source HIDS”

[12] Hwang Oun Seong, “APT Detection with Host-Based Intrusion Detection System and Intelligent Systems”, FCTA, Future Computational Technologies and Applications, 2017.

[13] Argyroudis, Patroklos, Paraskakis Iraklis, “Distributed Intrusion Detection using Mobile Agents”, 2002.

[14] Carey Nathan, Clark Andrew, Mohay George, “IDS Interoperability and Correlation Using IDMEF and Commodity Systems”, pp. 6-22, 2002.

[15] Hay Andrew, Bray Rory, Cid Daniel, Northcutt Stephen, “OSSEC Host-Based Intrusion Detection Guide”, Syngress, 2008.

[16] Diego Raphael, Cesar Henrique, Francisco Aparecido, "A Systematic Comparison of Supervised Classifiers", PLOS ONE, Public Library of Science, Vol 9, Issue 4, pp. 1-14, April 2014.

# An Expert Comparison of Accreditation Support Tools for the Undergraduate Computing Programs

Abdallah Namoun[1], Ahmad Taleb[2], Mohamed Benaida[3]
Faculty of Computer and Information Systems
Islamic University of Madinah
Madinah, Saudi Arabia

***Abstract*—Realizing continuous quality improvement within educational programs is a challenging task. However, there exist various assessment tools and models that help in this regard. This paper explores the features and capabilities of three major international accreditation support tools and compares their strengths and weaknesses. The investigated tools include EvalTools, CLOSO, and WEAVEonline. Two education quality experts performed a thorough comparison of the three tools across a range of criteria including coverage of the continuous quality improvement cycle, usability of the system, learning curve of faculty, data entry, data protection and privacy, among others. The paper highlights the advantages offered by each tool and identifies the gaps in respect to the continuous quality improvement cycle.**

***Keywords—Component; accreditation support tools; continuous quality improvement; undergraduate programs; assessment; student outcomes; software***

## I. Introduction

Maintaining accreditation standards within educational programs is a major endeavor for international universities today to ensure competitiveness and credibility of their quality [6]. Moreover, the assessment and accreditation of a certain educational program are two intertwined processes and are usually concurrent. Acquiring accreditation often means that the educational program must meet and satisfy certain requirements and criteria established by a dedicated world accreditation body [11]. For instance, the accreditation of engineering and computing programs by ABET necessitates the fulfilment of nine different criteria concerning student, student outcomes, continuous improvement, faculty, and curriculum, to name a few [1]. Realizing and monitoring accreditation standards manually is a cumbersome and complex procedure where evidences must be collected and documented throughout the assessment life cycle [19]. Therefore, more universities and educational institutions are working towards developing automated software solutions to govern and facilitate the assessment and accreditation processes [20].

However, three major challenges hinder the exploitation of such tools. Firstly, the literature reports on a limited number of accreditation support tools which necessitates further research. Secondly, the modules, components and functions of these tools are still not well understood. Thirdly, selecting or developing an appropriate accreditation tool is not a straightforward task and requires the definition of appropriate requirements. The herein paper investigates the details of three world accreditation tools with the aim of empowering education quality experts to recognize and select the most appropriate tool that suits their needs and context of use. Overall, this research contributes in the following ways.

- Provide a detailed review of the capabilities and functionalities of three modern accreditation tools for academic programs, namely EvalTools [2], CLOSO [3], WEAVEonline [4].
- Identify the strengths and weaknesses of each accreditation tool. The findings will enable decision makers in educational institutions to select the most appropriate tool that fulfill their needs, requirements, and resources.

The remainder of this paper is organized as follows. Section two briefly reviews the models and tools supporting the continuous quality improvement (CQI) cycle and assessment of student outcomes. Section three presents the procedure of the comparison and the selected tools. Section four presents the features of the tools. Section five compares the three accreditation tools.

## II. Related Works

Managing and promoting quality within education has become a top priority for educational establishments and decision makers [6]. This process, however, is time consuming and resourceful. The benefits of total quality management are well documented in the literature, comprising the provision of better services, satisfaction of students, maintenance of a competitive edge, and demonstration of high accountability [6]. There exist different quality models for achieving excellence in non-profit contexts such as the European Foundation for Quality Management (EFQM) and Malcolm Baldrige National Quality Award (MBNQA). Comparing the two models revealed that the EFQM model outperforms the MBNQA model in respect to satisfying the core principles of total quality management such as leadership, customer focus, process management and continuous improvement [9].

At the heart of quality management in education is the continuous improvement [21]. A longitudinal study of an Irish university demonstrated the need for external reviews, internal expertise, as well as training and senior management commitment as key elements for the successful realization of continuous improvement in higher education institutions [7]. A recent improvement framework describes the ingredients for

transforming knowledge into practice to achieve sustainable change and improvement in education [5]. Particularly, the framework encompasses three questions addressing the aims, improvement teams and workforce, and ways for change to result in improvement. These questions are then tested using the plan-do-study-act cycle to verify whether a proposed change indeed results in an improvement.

Outcome based education (OBE) is the new trend in teaching and learning where more focus is placed on delivering the knowledge, skills, and attitudes that learners should gain at the time of graduation [23]. However, achieving outcome-based education is laborious and time consuming. Therefore, [19] suggest the use of ICT to overcome the drawbacks of outcome-based education and facilitate continuous improvement. Indeed, various accreditation software applications are implemented and reported in the literature. ACAT is a web-based tool that facilitates the collection of data and generation of assessment reports as requested by ABET [12]. Overall ACAT encompasses four modules namely the administration, data entry, report generation, and authentication of access module. Similarly, a web-based assessment tool was developed in [13] to automatically collect assessment data, evaluate the student attainment levels and monitor the execution of remedial improvement actions for ABET-programs. The portal includes three key components namely a course module, monitoring module, and assessment module [16]. The portal streamlines the efforts of different stakeholders of the assessment cycle. Moreover, the tool enables the direct mapping of course outcomes to program outcomes which enables the identification of deficiencies within the course contents.

In [15], OBACIS is introduced as a framework to fulfill the requirements of Canadian Engineering Accreditation Board (CEAB) and subsequently enhanced to expedite the reporting phase of assessment documents and resources [18]. In fact, the tool automates the creation of course information sheets by combining three modules, namely an office application for grading and OBE reports, a web application for faculty templates, and a Windows application for program and assessment reports. All modules communicate with a central database of multiple resources. The authors expect that the tool could save up to 70% of the required assessment efforts. To improve the attainment level of student outcomes by weak student groups, [20] suggest the inclusion of a learning analytics' module.

Other sophisticated accreditation support tools employ the multi-agent paradigm. For example, [17] utilized software agents to check the needs and status of the institution's program and accordingly provide advice and generate assessment reports. In [10], a case study is undertaken to showcase the benefits of exploiting the Internet of Things, through a dedicated web service, to enhance the educational learning process. Other frameworks that aimed at instilling quality within education introduced the use of six sigma to develop quality management systems [8]. A Moodle tool based on the socio formative competencies model is suggested in [22] where instructors can design and include assessments according to the competency level of their students. Satisfaction scores of the tool were favorable.

In relation to our research, [14] utilized CLOSO, which is developed in Visual Basic .NET, to demonstrate the assessment of course outcomes and student outcomes in an automatic fashion in a bid to reduce instructors' time and efforts. Finally, EvalTools is detailed in [21] as a web-based tool for implementing specific performance indicators at the course level rather than the program level to fulfil the requirements of ABET. The tool applies a weighted algorithm to measure different learning levels (e.g. beginner, intermediate and master) across three learning domains (e.g. cognitive, affective, and psychomotor).

## III. Evaluation Procedure and Selected Tools

Two education quality experts who have more than 5 years of experience in the domain of local and international accreditation of various computing and engineering programs performed an-in-depth analysis of three accreditation tools. The tools that were selected for the expert evaluation included EvalTools, CLOSO, and WEAVEonline. Each expert went independently through the tools and investigated their various functionalities in detail documenting the key functionalities, reports generated by the tools, and strengths and weaknesses. The results were then discussed and integrated as summarized in the results' section.

### A. EvalTools

EvalTools is a complete online learning management and comprehensive assessment tool developed by Makteam [2], as depicted in Fig 1. It integrates three features mainly: learning management (e.g. attendance, lessons, assignments, gradebook … etc.), learning outcomes assessment (e.g. direct and indirect assessment of learning outcomes), and academic administration (e.g. dean office, department teaching, … etc.). Until today, EvalTools has been used by few universities to attain the ABET accreditation of undergraduate engineering and science programs.

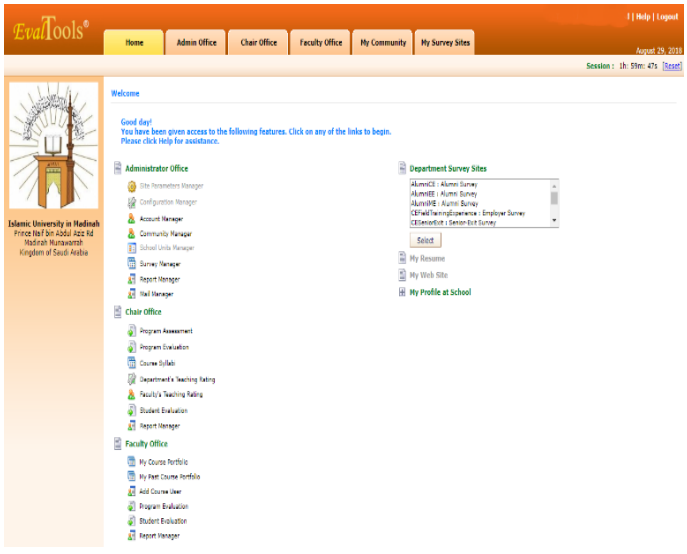



Fig. 1. EvalTools Admin Inteface.

### B. CLOSO

CLOSO is an accreditation tool developed by a research team in the United States in 2009 [3], as shown in Fig 2. This

tool incorporates one main module, a learning outcomes assessment, which automates the processes related to the continuous quality improvement and reduces the work load inflicted on faculty members and accreditation coordinators. As such CLOSO is designed to meet the requirements of ABET. Overall, the automated processes of CLOSO include:

- Creation of course folders;
- Evaluation of quantitative assessment data;
- Analysis of processes related to the continuous quality improvement; and
- Improvement of decision making for the program coordinators and policy makers.

Until today, CLOSO has been used by several universities to attain the international accreditation of science and engineering programs, specifically the ABET accreditation.


Fig. 2. CLOSO Home Page.

### C. WEAVEonline

WEAVEonline is a web-based tool for managing the assessment and evaluation of educational and non-educational outputs (e.g. outcomes or objectives) within various academic or non-academic units to assist universities and academic programs in the management of accreditation, evaluation, planning and quality improvement processes [4], as shown in Fig 3. This tool is designed to store evaluation and assessment data that ultimately improve the student learning and quality of services offered by universities and academic programs in various forms. WEAVEonline mainly focuses on the continuous improvement of all academic programs, administrative and student affairs units, and other services provided. In addition, this tool supports the implementation of improvement plans.

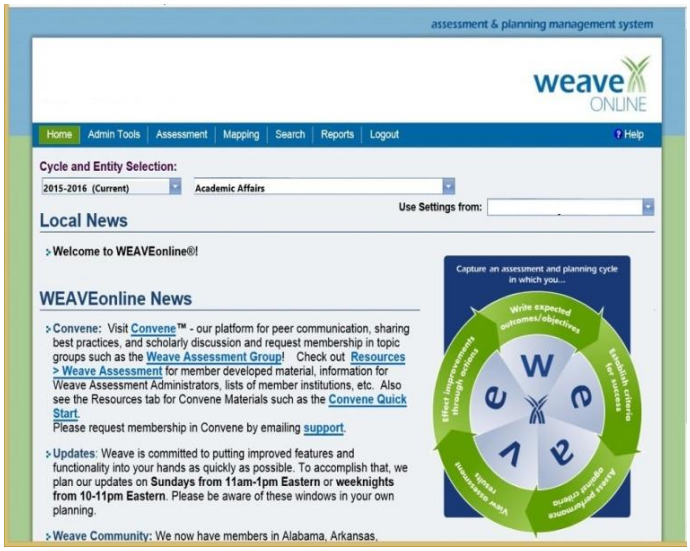



Fig. 3. WEAVEonline Home Page.

## IV. Results of the Expert Analysis

### A. EvalTools

A careful inspection of EvalTools demonstrated that there are four types of users who can access the different modules of EvalTools as detailed below.

- **Administrator**: the admin has full access and total privileges to all tool modules related to all academic programs of a faculty;
- **Program Coordinator**: the coordinator has access to the system modules that are related to the assessment of an academic program within a faculty;
- **Faculty Members**: instructors have limited access to the functions and modules of the system that are related to their courses and reports within a specific academic program; and
- **Students**: students have access to only the learning management module of their current and past courses.

Table I summarises the main functionalities and features made accessible to the different types of users of EvalTools.

TABLE. I. EVALTOOLS KEY FUNCTIONALITIES AND FEATURES

| Type of User Addressed | Key Feature / Function |
|---|---|
| All users | Login into EvalTools to access the functions of the system |
| | Edit profile information such as email address and photo |
| Administrator | Create new user profiles |
| | Manage or delete existing user profiles (e.g. update details, user types, access control … etc.) |
| | Create, manage and send automatic emails to the various users of the system |
| | Set up and manage the program assessment elements including adding and mapping PEOS, SOs, and PIs |
| | Set up, update and administer the surveys (e.g. course, exit, senior, employer, alumni, field experience … etc) used for the indirect assessment of programs |
| Administrator, Coordinator | Create, edit, and delete committees at the faculty and program level as needed |
| | Assign and add members to program committees and remove them as needed |
| | Access and view the courses syllabi of the program |
| | View rating of teaching performance of the program |
| | View rating of individual faculty of the program |
| | Download the program teaching rating report and faculty teaching rating report as word files |
| | Assign students to advisors through the advising system |
| | View and monitor the performance of all students of the program through the advising system |
| | View course assessment reports (indirect assessment achieved through surveys) for all courses of a program in different formats (tables and excel) |
| Administrator, Coordinator, Faculty | View and track progress of the faculty portfolios including their past and current courses |
| | Write own courses' description, and update the various sections of the courses |
| | View, add and edit existing course assignments and homeworks, along with their learning domains / levels |
| | Upload oe download teaching materials (e.g. lecture notes, research papers etc) in different formats |
| | Import existing course descriptions from a set of templates |
| | Export course descriptions into a word file |
| | Create or import rubrics for student assessment |
| | View and download students' submitted assignments |
| | Add students grades to the system directly through a table or an imported excel file |
| | Upload and download graded assignments for students |
| | Manage the grade book and specify the way grades are calculated for students with their percentages |
| | Assign grades to students for each assignment, as part of a table or as an imported excel file |
| | View the gradebook containing all students' grades as a table or excel file |
| | Publish and share notes by the faculty with their students |
| | View or export the results of the indirect assessments (i.e. surveys) as tables, word or excel files |
| | View the various elements of the program assessment including program educational objectives, student outcomes, and performance indicators, and their mappings |
| | View or export the elements of the program assessment including program educational objectives, student outcomes, and performance indicators, and their mappings as a word and excel file |
| | View or export program evaluation and executive summaries reports that are related to student outcomes evaluation and performance indicators evaluation |
| | View or download the objective evidence folders containing all evidences (direct and indirect assessments) that justify the performance levels for the student outcomes and performance indicators as word |
| | Add, update or delete reviews and improvement actions, about the program and courses, as a result of reflection about the achievement levels of student outcomes and performance indicators |
| | View action item matrix about the improvements and responsible entities that need to be taken care of in relation to the courses and program |

| | |
|---|---|
| | Generate and download curriculum outcome matrix that shows the links between course outcomes and specific student outcomes for a particular program |
| | View the committees of which he is a member and its members, documents and tasks |
| | Communicate with the members of designated committees directly using EvalTools |
| | Upload the relevant documents (e.g. minutes of meetings) to the designated committees |
| | Monitor and advise allocated students on their academic performance and provide regular feedback |
| | View and download course assessment reports (indirect assessment achieved through surveys) for their courses in different formats (tables and excel) |
| Student | Browse past and current courses to view their information (descriptions, assessment methods … etc.) |
| | Download the lecture materials /assignments |
| | Monitor academic achievement against a set of course outcomes and performance indicators per assessment |
| | Submit answers for a particular assignment |
| | View grades for various types of assessments |
| | Access students advising information through the student advising functionality |
| | Assess the quality of current courses by responding to various types of surveys |

Data entry is integral part of the continuous quality improvement where the stakeholders are required to enter different inputs into the accreditation tool to monitor the assessment cycle. The major data entry tasks performed by the administrator in EvalTools include creating new users of the tool, entering data related to the program assessment such as the student outcomes and performance indicators, creating a curriculum map and setting up the relevant indirect assessment surveys. The faculty member however is responsible about entering courses' information and uploading related materials (e.g. lecture slides and assignments), inserting students' grades, uploading graded assessments, and writing the reflections and actions about the performance of their students. Students, on the other hand, upload their assignments' answers to EvalTools and rate the quality of their courses by completing different types of surveys (e.g. course, senior, field experience etc.).

EvalTools automatically generates, based on the inputted data, a wealth of reports and documents that are considered useful to both the student learning and program assessment. Table II summarizes nine main reports that are created by EvalTools along with their inner sections.

TABLE. II. EVALTOOLS GENERATED REPORTS

| **Type of Report Generated** | **Description and Elements of the Report** |
|---|---|
| Course Syllabus | This report summarizes the key information of the courses offered within a program, including:<br>• Course identifier (code and name)<br>• Course Description<br>• Pre-requisites<br>• Course Outcomes<br>• Course Outlines / Topics<br>• Course Assessment Methods<br>• Assessment Methods vs. Course Outcomes (as a matrix)<br>• Textbooks<br>• Supporting Tools<br>• Course Policies<br>• Grading Policies<br>• Faculty Name<br>• Date |
| Gradebook | This report summarizes the grades of the students in respect to all courses' assessments and allocates a final grade. The report includes the following items:<br>• Student Name<br>• Student ID<br>• Final Grade<br>• Adjusted Score<br>• Total Score<br>• Various assessments of the course (e.g. homeworks, quizzes, exams, etc.) |
| Faculty Course Assessment Report (FCAR) | This is considered the most important and comprehensive report of EvalTools at the course level as it gives an overview of the performance of students during the term, with strengths and weaknesses clearly highlighted in relation to the student outcomes and performance indicators. It also includes the reflections and improvement actions planned for the next term. Graphs are used throughout the report to aid understanding. The report includes these sections: |

| | |
|---|---|
| | • Course Identifier (Code and Name)<br>• Pre-requisites<br>• Course Description<br>• Outcomes Section<br>• Grade Distribution<br>• Reflection on Course Delivery<br>• New Action Items<br>• Course Outcomes Assessment<br>• Assignment List<br>• Summary of Assignment Distributions<br>• Student Outcomes Assessment<br>• Performance Indicators Assessment |
| Student Survey Responses | This report provides a summary of the student evaluation of the courses using a dedicated survey that enquires about the quality of various aspects of the courses, field training and learning experience in general. The questions focus on the quality of the course contents, outcomes, teaching and assessment methods, and faculty. These results represent the backbone of the indirect assessment of the quality of the courses and program. The report includes the following items:<br>• Outcomes Section<br>• Course Items<br>• Course Syllabus<br>• Course Instruction<br>• Faculty Items<br>• Assessment Techniques<br>• Overall Evaluation |
| Program Executive Summary | This report provides a set of graphs that indicate the achievement levels of student outcomes of a program as well as the achievement levels of their performance indicators. The report includes the following items:<br>• Associated Student Outcomes<br>• Student Outcomes Summary Graph<br>• Performance Indicators Summary Graph |
| Detailed SO/PI Executive Summary | This is a detailed report covering all student outcomes of a program. The report summarizes the review of the program including the reflections and actions about students' performance and achievement levels. These are linked to the scores of the performance indicators (out of 5). The report includes the following items:<br>• Student Outcome<br>• Discussion and Action<br>• Detailed Executive Summary<br>• Average<br>• Classification / Review Date |
| SO/PI PVT Summary | This report details the performance indicators under each student outcome and summarizes the achievement levels of each performance indicator along with a classification of the level of learning. Students are distributed within a table across the achievement thresholds (E, A, M, U). The report includes the following items:<br>• Student Outcome<br>• Performance Indicators<br>• Performance Indicators Achievements Results (As a table)<br>• Overall Average<br>• Performance Indicators Graph |
| Course Reflections / Action Items | This report lists all reflection and action items of the program. These actions are usually planned for the upcoming term and are related to various courses of the program. The report includes the following items:<br>• Course Name and Code<br>• Reflection<br>• Action Items |
| Learning Domains Evaluation | This report summarizes the performance of students in the program in relation to various levels and domains of learning (e.g. based on Blooms taxonomy). The report includes the following items:<br>• Domains Learning Analytic<br>• Individual SO Learning Distribution Analytics<br>• Summary of Individual Domains Activities<br>• Cognitive Domain Learning Analytic<br>• Affective Domain Learning Analytic<br>• Psychomotor Domain Learning Analytic |

### *B. CLOSO*

The users of CLOSO can be classified into administrators or instructors. The administrators have full access to all tool's modules and functions that are related to the program assessment. However, the instructors have limited access to the functions and modules that are related to the assessment of their courses. Table III summarises the main functionalities and features made accessible to the different types of CLOSO users.

TABLE. III. CLOSO KEY FUNCTIONALITIES AND FEATURES

| Type of User Addressed | Key Feature / Function |
|---|---|
| All users | Download and install CLOSO into a local machine |
| | Activate CLOSO using a license |
| | Login into CLOSO to access the relevant functions of the system |
| CLOSO Administrator | Specify, customize and edit program specifications including program type, student outcomes, terminology |
| | Set up and customize faculty and student surveys used to collect indirect assessment data |
| | Manage, set up and edit the course syllabi of all courses (e.g. description, course learning outcomes, CLO-So map) of the entire academic program (normally these data are received from the instructors) |
| | Access and view the courses syllabi (including description, course learning outcomes, … etc.) of the entire program |
| | Set up courses – student outcomes mapping of the program (i.e. program assessment details) |
| | Upload CLOSO course syllabus and customization files to the server |
| | Create analysis reports of the student outcomes and their achievement levels and create relevant evaluation reports |
| | Summarize the feedback provided by the courses' instructors to consolidate evaluation reports in relation to course readiness, student weaknesses, and ways to improve the program. |
| | Perform student outcome wise comparison and produce the necessary statistics for the program evaluation (e.g. closing the loop, improvement, weaknesses analysis … etc.) |
| | Save CLOSO files such as program details, analysis results, syllabus and customization files |
| | Import existing CLOSO files |
| | Maintain and administer the database of CLOSO |
| Course Instructor | Import syllabus data file into CLOSO either through a download link or email |
| | Import customization file (e.g. terminology and surveys) into CLOSO either through a download link or email |
| | View course information including class detail, CLOS, CLO-SO map |
| | Update instructor details including, personal informal, office hours and class schedule |
| | Open, save and create files related to course information |
| | Create, update and maintain a list of students in their courses including their details such as student ID and student names |
| | View course syllabus including contents, CLOS, text book, CLO-SO map … etc. |
| | Modify syllabus details and submit the requested changes to CLOSO administrator |
| | Create and set up a detailed assessment plan for the courses by specifying the weekly teaching plan, instruction/teaching methods, assessment methods, and assessment distribution |
| | Set up an assessment plan for the measuring the attainment of the student outcomes |
| | Design and manage the assessments of the course by specifying the assessment questions, the addressed CLO, SO, bloom's level and contribution of the question towards the final score |
| | Enter the grades of the course assessments for each student |
| | View and analyze the performance of the students in the course assessments |
| | Curve the assessment results if necessary |
| | View and analyze the final grades of the students and their overall distribution |
| | Load students' assessment and work samples in the form of images and pdf files |
| | View scanned pdfs / images of students' work samples |
| | Complete faculty survey including six items: learning readiness, syllabus coverage, CLO satisfaction, weaknesses, improvement methods, and SO loop closing |
| | Input or copy student responses to the student survey |
| | View and analyze student responses about CLO and SO attainment |
| | View and analyze CLO and SO satisfaction scores of all students |
| | Print necessary accreditation forms and evidences (e.g. course file forms, student survey form) |

The major data entry activities and tasks that need to be completed by CLOSO administrator include setting options related to the list of student outcomes, terminologies, and target satisfaction criterion, inputting data about the departments and courses offered in the programs and setting up the courses-student outcomes mapping (i.e. the curriculum map). However, the instructors are required to input course information (e.g. syllabus, CLOS, and topics) and students' names, to write the assessment questions to measure the attainment of student outcomes, and to upload the assessment evidences (e.g. students' answers) as PDFs or images. Obviously, data entry in CLOSO is less complex than EvalTools. Similarly, CLOSO generates nine reports that serve the accreditation process and assessment of student learning outcomes as listed in Table IV.

TABLE. IV. CLOSO GENERATED REPORTS

| **Type of Report Generated** | **Description and Elements of the Report** |
|---|---|
| Course Syllabus | This report summarizes the key information of the courses offered within a specific program. It contains the following sections:<br>• Course identifier (number and name)<br>• Course type (core / elective)<br>• Theory credit hours<br>• Lab credit hours<br>• Theory contact hours<br>• Lab contact hours<br>• Course objective<br>• Catalog Description<br>• Contents<br>• Pre-Requisites<br>• Textbook<br>• References<br>• Prepared by<br>• Approved by<br>• Coordinator<br>• Date (Updated on)<br>• Table of course learning outcomes (CLO ID and CLO)<br>• CLO to student outcomes matrix (CLO ID vs SO number) |
| Course List | This report helps fulfill criteria 5 (i.e. Curriculum) of the self-study report of ABET. The report contains the following data:<br>• Course Number<br>• Course Name<br>• Course Category<br>• Department of the course |
| Course – SO Map | This report represents the curriculum map which is basically a simple mapping between the courses of the program and the student outcomes. In more detail, the matrix table contain:<br>• Course number<br>• Course Name<br>• Credit hours for each course<br>• Student outcomes |
| Student Handout Course Information | This report is given to the students and provides detailed information about the course and class at the beginning of the term. It contains the following sections:<br>• Instructor Information<br>• Course Syllabus<br>• Instruction Plan<br>• Assessment Plan<br>• Grade Distribution |
| Course Folder Part I | This report contains information related to the course and results of indirect assessment of student outcomes from the various surveys. In detail, this report incorporates the following items:<br>• Instructor Information<br>• Course Syllabus<br>• Student Outcomes<br>• Students Survey Analysis<br>• Faculty Survey Data |
| Course Folder Part II | This report contains information related to the course and results of direct assessment of student outcomes from the various course assessments. In detail, this report incorporates the following items:<br>• Assessment Design<br>• Assessment Data<br>• CLO Satisfaction Analysis<br>• SO Satisfaction Analysis<br>• Assessment Samples and Model Solutions |

| | |
|---|---|
| Cover Sheets for Students Work Samples | This report help organizes the objective evidences of the students' assessments into easy to read folders. All direct assessments that are used to measure the attainment of the student outcomes are collected in this report. In detail, this report contains:<br>• Question Paper (with SO to question to CLO mapping)<br>• Model Solution<br>• Student Work Sample (High)<br>• Student Work Sample (Medium)<br>• Student Work Sample (Low) |
| Student Survey Form | This report provides a summary of the student evaluation of the courses using a dedicated survey that enquires about the quality of various aspects of the courses and learning experience. The questions focus on the attainment of the course learning outcomes. The key elements of this report include the following items:<br>• Title of the course, instructor, semester number and academic year<br>• A list of course learning outcomes and a score (out of five)<br>• Student name and signature |
| Faculty Opinion | This report is embedded within CLOSO and cannot be exported as an external file. However, it contains valuable information about the opinions of the faculty in respect to the performance of the students and ways to improve student skills. In detail, it contains statistics about:<br>• Course readiness<br>• Student weaknesses<br>• Ways to improve<br>• SO-wise analysis |

### *C. WEAVEonline*

The roles that can be created using WEAVEonline include the administrator, entity administrator, entity-write and review, entity-review only, approver, group owner, map owner, external reviewer, and data uploader. Table V summarizes the main functions of WEAVEonline.

TABLE. V. WEAVEONLINE KEY FUNCTIONALITIES AND FEATURES

| **Type of User Addressed** | **Key Feature / Function** |
|---|---|
| All users | Access the main page and view the institution news |
| | View the institution and programs information |
| Administrator, Entity Administrator | Add users and update their profiles |
| | Edit and update information related to the university and programs |
| | Manage (add, update and delete) entity trees and their types (academic, administrative … etc.) |
| | Add and update information about new, present and past assessment cycles |
| | Add and update the information of the beneficiary e.g. university such as University name … etc. |
| | Create and add strategic plans for the institution, faculties, and programs |
| | Add and edit key information about the curriculum, and priorities of the institution and programs |
| | Add and edit the local news of the university |
| | Add, update, and delete the mission statement, goals, outcomes and objectives of the beneficiary |
| | Add and update the analytical questions which could be used in the assessment summary report |
| | Modify the sections of the annual report by adding or altering existing sections |
| Administrator, Entity administrator, Entity-write and review, Entity-review only, Approver, Group owner, Map owner, External reviewer, and Data uploader | Create and view different types of assessment reports including full assessment, audit, data entry status, and analysis question reports |
| | Upload or download the required and relevant documents into the document repository |
| Administrator, Entity Administrator, Group Owner | Create new groups and describe their key responsibilities |
| | Create the necessary standards and connecting them to the available groups |
| | Provide responses to the questions related to the annual and special reports |
| | Upload documents and mapping them to the various sections of assessment |
| | Modify or delete the documents within the tool's repository |
| Administrator, Entity Administrator, Entity-write and review, Entity-review only, | Add, update and delete assessment methods and threshold for measuring the attainment of objectives and outcomes |
| | Add and update the results of the assessment results for an assessment cycle |

| | |
|---|---|
| Approver | Add, update or delete an action plan along with details such as start and end dates, status of plan, priorities, and budget. |
| | Add or delete an analysis of assessment data for a cycle |
| | View the different courses / experiences, outcomes / objectives, courses and learning outcomes |
| | Establish the mapping between courses and different learning outcomes and learning experiences |

Nine types of reports are generated by WEAVEonline along with their sections as summarized in Table VI. Interestingly, these reports go beyond the assessment of learning outcomes and include the assessment of key performance indicators (KPIs). The reports also clearly link the KPIS to the mission and objectives of the institution. Unlike EvalTools and CLOSO, WEAVEonline places more emphasis on improving the strategy of the educational institution and its units.

TABLE. VI. WEAVEONLINE GENERATED REPORTS

| **Type of Report Generated** | **Description and Elements of the Report** |
|---|---|
| Full Assessment | This report summarizes the assessment results and includes the subsequent sections:<br>• Mission/Purpose, Goals Outcomes/Objectives<br>• Measures<br>• Assessment Results/Findings<br>• Action Plans<br>• Analysis Question responses<br>• Annual Report section responses |
| Audit | These reports determine whether additional data entry is needed regarding:<br>• Outcomes/Objectives<br>• Measures that need targets<br>• Measures that need assessment results / findings<br>• Findings that need action plans |
| Data Entry Status | These reports provide a summary of data entry progress for each major area including:<br>• Mission / Purpose<br>• Goals<br>• Outcomes<br>• Measures<br>• Assessment Results<br>• Action Plan<br>• Analysis Questions<br>• Annual Report<br>• Charts of Progress |
| Achievement Status | This report shows the achievement details of actions for all entities including<br>• Charts<br>• Summary of achievements |
| Planning and Budgeting | These reports provide a view of the action plan with the following details:<br>• Action Plan Profile<br>• Action Plan Tracking / Implementation<br>• Budgeting – Resources Requested<br>• Comprehensive Planning (With Dates and Priorities) |
| Profile and Association | These reports focus on various Outcome/Objective associations including:<br>• Standard Profile (Outcome / Objective Association)<br>• Strategic Plan Profile (Outcome / Objective to Strategic Plan Association)<br>• General Education/Core Curriculum by Association<br>• Institutional Priority by Association<br>• Standard by Association<br>• Strategic Plan by Association |
| Special Assessment | These reports display the assessment results of each entity in the form of charts focusing on:<br>• Assessment Summary Chart<br>• Student Learning Outcomes/Objectives<br>• Sources of Evidence (Measures) |

| | |
|---|---|
| Analysis Question | These reports display the answers to all analysis questions grouped by entity and include:<br>• Analysis Answers Report Profile<br>• Analysis Answers Report by Question<br>• Annual/Special Report Profile<br>• Annual/Special Report by Section |
| Mapping | This report displays the entities that are using curriculum mapping in the assessment. |

*D. The Comparison Matrix*

Table VII contrasts the three accreditation tools EvalTools, CLOSO, and WEAVEonline, and allocates a rating to each tool against 36 comparison criteria, with the following rating scheme:

- **Yes**: means that the criterion has been fully satisfied;
- **No**: means that the criterion has not be met at all; and
- **Partial**: means that the tool partially complies to the specified criterion

TABLE. VII. THE COMPARISON OF THE ACCREDITATION SUPPORT TOOLS

| **Comparison Criterion** | EvalTools | CLOSO | WEAVEonline |
|---|---|---|---|
| Type of application | Web-based (no installation is required) | Standalone application (requires installation) | Web-based (no installation is required) |
| General purpose of tool | Outcome based assessment, but not limited to ABET assessment philosophy | Outcome based assessment related to ABET accreditation | Management of accreditation, assessment, planning and quality improvement processes for universities |
| Learning and course management system (for students) | Yes | No | No |
| Coverage of all continuous quality improvement (CQI) cycle | No | No | Partial |
| CQI planning | Partial | Yes | Yes |
| CQI implementation | Yes | Yes | No |
| CQI analysis | Yes | Yes | Yes |
| CQI reporting | Yes | Partial | Yes |
| CQI improvement actions | Yes | Partial | Yes |
| Usability (ease of use) | Complex to use | Easy to operate | Easy to operate |
| Learning curve | Steep – Time consuming | Requires less efforts | Requires less efforts |
| Adaptability to different program assessment models | Partial | No | No |
| Access control and various profiles | Yes | No | Yes |
| Security of system and data privacy | Weak | Medium | Medium |
| Data entry | High volume | Low volume | Medium volume |
| Automation | Yes | No | No |
| Performance (execution time and overall speed) | Slow | Fast | Fast |
| Customization (e.g. Terminology, surveys … etc.) | Low | Medium | Low |
| Accountability (ability to track and follow stakeholders of the program | High | Low | High |
| Indirect assessment of teaching (through surveys) | Yes | Partial | Partial |
| Direct assessment of teaching (through assignments, examinations etc.) | Yes | Yes | Yes |
| Student advising module | Yes | No | No |
| Archiving of objective evidences | Yes | Yes | Yes |
| Reuse of existing modules (e.g. Importing of previous assignments, rubrics … etc.) | High | Low | Very High |
| Monitoring of faculty performance | Yes | No | No |

| | | | |
|---|---|---|---|
| Survey management | Yes | Yes | No |
| Use of committees to manage work load | Yes | No | Yes |
| Exploitation of social media to communicate with students | No | No | No |
| Encouragement of group work amongst students | Yes | No | No |
| Using the system at different levels in the university (academic programs, college, university, administrative units, etc.) | No<br>(Used at the program level only) | No<br>(Used at the program level only) | Yes |
| Give the external evaluators access to the tool | Yes | No | Yes |
| The tool is designed based on an academic accreditation body | Partial (Based on ABET but can be used for other accreditation agencies) | Yes (based on ABET) | No |
| The system supports the quantitative and qualitative analysis of the outcomes achievements | Yes | Yes | Yes |
| The tool supports internal and external benchmarking in the continuous quality improvement process | No | No | No |
| The tool is linked to various other electronic systems of the university | No | No | Partial |
| Approval of important elements in the assessment process can be done electronically | Yes | No | Yes<br>(Approver Role) |

## V. Discussion of the Strengths and Weaknesses

Based on the extensive review and analysis carried out with the three accreditation support tools, various strengths and drawbacks were extracted and are highlighted below.

EvalTools complies with almost a complete continuous quality improvement cycle and addresses many aspects of the CQI phases, ranging from planning, implementation, reporting, and closing the loop (i.e. implementation of improvement actions). EvalTools can be flexibly applied to satisfy the requirements of differing assessment philosophies and accreditation bodies including the use of simple generic performance indicators (i.e. ABET assessment philosophy) and complex specific performance indicators (i.e. at the course level) to assess the programs' quality and achievement levels of the students. Moreover, EvalTools uses an adjustable model for scoring and calculating the achievement levels of the performance indicators and course outcomes. The model can be changed as per the needs and goals of each educational program. EvalTools implements a strong reporting functionality and provides a myriad of reports including direct assessment (e.g. performance indicators measurement using examinations) results and indirect assessment (e.g. course level surveys) results utilizing graphs and charts whenever possible. The reports are presented at different levels, a fine-grained level (e.g. student or course) and holistic level (e.g. program) which enables the monitoring of student performance accurately. EvalTools enforces the mapping of program student outcomes to the courses, courses' outcomes and performance indicators creating a solid and clear curriculum map. EvalTools empowers the design and administration of multiple surveys to collect feedback from various stakeholders of the program (e.g. students, faculty, employers … etc.) to support the indirect assessment of the program continuously.

On the negative side, EvalTools is a rather complex and intimidating system for the inexperienced faculty members; for instance, data entry requires multiple and redundant clicks. Moreover, EvalTools introduces a steep learning curve to operate its modules and to get accustomed to its full features. It does not provide any manual or user guide for beginners. This usually leads to mistakes in the entry of data (e.g. mapping of COs and PIs) and comprehension of generated reports. EvalTools is designed mainly to satisfy criterion four of ABET (i.e. Continuous Improvement). The remaining criteria of ABET (e.g. students, faculty … etc.) are not addressed by the tool which means that more work is needed for the accreditation. Whilst EvalTools covers to a large extent the assessment cycle, it does adequately cover the planning phase of the continuous quality improvement. Performance wise, EvalTools often suffers from constant delays and server crashes especially during peak times and concurrent access. EvalTools does not implement strong privacy and security measures as it is easy to access the profiles of other EvalTools users.

CLOSO empowers customization; for example, the terminology can be modified by the administrator to suit any academic program, department or university. CLOSO is built on top of the ABET assessment model. Therefore, it organizes the documentations required by ABET in neat way and makes them ready to print through a click of a button. In particular, it prepares criterion four and five of the self-study report of ABET in a professional manner. CLOSO covers the planning phase of CQI reasonably well as the assessment plan can be created through a set of dropdown menus with pre-defined options reducing the need to write an assessment plan from scratch. CLOSO uses graphs to aid understanding of the performance of the students; for example, the scores input form displays various interesting charts. CLOSO also provides a set of readymade and easy to answer faculty and student surveys to assess the quality of educational programs. Moreover, it empowers the preparation and implementation of an improvement plan along with the necessary evidences. Overall, CLOSO supports a sustainable assessment philosophy that can

be managed easily by the program coordinators and course instructors.

As drawbacks, only two types of users (i.e. administrator and instructors) can use CLOSO. However, educational programs usually include coordinators and students. The assessment plan offered by CLOSO is quite rigid which means that policy makers cannot introduce major changes to the existing assessment plan. Accreditation work is typically organized through various committees; however, CLOSO does not support the creation and management of committees and work groups. CLOSO does not support student advising based on the identified weaknesses from the assessment process. The surveys used for indirect assessment cannot be customized to fit the needs of the program. Although CLOSO encourages instructors to reflect on their teaching by specifying the limitations and potential improvements, these reflections are performed at the course level instead of the program level which results in an incomplete improvement plan.

Finally, WEAVEonline can be used for the outcome assessment at all levels (e.g. university, colleges, academic programs, administrative units, etc.) through the establishment of units. WEAVEonline supports the creation of a plan for the continuous development of all units of the university based on the achievement levels of outcomes and objectives, being either educational or non-educational. In fact, WEAVEonline is used by numerous American universities to manage the evaluation and assessment processes to ensure the continuous development. One interesting feature of WEAVEonline is that it allows monitoring the data entry completion rate of various stakeholders through the Data Entry Status (DES) Reports. Moreover, the tool generates rich reports such as the full evaluation report, audit reports, data entry status reports, analysis questions reports, planning and budget reports, personal reports, and so forth. There is a review and approval feature for all parts of the assessment and planning that need to be approved by the decision makers, enabling a good level of control and accountability. The possibility of identifying the responsible person / group for the implementation of any improvement plan is done electronically and the implementation of such action plans can be monitored through the tool. WEAVEonline provides the external reviewers access to all reports and assessment evidences.

As limitations, WEAVEonline does not allow the implementation of assessment methods (e.g. exams, quizzes, etc.) directly nor their mapping to the student learning outcomes. Instead, assessments must be carried out independently and only then the results can be inserted into the tool. Moreover, the administration of surveys to collect indirect assessment data is not possible via the tool. Although the surveys can be selected as an assessment method in WEAVEonline, its administration and analysis must be performed externally. Students cannot view the results of their attainment level of the student outcomes using the tool. Although quantitative data can be entered in WEAVEonline, the improvement plans rely heavily on the descriptive analysis.

## VI. Conclusions

This paper provides detailed insights about the functions and features of three world accreditation support tools for educational programs. Developing or selecting an appropriate assessment tool that satisfies the accreditation requirements is a challenging job and this research aims to empower decision makers to make an informed decision about the tool to adopt.

The expert analysis, carried out by two education quality reviewers, included a comprehensive review of the main functions, data entry activities, and generated reports of the selected accreditation tools. The results showed that no single tool covers all phases of the continuous quality improvement. EvalTools is feature rich with a strong reporting mechanism; however, it is relatively time consuming to operate and less user friendly. CLOSO covers criterion four and five of ABET and generates the respective ready to use reports. However, it does not support group work and collaboration and is limited by a rigid assessment plan. WEAVEonline supports the assessment of educational and non-educational outputs and outcomes and uses a tracking component to monitor the progress of assessment tasks and improvement action plans. However, WEAVEonline lacks a survey management module to collect indirect assessment data and requires the preparation of assessments externally of the tool.

Future research plans will focus on inferring and proposing an architecture for an accreditation support system that combines the strengths of all three tools discussed in this paper. The architecture will be translated into a functional prototype that will be subsequently subjected to a user testing.

## References

[1] ABET, Criteria for Accrediting Engineering Programs, 2018 – 2019, http://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2018-2019/#4

[2] MakTeam, EvalTools, https://iu.myschooling.net/EvalTools6/EvalTools/

[3] Smart-Accredit, CLOSO, http://www.smart-accredit.com/

[4] Weave, WEAVEonline, https://app.weaveonline.com/login.aspx

[5] L. Cohen-Vogel, A. Tichnor-Wagner, D. Allen, C. Harrison, K. Kainz, A. R. Socol, and Q. Wang, "Implementing educational innovations at scale: Transforming researchers into continuous improvement scientists," Educational Policy, vol. 29, no. 1, pp. 257-277, 2015.

[6] E. Sallis, Total Quality Management In Education. Routledge, 2014.

[7] S. O'Reilly, J. Healy, T. Murphy, and R. O'Dubhghaill, "A continuous improvement journey in the higher education sector: a case study of a university in Ireland," 2017.

[8] P. Adina-Petruța, and S. Roxana, "Integrating six sigma with quality management systems for the development and continuous improvement of higher education institutions," Procedia-Social and Behavioral Sciences, vol. 143, pp. 643-648, 2014.

[9] O. Al-Tabbaa, K. Gadd, and S. Ankrah, "Excellence models in the non-profit context: strategies for continuous improvement," International Journal of Quality & Reliability Management, vol. 30, no. 5, pp. 590-612, 2013.

[10] C. E. Cornel, "The role of internet of things for a continuous improvement in education," Hyperion Economic Journal, vol. 3, no. 2, pp.24-31, 2015.

[11] S. Chandrasekaran, A. Stojcevski, G. Littlefair, and M. Joordens, "Accreditation inspired project oriented design based learning curriculum for engineering education," In Proceedings of the 2nd International Engineering and Technology Education Conference (IETEC 2013), Ho Chi Minh City, pp. 1-11, January 2013.

[12] E. Essa, A. Dittrich, S. Dascalu, and F. C. Harris Jr, "ACAT: A web-based software tool to facilitate course assessment for ABET accreditation," IEEE.In Information Technology: New Generations (ITNG), 2010 Seventh International Conference on pp. 88-93, April 2010.

[13] W. Ibrahim, Y. Atif, K. Shuaib, and D. Sampson, "A web-based course assessment tool with direct mapping to student outcomes," Educational Technology and Society, vol. 18, no. 2, pp. 46-59, 2015.

[14] M. H. Imam, and I. A. Tasadduq, "Evaluating the satisfaction of ABET student outcomes from course learning outcomes through a software implementation," International Journal of Quality Assurance in Engineering and Technology Education (IJQAETE), vol. 2, no. 3, pp. 21-33, 2012.

[15] M. A. Ismail, "OBACIS: Outcome based analytics and continuous improvement system," In Proceedings of the Canadian Engineering Education Association (CEEA), 2016.

[16] Y. Atif, W. Ibrahim, and K. Shuaib, "A portal approach to continuous improvement of outcomes-based curriculum," In Global Engineering Education Conference (EDUCON). Marrakech: IEEE, pp. 1-6, April 2012.

[17] A. M. Al-Bakry, "E-assessment system based on ABET criterion for computing programs," Journal of University of Babylon, vol. 22, no. 5, pp. 1482-1489, 2014.

[18] M. A. Ismail, "OBACIS Phase II: Catalogs and auto-generated course information sheets," In Proceedings of the Canadian Engineering Education Association (CEEA), 2017.

[19] J. Balasangameshwara, "Uncovering the value of ICT in time management for implementation of OBE courses," Journal of Engineering Education Transformations, vol. 29, no. 1, pp. 43-54, 2015.

[20] I. Aghav, and J. Aghav, "Learning analytics to quantize and improve the skills development and attainment in large classes," International Journal of Research in Engineering and Technology, vol. 3, no. 10, pp. 358-362, 2014.

[21] W. Hussain, P. E. Fong Mak, F. M. Addas, "Engineering program evaluations based on automated measurement of per-formance indicators data classified into cognitive, affective, and psychomo-tor learning domains of the revised bloom's taxonomy," In ASEEE 123rd Annual Conference & Exposition. New Orleans, LA, 2016.

[22] S. Cardona, J. Velez, and S. Tobon, "A tool for competence testing in Moodle," In Computing Colombian Conference (9CCC), 2014 9th, IEEE pp. 35-41, September 2014.

[23] I. L. An, "Impact of outcome-based education instruction to accountancy students in an Asian University," Asia Pacific Journal of Education, Arts and Sciences, vol. 1, no. 5, pp. 48-52, 2014.

# Development of a Novel Approach to Search Resources in IoT

Nisar Hussain[1], Dr. Tayyaba Anees[2], Muhammad Rehan Faheem[3], Momina Shaheen[4]
School of System and Technologies
University of Management and Technology
Lahore, Pakistan

Muhammad Imran Manzoor[5]
Department of CS and IT University of Lahore, Pakpattan Campus
Pakpattan, Pakistan

Aimen Anum[6]
Department of Computer Science
Comsats University Sahiwal Campus
Sahiwal, Pakistan

***Abstract*—Internet of Things (IoT) referred to interconnected the world of things like physical devices, cars, sensors, home appliances, actuators and machines embedded with software at any time, any location. The increasing number of IoT devices facing challenges which are registration, integration, describing sensor, interoperability, semantics, security, discovery and searching. The current systems are suitable for limited number of devices. Our ecosystem change day by day which means we have billions and trillions of devices connecting to the Internet in future. One major challenge in current system is searching of suitable Smart Things from a millions or even billions number of devices in IoT. For the purpose of searching and indexing, some discovery methods and techniques are discussed and compared. Those techniques and methods are studied and find out the limitations and issued of the current system. Another challenge to searching the Smart Things is a variety of description models for describing the Smart Things. In this piece of work, a novel search engine is proposed to search the Smart Things with variety of description models. A web interface is implemented in this research with HTML, JSON and XML formats. The description models of Smart Things SensorML, SensorThings API and W3C JSON-LD are implemented in the current proposed system.**

***Keywords*—IoT; IoT resources; search engine for IoT; SensorM**

## I. INTRODUCTION

The physical objects have the ability to connect with Internet and share data in respect of Internet of Things. In 2020 the connected devices that are able to communicate with Internet are more than 50.2 billion according to CISCO prediction [1]. These physical objects are called Smart Things. IoT provide a range of connectivity protocols, applications and mechanisms to interoperability with existing infrastructure.

In this regard, it‘s essential to propose a search engine to search the devices with the context of the user requests. There are some standardized sensors description technologies which are currently used in many systems. These technologies are SensorML [2], SensorThings API[3] and W3C JSON-LD [4]. These technologies used different languages and formats. In this research, we use all these technologies and parse into a common format for supporting our system. Indexer helps us to store and retrieve the data quickly than rational database systems. Indexer also have the abilities to store structure and non-structure data like simple text, relation Database data, PDF format, JSON format, XML format and other structured data.

Due to increasing significantly number of devices is a problem of find out a proper device which is fulfills the requirements of the user as well as machine transferable code. Only searching is not enough but also supports reliability and robustness with top ranking resources. It is also difficult as compare to searching traditional web documents because of traditional search on search limited attributes of web documents. Another problem is that there is not a signal standard of describing the devices meta-data. This problem creates another challenge which is lack of describing standard for describing functionality of the devices and resource.

### A. Problem Statements

IoT platform is a scalable network that have huge amount of resources which is needed to be fast search and store. Traditional technique to store the resource description and meta-data using database system has some limitations like storing documents, non-structured data and rank up the search resources. These limitations can be solved by the indexer which can store resource description and provide facility to ranking the using scoring technique. In this research we use different sensor description formats to index descriptions of resources and also ranking up the search results for high performance and reliability. As compare to database systems indexer is more flexible to storing and searching of large data. Only indexer is not enough to search the resources another requirement is user interface which can help user to search resources.

### B. Objectives

In this research different sensor description models are studied and find common descriptions. Different models have different properties and different technologies so we need to parse each model into a shared schema. The parsed description will used further process for a common model to store into indexer. Indexer store description of thing and supply interface to query and sorted its description. After storing the description we need an interface to query the stored information and apply different algorithms to rank the sensors.

- This research compares different sensor description models.
- Find similar fields between different models and descriptions of Smart Things.
- On similarity base set fields for storing in indexer under single core.

### C. Contribution

This research work has developed a search engine for IoT with following functionalities:

- Parsing Smart Things descriptions: Every description model uses different languages for describing the Smart Things so there is need to parse the description from a related parser.
- Indexing of variety of description models: OGC SensorML, SensorThings API and W3C JSON-LD description models are tested in this research.
- Collecting required meta-data from descriptions: Common fields are fetched from description models and bind into a common format without losing the data.
- Searching of Smart Things: This research can help to search the Smart Things using keywords, types of Smart Things, name of location and Geo-Spatial searching within defined range.
- Open API: RESTful API service is provided for indexing and searching the Smart Things.

## II. State of the Art

### A. Internet of Things

The term Internet of Things (IoT) was not officially named until 1999. The pioneer in this term is a British Kevin Ashton that describes a device in the physical world that can communicate with Internet using sensors [5]. After some time Ashton shows the connectivity of Radio-Frequency identity [6] tags which are used in industry to calculate and track items without human interference. Today, the IoT has emerge as a modern term for representing sensors Things connected to Internet with description of computing capability cover objects, devices, sensors, and other daily used items.

A variety of groups studies the wide range of forecasts approximately the potential impact of Internet of Things on the Internet and the economic system throughout the following five to ten years. Cisco, predicted that the number of IoT objects connected to the Internet is greater than 50.2 billion by means of 2020 [7]; Morgan Stanley, however, predicted 75 billion networked objects in 2020. Huawei predictions 100 billion IoT connected devices in 2025 [8]. McKinsey Global Institute indicates that the economic impact of IoT on the international economic system may be as tons as 3.9to 11.1 trillion with the aid of 2025 [9]. The large number of IoT object is grouped and define a term Smart which leads to smart home, smart transportation, smart grid station, smart vehicle etc.

A simple approach is discussed in [10] to discover IoT resources. Discovery is a technique to acquire the data and resources without the knowledge of the source of the data using some discovery applications.

The main characteristic of IoT is "heterogeneity" which means the devices, sensors, and actuators are diverse in nature. The IoT devices used different protocols, different hardware, different data rates etc. The seconds challenge is storage capabilities, energy capabilities and lastly, the format of data ("audio", "video", "streams", "numeric", "textual") producing are also diverse in nature and the standards also. This diversity poses the challenge to discover things. IoT is an important source of producing data that is termed as big-data. The diversity in IoT data creates a challenge to discover the required data for the specific organizations also has the challenge of store heterogeneous data. We need a common framework which covers machine readable representation of data from different formats and stores. There is also need an interoperable mechanism to interoperate devices. These all challenge are needed to be solved for discovery.

The current technique to discover resources is directly connected with the resources using some applications that are restricted only to publisher resources not to discover publicly available resources. The application only discovers the publisher resources which is limited area because of publisher implement only own resources technologies to discovery but in the real world, the discovery mechanism covers all devices and resources which are available publicly and also support to the diversity of the protocols, formats, standards etc. To solve the issue of discovery there is need to describe the resources and devices in a standard format and need to subscribe itself with the discovery engine. The description models solve the diversity problem and discovery engine can easily parse the description of the sensors and resources. So the Discovery Engine provides the interfaces to machines and also for humans to discover the resources. Discovery phase has multiple paths to discover the real world objects using some structured models and apply some kind of knowledge driven queries for context understanding which is possible using discovery mechanism.

### B. Resource Discovery Techniques

In [11] comparison analysis for different resource discovery mechanism that are currently available as well as for future aspects is discussed. Author also proposed a search engine based solution for discover resources. Following are the techniques that author compare.

*1) P2P and distributed resource discovery*: A layered architecture approach is applied in [12] using hash table data

structure for distributed resource discovery with aim of 3 features: range query, p2p routing, and multiple attributes indexing. According to the author, most present techniques do not have the capability of multi-attribute and range queries. A distributed resource discovery mechanisms is discussed.

This solution support large number of heterogeneous devices. This technique is applied for discovery and registration of resources. Author also implements the technique and verified with sampling data and then evaluate on time and response base. An automated service and resource discovery mechanisms are discussed in [13]. This solution is fully automatic and not need to human intervention for configuration.

*2) Centralized architecture for resource discovery*: Jara et al provided a technique for the global resources discovery of devices and sensors across numerous conditions. The technique is called discovery advanced which allows sensors to be registered into a centralized registry. Another service oriented framework is suggested based on RESTful [14] API and JSON. The proposed framework integrated with centralized registry that is responsible for every activity. The resources categorized base on domains and the indexing of the resources also based on the domains. This framework also has a problem because resources are only register on base of domain.

*3) CoAP-based resource discovery*: A service discovery technique is used in [15] by using a RESTfull web server "./wellknown/core" that is responsible for any client request for service discovery [16]. When a client request to the server the server reply with a list of available resources with the attribute that specifies the type of the meta- data of the resource. This may use in numerous areas of a network but have also some limitations such as first-time registration. CoAP cannot implement the first time registration or announced itself the resource [14].

*4) Semantic-based resource discovery*: In [17] proposed a framework for discover resources with the help of Service advertisement. This research provides a semantic enhanced service proxy framework for Internet of Things to service control, service creation, and service discovery and service invocation mechanism in IoT domain. For this purpose author proposed a framework using SOA [18][19] design an ontology for semantics in the resources and a query based service reroute mechanism, micro-formats for describe the resources and a service advertisement mechanism for easy registration and discovery.

In [20] a technique is purposed to discover resources by using micro formats and micro data for WoT named DiscoWoT. Author user the XML, JSON schema and RDF technology to describe the resources and discover the resources using HTTP protocol for query the resources using GET and POST method and the response of query return by Json.

In [21] proposed a computational method for semantic similarity based on ontology and concepts, which is used to discover the resource according to user requirements. This method calculated the semantic distance between two concepts and length between two concepts.

### *C. Overview of Search Engines in IoT*

The traditional search engine uses the crawlers to discovery and searches the documents on the Internet. The major problem with IoT search engine is the sensors, actuators, and other resources are mostly battery powered which is not available all the time. So the crawlers [22] have not collected the information for the search engine. Another problem with this search engine security, most of the sensors are deployed by the owner of a house, shop, industry etc. so they are not allowed to access the most resources like camera, door lock etc.

Following are the some proposed search engine for IoT from the literature.

*1) Keyword-based search engine*: A keyword base for physical object search engine [23] "Snoogle" is proposed. As the physical object has the ability to communicate that‘s mean they make IoT. The physical object is like sensors that have some description. The author claimed that this is the first such kind of Information Retrieval system for the physical objects. Secondly, the author compares his results with existing techniques for reduced data transmission overhead. The response system with bloom filter combine with the search engine is the new concept that is included in this research. Thirdly introduce an algorithm which is used for reducing the cost of the query and response time of the client. Fourthly privacy and security is resolved with cryptography is used. Fifth, develop a simulator to simulate the prototype and validate the proposed solution.

*2) Location-based search engine:* A SenseWeb infrastructure is introduced in [24] in which the private sensors shared their data on a public web. Owners of the sensor upload the data of the Sensor using a GUI on a web. The uploading data only for the private account is not enough if we shared our data with other to use by other applications and human for gain information. With sharing the data help to take actions before an incident as author use example of thunderstorm hit a cab and can automatically share data with other department or other cabs to prepare in advance. Another example of soil sensor for research for students about soil problems and features using the shared data set. SenseWeb is providing a mechanism to shared data among different applications. SenseWeb provides the facility for different applications for shared data but in a uniform way. This architecture is based on Coordinates and Data Transformers.

*3) Real-time search engine*: In [26] search engine Dyser is proposed that have the ability of scalability of Things and also support rapidly change contents or in another word real-time search engine using two approaches, A) Proactive which called Push approach in which sensor update the index itself and query response by the search engine. B) In Pull approach

when a query is initiated then request sends to the sensor for required data. We know that in future there are more sensors than the queries. So the Pull approach is better solution other the PUSH approach in which a huge amount of data produced at real-time. The main assumption of this research is there are many sensors that produced data periodically. So the sensors and entities are assigned a URL to access. In this research, the language for the query is not specifically used a predefined language like SPARQL [25] or SQL. Implementation of this research is using Java and PHP technologies. Sensor Gateways are implanted using SOAP [26] and Java technologies for fetching the information. Gateways also generate automatically pages for testing purposes using REST interface. This feature minimizes the cost of the reading and indexing of the pages instead of gateways generates its self-page and required only current states of the sensors. After that prediction model is mapped based on states.

*4) Hybrid search engine:* A hybrid search framework [27] proposed for searching the IoT elements in three different ways. Here this research will offer to solve this problem by combining the three different techniques in a search engine. The user can search the resources using key-word: user can search the resources using keywords, location base: user can search the resources using specific location and state-base: help the user to search resources using state or value based. This system also called real-time searching because of sampling of data from resources continuously updates [28].

*5) Things description models*: This section describes the details about the different description models for describing of Smart Things. The models which are discussed in this section are SensorML, SensorThings API and W3C JSON-LD.

*a) SensorML:* SensorML is OCG standard designed for the Sensors and actuators. The main features of this standard are interoperability between devices and also with web nodes. Sensor Web Enablement which is controlled by the OGC. In SWE provided the encoding and interface for Sensor Web which is used by applications and services for accessing sensors. SWE defines the following prototypes language which provides SensorML, O&M, SOS, SPS, and SWE-CDS [29]. The current version of SensorML is 2.0 which is defined by the XML language. This provided structure for the process and processing which is related to measurements and post-measurements of the observation. SensorML supports two types of the Physical objects, Physical System and the Physical Component. A SensorML provides the following descriptions for defining the Meta data and details description of the sensor or system.

Fig. 1. Sensorml Namespaces and Schema Tag.

- definition
- type Of
- configuration,
- features Of Interest
- inputs
- outputs
- parameters
- modes

XML heading or starting tag of SensorML with the following namespaces: SensorML (Sensor Model Language) "xmlns:sml" used to define the physical objects with a model called SensorML [30]. SensorML provides two types of models for defining the complete physical system and single component.

In figure 1 the GML (Geography Markup language) "xmlns:gml" name space is used in SensorML to provides the XML based grammar for defining the location of the Physical Components. SWE (Sensor Web Enablement) "xmlns:swe" provide facilities to developers for defining the physical components e.g. sensors, actuators, transducers.

*b) OGC Sensor Things API*: The OGC SensorThings API offers the open, geospatial-enabled and unique system to interconnect the Internet of Things (IoT) [31]. SensorThings has the capability of manage sensor data, devices, and applications associated with IoT. SensorThings API implements two main elements and each element is managed by a component. a) Sensing part provides the metadata and the observation of sensors with the capability of heterogeneous devices [32]. The second Tasking part is not implemented yet and this part is used for sensors and actuators parameterizing. SensorThing API used REST web services, JSON language for manipulation and MQTT protocol for communication [33].

Sensing part of this standard also provides the GET, PUT, DELETE and POST methods to create, delete and update the applications and sensors [34]. This is part is developed based on Observation and Measurement model. This part also provides the location of the Sensor or Thing which helps the application to identify the position of the sensor deploy. The following figure 2 displays the Entities of the OGC SensorThing API which is discussed in details.


Fig. 2. OGC SensorThing API Thing Entity.

Each entity manage two types of concepts a) attributes and b) relationship with other the common attributes that are defined the id and self-link of the entity which is used for processing and locate the actual data from the implemented server. Here are the entities which are offered from SensorThing API.

Thing entity: This Entity represents the Thing.

Observed property entity this defines the observation that is covered under a Thing, This entity main feature provides the definition of the observation of the sensor or Thing.

Observation entity describes the details of the observation, DataStream, phenomenon time and feature of interest in for understanding the context of the Thing as shown in figure 2.

Feature of interest defines the context which is required by the user for observation. This entity helps the user to understand the context or observations in a particular place or location.

*c) W3C JSON-LD*: W3C proposed a language format for interoperability with machines using JSON extends to JSON-LD. Linked Data provides serialization of links to different documents and sites using JSON for interoperability. JSON-LD used JSON based storage because of JSON-LD follows the same format of the JSON.

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you?

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

## III. Research Methodology

### A. Solr Indexing Tool

For the purpose of the storing sensors description in this system, we use the Solr [29] that is working on the top of Lucene. Solr not only provides the indexing also support for analyzing and searching using RESTfull web service. The documents which are supported by the Solr are PDF, Word documents, Text Files, Rich Text Formats, XML, and JSON. Solr also support multiple databases connectivity and index data from the database.

### B. Parser

SensorML uses XML technology and W3C JSON-LD uses JSON with Linked Data technology which is needed to be parsed for obtaining required data.

*1) Orchestra XML Parser*: Orchestra XML parser [30] version 3.0 is used which is an open source parser specially developed using PHP for web service. This Parser used for parsing the SensorML document which is complex as compared to the simple XML file.

SensorML used not only nested tags but also attribute so there is need a proper parser for avoiding the loss of the data which is complex to handle by the simple document reader.

*2) JSON-LD Parser:* Lanthaler [31] developed JSON-LD parser which is an open source parser that is officially used by the JSON-LD. This parser provides Expand, Compact, Frame, Flattern, RDF and String formats. By using this parser we use the Expand method to collect sensor description form the JSON-LD format. The version of Lanthaler JSON-LD 1.2 is used in this system.

### C. APIs

In this system, we use two API‘s which are Google Maps and Solarium which is providing connectivity between Solr server and Web service.

*1) Google Maps*: Google Maps [32] is a web service which provides maps of the world, street view, 360degree view of the globe and many other services. In this system, we used the Google Maps to fetch the locations using coordinates which are offered by the sensor. Location name provides also an easy search for searching sensors for a particular location.

*2) Solarium:* Solarium [33] is a client library for providing services to PHP frameworks by connected with Solr. Solarium provides setting parameters, Modifying query or expanded a query, create indexing documents, Provide CURD operation, building strings, hiding all this with an easy to use API, which is actually your business logic.

### D. Web Framework

Laravel Framework [34] is an open source MVC framework which provides integration of API‘s and also a RESTful Web service. The version of Laravel 5.4 is used in this system. Laravel not just providing web framework also provide such as authentication, routing, sessions, and caching.

## IV. Implementation

A conceptual model of the proposed architecture is provided. The proposed architecture is implemented using several different software/hardware technologies. This chapter describes a prototype implementation of the proposed layered architecture using APACHE, JAVA, XML, JSON, PHP and MVC framework.

### A. Proposed Architecture

In order to complete the proposed solution the layered approach is used to Index and Search the Smart Things in the IoT using a search engine technique.

Figure 3 shows the proposed architecture which support indexing and searching large number of Smart Things, this system provides the following functions.

- Indexing Smart Things: The description of the Smart Things can be indexed using the WEB and RESTful API interface.
- Parsing Description Model: This system automatically detect sensor model language and select parser to collect information.

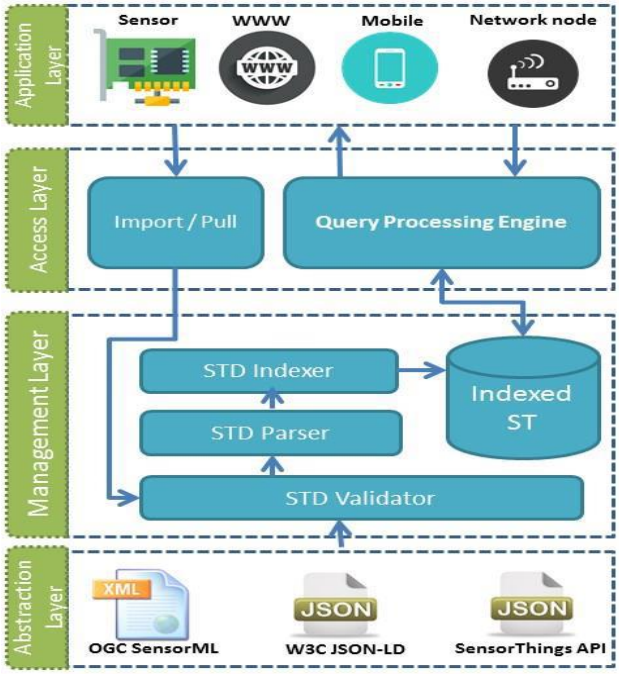



Fig. 3. Purposed Architecture.

- Analyzing Data: This system also analyze the data from the parser and convert it to the proper format under a common schema.
- Collect Structured Information: When the model is parsed this system collected selected fields which are required by the system for searching and indexing.
- Convert coordinates to location: Smart Things are deployed on different locations which are support to Geo-Spatial search. The location of sensors are in GPS coordinates latitude and longitude. The coordinates are needed to further process to fetch name of location.
- Creating Indexes: After collecting and processing the required data this system creates indexes and store them to the indexer.
- Searching Smart Things: The proposed solution provides WEB interface and RESTful API for searching Smart Things with HTML, JSON and XML format.

The proposed architecture consists on four layers Abstraction layer, Management layer, Access layer and Application layer.

*1) Abstraction layer*: This layer covers the sensors description that supports different models of sensors and Things. Different sensors descriptions models has different capabilities to describe the sensors. The different description technologies which are used in this research are OGC sensorML, OGC sensorThings API, and W3C JSON-LD.

*2) Management layer*: In this layer, the description of the sensors parses and index to indexer. First we check "Sensor Description" against the schema of the description models. If Description passes the Schema Model then the required and common attributes of the sensor mapped to our required schema attributes. Using this technique the Sensors Description is now in a common format that is matched our model next step is to store the Sensor Description to the indexer. Because of different description models converted into a common format so there is no need to indexing the Sensor Description separately instead of single indexer.

*3) Access layer*: This layer supports the connectivity of the Application Layer and the Parsing and Storage Layer. Query Parser, Optimizer and Things.

*4) Application layer:* This layer provides the interface to the user for registering their Things and also querying to search engine. This layer support multiple interfaces for web, sensor, machine, and mobile also support open API for query and register things using REST API Service with different methods.

### *B. Abstract Languages and Model*

Description languages of the Smart Things which are used as a standard, have the variety of sections, fields, and attributes. Description models provide identification as well as observation of Smart Things but our system not required all the details from the model but only a few important fields and attributes that are common in all the description models. Follow table describes the fields and attributes with datatype which are required from all implemented Description language models.

*1) OGC SensorML*: SensorML describe the Smart Things using XML language.

ID: Tag <sml:PhysicalComponent gml:id= MY_SENSOR‖> provides the attribute of "gml:id" for ID of the sensor.

Name: Tag <gml:identifier> describe the id of the sensor.

TABLE I. COMMON REQUIRED FIELDS FOR SEARCH ENGINE

| # | Fields | Type | Example |
|---|---|---|---|
| 1 | id | string | Davis7817 |
| 2 | name | string | urn:davisweather:7817 |
| 3 | description | string | a simplet thermistor sensor |
| 4 | Coordinates | double | 47.8 , 88.56 |
| 5 | uri | url | http://www.opengis.net/def/crs/EPSG/0/4326 |
| 6 | type | string | weather station |

Description: SensorML use <gml:description> to provide the short description of sensor. Coordinates: Tag <gml:coordinates>47.8 88.56</gml:coordinates> in

<sml:position> provides the latitude and longitude for the sensor.

Uri: Document itself.

Type: Tag <swe:elementType name=temprature> in <swe:DataStream> provide the type of the sensor.

*2) OGC sensorthings API*

```
{
   "Thing":{
      "@iot.id": 252560,
      "@iot.selfLink": "http://sensors.example.com/v1.0/Things(252560),
            "name": "Humidity Monitoring In a Mall",
            "description": "Collected the Humidity values of different
     shops in a mall"
   }
   "Location": {
      "location": {
         "type": "Point",
                  "coordinates": [30.133, 49.08]}}
```

ID: in Thing, object "@iot.id" provides the id of the thing.

Name: The name attribute also provided by the Thing object with name attribute. Description: Thing attribute also contains the description attribute for the short description of the sensor.

Coordinates: Location object describes the location object with the array of latitude and longitude under the coordinates array.

Uri: Under the Thing object the attribute @iot.selflink provides the URI of the thing. Type: The type of the Thing is described under the object of DataStream with name attribute. We tokenized the name to get the original type of the sensor.

*3) W3C JSON-L*

```
{ "iot": "https://iotdb.org/pub/iot#",
"iot-attribute": "https://iotdb.org/pub/iot-attribute#",
"iot-unit": "https://iotdb.org/pub/iot-unit#",
"vocab": "http://www.example.org/vocab#",
"sensorID": "iot:uuid",
"sensorName": "iot:name",
"sensorDescription": "iot:description",
"sensorValue": {
        "@id": "iot-attribute:sensor.chemical",
"sensorUnit": "iot:unit",
"timestamp": "iot:datetime",
"sensorType": "iot:type",
"sensorUrl": "iot:uri",
"sensorGeo": "http://schema.org/geo",
"latitude": {
        "@id": "http://schema.org/latitude",
```

ID: sensorID describe the ID of the sensor as "https://iotdb.org/pub/iot#uuid": 5484 Name: "sensorName" describe the name of the sensor as "https://iotdb.org/pub/iot#name":

"Flam Sensor with Servo Motor" Type: sensorType describe the Type of the sensor as "https://iotdb.org/pub/iot#type":

"Flam"

URI: sensorUrl describe the URI of the sensor as "https://iotdb.org/pub/iot#uri": http://www.sensors.

Coordinates: http://schema.org/geo": latitude and longitude attributes defines.

### C. Management Layer

In this Layer there is discussion about how an Description Model is dispatched with the Schema and pass the Schema for successful parsing.

*1) Things modeling languages validator*: A schema is a language for describing the information about the tags and attributes which are used in the XML or other languages. There are different schema languages for XML like DTD and XSD for defining the schema of the XML. The XML its self not enable to provide meanings so we need to use the schema for defining the meanings to the tags and attributes, associate the attributes or tags to data types, control the appearance of the tags or positions of the tags and attributes, provide documentation for machine readable and for human readable and proving the formal definition to one or more documents.

As like XML schema validator we have JSON schema validator also which validate the JSON against JSON Schema. Before processing further to parse or obtain data from the file or from the server we need to verify the attributes against the schema. JSON-LD also has the capability of multiple and complex data types which are described under the @context object to validate each attribute with data types.

Schema validator also helps us to verify the uniqueness of the data from different resources. This mechanism helps to produce a compatible system with all other IoT resources which are wishing to connect with our system. If the validator passes the data then process further otherwise discard the processing and throw an exception of error a code for response to the requester.

*2) Smart things description language parser*: This module parses the Things modeling; Languages after validating the schema of the related Model. When the schema validates the document of Things then the next procedure is started which is parsing. The main purpose of this module is parsing. In simple this module collected the related information from the document which is required. This module collects all the attributes step by step procedure which is defined by internal logic. We need three different model parser because of we implanted three different Molding Languages.

- SensorML used XML language to describe the sensor meta-data and description. Our proposed work works on WEB, so that we use a PHP language to parse the

model. XMLReader is used to read the document. Because of XML has the capability of nesting tags. So we have a lot of processing to collect required attributes. As mention early in Background Study chapter under the section of SensorML, The XML document has multiple namespaces so we need to examine all the namespaces for collected related data, not unrelated data.

- SensorThings API used JSON for describing the Sensors and Things. This Modeling Language has also other related objects like observations, locations, data streams etc. We need only required attributes which are in the entities of Things, Sensors and Location. So we need recursive processing to retrieve the all required attributes because of the entities have different URL's.
- W3C JSON-LD is based on JSON language for describing the Smart Things. All required attributes are defined in same file. After collecting the required attributes from these description models the next step is to convert into same schema called Thing Schema for removing and refining the attributes and for further processing.

*3) Smart things indexer:* In this module, the collected attributes are mapped to the internal schema and refined. The refinement of the collected attributes is necessary because of the use of the different Modeling languages. In this module, the collected attributes are formatted in common data-types because of the internal indexing system. If the format system is not used then we have another problem which is the heterogeneity of the Modeling Languages. To solve the heterogeneity problem there is need of different indexer and when we query the data we need to combine the indexer on the query. So one problem creates another problem to solve this problem the best way is to convert the data types into same data types without losing the data. One attribute which is URI is used for getting original data from the sensor.

Another requirement for location-based query needs because of human search with an area. But in the Modeling Language cannot provide the location names but only coordinates latitude and longitude. We have another scenario which is query base on location. To solve this issue we have Google Map API which is helping us to convert coordinates to location. Google API returns the Country, State, City, Postal code and street address. This information is enough for searching sensors based on specific locations. The attributes which are required from modeling language are mentioned in 4.1 tables. Required processing done thing information sends to indexer to store.

*4) Indexing descriptions of smart things*: This module is the major part of the search engine which holds meta-data of the Things. For index meta-data, we use lucene Solr which provides RESTfull API for selecting, updating and inserting structure data. Solr provide the customized schema to store structure documents. For meeting the system requirements, we customize the schema.

Solr work as a cloud service and use JAVA language. So we need an API to interact with Solr because of this research deployed on HTTP server. A plug-in named Solarium PHP is used for connectivity with the Solr and with our server.

Solarium PHP provides connectivity and also the mechanism to query and updating the index, customized schema and other facilities. Following are the steps to index the sensor.

- After parsing the sensor to the schema, the converted schema transforms as Solr document (Thing) for storing.
- Initialized a client and open connection with Solr using Solarium PHP API and send the data to the Sorl.
- Solr check income data with the ID. If the ID already exists then update the data and if ID does not exist then new document tokenized and store in Solr Indexer.
- After updating the Things Solr response back to the client with Code and Query Time.
- If the Thing is not updating within a given time period a service delete the document from the indexer.

Solr has the capability of handle multiple clients and queries at a time by distributed mechanism.

*D. Access layer*

This layer provides query, optimization, results, ranking, importing and pulling things from servers and our system. This layer also helps us to connect with the user interface using HTTP requests.

*1) Importing things*: In this module importing mechanism is implemented in which user can import things and sensors using Modeling Languages which are used in our system. Using RESTful Web Service ./import/ we can import Things and Sensors.

- SensorML ./import/sensorML/ used to import sensorML document which is written in XML language.
- SensorThings ./import/sensorThings/ used to import sensorThing document which is written in JSON language.
- W3C JSON-LD ./import/W3CJson/ used to import W3C statandard JSON-LD document which is written in JSON-LD language. These documents are sent to validator Module in Parsing and Storage Layer. After completing the Parsing process the document store in indexer and response back to the user using back chain process. After successfully storing the document to indexer the document its self "file" is deleted from the server.

*2) Pulling smart things from sandboxes*: In this module, we can pull Sensors or Things from Sandboxes. SensorThings API implemented by other service providers so we have the capability of pulling data from that service provider which publicly available. E.g sensorup.com implements the

sensorThing server and also provides the open access to the server for working with the sensorThing server. We need only the URL of Things our system provides the facility to pull all data from the server in the recursive fashion.

Here is the implemented service http://toronto-bike-snapshot.sensorup.com/v1. 0/Things by using this URL in our system all Things that are available for public fetched and other attributes like "Location" are also fetched by using internal logic.

*3) Query processing engine*: Query parser accepts the query from the application layer and sends it to Optimizer for optimization for query for better results. Query parser following the step to complete the process these are following:

- The Application Layer has multiple techniques implemented so there is need to identify the query first then send to query parser.
- All application use HTTP protocol so we can easily understand the query because of defining a RESTful service.
- All search requests are created using GET method and send to /search/.
- Our system supports two types of queries keyword and Geo-Spatial.
- When query receives it evaluate the types of the query and dispatch the query to the optimizer to optimize the query for better results.

Here are the implanted query types in this system. This search engine support two types of query a) Keyword Based Query and b) Geo-Spatial Based Query.

*4) Keyword query*: Keyword Query support full text based query for search Things. The user can query with the keyword and this system fully supports all the attributes for the specific keyword. RESTful URL "./search/keyword" accepted the request of the keyword query. "Keyword" is the text which is requested to search. The user initiated the query using HTTP protocol. If the keyword based on more than one word then the words tokenized based on space and make the search for each token.

*5) Location/geo-spatial query*: The second approach is Geo-Spatial Query which is a complex query. The user can search for thing using a specific location and also within a range of the specific location. Geospatial Query needs three attributes latitude, longitude, and distance.

- "Latitude" which define the angular distance of the earth from north or South Pole.
- "Longitude" defines the distance of the earth from the east and west.
- "Distance Attribute" defines the distance from latitude and longitude in KMs.
- "Format" attribute defines the results format current supported "JSON and XML". So the user replaces format with JSON or XML.

Using all these three attributes query for Geo-Spatial Query use the RESTful URL as

/search/lat/?/long/?/dist/?/format. lat defines the latitude and first "?" is latitude value, long defines longitude and second "?" is the value of the longitude and finally, dist. defines distance in km third "?" is the value of the distance. If the user wants distance in miles then use dist.*0.621371.

*6) Query and results optimizer:* This module helps to optimize the query and search results and control data between Indexer in Processing and Storage Layer and with Core Layer.

*7) Query optimizer*: The raw query may use the more resources that not suitable for the huge amount of data. so there must be a mechanism to sort out this problem. We need an optimizer minimized the cost of the resources because of in future the data is generated more and more that cause the system speed down. The query optimizer also supports for multiple indexers. Optimizer select different query planner for the same query by the internal logic and select best query to minimize the overhead and resources.

*8) Results optimizer*: While the results against the query are retrieved from the Indexer there is need to optimize the results. All the results may not require by the user there is need a ranked results base on frequent results or most search Things. So the optimizer sorts out the results and rank based on the score of the Things. After ranking and sorting the results send to Search Results module.

*9) Search results*: Optimizer sends the results to this module. This module first converts the index documents which is return against the query to the required format. As mentioned earlier in the Query Parser format tells the system to in which language output is generated. The following attributes are defined in each of the formats. Id: defines the unique id for identifying the sensor. Name: defines the name of the according to Modeling Languages. Description: provides the short description of the sensor. lat-long: provide the latitude and longitude in "31.56, 65.66" format separated by ",". Type: defines the sensor type e.g "temperature, motion, vibration" etc.

### *E. Application Layer*

Currently, Application Layer support only HTTP protocol for displaying results and query interfaces using HTML and CSS technologies to interact with the system. The Layer also provides features to another system to directly interact with the system by using open REST API with an authentication token. WEB interface used to directly interact with the system by the human. Other network devices also interact with the system using RESTful Web Services. This layer also provides the service to other systems to creating, updating, querying and resulting with the system. This layer also supports multiple languages responses back to the requester. This system helps the user to search and store their sensors and things using multiple Modeling Language in the web browser.

Commonly IoT devices connected to the centralized hub for accessing and storing data. In this system, there is no need of centralized hub system we use the web browser to handle the things. The web browser connects to the cloud using REST APIs so there is no need of getting access to the server or hub.

New Thing and Sensor also added to the system using API which is provided in this system for achieving the desired operation can be solved under the following modules.

*1) URL resolver*: This module translates the URL and solves the protocols issues. Currently, support HTTP protocol only so the protocol parameter set defaults to HTTP. This module also understands the request and map over the correct model for processing further.

*2) User interfaces:* This module provides the UI for interaction with the system using HTML, CSS, and Javascript language. This module works under two different conditions one for guest and one for the system user.

*a) Guest*: This UI for available for the public for searching the sensors, get the description in different formats. The formats that are used fo display HTML and CSS used, the user also has the ability to get the description in XML and JSON format.

*b) System User***:** This user need the login to the system if he uses the WEB, otherwise the user can also send the request to the system using authentication token. Authentication token provides to the user after registration with the system. System user has also the ability to register and update Thing directly from any application or language using POST and PUT method.

*3) Inserting thing*: System user has the ability to insert the Things using the authentication token. For inserting new Things with the system user can use POST method to the REST URL /thing/insert. Parameters that are required to create the new Thing are defined in Table 4.1. An extra parameter that is required for authenticating the request for security purpose is "Token". "Token" is string type and replaces the value of authentication token which provides the system on registration. When request successfully accepted a response will get in JSON format with the id of the Thing. This id uses next time to update the Thing.

*4) Updating thing*: System user has also the ability to update the Things after creating the Thing by using the authentication token. For updating the Things with the system user can use PUT method to the REST URL http://www.example.com/thing/update/id. Parameters that are required to update the Thing are defined in Table 4.1 table. An extra parameter that is required for authenticating the request for security purpose is "Token". "Token" is string type and replaces the value of authentication token which provides the system on registration.

*5) Query interface:* The UI also helps to query the system with Query Interface. Two types of queries are supported currently keyword and Geo-Spatial. The user can switch from keyword to Geospatial based query by selecting radio button on the page of Search Things.

*6) Result interface*: After the query to search engine results are returned from the system which is displayed on the results page on the system.

## V. Results and Discussion

Few experiments were conducted to validate the Indexing and Searching of the prototype. Indexing technique implemented using Solr by providing the WEB interface and also searching using the same interface. A REST API is also tested and validated by Indexing and Searching of Smart Things for machines. The proposed solution also support for indexing Smart Things from different sandboxes.

### A. Results

This research helps the sensors description model to parse, index and searched. Following are the results which are performed in this system.

*1) Indexing*: Figure 4 provide the GUI to index the description of the Smart Things. This interface has the ability to select the variety of description models that are selected more than one at a time.

The selected models are parsed using different parsers and then converting into the shared schema that is shown in the figure 5 the common fields that are same in models are fetched and converting the fields into same data types.

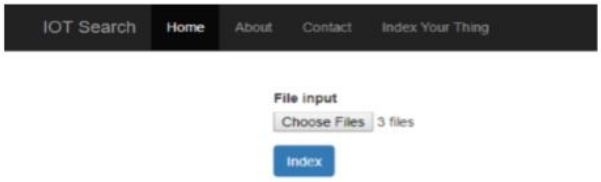



Fig. 4. Indexing Smart Things by Uploading Descriptions.


Fig. 5. Smart Things Description Parsed into Shared Schema.

```
[
  {
    "id":"sensorThing:3971",
    "description":"Ita nemo beato beatior....",
    "name":"Flam Sensor with Servo Motor",
    "latlong":"30.828,72.038",
    "location":"address:Gharh Mahraja Road,city:Shorkot,district:Jhang District,state:Punjab,country:Pakistan"
    "uri":"http://toronto-bike-snapshot.sensorup.com/v1.0/Things(3971)",
    "type":"Flam",
    "_version_":1582571829369765888
  },
  {
    "id":"urn:tr:21449",
    "description":"Primum quid tu dicis breve...",
    "name":"IR Sensor on High way raod",
    "latlong":"30.431,71.744",
    "location":"address:Unnamed Road,district:Khanewal District,state:Punjab,country:Pakistan",
    "uri":"http://www.opengis.net/def/crs/21449",
    "type":"IR",
    "_version_":1582571829444214784
  },
  {
    "id":"JsonLD:30474",
    "description":"Ita nemo beato beatior.....",
    "name":"Gas Sensor with available actuators",
    "latlong":"31.141,71.261",
    "location":"address:Unnamed Road,district:Layyah District,state:Punjab,country:Pakistan",
    "uri":"http://www.sensors.org/sensor/30474",
    "type":"Chemical",
    "_version_":1582571829449457664
  }
]
```


Fig. 6. Smart Things Descriptions Indexed in Solr.

**Indexing query executed**
Number of Smart Things Index : 30
Query status: 0
Query time: 0.791 seconds

Fig. 7. Smart Things Description Index Response from Solr.

After parsing the models into shared schema the collection of schema convert the model into the Solr document. Solarium provided the connectivity between Solr and web framework.

Using Solarium the documents transferred to Solr from the web figure 6.

When the documents are indexing into the Solr the response of the indexed files with the status of indexing and time is the return which is shown in the figure 7.

The figure 8 compared the response time of indexing the description of the Smart Things. The figure shows that the indexing of description from 5 to 500 with time in seconds. All the description models have the same time to indexing with a little difference.

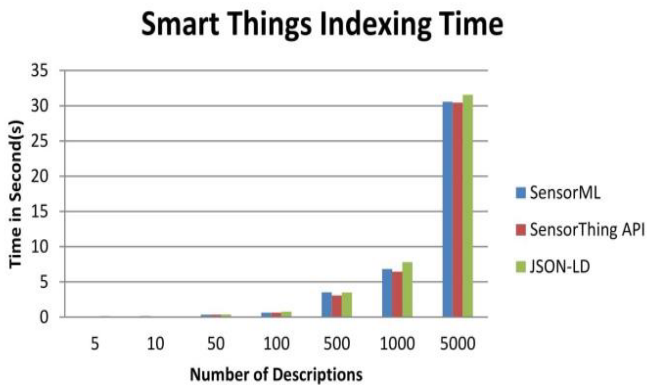



Fig. 8. Response time for Indexing of Descriptions.

*2) Searching:* After indexing the description of Smart Things the next step is to search. Our system provides two types of the query for searching Smart Things first Keyword-based searching and second one is Geo-Spatial using coordinates of location with range. In figure 8 keyword-based searching interface is shown in which user type the keywords like vibration, temperature, flam, fire etc.

Another keyword technique is implemented using types of the sensor. As shown in figure 9 all sensors which are register with this system dynamically load all the types of the Smart Things with a number of register Things.

Under the keyword-based search, the second functionality which is implemented is the location as a keyword. As shown in figure 10 users can also type the name of the location and search Smart Things which are deployed on given location.

The second approach for searching Smart Things is Geo-Spatial which used coordinated along with latitude and longitude and distance which is depicted in figure 11. The user can search the Smart Things for a specific location with distance kilometers.

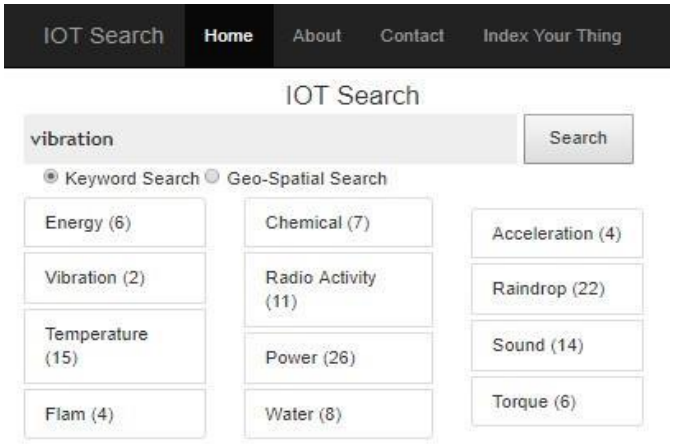



Fig. 9. Keyword based Search WEB Interface.

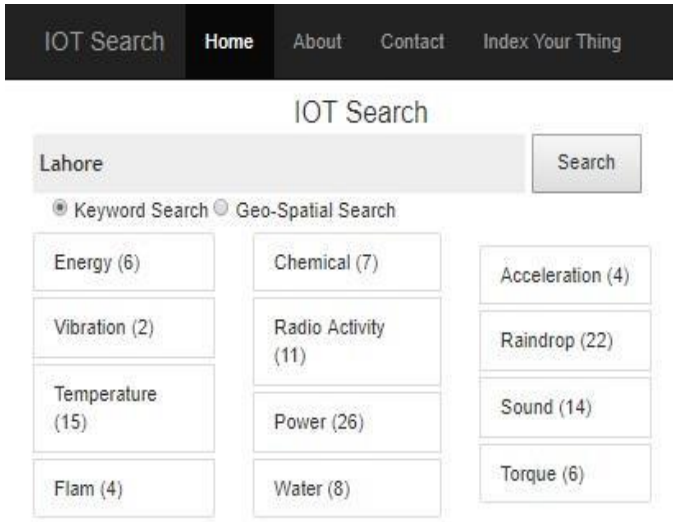



Fig. 10. Searching Smart Things with Name of Location.

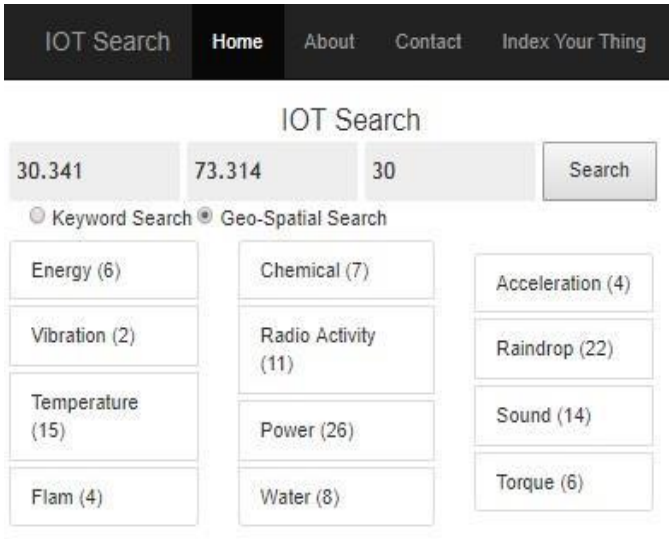



Fig. 11. Geo-Spatial Search of Smart Things with Coordinates.

*3) Search results:* The results of the search responses by the interface which contain the following fields id, description, name, lat-long, location, URI and type of the sensor as shown in the figure 12.

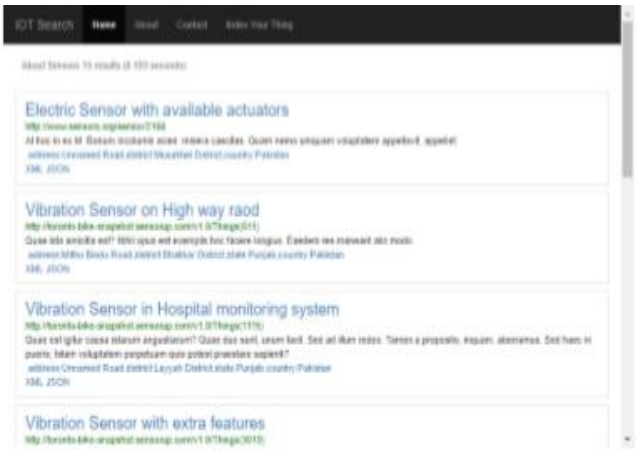



Fig. 12. Search Results for Keyword.

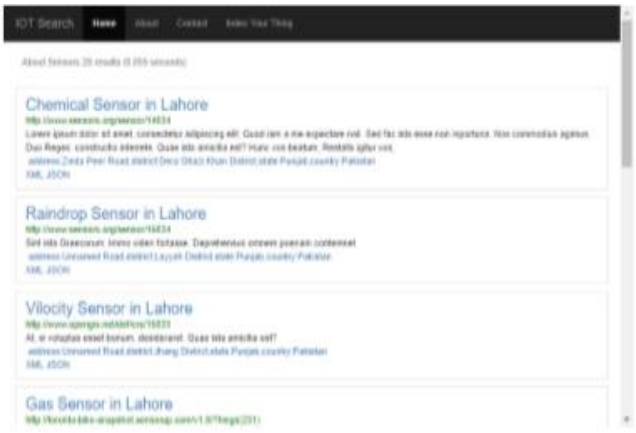



Fig. 13. Search Results for Name of Location.

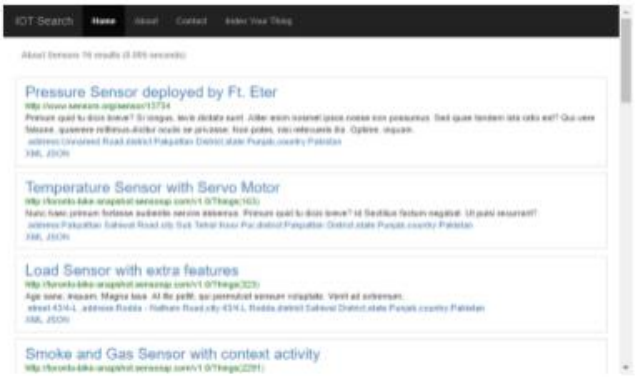



Fig. 14. Search Results for Geo-Spatial Query.


```
{
    "time": 0.002,
    "num": 1,
    "results": [
        {
            "id": "JsonLD:14034",
            "description": "Lorem ipsum dolor sit amet, c
            "name": "Chemical Sensor in Lahore",
            "latlong": "30.441,70.438",
            "location": {
                "address": "Zinda Peer Road",
                "district": "Dera Ghazi Khan District",
                "state": "Punjab",
                "country": "Pakistan"
            },
            "uri": "http://www.sensors.org/sensor/14034",
            "type": "Chemical"
        }
    ]
}
```


Fig. 15. Search Results for Machine with JSON Format.

The figure 13 depicts the results of the location-based search as a keyword. It acts like a keyword search on indexed Smart Things.

The figure 14 showed the results of the Geo-Spatial search approach with a range of distance. Using this technique user can collected data for the specific area.

This system cannot provide only the GUI for the user but also interpretable between machines. The machine can also query and index for the Smart Things over HTTP and using JSON and XML formats. In figure 15 shown the results of the search by machine in JSON format. The machine specifies the format of results at the time of the query.

## VI. Conclusion

The searching of physical objects is difficult as compare to traditional searching of documents for web. The traditional systems used database to store and used crawlers to retrieve the meta-data form the websites. On the other hand description of physical objects has lack of standardization to describe itself. Traditional document search engines provides only keyword-based search which is not enough for searching of the physical objects. Databases seem not feasible to handle the exponential growth of documents. The crawlers are facing a challenge for retrieving the physical objects in IoT. These Smart Things increasing significantly and have a variety of descriptions models. These models are using different languages that pose a challenge of fetching useful meta-data required for searching. The area of searching and indexing of

Smart Things in IoT is not fully explored. The variety of the description models is a barrier to searching and indexing of Smart Things. The technique proposed in this research can helps to index and search Smart Things with a variety of description models. The major functionalities provided by the proposed system are retrieving Smart Things descriptions parsing of descriptions, analyzing meta-data, collecting structured data, converting coordinates to locations and develop index of Smart Things. Two interfaces first web interface and second for machine with XML and JSON formats. The developed solution currently supports SensorML, SensorThings API and W3C JSON-LD standards which are used to describing Smart Things. These standards are automatically analyzed by this system with internal logic and indexed them into Indexer.

The current system provides interfaces for the machine and for the human to search Smart Things using two types of queries. First keyword based and second one Geospatial. Keyword-based query not only provides the searching description of the Smart Things but also supports the location as a keyword either city, district, province or country. The second search is Geo-Spatial which helps the user to search Smart Things using coordinates with latitude, longitude, and distance. Our findings showed significant search results that can lead to an innovative search engine for Internet of Things.

## VII. Future Work

For future looking for Optimizing search results on most frequent searches, Optimizing indexing process, Develop more rich queries such as values based and real time, Handling mobility of Smart Thins, Extending search results to a variety formats and Solution can be moved to cloud base.

## References

[1] Tapscott D, Williams AD. Don Tapscott, Anthony D. Williams n.d.

[2] Botts M, Percivall G, Reed C, Davidson J. OGC (R) Sensor Web Enablement: Overview and High Level Architecture. Lect Notes Comput Sci 2007;4540:175–190. doi:10.1007/978-3-540-79996-2.

[3] Mitton N, Petrolo R, Loscr V, Mitton N. Towards a smart city based on cloud of things , a survey on the smart city vision and paradigms the smart city vision and paradigms 2015. doi:10.1002/ett.

[4] Rose K, Eldridge S, Chapin L. THE INTERNET OF THINGS: AN OVERVIEW. Understanding the Issues and Challenges of a More Connected World. Internet Soc 2015:80. doi:10.5480/1536-5026-34.1.63.

[5] Gope P, Amin R, Hafizul Islam SK, Kumar N, Bhalla VK. Lightweight and privacy-preserving RFID authentication scheme for distributed IoT infrastructure with secure localization services for smart city environment. Futur Gener Comput Syst 2018;83:629–37. doi:10.1016/j.future.2017.06.023.

[6] Kilper DC, Bergman K, Chan VWS, Monga I, Porter G, Rauschenbach K. Optical Networks come of Age. Opt Photonics News 2014:51–7. doi:10.1364/OPN.25.9.000050.

[7] Riggins FJ, Wamba SF. Research Directions on the Adoption, Usage, and Impact of the Internet of Things through the Use of Big Data Analytics. Syst Sci (HICSS), 2015 48th Hawaii Int Conf 2015:1531–40.

[8] Friess P, Vermesan O. Building the Hyperconnected Society. 2015. doi:978-87-93237-99-5.

[9] Guinard D, Trifa V, Karnouskos S, Spiess P, Savio D. Interacting with the SOA-based internet of things: Discovery, query, selection, and on-demand provisioning of web services. IEEE Trans Serv Comput 2010;3:223–35. doi:10.1109/TSC.2010.3.

[10] Arnold AS, Wilson JS, Boshier MG. Arnold , A S and Wilson , J S and Boshier , M G ( 1998 ) A simple extended- cavity diode laser . Review of Scientific Instruments , 69 ( 3 ). pp . 1236- This version is available at https://strathprints.strath.ac.uk/33635/ A simple extended-cavity diode l 1998;69:1236–9.

[11] Paganelli F, Parlanti D. A DHT-Based Discovery Service for the Internet of Things. J Comput Networks Commun 2012;2012:1–11. doi:10.1155/2012/107041.

[12] Tuna G, Kogias DG, Gungor VC, Gezer C, Taşkın E, Ayday E. A survey on information security threats and solutions for Machine to Machine (M2M) communications. J Parallel Distrib Comput 2017;109:142–54. doi:10.1016/j.jpdc.2017.05.021.

[13] Alghamdi TA, Lasebae A, Aiash M. Security Analysis of the Constrained Application Protocol in the Internet of Things Security Analysis of the Constrained Application Protocol in the Internet of Things. Futur Gener Commun Technol (FGCT), 2013 Second Int Conf 2013:163–8. doi:10.1109/FGCT.2013.6767217.

[14] Li S, Xu L Da, Zhao S. 5G Internet of Things: A survey. J Ind Inf Integr 2018;10:1–9. doi:10.1016/j.jii.2018.01.005.

[15] Jara AJ, Lopez P, Fernandez D, Castillo JF, Zamora MA, Skarmeta AF. Mobile Digcovery: A Global Service Discovery for the Internet of Things. 2013 27th Int. Conf. Adv. Inf. Netw. Appl. Work., IEEE; 2013, p. 1325–30. doi:10.1109/WAINA.2013.261.

[16] Fielding RT, Taylor RN, Erenkrantz JR, Gorlick MM, Whitehead J, Khare R, et al. Reflections on the REST architectural style and "principled design of the modern web architecture" (impact paper award). Proc 2017 11th Jt Meet Found Softw Eng - ESEC/FSE 2017 2017:4–14. doi:10.1145/3106237.3121282.

[17] Wollschlaeger M, Sauter T, Jasperneite J. The future of industrial communication. IEEE Ind Electron Mag 2017;11:17–27. doi:10.1021/ie50124a022.

[18] Soylu A, Giese M, Jimenez-Ruiz E, Vega-Gorgojo G, Horrocks I. Experiencing OptiqueVQS: a multi-paradigm and ontology-based visual query system for end users. Univers Access Inf Soc 2016;15:129–52. doi:10.1007/s10209-015-0404-5.

[19] Ejaz W, Naeem M, Shahid A, Anpalagan A, Jo M. Efficient Energy Management for Internet of Things in Smart Cities. IEEE Commun Mag 2017:84–91. doi:10.1109/MCOM.2017.1600218CM.

[20] Zhou M, Ma Y. A web service discovery computational method for IOT system. 2012 IEEE 2nd Int. Conf. Cloud Comput. Intell. Syst., IEEE; 2012, p. 1009–12. doi:10.1109/CCIS.2012.6664533.

[21] Durumeric Z, Adrian D, Mirian A, Bailey M, Halderman JA. A Search Engine Backed by Internet-Wide Scanning. Proc 22nd ACM SIGSAC Conf Comput Commun Secur - CCS '15 2015:542–53. doi:10.1145/2810103.2813703.

[22] Yi S, Li C, Li Q. A Survey of Fog Computing: Concepts, Applications and Issues. Proc 2015 Work Mob Big Data - Mobidata '15 2015:37–42. doi:10.1145/2757384.2757397.

[23] Guo B, Wang Z, Yu Z, Wang Y, Yen NY, Huang R, et al. Mobile Crowd Sensing and Computing: The Review of an Emerging Human-Powered Sensing Paradigm. ACM Comput Surv 2015;48:1–31. doi:10.1145/2794400.

[24] Nitti M, Pilloni V, Colistra G, Atzori L. The Virtual Object as a Major Element of the Internet of Things: A Survey. IEEE Commun Surv Tutorials 2016;18:1228–40. doi:10.1109/COMST.2015.2498304.

[25] Khan MW, Abbasi E. Differentiating Parameters for Selecting Simple Object Access Protocol ( SOAP ) vs . Representational State Transfer ( REST ) Based Architecture. J Adv Comput Networks 2015;3. doi:10.7763/JACN.2015.V3.143.

[26] Datta SK, Da Costa RPF, Bonnet C. Resource discovery in Internet of Things: Current trends and future standardization aspects. IEEE World Forum Internet Things, WF-IoT 2015 - Proc 2015:542–7. doi:10.1109/WF-IoT.2015.7389112.

[27] Pydipaty R, Saha A. On Using Non-Volatile Memory in Apache Lucene n.d.

[28] Chang B-R, Tsai H-F, Hsu H-T. Secondary index to Big Data NoSQL Database – Incorporating solr to HBase approach. J Inf Hiding Multimed Signal Process 2016;7:80–9.

[29] Yu HR. Design and implementation of web based on Laravel framework. Atl Press 2015:301–4. doi:10.2991/iccset-14.2015.66.

[30] Sporny M, Bazaar D, Kellogg G, Associates K, Lanthaler M, Sporny M, et al. A JSON-based Serialization for Linked Data 2018.

[31] Kingston R. Public Participation in Geocomputation to Support Spatial Decision-Making. Geocomputation A Pract Prim 2015:301–19.

[32] Smiley D, Pugh E. Apache Solr 3 Enterprise Search Server. Search 2009:7–10.

[33] Laravel: Up and Running - O'Reilly Media n.d. http://shop.oreilly.com/product/0636920044116.do (accessed June 14, 2018).

[34] Medina - Santiago A, Cisneros - Gómez A, Melgar - Paniagua EM, Gutierrez T, B Nango - Sólis MG, Moreno - López EA, et al. Web Application Development by Applying the MVC and Table Data Gateway in the Annual Program Budget Management System. IJACSA ) Int J Adv Comput Sci Appl 2017;8:250–5. doi:10.14569/IJACSA.2017.080233.

# Automatic Pavement Cracks Detection using Image Processing Techniques and Neural Network

Nawras Shatnawi
Department of Surveying and Geomatics Engineering,
Al-Balqa' Applied University, Al-Salt, Jordan

***Abstract*—Feature extraction methods and subsequent neural network performances were used in this research to impose proper assessment for distressed roads for a case study area in the North of Jordan. Object recognition method was used to extract roads cracks from airborne images acquired by drones. After images has been thresholded and the noise removed, digital image processing algorithms were applied to detect the presence of different crack types in the surface of pavement. In addition to that, the process was capable to automatically determine the length and the orientation of the cracks which were used as input for a neural network pattern recognition function designed for this purpose. Artificial Neural Network was used, tested and verified for cracks extraction. Different patterns and numbers of hidden layers were also investigated. The results revealed that using image processing techniques and neural network could detect pavement cracks with high accuracy.**

***Keywords*—Artificial neural network (ANN); feature extraction; image processing; pavement crack***

## I. Introduction

Roads are one of the most important elements in the civilized communities. They play vital role in connecting villages, cities, districts, states, production lines, and even countries together. Due to frequent and high load of vehicles on roads, it is normal to have failures in the top surface (pavement) of the roads. Standard failures of pavement are called distresses which are divided into several types that affect the performance of the road; these distresses can be categorized as: Cracking, which is the major distress type in main roads, while in secondary roads, potholes, patches or rutting are often found [1, 2]. Mainly, nineteen distresses do exist in flexible pavement. However, this study will focus on cracks evaluation and assessment. In order to maintain good performing roads, proper maintenance should be followed frequently based on accurate pavement inspections and surveys. This kind of inspections can be done by specialists who can monitor, acquire images, collect data and assess roads distresses. This traditional way of data collection has many drawbacks such as: labor intensive, time consuming, dangerous specially in highways and prone to subjectivity [3], [4] and [5]. Therefore; Automatic distresses data collection using digital imaging technology has been used by researchers since more than 50 years, this method reduces disturbance to the public traffic and road hazard to human inspectors during the survey [6], [7]. Different techniques were used for automatic cracks detection in roads pavements such as [15] and [16], Medina et al, [8] used Gabor filter to detect concrete tunnels cracks, the new used method was capable to detect cracks in any direction with an accuracy of 95.27%. Zhang et al. [17] automatically detected cracks in subway tunnel with the application of high-speed complementary metal-oxide-semiconductor (CMOS) industrial cameras, using morphological image processing techniques and thresholding operations. Oliveira and Correia [9] applied a modified Otsu thresholding algorithm on digital images after dividing them into different zones and compared each pixel with the surrounding histogram. Fukuhara et al. [14] provided methods that could only judge whether a crack existed and the results were not good enough, while Zhang et al. [10] used neural networks for cracks detection in roads pavements. The used methodology of the current research has an accuracy of 82.5 %, which is much less than Zhang et al. [17] for example. However, it can be considered good considering the high sensitivity and precision which were 90% and 87 % respectively. The previous works used different methodologies for automatic pavement cracks detection. Up to the author knowledge, the combination of bilateral filter, Canny algorithm, k-means clustering and a morphological filter with ANN for pavement cracks detection from airborne images is recent.

## II. Work Methodology

The followed methodology in this work consists of four main parts: Image enhancement and pre-processing, feature extraction, Neural network modeling and validation process as shown in Fig. 1.

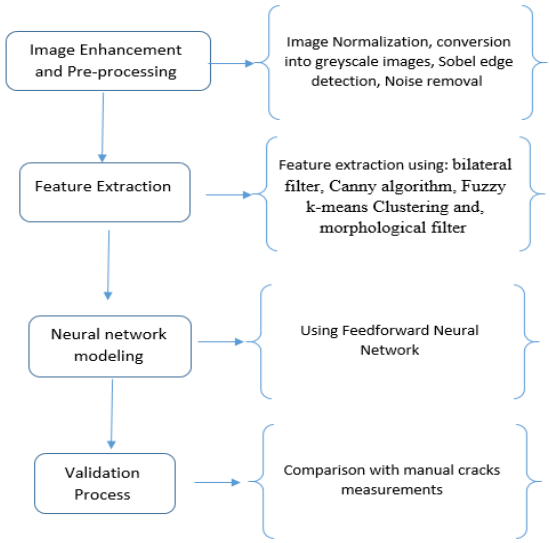



Fig. 1. Flowchart Showing the Methodology Followed in Conducting the Study.

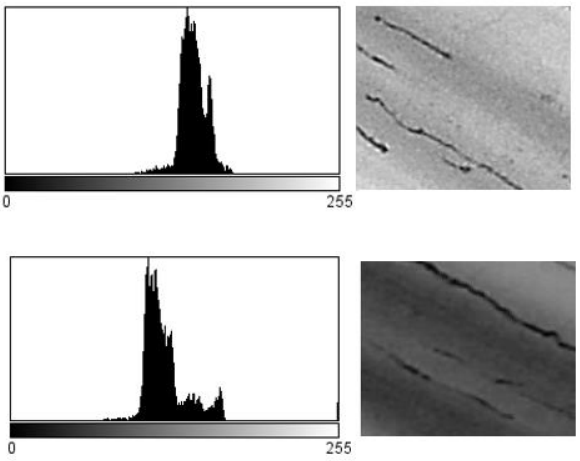



Fig. 2. Histogram of Road Pavement with Cracks.

In the first step; the acquired images by a drone were collected and preprocessed , the collected images were running through normalization function to average the intensity of the images, the images were then converted to grayscale, after that Sobel filter was used for edges detection followed by Noise removal filter. Thresholding was then applied to observe features. In fact, cracks can be identified by visual inspection which appears darker and different compared with road pavement pixels in the histogram as shown in Fig. 2. But unfortunately, the gray-level distribution of cracks is impossible to be separated from the gray-level distribution of the road pavement, therefore, image processing is required to support solving such problems.

Different extraction methods such as: bilateral filter, Canny algorithm, k-means clustering and a morphological filter were used, tested and compared. Among these methods, morphological filter 'closing' -which is simply a dilation followed by an erosion filter- this compound morphological operation showed the best results according to the minimum discrepancies between the actual and measured number of cracks and between the actual and measured length of cracks. The extracted features from the previous step were then used for pattern analysis and classification as input to the ANN module. In this research Feedforward Neural Network was used to simulate the extracted features. Finally, the extracted cracks values compared to manual cracks measurements on the road surface for model validation.

## III. Database Preparation

Sub-images were prepared from the captured aerial images, 800 image sets were used as source of data of cracked pavement and crack-free pavement as shown in Fig. 3.

Cropped images of the same crack were rotated to provide four more different directions for the ANN modeling as shown in Fig. 4.

In order to classify the images using ANN, data was divided as follows: Training set consists of 80 % of data to build the model and determine the parameters (weights and biases), validation data set includes 10 % of data to measure the performance of the network, and 10 % of data is used to increase the robustness of the model in the test phase. The structure of the used ANN is shown in Fig. 5, in which Variable Learning Rate Backpropagation learning function is used to train the network.

Fig. 3. Database Sampling from Images Subsets.

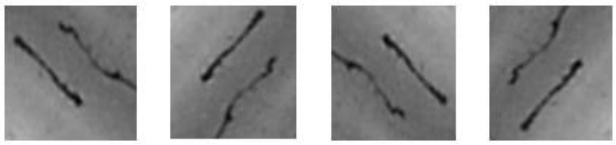


Fig. 4. Cropped Image of the Same Crack Rotated to Provide Four Different Directions.

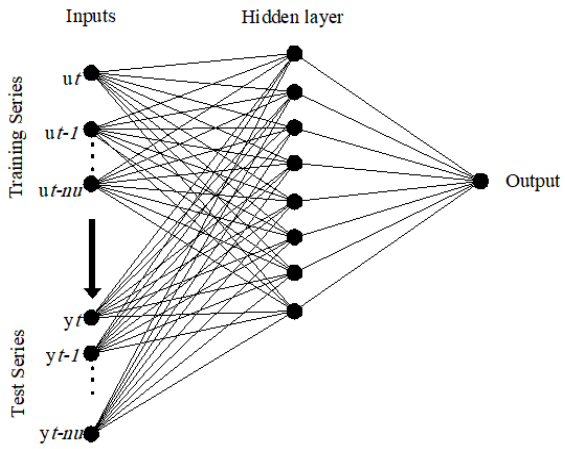



Fig. 5. Structure of the used ANN Model for Cracks Detection Model.

TABLE I. Optimal Number of Hidden Neurons in the Hidden Layer

| Number of Neurons | Mean Squared Error (MSE) |
|---|---|
| 1 | 0.0324 |
| 2 | 0.0281 |
| 3 | 0.0216 |
| 4 | 0.0201 |
| 5 | 0.0171 |
| 6 | 0.0113 |
| 7 | 0.0101 |
| 8 | 0.0067 |
| 9 | 0.0078 |
| 10 | 0.0092 |

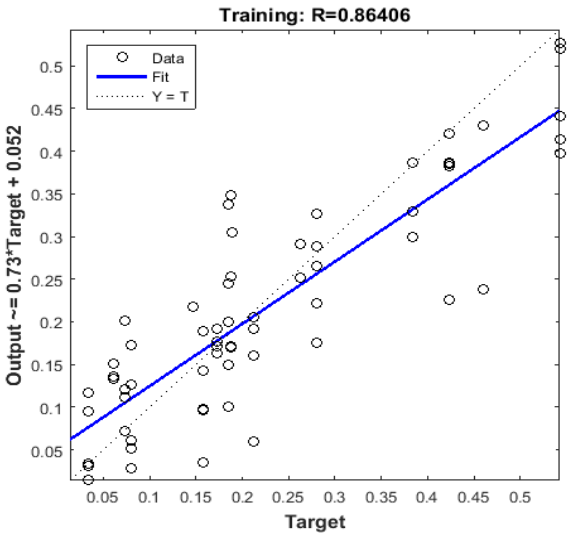



Fig. 6. Performance of the used ANN Model.

One hidden layer of 8 neurons is used in this network. The optimal number of neurons was identified after performing many trials based on the minimum value of mean squared error (MSE) as shown in Table 1.

Fig. 6 shows the performance of the used model, it presents the correlation between data points and curve fitting plots for training. It can be clearly noticed from this figure that the regression is approximately 0.864 which can be considered high enough to reflect a good relationship between the targets and the outputs.

## IV. Results and Discussion

Four measures in this analysis were carried out; the true positives which are the samples that are correctly classified as cracks, true negatives are samples that are correctly classified as non-cracks, where, false positives are the samples that are not cracks but wrongly classified as cracks by the network and false negatives are crack samples but wrongly classified as non-cracks by the network [11]. Table 2 summarizes these four measures for a subset of 80 images.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True positive} + \text{False positive}}$$

TABLE II. TEST PERFORMANCE PARAMETERS

| Test Results | Present | Absent | Total |
|---|---|---|---|
| Positive | True Positive (54) | False Positive (8) | 62 |
| Negative | False Negative (6) | True Negative (12) | 18 |
| Total | 60 | 20 | 80 |

In order to determine how well this cracks detection method performs; sensitivity, specificity, precision and accuracy were computed for the same subset of images. The sensitivity (recall) of a test quantifies its ability to correctly identify subjects with the condition [12], in other words, it is the percent of crack samples that are identified by the network out of the total number of cracks in the dataset [11], given by:

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

Specificity is the ratio of true negatives divided by summation of true negative and false positives. It represents the probability of a test without giving false-positive results, given by [13]:

$$\text{Specificity} = \frac{\text{True negative}}{\text{True negative} + \text{False positive}}$$

A precision is the percent of predicted cracks that were actually cracks, given by:

Accuracy is another factor for performance measurement, it represents the proportion of true results (both true positive and true negative) in the selected population.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True negative}}{\text{Total}}$$

A summary of the mentioned factors for the tested subset is shown in table 3.

TABLE III. SENSITIVITY, SPECIFICITY, ACCURACY AND PRECISION VALUES OF THE PERFORMED MODELING

| PARAMETER | RESULT % |
|---|---|
| Sensitivity | 90 |
| Specificity | 60 |
| Accuracy | 82.5 |
| Precision | 87 |

## V. Conclusions

Extraction of cracks in roads pavement from airborne images was automatically detected using artificial neural network. The used ANN model showed a suitability for pattern recognition for roads cracks based on the high values of performance parameters. The proposed method opens the door for numerous applications of ANN on distresses classifications and measurements. It has been shown that the performance of the used feedforward neural network increased up to a certain number of neurons in the hidden layer. In contrast, some samples were predicted wrongly as false negative and false positive which represents fake cracks in the realty, such problems might be related to image noise and image interference like: shadows, lane line and trees branches. It is recommended that future research should cover the issue of image acquiring and accuracy assessment of the system and its effect on the extracted cracks.

It is expected that the proposed technology, setup configuration and procedures of automatically extracting

pavements cracks will open the door for other applications in the domain of transportation engineering, road construction, and other domains. Moreover, we are looking forward to having hand-held cell phones applications of this technology on the smart phones themselves.

## VI. Conflict of Interest Disclosure

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

## References

[1] H. Oliveira, "Crack detection and characterization in flexible road pavements using digital image processing" Thesis in electrical and computer engineering, Universidade Técnica De Lisboa, Instituto Superior Técnico. 2013

[2] T. Al-Suleiman, M. Kheder, and H. Al-Masaeid, "Development of Pavement Performance Models for Rural Roads," Journal of Road and Transport Research, Australian Road Research Board, Vol. 1, No. 4. 1992.

[3] D. Meignen, M. Bernadet, and H. Briand, "One Application of Neural Networks for Detection of Defects Using Video Data Bases: Identification of Road Distresses," IEEE 8th International Workshop on Database and Expert Systems Applications - DEXA, 1-2 September 1997, Toulose, France, pp. 459-464.

[4] Cheng, H. and Miyojim, M. "Automatic Pavement Distress Detection System," Information Sciences, vol. 108, no. 1-4, July, pp. 219-240. 1998

[5] P. Subirats, O. Fabre, J. Dumoulin, V. Legeay, and D. Barda, "A Combined Wavelet-based Image Processing Method for Emenrgent Crack Detection on Pavement Surface Images," 12th European Signal Processing Conference - EUSIPCO, 6-10 September 2004, Vienna, Austria, pp. 257-260.

[6] Wang, K. "Design and Implementation of Automated System for Pavement Surface Distress Survey," Journal of Infrastructure System, 6 (1), 24-32. 2000

[7] J. Zhou, P. Huang, and F. Chiang, "Wavelet-Based Pavement Distress classification," Journal of the Transportation Research Board, no. 1940, pp. 89-98, 2005.

[8] R. Medina, R., Llamas, J. Bermejo, E. Zalama, and M. Segarra, "Crack Detection in Concrete Tunnels Using a Gabor Filter Invariant to Rotation," Sensors, 17 (1), pp1670,2017.

[9] H. Oliveira, P.L. Correia, "Automatic road crack segmentation using entropy and image dynamic thresholding," In Proceedings of the 17th European Signal Processing Conference (EUSIPCO'2009), Glasgow, UK, pp. 622–626.

[10] L. Zhang, F. Yang, Y.D. Zhang, Y.J. Zhu, "Road crack detection using deep convolutional neural network," In Proceedings of the IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3708–3712.

[11] R. Fuentes, L. Pauly, and D. Hogg, "Deeper Networks for Pavement Crack Detection," 34th International Symposium on Automation and Robotics in Construction (ISARC 2017).

[12] H. Wong, and G. Lim, "Measures of Diagnostic Accuracy: Sensitivity, Specificity, PPV and NPV," Proceedings of Singapore Healthcare, 2011, 20 (4), 316-318.

[13] H. Brenner, O. Gefeller, "Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence," Statistics in Medicine, 16: 981–991. 1997.

[14] T. Fukuhara, K. Terada, M. Nagao, A. Kasahara, and S. Ichihashi, "Automatic Pavement-Distress-Survey System," J. Transportation Engineering, ASCE 116 (3), 1991, pp 280-286,.

[15] L. Sjogren, (2002) "State of the art: Automated Crack Measurement of Road Surface," research project report ,Swedish National Road and Transportation Research Institute (VTI). 2002.

[16] B. Xu, and Y. Huang, "Development of an automatic pavement surface distress inspection system," research project report, Center for Transportation Research at The University of Texas at Austin. 2005.

[17] W. Zhang, Z. Zhang, D. Qi, and Y. Liu, "Automatic crack detection and classification method for subway tunnel safety monitoring," *Sensors*, vol. 14, no. 10, pp. 19307–19328. 2014.

# Study of Routing Protocols on CBR and VBR Applications in VANET Scenario

Pooja Sharma
School of Computer Science & Engg,
Shri Mata Vaishno Devi University Katra (J&K)- 182320.India

***Abstract*—Vehicular Adhoc Networks (VANETs) are special type of Mobile Adhoc networks (MANETs) where node movement is in pre ordered fashion but with high velocity in comparison to MANETs where nodes move in random manner. Due to high mobility of nodes, reliable data streaming in vehicular networks is a complex and challenging task. Moreover, transmission of data is difficult because of varying requirements of different applications in terms of various resources like time, energy and bandwidth. This paper gives an overview of performance evaluation of four types of routing protocols on CBR and VBR applications. This paper emphasizes on packet delivery ratio, packet loss and packet loss ratio for CBR and VBR applications in different scenarios like varying node density, varying speed of nodes, pause times and packet size. The effectiveness of various routing protocols shows variation in different conditions. The performance evaluation of different applications in terms of Quality of service (QoS) parameters like packet delivery ratio, packet loss and packet loss ratio has been studied by varying different conditions of CBR traffic and VBR traffic which has gives an insight to improve packet delivery ratio which in turn can be utilized to improve performance of an application in future.**

***Keywords—VANETS; routing protocols; qualnet; traffic types introduction***

## I. INTRODUCTION

There are various applications like constant bit rate traffic (CBR), variable bit rate traffic (VBR), File Transfer protocol(FTP) , Voice Over Internet Protocol(VOIP) . All these applications can be evaluated on various parameters in terms of throughput, jitter, average end to end delay. Internet traffic is broadly classified into two categories i.e real time applications or time critical traffic and the non real time applications or elastic applications. Real time traffic is further divided into two forms like real time streaming applications and real time control applications eg machine control, games. Examples of non real time applications are web browsing, email, FTP and telnet. Few of the real time and asymmetric applications are audio broadcasting, video broadcasting, interactive audio on demand , interactive video demand. Real time and symmetric applications are those which are conversant in nature such as teleconferencing (video and audio conferencing) and Voice over Internet protocol (VOIP).

There are various factors which affect the performance of different protocols along with applications like node density, velocity of nodes, mobility patterns, traffic types which will be implemented in this study. Effect of single hop routing and multi hop routing protocols can be easily observed in this research work. This research study basically analyzes the Quality of Service (QoS) metrics for different applications as mentioned on different protocols like AODV, DYMO, DSR and LAR so as to focus on some parameters where improvement can be done in order to optimize the performance of applications . This paper is divided into four different sections. First one is introduction followed by literature review and simulation and results and conclusion..

## II. LITERATURE REVIEW

The existing literature has been reviewed regarding different requirements of various applications gathered from the various sources available. Some of the data has been taken from [1] and [2] regarding bandwidth, delay and jitter requirements for video streaming, VOIP, Interactive video etc. In continuous media, especially audio and video data has spatial relationships that must be taken care of. The requirements of time critical applications are commonly expressed as a set of values representing bandwidth, delay, and jitter and loss rate. Continuous streaming applications can cope with QoS which is significantly lower than real time streaming applications. Various voice encoding techniques such as G.711, G.726, G.727, G.728 are available, similarly several encoding or compression technique for CD quality sound have been developed such as MPEG Layer-1, MPEG Layer-2, MPEG Layer-3. The important building blocks of Prediction Based Routing ( PBR) protocol are obtaining location and velocity information of vehicles on the route to the gateway. Its basic operation is to create routes in the same way as reactive protocols. When a node wants to communicate, it sends route request (RREQ) packet with a TTL (time to live) value that specifies the number of hops to search for a gateway. In PBR, the lifetime of link between two nodes i and j is predicted as

$$\text{Lifetime}_{\text{link}} = \frac{R - |dij|}{|v_i - v_j|}$$

where R is communication range of WLAN technology, $|d_{ij}|$ absolute distance between nodes i and j, $v_i$ and $v_j$ are velocities.

Since a route comprises of one or more links, the route lifetime is the minimum of all its link lifetimes as done by Vinod Namboodri in [3]. With low vehicle density in the forward direction using routes through oncoming vehicles has an effect similar to doubling the vehicle density in the forward direction in terms of connectivity.

In one of the research study by shaily mittal [4] in "performance comparison of AODV, DSR and ZRP routing

protocols in MANETs" analysis has been done for different routing protocols. One more research study has been done by Mrs. Vaishali D and Dr Ketan [5] in their study title "Simulation based performance evaluation of routing protocols in Vehicular Adhoc Networks". The above mentioned studies emphasize on effectiveness of different routing protocols in different situations.

Different protocols have different impact on throughput, jitter, delay and average end to end delay. Moreover, the type of application to be simulated matters a lot. There is a change observed in metrics for different applications like CBR applications, VBR applications, Telnet, FTP applications. Different applications in scenarios like varying node density, node speed, pause time , packet size are analyzed and results are observed keeping in consideration of vehicular adhoc networks.

### A. Adhoc on Demand Vector Routing Protocol (AODV)

It is purely on –demand routing protocol. In order to find the path between source and destination a RREQ message is broadcasted to all the neighbor nodes who again continue to send messages to their neighbors until the destination is reached. Every node maintains two variables sequence number and broadcast ID in order to have loop free and maintain most recent information. Path discovery process and route maintenance is done in AODV as mentioned in [6] and [7].

### B. Dynamic Source Routing (DSR)

Is also an on Demand routing protocol in which sequence of nodes through which packets will travel is calculated and information is stored in packet header. The source nodes sent request packets to all the neighbors in the network containing the address of the destination node and a reply is sent back to the source nodes with the list of network nodes through which it should move forward in the process. Route maintenance can be done either by hop by hop acknowledgement at the data link layer and end to end acknowledgements. Hop by Hop method allows early detection and retransmission of lost or corrupt packets in the data link layer. When the wireless transmission between two nodes do not work well then end to end replies on the application layer or transport layer may be used to indicate the status of the route from one host to the other. All intermediary nodes along the path simply forwards the packet to the next node as specified in the packet header[8].

### C. Location aided Routing (LAR)

Protocol given by Young Br Ko et al [9] suggests an approach to utilize location information ( for instance , obtained by global positioning system) to improve performance of routing protocol for adhoc networks. By using location information, proposed protocol limits the search for a new route to a smaller request zone of the adhoc network. This results in significant reduction in number of messages. Location information is used for route discovery.

### D. Dynamic MANET on Demand Routing Protocol (DYMO)

Enables dynamic , reactive , multi hop routing between participating nodes wishing to communicate. The basic operations of the protocol are route discovery and management. Using adhoc on demand distance vector (AODV), DYMO borrows "Path Accumulation" from Dynamic source routing and removes unnecessary route reply(RREP), precursor lists and Hello messages(Route exploration messages) thus simplifying AODV[10]. It retains sequence numbers and Route error messages from AODV [11]

The above mentioned protocols have been implemented on MANETs very well but their performance on VANETs is still not done as VANET nodes have high mobility which is covered in this paper. Moreover , these set of four protocols have been implemented on CBR and VBR applications to see that how these four protocols will satisfy most QoS parameters and show variations in packet delivery ratio , packet loss and packet loss ratio.

## III. Simulation and Results

Qualnet version 5.0.1 is a discrete event simulator which is being used here. In this paper Qualnet is configured by having area 1500x1500 having different scenarios. Four protocols are taken into consideration i.e AODV,DSR,DYMO and LAR with varying network sizes , mobility of nodes by varying speeds of the nodes, pause time and vary size of packet ,simulation time. Different scenarios are made keeping some of the factors constant and change one at a time to see the impact of that particular factor on QoS metrics under observation. Very important, different traffic types will have different set of results keeping the same protocols when simulated in one of the scenario. In this study different formulae have been used for calculation of packet delivery ratio, packet loss and packet loss ratio.

Perform

ance Metrics

Packet Delivery Ratio is important metric to measure the performance of routing protocol .Packet Delivery Ratio is defined as the ratio of the total packets received by all destination nodes and the total packets sent by all source nodes[5].

$$\text{Packet Delivery Ratio} = \frac{\sum (\textit{Total packets received by all destination nodes}}{\sum (\textit{Total packets send by all source nodes}}$$

Packet Loss is the ratio of the number of packets that never reached the destination to the number of packets originated by the source. [5]

$$\text{Mathematically, PL} = \frac{nsentpackets - nreceivedpackets}{nsentpackets}$$

$$\text{Packet Loss Ratio} = \frac{nsentpackets - nreceivedpackets}{nsentpackets} * 100$$

TABLE I. VARIOUS PARAMETERS ARE USED WHILE VARYING NUMBER OF CONNECTIONS

| Parameter | Value |
|---|---|
| Protocols | AODV,DSR,DYMO, LAR |
| Number of Nodes | 30,50,80,120 |
| Pause Time | 30,60,100s |
| Simulation time | 30s |
| Traffic Type | CBR, VBR |
| Transmission Range | 250 m |
| Mobility Model | Random Way point Model |
| Simulation Area | 1500x1500 |
| Node Speed | 0,10,20,60,90 km/h |
| Interface Type | Queue |
| MAC Protocol | 802.11 Ext |
| Packet Size | 512 ,1024 |
| Radio Propagation Model | Two Ray Ground |

Table 1 is giving details of all the parameters used while varying number of connections.In this paper four scenarios are simulated on the basis of which certain results are obtained and some outcomes and inferences are drawn. The effect of all the four protocols namely AODV,DSR,DYMO and LAR are studied on both type of traffic like constant bit rate traffic and variable bit rate traffic in terms of packet delivery ratio, packet loss and packet loss %.

**Scenario I CBR Traffic and VBR traffic with varying node density in terms of PDR, Packet Loss ( PL), Packet Loss%**

TABLE II. VARYING NODE DENSITY

| Node Density | 30 | 30 | 30 | 50 | 50 | 50 | 80 | 80 | 80 | 120 | 120 | 120 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PDR | PL | PL% | PDR | PL | PL% | PDR | PL | PL% | PDR | PL | PL% |
| AODV | 0.82 | 0.19 | 19.17 | 0.83 | 0.18 | 17.50 | 0.81 | 0.19 | 19.17 | 0.875 | 0.125 | 12.500 |
| DSR | 0.64 | 0.36 | 35.83 | 0.65 | 0.35 | 35.00 | 0.66 | 0.34 | 34.17 | 0.675 | 0.325 | 32.500 |
| DYMO | 0.83 | 0.17 | 16.67 | 0.78 | 0.22 | 21.67 | 0.78 | 0.23 | 22.50 | 0.858 | 0.142 | 14.167 |
| LAR | 0.88 | 0.12 | 11.67 | 0.88 | 0.13 | 12.50 | 0.89 | 0.11 | 10.83 | 0.858 | 0.142 | 14.167 |

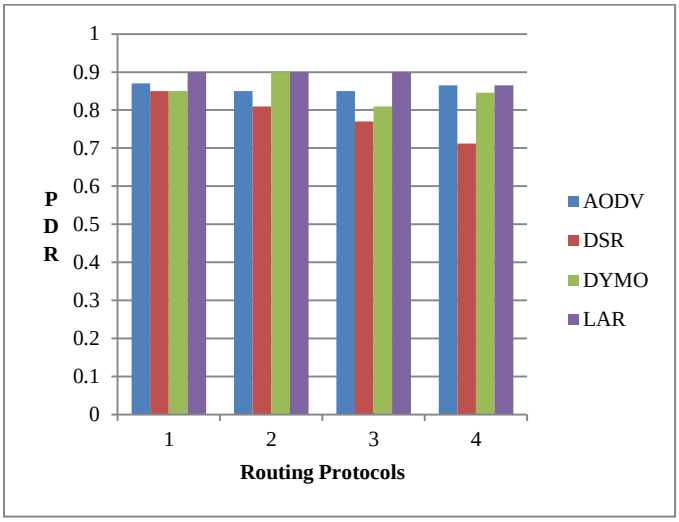



Fig. 1. Packet Delivery Ratio for CBR Traffic with varying Node Density

TABLE III. VBR TRAFFIC WITH VARYING NODE DENSITY

| Node Density | 30 | 30 | 30 | 50 | 50 | 50 | 80 | 80 | 80 | 120 | 120 | 120 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PDR | PL | PL% | PDR | PL | PL% | PDR | PL | PL% | PDR | PL | PL% |
| AODV | 0.87 | 0.13 | 13.46 | 0.85 | 0.15 | 15.38 | 0.85 | 0.15 | 15.38 | 0.865 | 0.135 | 13.462 |
| DSR | 0.85 | 0.15 | 15.38 | 0.81 | 0.19 | 19.23 | 0.77 | 0.23 | 23.08 | 0.712 | 0.288 | 28.846 |
| DYMO | 0.85 | 0.15 | 15.38 | 0.90 | 0.10 | 9.62 | 0.81 | 0.19 | 19.23 | 0.846 | 0.154 | 15.385 |
| LAR | 0.90 | 0.10 | 9.62 | 0.90 | 0.10 | 9.62 | 0.90 | 0.10 | 9.62 | 0.865 | 0.135 | 13.462 |

It is observed here in table 2 and table 3 that Packet Delivery Ratio has been decreasing with increasing node density truly for the VBR traffic for AODV, DSR , DYMO whereas not much effect has been seen for LAR with increasing node density. For DYMO, there is no fixed pattern observed for packet delivery ratio for VBR traffic. We can observe in Figure 1 and Figure 2,CBR traffic, AODV and DSR there has been reduction in packet delivery ratio but when nodes are increased to 120, a gain in PDR has been observed. Similar is the case for DYMO but LAR keeps a steady pace for packet delivery ratio in turn keeping packet loss ratio more or less same.

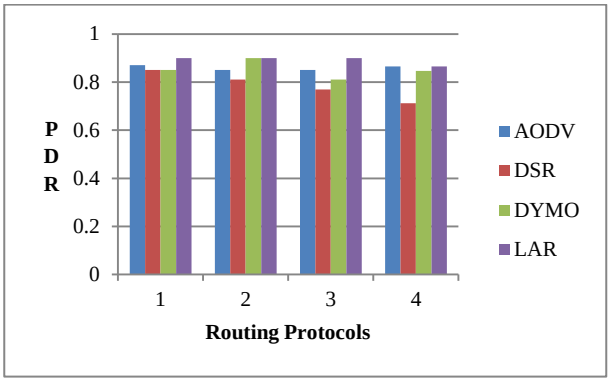



Fig. 2. Packet Delivery Ratio for VBR Traffic with varying node density

**Scenario II Node Density Fixed, speed of the nodes varying CBR (speed in mps)**

TABLE IV. EFFECT OF MOBILITY OF NODES ON CBR TRAFFIC IN TERMS OF PDR, PL, PL RATIO

| Node Speed | 10-20 | 10-20 | 10-20 | 20-60 | 20-60 | 20-60 | 20-90 | 20-90 | 20-90 |
|---|---|---|---|---|---|---|---|---|---|
| | PDR | PL | PL% | PDR | PL | PL% | PDR | PL | PL% |
| AODV | 0.92 | 0.08 | 8.33 | 0.81 | 0.19 | 19.17 | 0.82 | 0.18 | 18.33 |
| DSR | 0.94 | 0.06 | 5.83 | 0.64 | 0.36 | 35.83 | 0.69 | 0.31 | 30.83 |
| DYMO | 0.92 | 0.08 | 8.33 | 0.83 | 0.17 | 16.67 | 0.86 | 0.14 | 14.17 |
| LAR | 0.93 | 0.07 | 6.67 | 0.88 | 0.12 | 11.67 | 0.88 | 0.13 | 12.50 |

TABLE V. EFFECT OF MOBILITY OF NODES ON VBR TRAFFIC IN TERMS OF PDR, PL, PL RATIO (SPEED IN MPS)

| Node Speed | 10-20 | 10-20 | 10-20 | 20-60 | 20-60 | 20-60 | 20-90 | 20-90 | 20-90 |
|---|---|---|---|---|---|---|---|---|---|
| | PDR | PL | PL% | PDR | PL | PL% | PDR | PL | PL% |
| AODV | 0.92 | 0.08 | 7.69 | 0.87 | 0.13 | 13.46 | 0.81 | 0.19 | 19.23 |
| DSR | 0.96 | 0.04 | 3.85 | 0.85 | 0.15 | 15.38 | 0.77 | 0.23 | 23.08 |
| DYMO | 0.90 | 0.10 | 9.62 | 0.85 | 0.15 | 15.38 | 0.85 | 0.15 | 15.38 |
| LAR | 0.92 | 0.08 | 7.69 | 0.90 | 0.10 | 9.62 | 0.87 | 0.13 | 13.46 |

As observed in Table 4 and 5, Packet delivery ratio has been perfectly decreasing with the increasing velocity of nodes for VBR traffic for all the four protocols but CBR traffic is showing irregularities in its behavior in terms of packet delivery ratio. DSR is hugely reducing packet delivery ratio when increasing speed of nodes for CBR traffic whereas AODV, LAR and DYMO are more or less behaving in same manner. In figure 3 and 4 ,it is observed that LAR is giving maximum packet delivery ratio even with varying and highest speed of minimum 20 and maximum 90 mps.

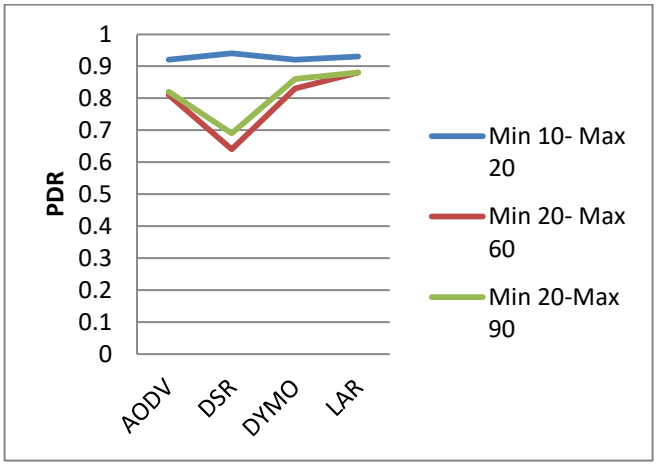



Fig. 3. Packet delivery ratio for CBR traffic with varying speed of nodes

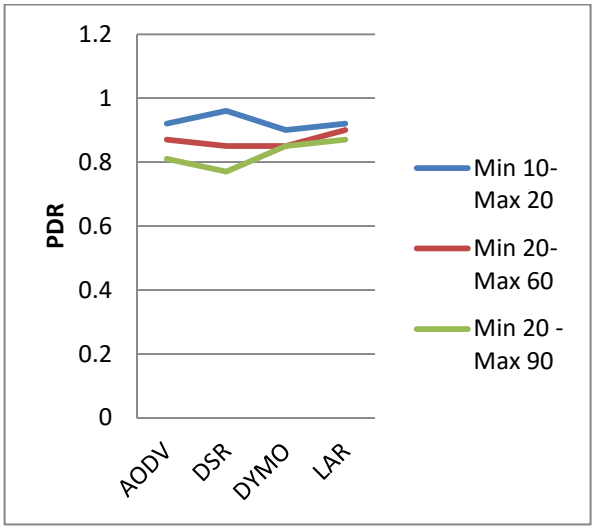



Fig. 4. Packet Delivery Ratio of VBR Traffic with varying speed of nodes

**Scenario III Varying pause time on CBR and VBR Traffic**

TABLE VI. CBR TRAFFIC WITH DIFFERENT PAUSE TIMES

| Pause Time | 30s | 30s | 30s | 60s | 60s | 60s | 100s | 100s | 100s |
|---|---|---|---|---|---|---|---|---|---|
| | PDR | PL | PL% | PDR | PL | PL% | PDR | PL | PL% |
| AODV | 0.92 | 0.08 | 8.33 | 0.90 | 0.10 | 10.00 | 0.88 | 0.12 | 11.67 |
| DSR | 0.94 | 0.06 | 5.83 | 0.95 | 0.05 | 5.00 | 0.93 | 0.08 | 7.50 |
| DYMO | 0.92 | 0.08 | 8.33 | 0.93 | 0.08 | 7.50 | 0.92 | 0.08 | 8.33 |
| LAR | 0.93 | 0.07 | 6.67 | 0.83 | 0.17 | 16.67 | 0.96 | 0.04 | 4.17 |

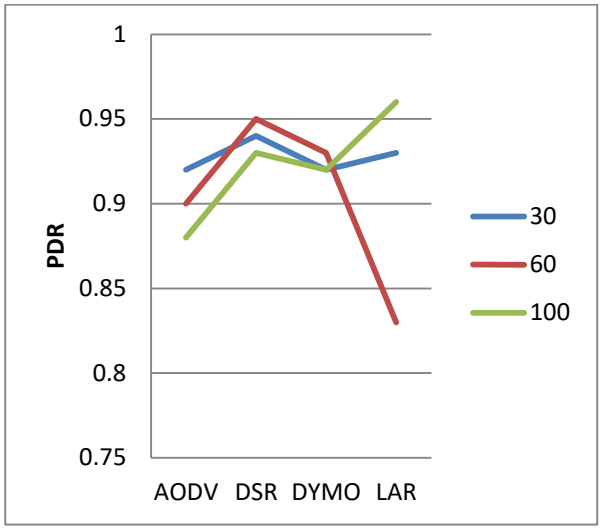



Fig. 5. Effect of changing pause times on CBR in terms of packet delivery ratio

TABLE VII. VBR TRAFFIC WITH CHANGING PAUSE TIMES IN TERMS OF PDR, PL, PACKET LOSS%

| Pause Time | 30s | 30s | 30s | 60s | 60s | 60s | 100 s | 100 s | 100s |
|---|---|---|---|---|---|---|---|---|---|
| | PDR | PL | PL % | PDR | PL | PL % | PDR | PL | PL% |
| AODV | 0.92 | 0.08 | 7.69 | 0.98 | 0.02 | 1.92 | 0.83 | 0.17 | 17.31 |
| DSR | 0.96 | 0.04 | 3.85 | 0.96 | 0.04 | 3.85 | 0.94 | 0.06 | 5.77 |
| DYMO | 0.90 | 0.10 | 9.62 | 0.94 | 0.06 | 5.77 | 0.92 | 0.08 | 7.69 |
| LAR | 0.92 | 0.08 | 7.69 | 0.94 | 0.06 | 5.77 | 0.96 | 0.04 | 3.85 |

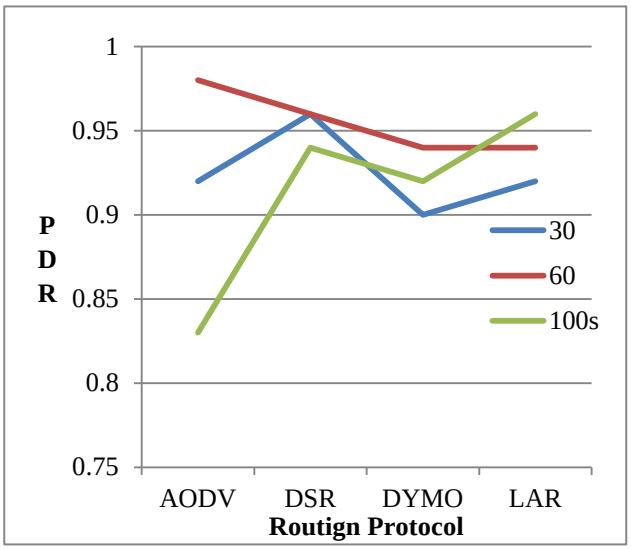



Fig. 6. VBR Traffic with changing pause times in terms of packet delivery Ratio

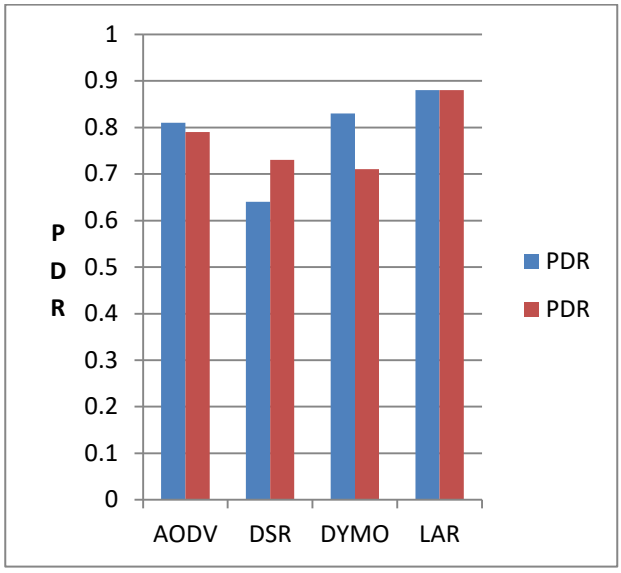
When pause times are varied, as seen in Figure 5 and 6 along with table 6 and 7,significant decrease has been seen in packet delivery ratio for CBR traffic especially when pause time is made 60s whereas for pause time 100s, there has been reduction in packet delivery ratio for AODV and DSR. DYMO doesn't change packet delivery ratio but LAR increases PDR. For VBR traffic, there has been an increase in PDR for all the protocols except DSR which is not making any change in PDR when pause time is 60s but when talking of 100s, there has been an decrease in PDR for AODV and DSR but DYMO and LAR increases PDR and so reduces packet loss ratio.

**Scenario IV Vary packet size of the traffic for CBR and VBR**

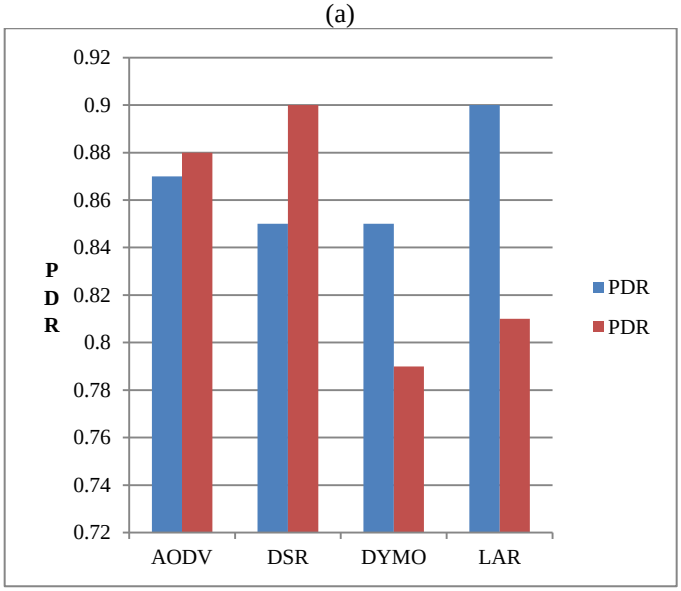
TABLE VIII. EFFECT OF PROTOCOLS ON CBR AND VBR IN TERMS OF PDR, PL, PL%

| Packet Size | 512 | 512 | 512 | 1024 | 1024 | 1024 | 512 | 512 | 512 | 1024 | 1024 | 1024 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CBR | | | | | | VBR | | | | | |
| | PDR | PL | PL % | PDR | PL | PL % | PDR | PL | PL % | PDR | PL | PL % |
| AODV | 0.81 | 0.19 | 19.17 | 0.79 | 0.21 | 20.83 | 0.87 | 0.13 | 13.46 | 0.88 | 0.12 | 11.54 |
| DSR | 0.64 | 0.36 | 35.83 | 0.73 | 0.28 | 27.50 | 0.85 | 0.15 | 15.38 | 0.90 | 0.10 | 9.62 |
| DYMO | 0.83 | 0.17 | 16.67 | 0.71 | 0.29 | 29.17 | 0.85 | 0.15 | 15.38 | 0.79 | 0.21 | 21.15 |
| LAR | 0.88 | 0.12 | 11.67 | 0.88 | 0.13 | 12.50 | 0.90 | 0.10 | 9.62 | 0.81 | 0.19 | 19.23 |

Figure 7 is giving an information as Packet size is varied in fourth scenario by making 1024 bytes for CBR and VBR traffic instead of 524 bytes. The impact of four protocols has been studied like AODV, DSR, DYMO and LAR. In table 8,There has been an articulate decrease seen in packet delivery ratio for all protocols except DSR where an increase is seen for CBR traffic. As far as VBR traffic is concerned, there has been an increase in packet delivery ratio for AODV and DSR, but for DYMO and LAR there has been a decrease in packet delivery ratio




(a)




(b)

Fig. 7. CBR Traffic showing PDR with varying packet size for four protocols

## IV. Conclusion

The above qualitative results show that AODV has been found to be better in most of the scenarios followed by LAR. AODV and LAR have shown showing little variations in PDR with reference to node density whereas DSR has shown large variation in PDR. With reference to the change in mobility of nodes, all the four protocols have shown a decline in packet delivery ratio, packet loss and packet loss ratio for both CBR and VBR traffic. For VBR traffic, AODV has hardly shown any change for PDR, PL and PL ratio whereas DSR has shown an increase in PDR and DYMO and LAR are showing a reduction in PDR. In future, we propose to consider metrics like delay and jitter on applications like FTP, Telnet and VOIP. The results of the study has indicated that by combining two or more protocols the packet loss ratio can further be reduced with improvement in PDR by mitigating error propagation in data transmission

## References

[1] Residential Broadband Internet Services and Application Requirements" T K Wok IEEE Communications Magzine June 1997.

[2] www.ciscopress.com/articles.

[3] Vinod Namboodri , L Gao IEEE Transactions 2007 " Prediction based routing Protocol for vehicular adhoc networks

[4] Shailey Mittal "Performance comparison of AODV, DSR and ZRP routing protocols in MANETs"

[5] Mrs Vaishali and Dr ketan Katocha "Simulation based performance evaluation of routing protocols in Vehicular Adhoc Networks" International Journal of scientific and research publications Volume 3 Issue 10 Oct 2013

[6] Asma Tuteja, Rajneesh Gujral "Comparative performance analysis of DSDV,AODV and DSR in MANET using NS2

[7] N Suryati, A Abdullah and A Faisal "Performance evaluation of AODV, DSDV and DSR routing protocol in Grid environment" International Journal of Computer Science and Network Security Vol 9 No 7 July 2009

[8] Amitava Mukherjee, S Bhandyopadhyay, D Saha "Location Management and Routing in mobile Wireless Netowrks

[9] Y Be Ko, Nitin H Vaidya "Location aided routing protocol in Mobile Adhoc Networks" Wireless Networks 6 (2000) 307-321

[10] Ian D Chakeres and E M Royer AODV routing protocol implementation design. In ICDCSW '04 Proceedings of 24th international conference on distributed computing system workshops- W7 EC(ICDCSW'04) pages 698-703 Washington DC, US

[11] Ian D Chekeres and C E Perkins Dynamic Manet on Demand (DYMO) routing protocol Internet-Draft Version 06, IETF, Oct 2006

# A Survey on Smartphone-Based Accident Reporting and Guidance Systems

Alexandra Fanca[1], Adela Puscasiu[2], Honoriu Valean[3], Silviu Folea[4]
Department of Automation
Technical University of Cluj-Napoca
Cluj-Napoca, Romania

***Abstract*—Every day, around the world, a large percentage of people die from road accidents and falls. One of the reasons for a person's death during accidents is the unavailability of first aid, due to the delay in informing about the accident. Thus, in the case of incidents involving vehicles or falls, response time is crucial for the timely provision of emergency medical services. An effective approach intended to reduce the number of traffic-related deaths is: the use of a system for detecting and reporting the occurred accidents, as well as reducing the time between the occurrences of an accident and sending the first emergency respondents to the scene of the accident. This paper presents a recent study on mobile terminal solutions (smartphones) for detecting and preventing accidents (road, falls, bicycles) and systematic comparisons of existing solutions.**

***Keywords*—Smartphone; accident; detection algorithm; reporting accident; mobile application; sensors**

## I. Introduction

The mobile phone has become part of the daily routine of the modern man, a non-detachable assistant, a gateway to news, social media, entertainment and all forms of communication. Clearly, smartphones have penetrated every aspect of our life. Smartphone is one of the most fascinating object that compared to PC or laptop is much smaller, but with a similar processing power, easy to customize and sometimes they work as an extension of the individual [1].

Introducing operating systems on smartphones has led to the emergence of so-called applications, which are installed programs such as Google Chrome, Games, and more. Thus, each individual can install the applications they need to navigate the web, edit and share documents, access e-mail, social media account, order food and drink, find information about weather, health and fitness, e-banking, or shop online, from anywhere.

Due to the competition between smartphone manufacturers, over the years, a wide range of sensors have been built to record large amounts of data, powerful computing units and powerful communication modules. Thus, the smartphone has become a mobile platform for collecting and retrieving data. Today's smartphones include: high-resolution camera, microphone, compass, accelerometer, 3-axis gyroscope, GPS receiver, as well as high-speed Wi-Fi and 3G communication capabilities. The new features added to the smartphone allow application developers to determine geographic location and user moves.

Big processing power, motion detection, wireless communications, open source operating systems, popularity, relatively low cost and, in particular, the ability to obtain a rich variety of data, make smartphones a good tool for mobile applications in the medical and safety fields.

By introducing the concept of "Internet of Things" into modern electronics, embedded sensors in smartphone have been increasingly used by mobile application developers for various types of systems such as monitoring, detection, guidance and / or reporting of human behavior in various situations such as road accidents, falls, driving inadequate, etc. [2].

Due to daily use of mobile phones, this device is perfect for detecting the fall or other types of accidents with trauma and helping the injured user to receive first aid in the shortest time.

With the increase in sales of mobile devices, the smartphone application market has grown considerably. Mobile devices, especially smartphones, have been used to monitor, detect, guide and / or report accidents in many applications, both academically and commercially. The reason why smartphones are becoming more and more used in such applications is that they can track the vehicles, determine their speed and offer traffic at a lower cost compared to the traditional approach using loop detectors. In addition, smartphones are immediately available to most users, which means there is no need for additional specialized detection hardware.

Another very important aspect discussed by researchers and application developers is to minimize false positives in smartphone detection and reporting systems. Every year a large number of accidental calls are reported by emergency services, so reducing the false rate of detection and reporting of accidents is important [3].

### *A. The Benefits of using Smartphone in Accident Detection*

In recent years, researchers have presented in various papers and articles the benefits of using smartphones in accident and fall detection and notification systems. The main benefits of smartphone are:

- They are updated more frequently in software than vehicles and their cost is much lower;
- Using integrated sensors (such as accelerometer, GPS, gyroscope, etc.) to get a rich variety of data;

- Smartphones are always with their owners, indicating an accident, even if the vehicle they are in is not involved in an accident (witnessing an accident);

### B. Major Objectives of SARGS

The most important objective of any SARGS is the automatic detection of any accidents by using sensors embedded in the smartphone. Another important objective is to report the occurrence of an accident as soon as possible by using the real-time system, the correct records, etc.

Other objectives could be to send automatically warnings in case of accident and the possibility to cancel reporting in case of false accidents.

The rest of the paper is organized as follows: Section II describes the integrated systems in vehicles, section III presents a series of smartphone based systems and the main advantages and disadvantages of the systems, and the proposed idea is presented in Section IV, and in the section V are presented the conclusions.

## II. Accident Detection Integrated Systems for Vehicles

Various types of accident prevention and detection systems have been studied and implemented over the years. Thus, various integrated automatic detection and notification systems have been developed at the time of vehicle traffic accident.

One of these systems is OnStar [4] being available only to Opel or Vauxhall car owners. This system has been running since the 1990's in the United States and since 2016 it has been introduced in Europe in thirteen countries. This system offers free trial for 3 months and is available in 8 languages. The system provides assistance for all emergencies as well as other assistance (location, car data, vehicle diagnosis, etc.). A disadvantage of this system is the impossibility to cancel the call to the emergency service if the SOS button has been accidentally pressed.

Even if the integrated systems in the vehicle are working well, they are expensive and not available for monitoring older vehicles.

## III. Mobile Applications for Detecting and Reporting Accidents

In the attempt to eliminate the false positive of an accident, various methods are integrated, such as: combining the data provided by the smartphone sensors (accelerometer, microphone, GPS, gyroscope, camera, etc.) with vehicle data collection (airbags, GPS position, etc.) by wireless communication (Wi-Fi, Bluetooth, OBD-II platform, and VANET) or the additional use of a Kalman filter.

In this section, we offer some existing applications which can be used in the event of an accident (road accidents, fall, etc.).

The paper [5] describes how smartphones can automatically detect a road accident, relying on data processing capability and built-in sensors without direct interaction with vehicle sensors. For this purpose, a mobile application called WreckWatch has been proposed that uses the accelerometer and microphone for accident detection and defines a formula based on which an accident can be assumed. WreckWatch records different types of data such as: accelerometer trajectory, acceleration and acceleration forces, images, videos, acoustic data and GPS coordinates.

Chalermpol Saiprasert and Wasan Pattara-Atikom [6] are proposing a system for reporting abnormal driving using smartphone as a terminal, due to its capabilities to collect, store and send real-time data. By collecting data from the built-in GPS of the smartphone, the user can get a driving report by analyzing the vehicle's speed over a route's distance. An algorithm capable to detect speed profile anomalies has been proposed.

In the paper [7], Hamid M. Ali and Zainab S. Alwan propose a road accident detection and notification system called CADANS. The accident detection algorithm used by Hamid M. Ali et al is based on data collection continuously from various sensors integrated into the smartphone such as: accelerometer (used to record acceleration forces (force G)), the GPS receiver (used to determine the vehicle speed, speed used to increase the probability of detecting an accident based on accelerometer data), the microphone (which determines the high acoustic events such as the airbag triggering sound ) as well as the camera / video camera (used to send videos to the emergency help center for any further analysis). The notification sent to the "help center" contains, in addition to the location of the accident, other data such as vehicle speed, acceleration force G, airbag status, and the occurrence of the accidents.

F. J. Bruwer and M.J. (Thinus) Booysen propose in the paper [8] a way to enable reckless driving detection, using sensors in current smartphones (accelerometer, gyroscope and magnetometer), relying primarily on vehicle acceleration. The main objective pursued in this paper is to eliminate the effects of the gravity vector from the acceleration measured by smartphone in a vehicle. In addition to the smartphone's built-in sensors, a Kalman filter was used to combine accelerometer, gyroscope, and magnetometer and vehicle dynamics data in a way that maximizes the probability of accurately estimating the orientation of the sensors to the Earth's axes. After estimating the gravitational acceleration, it is removed from the acceleration measured to obtain vehicle acceleration.

Bruno Fernandes et al, presents in paper [9], HDyCopilot, an Android application developed for automatically detecting road accidents with eCall and IEEE 802.11p. For the accident detection algorithm, it was chosen to combine the data collected from the vehicle (via ODB-II) with the data obtained from smartphone built-in sensors, namely accelerometer, magnetometer and gyroscope. The GSM / GPRS capabilities of the smartphone are also used to implement eCall. The proposed algorithm detects the occurrence of an accident when one of three situations occurs: triggering the airbag, overturning the vehicle or collision.

Upon detection of an accident, an alarm signal is transmitted via the IT2S platform. A road hazard warning message is automatically sent to all neighboring cars. The second alarm signal is transmitted to the EMS (Emergency

Medical System) by making an eCall, and sending an SMS containing the minimum eCall data set.

The HDyCopilot application runs on the Android operating system, specifically chosen for allowing access to hardware components through its open source APIs.

Cano et al [10] describe an app running on smartphones with the Android operating system that connects to the car's OBD-II system to detect road accidents. The proposed solution is based on Bluetooth technology in terms of the connection between the mobile application and the OBD-II interface, eliminating the cable connection, making it robust against car accidents.

The proposed application combines vehicle data (such as the airbag condition) with those provided by the smartphone (accelerometer, GPS) to improve the effectiveness of emergency services by automatically detecting the accident.

Akshay Agrawal et al. presents in the proposed paper [11] an idea of creating a car accident detection system by implementing an adaptive algorithm. The purpose of such an application is to help save lives by alerting you to the accident. The authors proposed to use an adaptable algorithm to detect accidents in any type of machine. Being an Android application, the system will be available at a low price, also available for low-end vehicles and usable in any country due to Android Market app uploading. When an accident occurs, the proposed system will detect it, then with the front and main camera, events are taken, and an SMS will be sent to the predefined contact numbers.

In the paper [12], the authors Yue Shi et al. propose a new accident detection concept using the features of a five-phase model to change the status of the user's movement during the fall. The concept of fall detection uses acceleration data provided by smartphone sensors (such as accelerometers and gyroscopes). Apply new features derived from a five-phase model describing the multi-stage drop process. A demonstration application called uCare, running on the Android operating system, is designed and implemented with the main purpose of helping people prevent and detect the fall.

Huang, Chuen-Min et al. propose a mechanism for detecting and tracking bicycle accidents based on mobile terminals [13]. In case of an accident, the proposed system uses different smartphone sensors such as accelerometer (G sensor) for detection, GPRS, Google Map and GPS for tracking and notifying the crash. To test and validate the performance of the proposed system, various hypotheses were studied, and an implementation of an Android mobile application called GoGoBike was performed.

Dr. Apps [14] launching a free mobile application that runs on smartphones with Android operating system and is called Accident Report. This is a crash application that collects different types of data to create a complex report like a PDF file, which can be send via e-mail. Creating a report involves following steps such as: drawing of the damage to the vehicle, photographing the vehicles, obtaining the location, filling in the personal details, details of the accident and information on the witnesses involved and information of the other party involved in the accident.

Beat the Traffic [15] is another Android app, developed later for the iOS operating system, which is used to report an accident automatically using the GPS. This is available in the US and Canada, with free access to camera images. An important feature of this app is "shake to report", which allows an accident to be reported by simply shaking the smartphone. To access the full functionality of the app, the user must sign up for a free account and customize the app according to his preferences.

SOSmart is a mobile app available in Google Play and the App Store for automated car accident detection using smartphone's integrated sensors. This application can also be used manually (by pressing the panic button). According to the official site [16], the application is able to automatically detect an accident using smartphone sensors such as GPS and accelerometer and can notify predefined contacts. The application also uses an algorithm that is based on real-time data provided by the National Highway Traffic Safety Administration [17], making it possible to distinguish between a landline phones, heavy braking, or a minor accident. Other features offered by the application are historical trips, as well as viewing the nearest hospitals.

Another important aspect is the ability to configure the application, depending on the users. If we talk about users who frequently travel by car, then the application can be set to automatic mode and it will start without the user intervening, running the application in the background. For people traveling less frequently, it is recommended to use manual mode, where start of race tracking is done from the application.

INRIX Traffic [18] is another application for reporting an accident and routing automatically. The application collects data from users, which are sent through real-time notifications. Reporting an accident is done through the application (at the button's push), not automatically. This application uses automated learning technologies for vehicle driving by users and personalizes routes to avoid traffic incidents. It also recommends routes and offers automatic alerts on changing road conditions such as accidents, roadblocks, etc. This application is free and available for smartphones with Android and iOS operating systems.

Another similar application is the Crash Notifier developed by R Systems International Limited [19] that offers automatic detection of the accident using the accelerometer and emergency notification functions after detection. For this application, the crash detection is performed using the existing accelerometer in the smartphone that is set to the 6G threshold. If this threshold is reached, based on the implemented algorithm, the timer starts and after 10 seconds a notification is sent to the previously established contact persons. In the case of false alarms, this can only be canceled manually in the ten seconds. If the notification has been sent and the accident is minor, or if it is not necessary to warn the contact, you can use the False Alarm button, which sends another notification and ignore the previous message. To improve the detection algorithm, the threshold value G can be manually configured from the application.

AxiKit is a step-by-step accident tracking application that stores different types of data, voice recordings, mapping the

location of the accident (date and time) automatically, capturing details with the camera, making a detailed report which can be easily transmitted by email [20]. The application also provides an emergency button for the user to call emergency services and a predefined contact person. It was originally only available for the iOS operating system, being optimized for the iPhone 5, but behind some updates, it also runs on the Android operating system.

CRADAR is a free Android application offered by ActionXL, which is used in case of an accident [21]. This application uses the phone's accelerometer to detect a fall and GPS to identify the location of the crash. The algorithm analyzes the data collected from the accelerometer and detects when the user falls. After detecting such an accident, the phone starts ringing and vibrating, and a counter will start. If the phone notices movements or the "I'm OK" button is pressed before the time expires, the alert is canceled, and the application goes back to the monitoring mode. After the default time has elapsed, a message containing accident location data (provided by the GPS) is sent to the established emergency contact.

Annex 1 lists the main advantages and disadvantages, the operating system and the sensors used for each system described above.

## IV. Proposed Idea and Further Research

People's safety and health is a lasting global concern. Every day, many human lives are lost due to accidents and delays in calling for rescue services [22]. Researchers are looking for a solution to reduce the loss of these lives. This delay is caused by various reasons. The most common is the lack of proper communication to emergency services.

An effective system is proposed to automatically notify these services of the accident and to guide them on the spot. In addition to other accident report systems, the proposed system will be able to detect different types of accidents, starting from simple falls to traffic accidents with serious injuries.

As shown in this paper, the smartphones can detect the occurrence of an accident, but the main problems remain the high rate of false detection. Different from other systems, our system will learn the user's daily behavior. For this we can use different machine learning algorithms such as neural networks [23], support vector machine (SVM) or decision tree. In the beginning, the system will work in the learning phase and after that it can be used being able to minimize false detection.

Based on the research presented, it has been found that the Android operating system is open source, more permissive with the use of embedded sensors, and phones with this operating system are much cheaper than iPhone smartphones.

Therefore, Android will be used as the main platform for designing and deploying the demonstration application, for implementing services, accident detection, emergency event reporting, and testing the functionality of the proposed system.

## V. Conclusion

In this paper are presented a series of systems used in case of accident based on mobile terminals such as smartphone, and the importance of their use in emergency situations. For each system it is highlighted how it works, the operating system on which applications are running, and the main advantages and disadvantages.

Chapter III present some of the existing applications for detecting an accident using smartphone built-in sensors. The user can choose the desired application, considering several features such as: the phone's operating system, the sensors used for detection and the additional features offered by each application.

## References

[1] The ecomonist (2015), Planet of the phones, Smartphone ubiquitous addictive and transformative plantet phones, Available:https://www.economist.com/news/leaders/21645180-smartphone-ubiquitous-addictive-and-transformative-planet-phones.

[2] Fortino, G, Trunfio, P., "Internet of things based on smart objects: Technology, middleware and applications", 1 January 2014, Pages 1-198, ISBN: 978-331900491-4;

[3] Erwin A. Blackstone, Andrew J.Buck, Simon Hakim, "Evaluation of alternative policies to combat false emergency calls", Evaluation and Program Planning, Volume 28, Issue 2, May 2005, Pages 233-242.

[4] OnStar System, Official site Opel, Available: http://www.opel.ro/onstar/onstar.html

[5] Jules White, Chris Thompson, Hamilton Turner, Brian Dougherty, and Douglas C. Schmidt, "WreckWatch: Automatic Traffic Accident Detection and Notification with Smartphones", Mobile Networks and Applications, DOI: 10.1007/s11036-011-0304-8, June 2011.

[6] Chalermpol Saiprasert and Wasan Pattara-Atikom, "Smartphone Enabled Dangerous Driving Report System", 46th Hawaii International Conference on System Sciences, DOI: 10.1109/HICSS.2013.484, 2013.

[7] Hamid M. Ali et al, "Car accident detection and notification system using smartphone", International Journal of Computer Science and Mobile Computing, Vol.4 Issue.4, April- 2015, pg. 620-635, ISSN 2320–088X.

[8] F.J. Bruwer and M.J.(Thinus) Booysen, "Vehicle acceleration estimation using smartphone-based sensors", Conference: South African Transport Conference, At Pretoria, July 2015, DOI 10.13140/RG.2.1.3244.4640.

[9] Bruno Fernandes, Vitor Gomes, Joaquim Ferreira and Arnaldo Oliveira, "Mobile application for automatic accident detection and multimodal alert", IEEE Vehicular Technology Conference, At Glasgow, July 2015, DOI: 10.1109/VTCSpring.2015.7145935.

[10] J. Z. C. T. C. Juan Carlos Cano, Pietro Manzoni, "Providing Accident Detection in Vehicular Networks Through OBD-II Devices and Android based Smartphones," in Proceedings of the 5th IEEE Workshop On User Mobility and Vehicular Networks, December 2011, DOI: 10.1109/LCN.2011.6115556.

[11] Akshay Agrawal, Anand Khinvasara, Mitali Bhokare, Sumit Kaulkar, Prof. Y. K. Sharma, "Accident detection system application", International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS), Issue 6 Volume 5, pp. 425-428, ISSN : 2279-0055.

[12] Yue Shi, Yuanchun Shi, Xia Wang, "Fall Detection on Mobile Phones Using Features from A Five-phase Model", Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC), October 2012 , DOI: 10.1109/UIC-ATC.2012.100.

[13] Chuen-Min Huang, Yen Chang. "A mobile-based novice accident detection and tracking system for bicycle", Available:https://www.researchgate.net/publication/267977300_A_MOBILE-BASED_NOVICE_ACCIDENT_DETECTION_AND_TRACKING_SYSTEM_FOR_BICYCLE.

[14] Accident Report Application ( 2015), Available online: https://play.google.com/store/apps/details?id=com.rany.accidentreport.en, Developer of the application Drapps, Available online: http://drapps.info/.

[15] Beat the Traffic application (2017), Available online: http://www.appsapk.com/beat-the-traffic/.

[16] SOSmart application (2017), Available online: http://www.sosmartapp.com/.

[17] National Highway Traffic Safety Administration, Available online: https://www.nhtsa.gov/.

[18] INRIX application (2016), Available online: http://inrix.com/mobile-apps/.

[19] Crash Notifier (2017), Developer of the application: R Systems International Limited, Available online: http://www.rsystems.com/.

[20] AxiKit application, Available online: http://www.axikit.com/fleet#AxiKit-Mobile-App.

[21] Cradar Application, Available online: http://actionxl.com/CRADAR.html.

[22] 10 Leading Causes of Death, West 2000 - 2015, Available online: https://webappa.cdc.gov/cgi-bin/broker.exe.

[23] M. Soltane, M. Ismail, ZAA Rashis. "Artificial Neural Network Approach to PPG Signal Classification", International Journal of Computing and Information Science, Vol 2. No 1 2004, pp. 58-65.

ANNEX 1

| No | Solutions | | | | |
|---|---|---|---|---|---|
| | ***Name of system*** | ***Operating system*** | ***Sensors used*** | ***Advantages*** | ***Disadvantages*** |
| 1 | WreckWatch | It's not specified | AccelerometeMicrophone GPS | Sensor smartphones can measure forces closer to those of victims;<br>Reduce the complexity of software maintenance through smartphone update mechanisms. | Accident detection systems consume a significant amount of battery power;<br>Low speed traffic can trigger WreckWatch service deactivation. |
| 2 | Smartphone Enabled Dangerous Driving Report System* | It's not specified | GPS | The values obtained by the three smartphones used in the experiments are the same as those provided by the vehicle's speedometer, resulting in a high GPS accuracy; | The way the data is received, the reception condition affects the accuracy and sensitivity of the recording data; The position of the smartphone in the vehicle has an effect on the correctness of the recorded data; |
| 3 | CADANS | It's not specified | AccelerometerGPS<br>Microphone Camera | Combining the data provided by the sensors and determining whether the occupant is inside or outside the vehicle helps to minimize falsities;<br>CADANS offers the possibility of canceling the alarm and notifications if a false accident has been detected by pressing a cancel button within 10 seconds of the accident detection; | Testing the system in a real environment (car crash) is not realistic and possible;<br>Microphones built into smartphones differ from brand to brand, some have set a maximum threshold of 140db, making it hard to distinguish the sound produced by the airbag and the maximum volume of the radio; |
| 4 | Vehicle acceleration estimation using smartphone-based sensors* | It's not specified | AccelerometerMagnetometerGyroscope<br>Kalman filter | The proposed system is capable of detecting low acceleration values, which is beneficial for the development of mobile applications where low-speed acceleration is of interest. | A disadvantage discovered after comparing measured results with those calculated theoretically show that in case of an accident in which the vehicle overturns, Kalman filtering operation is adversely affected. |
| 5 | HDyCopilot | Android | Accelerometer Magnetometer<br>Gyroscope vehicle sensors | To avoid reporting a fake accident, alert notifications can be canceled within a certain amount of time via the mobile application | The disadvantages of the system are not presented in this paper. |
| 6 | Providing Accident Detection in Vehicular Networks Through OBD-II Devices and Android based Smartphones* | It's not specified | Accelerometer, GPS<br>vehicle sensors (airbag) | To reduce false alerts, the user can interrupt sending an accident report within a set interval (one minute) via the mobile application. | Using Bluetooth mode continuously affects the battery life |
| 7 | Accident detection system application[1] | Android | Camera, Accelerometer<br>GPS | The adaptive algorithm has been chosen because each machine has different vibration inputs, being able to set accident thresholds | For the proposed system, the smartphone has to be positioned in a phone holder to detect machine vibrations through the built-in sensors |
| 8 | uCare | Android | Accelerometer Gyroscope | It takes into account the distinction of five types of fall (falling on the hands, knees, back, left and right of the body) | The disadvantages of the system are not presented in this paper. |
| 9 | GoGoBike | Android | GPS<br>Accelerometer | The advantages of the system are not presented in this paper. | Continuous use of GPS mode affects battery life;<br>Reporting of the accident occurs only in the adhered group; |

*In these papers, the proposed system does not have a predefined name.

| | | | | | |
|---|---|---|---|---|---|
| 10 | Accident Report | Android | GPS<br>Camera | Free application;<br>Available for Android 3.0+; | To complete a complex report you must go through many steps, and in emergency situations it is difficult to achieve; |
| 11 | Beat the Traffic | Android and iOS | GPS<br>Accelerometer | The advantages of the system are not presented in this paper. | This app is only usable for minor accidents to inform drivers involved in traffic, so an accident is reported only by marking it on the map; |
| 12 | Sosmart | Android and iOS | GPS Accelerometer | Automatic accident detection;<br>Quick answer in case of accident;<br>Detects the crash in the absence of the signal but does not report it; | On iOS, the application must also be installed by predefined contacts to receive crash notifications;<br>Does not work without GPS mode turned on;<br>Uses a lot of energy (battery discharge); |
| 13 | INRIX Traffic | Android and iOS | GPS | INRIX uses a cloud platform, Autoelligent;<br>In addition to the main features, this app also offers some additional features; | A major drawback of the application is the continued use of GPS in the background that drastically reduces battery life; |
| 14 | Crash Notifier | Android | Accelerometer | Application is free and available on Google Play | Setting the manual threshold inside the application can result in false detection of an accident |
| 15 | AxiKit | Android and iOS | Camera, Microphone<br>GPS | AxiKit can be downloaded for both operating systems for free; | Is also a personalized application that involves paying costs, and is generally intended for large companies; |
| 16 | Cradar | Android | Accelerometer GPS | In order not to affect the life of the battery, CRADAR uses the GPS only when a fall is detected. | The disadvantages of the system are not presented in this paper. |

# Printed Arabic Script Recognition: A Survey

Mansoor Alghamdi
Department of Computer Science
Community College
University of Tabuk
Tabuk, Saudi Arabia

William Teahan
School of Computer Science
Bangor University
United Kingdom

***Abstract*—Optical character recognition (OCR) is essential in various real-world applications, such as digitizing learning resources to assist visually impaired people and transforming printed resources into electronic media. However, the development of OCR for printed Arabic script is a challenging task. These challenges are due to the specific characteristics of Arabic script. Therefore, different methods have been proposed for developing Arabic OCR systems, and this paper aims to provide a comprehensive review of these methods. This paper also discusses relevant issues of printed Arabic OCR including the challenges of printed Arabic script and performance evaluation. It concludes with a discussion of the current status of printed Arabic OCR, analyzing the remaining problems in the field of printed Arabic OCR and providing several directions for future research.**

***Keywords*—Optical character recognition; arabic printed OCR; arabic text recognition; arabic OCR survey; feature extraction; segmentation; classification***

## I. Introduction

Optical Character Recognition (OCR) is a technique that transforms a printed or handwritten text image into an electronic format. OCR development is considered a challenging task in the field of pattern recognition. Many OCR approaches have been proposed for Latin and non-Latin scripts. However, printed Arabic OCR still poses great challenges because of the special characteristics of Arabic script [1].

Arabic OCR is highly desirable in various real-world applications, such as digitising learning resources to assist visually impaired people, bank cheque processing and mail sorting[2], [3]. Furthermore, there are many initiatives for Arabic digital content enrichment [4]. One of these initiatives is King Abdullah's Initiative for Arabic Content. Therefore, a robust and efficient Arabic OCR is required to support this initiative by increasing Arabic content on the Internet.

Numerous methods have been proposed for recognising printed Arabic script from an image, yet we are unaware of comprehensive surveys of printed Arabic OCR during the last fifteen years. Two surveys have been conducted on printed Arabic OCR [2], [5]. However, these reviews do not reflect the current progress in printed Arabic OCR. Therefore, establishing a guide and baseline for future directions remains important for Arabic OCR researchers.

This work will establish this guide and baseline for Arabic OCR researchers by providing a comprehensive literature review of printed Arabic text recognition research. It reviews techniques that have been utilized for developing printed Arabic OCR with emphasis on the issues related to Arabic script. It also highlights the current status of printed Arabic OCR and provides several directions for future research.

This paper is organised as follows. In section 2, Arabic script characteristics and challenges are discussed. Section 3 presents the methodologies of printed Arabic OCR, with subsections that review the five stages of the development of printed Arabic OCR: preprocessing, segmentation, feature extraction, classification and post-processing. Section 4 discuses performance evaluation issues of printed Arabic OCR. Section 5 concludes with a discussion about open problems and future directions.

## II. Arabic Script Characteristics and Challenges

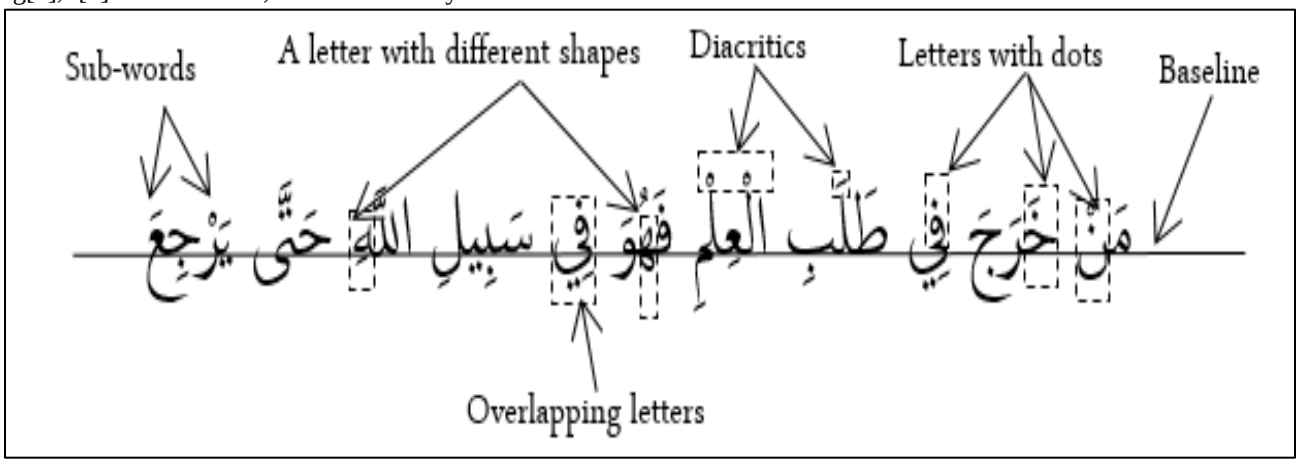



Fig. 1. Arabic script characteristics

TABLE I. ARABIC CHARACTERS WITH DIFFERENT POSITIONS AND SHAPES

| Isolated | Initial | Middle | End |
|---|---|---|---|
| ا | ا | ـا | ـا |
| ب | بـ | ـبـ | ـب |
| ت | تـ | ـتـ | ـت |
| ث | ثـ | ـثـ | ـث |
| ج | جـ | ـجـ | ـج |
| ح | حـ | ـحـ | ـح |
| خ | خـ | ـخـ | ـخ |
| د | د | ـد | ـد |
| ذ | ذ | ـذ | ـذ |
| ر | ر | ـر | ـر |
| ز | ز | ـز | ـز |
| س | سـ | ـسـ | ـس |
| ش | شـ | ـشـ | ـش |
| ص | صـ | ـصـ | ـص |
| ض | ضـ | ـضـ | ـض |
| ط | طـ | ـطـ | ـط |
| ظ | ظـ | ـظـ | ـظ |
| ع | عـ | ـعـ | ـع |
| غ | غـ | ـغـ | ـغ |
| ف | فـ | ـفـ | ـف |
| ق | قـ | ـقـ | ـق |
| ك | كـ | ـكـ | ـك |
| ل | لـ | ـلـ | ـل |
| م | مـ | ـمـ | ـم |
| ن | نـ | ـنـ | ـن |
| ه | هـ | ـهـ | ـه |
| و | و | ـو | ـو |
| ي | يـ | ـيـ | ـي |

There is no doubt that printed Arabic OCR faces a number of challenges and there is still an intensive need for more research [6]. However, most challenges facing the development of Arabic OCR are due to the characteristics of Arabic script. Arabic script has some features that distinguish it from other languages. Compared to English, the most obvious feature of Arabic script is that it is written cursively from right to left in both printed and handwritten. The greatest challenges are due to the more complex characteristics of Arabic script. In the following section, the characteristics of Arabic script that may complicate recognition will be discussed:

### A. Shapes and Positions

The Arabic alphabet has 28 basic letters (see Table 1). However, an Arabic letter may contain four dissimilar shapes in relation to its location inside a word: whether it is an isolated letter, an initial letter (in which a letter is inked from the right side, an ending letter (in which a letter is linked form the left side) or a middle letter (in which a letter is linked from the right and left sides). Thus, the number of letters to be recognized will increase from 28 letters to 125 letters.

### B. Overlapping characters and Ligatures

Characters in an Arabic word might be overlapped vertically with or without touching each other (see Figure 1). In particular, some characters are combined and written as a ligatures such as (لا) which is a combination of two letters Lam (ل) and Alf (ا). However, ligatures occurs in Arabic script depending on the type of fonts being used. For instance, in Traditional Arabic font, there are about 220 ligatures whereas Simplified Arabic incorporates about 150 ligatures, [7].

### C. Diacritics

Characters in an Arabic word can exist with diacritics or short vowels such as Fat-hah, Dhammah, Mada'ah, Kasrah and Sukkun, as illustrated in Figure 1. These can be placed either over or below the letters as strokes. In addition, Tanwen is considered as a diacritic which is indicated by double Fat-hah, double Dhammah and double Kasrah. One more diacritic that Arabic script has is Shaddah which is similar to the number 3 as it is rotated 90° clockwise.

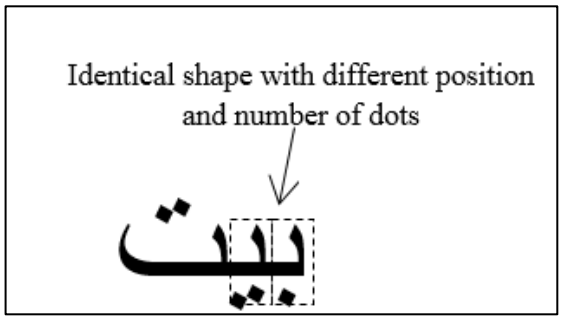



Fig. 2. Two characters (*Ba* and *Ya*) with an identical shape and a different number of dots.

### D. Cursive

As mentioned above, Arabic script is a cursive script which means that a word is composed of connected characters. However, six characters (ا, د, ذ, ر, ز, و) of the Arabic alphabet are not linked with succeeding letters. This can present a challenge because these characters can divide a word into one or more units as sub-words (see Figure 1).

### E. Presence of dots

The Arabic alphabet relies on number and position of dots in order to differentiate between similar letters (see Figure 2). Fifteen characters in the Arabic alphabet have dots. They can be placed below the character, above it or in the middle. Ten of these characters are dotless, three have two dots and two have three dots, as shown in Table 1.

## III. GENERAL ARABIC OCR METHODOLOGY (MODEL)

This section will focus on the methodologies used by printed Arabic OCR systems. Published approaches and systems for Arabic OCR indicate that the process of implementing Arabic OCR consists of five phases: (1) pre-processing; (2) segmentation; (3) feature extraction; (4) classification and (5) post-processing, see figure 3.

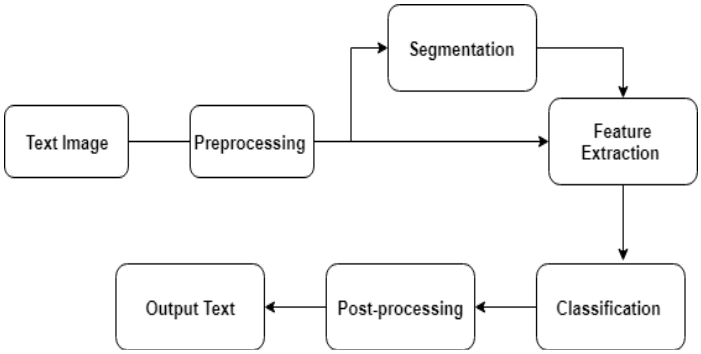



Fig. 3. General printed Arabic OCR methodology.

### A. Preprocessing Phase

This is the first phase of OCR methodology which is responsible for enhancing the readability of the input image. Preprocessing is a combination of algorithms that are applied to the input image in order to reduce noise and alterations, thus simplifying the subsequent phases of OCR methodology [2]. There are various factors that affect the quality of the input image. A study lists the history of image, the printing process, the kind of font, the quality of paper, the condition of the image and the image acquisition as the vital factors that influence the input image quality [2].

Researchers emphasize that the downstream OCR accuracy relies on the quality of the input image [8]. Furthermore, a study states that OCR systems, which report high recognition accuracy on some input images, will report less recognition accuracy on input images that are poor in quality [9]. Thus, the preprocessing phase is a critical stage in OCR development that simplifies the data for the subsequent phases to operate accurately. Generally, several preprocessing operations are employed on the input image: binarization, layout analysis, thinning, smoothing and filtering, size and slant normalization, slant detection, skew detection and baseline detection. However, the selection of these operations, to be applied in the preprocessing, relies upon the conditions of the input image, such as the amount of noise and skew in the input image [10]. In the following section, the preprocessing techniques which are applied in Arabic OCR, will be clarified.

#### 1) Binarization

For character recognition, the binarization (sometimes called thresholding) process involves converting an input gray scale image into a binary image, in which a pixel has only two values 0 and 1. The binary image has the critical information, such as the shape of characters. It has been found that increasing processing speed and reducing storage capacity are the key benefits of binarization technique [7], [11]. Researchers suggest selecting the most appropriate method for binarization might separate connected objects or joining isolated objects [12]. A number of studies have confirmed the efficiency of computing the histogram of the gray scale of an image and then detecting a cut-off point as the binarization method [13] [12]. However, some researchers work on recognition without applying binarization methods, such as [14], [15].

#### 2) Size Normalization

Since Arabic characters differ in size, as described earlier, size normalization is commonly applied to characters or words by scaling the characters or the words to an adjusted size. This process is crucial for the recognition or classification phase, since some recognition methods are sensitive to dissimilarity in size and position, such as template matching and correlation approaches [16]. A study classified normalization methods into two approaches: moment-based normalization; and nonlinear normalization [17]. It is argued that normally a character is normalized to a standard size for classification [18]. However, in terms of word normalization, applying normalization to a word instead of a character will result in losing critical information [18], [19].

#### 3) De-noising

Noise may be presented during the acquisition process via scanners which results in distortions and variations in the input text image. Besides this, very small items in the text image can be reflected as noise [11], which are byproducts of image scanning or binarization and which are not parts of the text. Such noise may has a major impact on the performance of OCR systems [20]. Noise removal is an operation for enhancing the visual quality of the input image [21].

As a solution, several techniques have been introduced that are considered as noise removal methods [22]. These methods include filtering and morphological operations (smoothing) which are conditioning processes in terms of OCR development [2], [23]; for instance, dilation algorithms, which are applied to broken letters, and erosion algorithms which are applied to text images with touching letters [17]. In addition, the median filter approach is commonly used in both printed text images and handwritten text images. For example, a study apply this approach for removing noise in printed Arabic text images [24]. Another example of a study that applies a median filter algorithm in handwritten Arabic text images is in [25].other researchers [26] applied a morphological noise removal method for Arabic printed OCR proposed in [27]. However, a study discovered that letter holes could be filled while applying this method, with lower thresholds, to Arabic text images [26].

In fact, the review suffers from the fact that some printed Arabic OCR studies applied noise removal algorithms without providing information of the applied algorithm, for instance, in [28]. Such approaches, however, should be selected carefully when considering OCR systems. That is, because of the similarity between Arabic letters, any alteration of a letter might change it to another letter. Thus, a perfect noise removal method is able to eliminate noise while preserving the shape of the character [20].

#### 4) Skew Detection and Correction

Initially, a text image has zero rotation, yet when physically scanning the image manually, rotation of images up to 20º might occur [5]. This rotation is called skew which results in non-zero skew text images (see Figure 4). The skew can lead to incorrect recognition and baseline detection [29]. It is impossible to segment a text if the text is rotated [30]. As a result, detecting and correcting the skew is critical to OCR applications that rely on segmentation approaches to recognize characters.

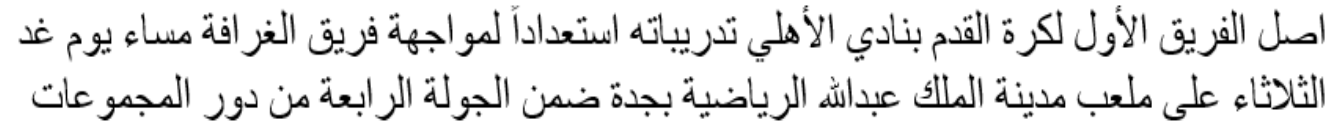



Fig. 4. An example of Arabic text skew.

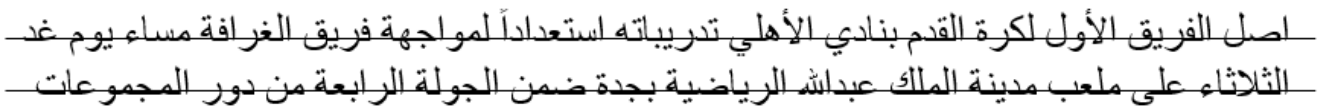


Fig. 5. Baseline detection of a printed Arabic text image.

The process of estimating the skew angle is known as skew detection, whereas the process of rotating the image with the purpose of correcting the skew is called skew correction. A wide variety of skew detection and correction methods have been proposed. A study groups these methods into five groups: projection profile, Hough transform, Fourier transform, nearest neighbor clustering and correlation [31]. The Hough transform is the standard approach for detecting the skew [32]. A method based on the projection profile was introduced in [33]. A researcher has provided a comprehensive review of twenty–five skew detection and correction approaches [34]. The author concludes that further work on more sophisticated methods is still required. The Radon transform method has shown its efficiency for skew correction [35]. Some methods are designed for specific applications and image type. For example, a new method has emerged for Arabic text images in [36]. One study concludes that selecting a skew detection and correction method relies on the image type [37].

*5) Baseline Detection*

As described in the previous section, Arabic characters are joined through a horizontal line called the baseline (see Figure 5). Graphically, the baseline can be described as the line which has the maximal amount of black pixels [38]. This line contains critical information about the text, such as text orientation and position of connection points between Arabic letters [2]. Thus, detecting the baseline is beneficial for many OCR stages, for instance, skew normalization [39], segmentation [40], [41] and structural features extraction such as the character's dots [42].

It has been reported that most Arabic OCR has applied baseline detection methods as a preprocessing step [25] .The baseline detection techniques for Arabic script has been classified into four groups in [36]; namely, horizontal projection methods, the word skeleton method, contour tracing and principle component analysis. Among these, the horizontal projection technique is widely implemented for determine the baseline in Arabic OCR, such as in [43], [2]. Several studies implement a horizontal projection approach in OCR systems for detecting the baseline, such as in [26], [42], [44], [45]. It has been emphasized that the horizontal projection method is simple and efficient for Arabic printed text [43], [46]. However, this method is applicable only for noise-free images, as it fails for unclean images [47].

Another baseline detection approach is the x-y cut proposed in [48] which is based on a horizontal projection method. This method works well for Arabic noisy images, though it fails in the presence of large amounts of noise and skew [47]. Consequently, researchers proposed using a ridge-based text line detecting approach for Arabic text [47]. The former method's efficiency has been tested and recommended for different types of Arabic text images, since it was found to achieve above 96% text line detection accuracy [47].

Researchers summarize the state of the art of baseline detection methods in Arabic script [49]. In summary, for printed Arabic text, the standard horizontal projection method is sufficient for detecting the baseline, since the baseline in printed text is straight. Whereas for handwriting, the baseline is not straight, thus more sophisticated approaches should be considered [50].

*6) Thinning and Skeletonization*

Thinning " skeletonization" can be defined as "the process of peeling off a pattern as many pixels as possible without affecting the general shape of the pattern" [18]. In other words, it involves operations that can be implemented in order to produce the skeleton of text images. Thinning is a crucial processing step for text recognition, in particular for such OCR applications in which extracting the skeleton of a character is essential [2], [32]. However, in terms of the obtaining the skeleton, it must be as thin as possible, connected, and centered [18]. Thinning simplifies the process of the segmentation, future extraction and classification phases as a result of reducing the amount of data that needs to be considered in the input image [25].

Most of the existing thinning algorithms have been designed for general purpose or other text languages [51]–[53]. However, when applying thinning algorithms to Arabic scripts, various obstacles are encountered [49]. One problem is the reduction in the number of dots in some Arabic characters as a result of the thinning process for which the number of dots is a crucial aspect in differentiating between these characters [54].


Fig. 6. Example results of different thinning algorithms: (a) original word, (b), (c) and (d) thinned word.

Fig. 7. An example of line segmentation.

Also, dots in Arabic characters are likely to be vulnerable to noise. However, some researchers extract dots of Arabic characters before applying thinning algorithms, in order to overcome this problem [55], [12]. Another problem of thinning algorithms when considering Arabic script concerns preserving the connectedness of Arabic text. Some thinning approaches may not cope well with Arabic text due to its connectivity characteristic [49]. Thus, this should be taken into consideration, when selecting thinning algorithms for Arabic text. Also, since Arabic characters consist of different shapes such as loops and lines, the selected thinning algorithm must be capable of preserving these different shapes.

Therefore as a consequence of specific characteristics of Arabic script detailed above, direct adoption of thinning algorithms, which have been developed for other languages, may not be as effective [24]. As a result of these difficulties, there is comparatively little published work on developing thinning algorithms for Arabic [25], [24].

Some studies introduce thinning algorithms for Arabic letters [56], [57]. However, the proposed algorithms can only deal with isolated Arabic characters. One study provide a thinning algorithm which is designed specifically for printed Arabic script recognition to overcome dis-connectivity and loss of information [58]. This algorithm is applied on Arabic text to illustrate the efficiency of reducing the outline of each word's characters (its number of pixels) thereby overcoming the challenges of Arabic script. Also, the authors propose an experimental framework with new performance measures for the evaluation of thinning algorithms. Figure 6 shows the output of three different thinning algorithms.

### *B. Segmentation Phase*

After the preprocessing phase, an enhanced text image in the sense of low noise and variation, and a necessary amount of character information [2], has been produced. During the segmentation phase, the text image is segmented into small components, with a page being segmented into lines, a line into words and a word into letters [59]–[61]. Segmentation is a crucial step in Arabic OCR system development because of the fact that it plays a vital role in ensuring the success of the subsequent feature extraction and classification stages [3], [46]. However, the author in [46] stresses that misrecognition can arise by applying a poor segmentation method. As a result, this stage will have a critical impact on the recognition rate of the text [7].

As explained previously, one of the main challenges facing Arabic OCR development is the cursiveness of Arabic script. Segmentation of Arabic text thus can be more difficult and time consuming for the development of Arabic OCR systems [3]. Correspondingly, segmentation has been considered as the main contribution for increasing the recognition error rate in Arabic OCR systems [46], [62], [63].

Generally, segmenting a text image can be graded into two types: external segmentation; and internal segmentation [64]. While the former type deals with the isolation of different writing objects such as, paragraphs, sentences and words, the latter deals with the isolation of characters [64], [65].

#### *1) External Segmentation*

External segmentation refers to the document layout analysis, in particular page decomposition. Document layout analysis is accomplished in order to identify the physical structure of a page [66]. As far as offline OCR development is concerned, page analysis is a basic step which segments the image into its different logical parts with the identical type of information, such as graphs, text and tables. Page layout analysis is performed in two approaches: structural analysis by which a page is decomposed into blocks of the page elements, such as paragraphs and words; and functional analysis by which a page is decomposed into functional elements such as title and abstract [41], [65], [66].

With respect to Arabic document processing, page decomposition refers to the isolation of text lines of a texture region and the segmentation of words and sub-words [5], [7], [67], since it is restricted to text images [5]. Applying a fixed threshold to Arabic text documents to determine text lines is the standard method [5], [68], [69]. However, this method fails with a skewed text image [40].

Methods based on histogram projection are considered as conventional approaches for isolating lines and words in Arabic text documents [68], [70]. Several studies have relied on horizontal projection techniques for segmenting Arabic text images into lines, such as in [71]–[76], [28]. [72] It is recommended horizontal projection be applied for text images because of its advantages in reducing computational load and its simplicity of implementation [71]. Moreover, horizontal projection is an appropriate method for locating text lines in Arabic printed text, since the text lines in printed text are straight [50].

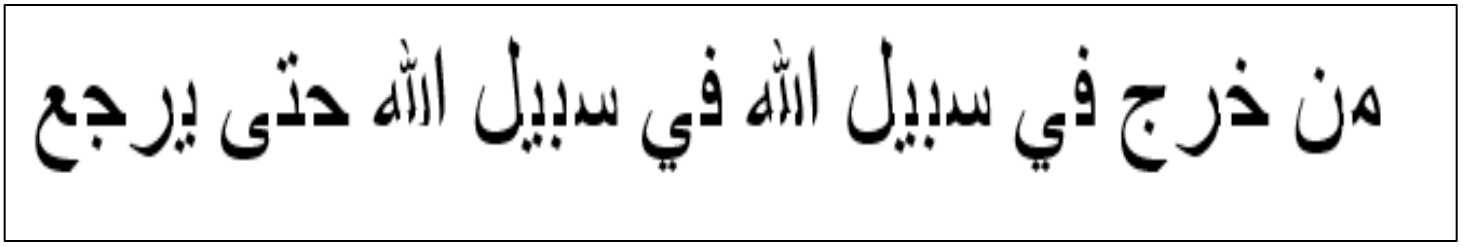



Fig. 8. Segmenting Arabic words into their characters.

For line segmentation, researchers in Arabic OCR determine words in a line of text by inspecting the vertical projection [14], [28]. (See Figure 7). This method depends on the estimation of the minimum space between words. However, it was pointed out in the Arabic script characteristics section above that some Arabic characters are not linked with succeeding letters, thus this results in a word having with one or more connected components (sub-words), as shown in

Figure 7. To overcome this issue, methods based on vertical projection consider that the width of spaces between sub words is smaller than the width of the spaces between words [14].

Generally, it is relatively easy to segment a text line into words in printed text images, compared to handwritten text images which involve overlapping and touching characters by using vertical projection histogram profiles [37], [59], [25]. However, some Arabic fonts contain characters that vertically overlap, such as the Traditional font type. Thus, Arabic script even in printed form can contain touching and overlapping characters, so algorithms that have been designed to overcome this challenge for handwritten script may be utilized for printed Arabic. For example, the authors in [77] have developed a method based on the connected components that analyses the distance between connected components in order to segment handwritten words.

*2) Internal segmentation*

Internal segmentation deals with segmenting a word into characters. When reviewing segmentation methods in the literature, a major complication arises concerning the classification of word segmentation approaches. For instance,[5] a study classifies Arabic OCR systems based on word segmentation into 'segmentation based systems', which is based on analytical techniques where a word is segmented into characters, and 'segmentation–free systems', which is based on recognizing a word as a unit without segmentation [72]. Some researchers discuss word segmentation in terms of implicit and explicit segmentation [73], [78]. Others classify word segmentation in terms of techniques which have been applied to segmenting a word, such as [59], [46], [3]. Researchers organize segmentation methods for Arabic script into holistic approaches and analytical approaches [25].

Mostly, Arabic OCR systems have been developed by two main paradigms: holistic approaches (segmentation–free) which require a large lexicon of Arabic words, and analytical approaches (segmentation based) where a word is segmented into units and each unit is recognized separately.

### *C. Holistic Approach*

Segmentation-free or holistic Arabic OCR systems perform the recognition on the entire word as a unit without segmenting the word or recognizing characters separately [2]. Several studies have investigated the holistic approach for printed Arabic scrip OCR such as in [79], [16], [8]. OCR systems based on a holistic approach require tracing the feature of the entire word and dealing with words instead of characters. As a result, this approach is restricted to recognizing a word against a lexicon [2]. Moreover, this approach has the challenge of how to deal with the large lexicon size of Arabic words. It is claimed that systems based on this type of segmentation are not useful for general text recognition. A study suggests this approach for systems in which a lexicon is statically defined, such as bank cheque recognition where vocabulary is limited [80].

### *D. Analytical Approach*

For the analytical or segmentation based approach, Arabic OCR systems segment words into smaller units like characters (see Figure 8). In the typical Arabic OCR system, the analytical approach is divided into two approaches: explicit segmentation and implicit segmentation.

*1) Explicit Segmentation*

The explicit segmentation approach, which is also called dissection segmentation, attempts to segment a word into smaller units. These units could be characters, strokes or loops. Researchers argue that there are two classes of explicit segmentation, which are: direct segmentation and indirect segmentation [81]. In the former, a word is directly segmented into characters exploiting a set of heuristics, while in the latter, a word is divided into smaller segments which can be characters or marks that over segmented characters, such as strokes.

Projection analysis is considered as one of the earliest applied dissection methods on Arabic character segmentation [46], [68], [70]. The projection method of the text image aims to reduce 2D information into 1D in order to simplify the character segmentation process. A method based on a modulated histogram of the image has been proposed in [82]. However, this method has been tested on specific Arabic fonts which do not contain overlapping and ligatures. Consequently, this method would not be appropriate for Arabic fonts that have ligatures, such as traditional Arabic font [3], [62].

Another histogram projection method is presented for printed Farsi word segmentation in [83] which is also applicable to Arabic script, as Arabic script is similar to Farsi script [3]. However, this method is font dependent and ineffective in segmenting small font sizes. Although many of the other techniques based on projection analysis have been devolved for Arabic script such as in [84], [62], [85], it seems that no projection based segmentation algorithm is accurate in segmenting Arabic text [50].

Instead of applying projection analysis methods, contour–based algorithms, which are used for dissection segmentation that rely on the skeleton or contour of Arabic words, are used to simplify the Arabic word segmentation such as in [78]. Other methods rely on white space and pitch finding techniques for segmenting Arabic words [46], [74]. However, a major criticism of the explicit approach is that it is expensive because of the requirement of finding the optimum word from the arrangement of segmented units [81]. The researchers in [71] conclude that an accurate segmentation may not be acquired by relying on dissection segmentation approaches.

*2) Implicit Segmentation*

In OCR systems based on implicit segmentation, the segmentation phase and recognition phase are performed simultaneously [3]. In other words, a word is segmented into characters while being recognized without segmentation in advance [46]. Straight segmentation and recognition based segmentation are also referred to as implicit segmentation [46]. This segmentation approach searches the text image for components that match predefined classes. The principle of implicit segmentation is to utilize a sliding widow to segment the word image into frames of fixed width on which classification relies to make a decision [86]. Owing to challenges in segmentation of cursive scripts such as Arabic, researchers use the implicit segmentation approach in order to overcome the problems of word segmentation [80]. In

principle, by applying this type of segmentation, there is no need for a specific dissection algorithm for Arabic script segmentation and the accuracy performance relates to the classification performance [87]. Thus, some researchers implement techniques based on implicit segmentation in order to improve recognition accuracy of Arabic OCR, such as in [88], [12], [89].

### *E. Feature Extraction Phase*

Once the text image is segmented into isolated regions (such as character, part of character), the next step is feature extraction which is the process of obtaining distinguishing attributes of the segmented character to be utilized by the next phase which is classification [90]. Feature extraction is the most significant level that heavily influences overall OCR performance [11], [25], [60]. The feature extraction stage is correlated with other OCR stages, such as preprocessing and classification stage. In other words, the authors in [91] point out that the selection of feature extraction methods depends on the output of the preprocessing stage. For instance, some techniques for feature extraction work on skeletons, whereas others work on grayscale images. Moreover, the set of features extracted must match the specification of the selected classifier [2].

In terms of OCR performance, feature extraction plays a critical role in achieving high accuracy performance [11], since the feature extraction stage has the contributes to the success of the classification step [60]. However, selection of feature types is a major issue in OCR development [92]. Researchers recommend that the feature extraction methods should be independent of scalable font characteristics such as font styles, font types, font sizes and should be able to describe and distinguish different patterns effectively [92], [93]. In other words, a study emphasizes that the key purpose of selecting good features is to maximize the effectiveness and the efficiency of the OCR system minimizing the complexity and processing time simultaneously [94].

Among OCR system development, researchers propose various types of features. Such features can be categorized into three groups: structural features; statistical features; and global transformation feature [5]. In the following, these features will be discussed in the context of recognizing Arabic script.

#### *1) Structural Features*

Structural features illustrate a text image in terms of its topological and geometrical characteristics by using its local and global properties [2], [92]. In case of Arabic script, lines, dots, loops, holes, strokes and zigzags are some structural features [92]–[99]. Considering Arabic script characteristics, some characters have common primary shapes and they can only be differentiated by the number and location of their dots. Thus, the researchers in [59] claim that structural features have been commonly used for Arabic script in order to capture the dot information of characters explicitly.

On the other hand, A study argues that structural feature methods are not capable of discriminating between characters having similar shapes [100]. Similarly, a study reports that relying on the structural features of Arabic script may result in misrecognition, owing to the small difference between Arabic letters [92]. It is mentioned in [5] that extracting structural features of Arabic characters is a challenging task. Furthermore, it is claimed that Arabic OCR systems implementing structural feature methods are processed exhaustively [101]. Likewise, various studies have reported that another complication of applying structural features is that it involves expensive preprocessing techniques, such as skeletonization which may result in character shape distortion and loss of structural feature data [60], [59], [102]. Therefore, research on Arabic OCR has been carried out on other feature extraction approaches, as will be discussed below, that are effective in reducing process time and improving performance accuracy [101].

#### *2) Statistical Features*

Statistical features are derived from statistical representation of patterns which provide a measurable event of interested patterns. Researchers in Arabic OCR systems adopt different approaches to produce statistical features. Some examples of the approaches, which have been applied for representing Arabic characters, are zoning, moments, characteristic loci, histograms and crossing [92], [5], [2], [25].

The zoning method divides the character image into serval overlapping and non-overlapping regions. Then, the density of each region pixel is analysesd and used as a feature [92], [103].

The moment method is a common statistical feature approach that has been applied in patter recognition applications [26]. Moments, including Legendre moments, Zernike moments, central moments, pseudo-Zernike moments and Hu moments, extract geometric features in an image, such as, the shape area of a pattern and the center of the mass [17], [104] and [105]. Several studies in printed Arabic script, such as, [106], [107]–[109], have applied moment invariants as a feature vector.

In short, it is claimed that statistical features for pattern representation are easy to extract [92], [2]. Moreover, such features can be effective in recognition systems and providing high speed and low complexity implementation [60], [110]. However, special attention to the prepossessing techniques should be given, since misrecognition may accrue due to poor prepossessing techniques [5]. Nevertheless, the fundamental issue is to determine a set of statistical features, which need to be the most representative data of a pattern, maximizing the performance accuracy and minimizing the processing time simultaneously.

As a result, researchers call for investigating other statistical features which maximize the performance accuracy and minimize the processing time [2], [64].

#### *3) Global Transformation Feature*

The global transformation method is applied to convert a skeleton or contour of a pattern by a linear transform into a form that reflects the most relevant features of the transformed pattern [64]. Numerous global transformation methods have been used in developing Arabic OCR systems. An example of such methods is the Fourier descriptor which represents the characteristic of a pattern in a frequency domain [111]. The Fourier descriptor has been applied to Arabic script, such as in [8], [27]. Another method is the Hough transform which

detects lines in binary images and then define the parameter of the lines [18]. Other, such as in [112], [113], utilized the Hough transform for extracting features from Arabic script. Also, some other global transformation methods that have been applied for Arabic OCR for feature extraction are the direction codes method such as Freeman's chain code in [28], Wavelets in [114] and Walsh transformation in [107].

Overall, it is claimed that global transformation feature techniques have several advantages over structural and statistical approaches. For example, they are applicable for new fonts and easily implemented. Another advantage is that they are robust to noise and variation. However, they might require the implementation of other features in order to obtain high accuracy performance.

In conclusion, the feature extraction stage plays a critical role in Arabic OCR development in which distinguishing attributes are extracted and it is clear that each Arabic OCR developer needs to apply different feature extraction approaches. Still, good features are required, which assist in distinguishing a character from other characters and maximize the accuracy performance simultaneously. Furthermore, these features must be selected specifically for a selected classifier. Some researchers apply different feature extraction methods in combination. However, this may cause extra complications for the implementation [8].

### *F. Classification Phase*

The classification phase has the responsibility for assigning a pattern into a pre-classified class based on the features of the pattern which have been extracted in the previous phase [18]. The pre-classified classes can be words, sub-words, characters or strokes, based on the OCR approach used [6]. There are a number of different classification approaches that have been applied for Arabic OCR, such as Hidden Markov Models (HMM), Support Vector Machines (SVM), K-nearest neighbour.

SVM, which is a binary classifier, has been used in the implementation of printed Arabic OCR systems [106], [95], [115]. (For a comprehensive review of applying SVM to Arabic OCR, refer to [116]). However, classifiers based on SVM are mostly applied to a small set of data due to the high complexity of training and processing time [117], [118]. Another classification technique that has been applied to printed Arabic OCR are Hidden Markov Model (HMM) based techniques. HMMs are statistical models that are considered as being one of the most efficient for recognition applications especially for speech recognition [17]. Therefore, researchers in OCR have implemented HMMs for OCR in order to obtain high performance OCR systems, such as in [72], [119]–[122].

### *G. Post-processing Phase*

Post-processing is the final stage of the development of Arabic OCR. The objective of this step is to enhance the recognition accuracy by detecting and correcting linguistic misspellings in the produced OCR text without human intervention. Research studies on Arabic OCR have implemented post-progressing methods in order to improve the output, such as [123], [124]. It is worth mentioning that three main elements should be considered in correcting OCR output: non-word errors correction; isolated word errors correction; and context–based word correction [50]. Generally, post-processing methods can be categorized into two main approaches: lexicon-based methods; and context-based (statistical) methods [125].

The typical technique for correcting the mistakes of Arabic OCR outputs is the lexicon-based method which requires the utilization of an Arabic dictionary, such as in [126], [127]. This technique corrects errors without considering any contextual information in which the errors appear. Therefore, a problem might occur with using this approach when a word is misrecognized by an OCR system and is also in the lexicon (these are called real-word errors) such as, *Fear* for *Tear*. This occurs in many languages such as Arabic in which a large fraction of three characters sequences are corrected words. Consequently, only non-word errors can be corrected, since this method is comparing the recognized words with the words that are in the dictionary. Also, this approach requires a wide-ranging lexicon that consists of all single words. However, the Arabic language has various dialects and it is also a triglossic language with three forms – modern standard Arabic and classical Arabic [128] and mixtures of the two. Therefore, this approach is less appropriate for Arabic language since building a single lexicon for Arabic language is more complicated.

On the other hand, context-based (statistical) methods take into account the contextual information in which the misrecognized words appear. A few studies have implemented statistical language models for improving the recognition accuracy of Arabic OCR systems, such as in [129], [130]. Using such methods will help overcome the problem of correcting real-word errors. Moreover, they are also useful in correcting word errors that might have several potential corrections, since these techniques can correct word errors based on grammatical concepts and semantic context [131].

Recently, there have been several attempts to provide systems for correcting Arabic OCR output. For instance, the authors in [123] propose a system for Arabic OCR output correction based on Google online suggestions within Microsoft Office Word. On the other hand, the authors in [131] describe a context-based technique for detecting and correcting Arabic OCR errors. Although there are some studies on applying context-based methods for correcting Arabic OCR output, more research is needed on investigating the use of Arabic contextual information for OCR output correction [132].

## IV. Performance Evaluation

OCR performance evaluation can be classified into two types: black-box evaluation and white-box evaluation. In the former, an entire OCR engine is treated as an indivisible unit, so the submodules of the OCR system are not known to the evaluator, whereas with the white-box evaluation, each submodule of the OCR system is evaluated if the submodules are accessible [133]. Performance evaluation of OCR systems is essential for monitoring progress of OCR systems development, assessing the effectiveness of OCR algorithms, identifying open areas for further research and providing scientific justification for the performance of OCR systems [134], [135]. Although the performance evaluation of OCR

systems is important, there has been very little work focus on empirical evaluation of Arabic OCR systems, such as in [135], [136] and a recent study in [1]. Furthermore, these evaluation studies have conducted a black-box evaluation on Arabic OCR systems as the submodules are not accessible. Thus, only the overall performances of Arabic OCR have been reported.

Performance evaluation in research areas of pattern recognition is facing several obstacles [137]. For Arabic OCR, conducting performance evaluation is challenging as no standard dataset is available [138], [94]. Moreover, most Arabic OCR systems are evaluated in terms of character accuracy which is a general metric, such as performance results reported in [138], [135], [136]. This accuracy metric is insufficient to assess how Arabic OCR systems are overcoming the challenges of Arabic text. However, a study suggests a new set of objective performance metrics for evaluation Arabic OCR with respect to the challenges of Arabic script which are character accuracy based on character position, dot character accuracy, zigzag-shaped character accuracy, loop-shaped character accuracy and diacritics accuracy [139].

## V. Discussion and Future Directions

This paper has overviewed the main stages used in printed Arabic OCR. It main aim is to reveal the current status of printed Arabic OCR. Although there are various attempts to solve the problems of Arabic text recognition, there is still a crucial need for more research.

In an attempt to evaluate the status of printed Arabic OCR and support the claim that more research is needed in many areas, we used Google scholar to search for scientific research publications using phrases that are related to Arabic text recognition. The findings are summarized in Table 2. The table shows the search phrases used and the search results returned by Google Scholar. It is apparent from the table that there is a lack of Arabic OCR research as comparatively very little research has focused on Arabic OCR compared to studies in OCR for other languages. For example, there 322,000 results were returned for the more general search query 'OCR', whereas there were only 956 results returned for the more specific search query 'Arabic OCR'.

In order to provide a measure of the coverage of research in a particular area, we can estimate the probability that a particular research paper will be in a more specific topic area compared to the more general topic area. For example, we may be interested in the general topic area "single font OCR" and wish to see how much research has been published in the more specific topic area "single font Arabic OCR" in comparison. We can estimate the probability $p$ that the more general topic will be concerned with the more specific topic as $p = s / g$ where $s$ is the count of the number of papers found for the specific topic compared to the count $g$ of the number of papers found for the more general topic. Then we can define the 'Information Coverage' $I$ associated with the specific topic in relation to the more general topic as $I = -\log p$. If this value is high compared to other specific vs. general comparisons for the same overall topic (e.g. in relation to Arabic OCR vs. OCR in general), then this reflects that research may under-represented in this area.

This analysis has been done using the values from Table 2 and graphed in Figure 9. In the Figure, we see that except for papers on Arabic OCR concerning easy fonts and diacritics, the remaining topics have higher Information Coverage values meaning that there have been less papers published in these areas proportionately compared to papers published in the more general (non-Arabic) areas. We can use Figure 9 to help gauge the present status of printed Arabic OCR research as it highlights some open areas which need more research. This is based on the number of publications for single, omni and multi font OCR concerning various elements that are related to text recognition concerning easy fonts, complicated fonts, diacritics, page layout, multi-language and noisy documents. In particular, for Arabic text images which contain complicated fonts, there are still many gaps in the research. Furthermore, for single, omni and multi font Arabic OCR on multi-language text images, intensive further research is needed.

Figure 10 plots the number of papers per 5-year period for the top 100 Google Scholar searches using the Arabic OCR related phrases. A number of striking results are apparent in Figure 10. For example, publications for Arabic OCR peak since 2005. Also, the numbers of papers for printed Arabic OCR decrease since 2005 (this could be because researchers have focussed on handwritten Arabic OCR). Furthermore, it is apparent that the smallest numbers of papers in the period of Arabic OCR research are papers related to noisy documents.

Note that the earliest research papers for the 'Arabic OCR' search query are from 1985, which is a result of the top-100 ranking returned by Google Scholar. If, however, we restrict the range of years for which we search, we find that the first papers returned by Google Scholar appear in the 1970 to 1980 period. In contrast, the author in [9] states that text recognition research first originated in 1940, and papers related to the 'text recognition' query appear in Google Scholar from the 1960s.

TABLE II. GOOGLE SCHOLAR SEARCH RESULTS FOR ARABIC OCR RELATED PHRASES

| Google Scholar search phrase | | Number of papers |
|---|---|---|
| 1 | +"Arabic OCR" | 956 |
| | +"OCR" | 322,000 |
| 2 | +"OCR" + "Arabic printed text" | 247 |
| | +"OCR" + "printed text" | 7,190 |
| 3 | +"Arabic OCR" + "diacritics" | 302 |
| | +"OCR" + "diacritics" | 1,800 |
| 4 | +"Arabic OCR" + "page layout" | 54 |
| | +"OCR" + "page layout" | 3,360 |
| 5 | +"Arabic OCR" + "multi-language " | 20 |
| | +"OCR" + "multi-language" | 866 |
| 6 | +"Arabic OCR" + "Omni font" | 60 |
| | +"OCR" + "Omni font" | 380 |
| 7 | +"Arabic OCR" + "single font" | 61 |
| | +"OCR" + "single font" | 717 |
| 8 | +"Arabic OCR" + "multi-font " | 149 |
| | +"OCR" + "multi-font" | 1270 |
| 9 | +"Arabic OCR" + "noisy document" | 12 |
| | +"OCR" + "noisy document" | 515 |
| 10 | +"Arabic OCR" + " Simplified Arabic " + "single font" | 20 |
| | +"Arabic OCR" + " Simplified Arabic " + "Omni font" | 22 |
| | +"Arabic OCR" + " Simplified Arabic " + "multi font" | 48 |
| 11 | +"Arabic OCR" + " Advertising Bold " + "single font" | 5 |
| | +"Arabic OCR" + " Advertising Bold " + "Omni font" | 5 |
| | +"Arabic OCR" + " Advertising Bold " + "multi font" | 15 |
| 12 | +"Arabic OCR" + "diacritics" + "single font" | 30 |
| | +"Arabic OCR" + "diacritics" + "Omni font" | 36 |
| | +"Arabic OCR" + "diacritics" + "multi font" | 62 |
| 13 | +"Arabic OCR" + " page layout " + "single font" | 8 |
| | +"Arabic OCR" + " page layout " + "Omni font" | 9 |
| | +"Arabic OCR" + " page layout " + "multi font" | 11 |
| 14 | +"Arabic OCR" + " multi-language " + "single font" | 17 |
| | +"Arabic OCR" + " multi-language " + "Omni font" | 9 |
| | +"Arabic OCR" + " multi-language " + "multi font" | 29 |
| 15 | +"Arabic OCR" + " noisy document " + "single font" | 1 |
| | +"Arabic OCR" + " noisy document " + "Omni font" | 2 |
| | +"Arabic OCR" + " noisy document " + "multi font" | 3 |

From the review of each stage used in Arabic OCR, the following observations have been noted for possible research directions:

- All the reviewed research of printed Arabic OCR have used the general OCR methodology which involves the five stages; pre-processing; feature extraction; segmentation; classification and post-processing. However, the following are still open questions: 'Is the current OCR methodology the most effective for designing Arabic OCR?' and 'Are there alternative methodologies that might yield better results for Arabic OCR'?
- From the review of each OCR stage, it is apparent that the most challenging task in the development of Arabic OCR is the segmentation task. Although previous studies have presented different segmentation techniques for Arabic OCR, these studies have not provided the accuracy performance of these techniques. Since only the overall OCR performances have been reported, it is difficult to gain an insight into which segmentation techniques perform better for printed Arabic OCR. A performance evaluation tool should be developed to assess the different segmentation techniques.
- The pre-processing stage review given in this study reached the conclusion that direct adoption of pre-processing methods which are designed for general purposes might be not applicable for Arabic script. Thus, developing pre-processing methods that consider the specific characteristics of Arabic script is needed.
- Most of the proposed methods for feature extraction in Arabic OCR have been adopted from methods that have been developed for other languages without considering the characteristics of Arabic script. Such

methods may not be the most appropriate for accurate recognition. The characteristics of Arabic script should be taken into consideration when selecting a feature extraction method that is able to distinguish between Arabic characters.

- The studies on performance evaluation of printed Arabic OCR have used a black-box evaluation method which can only provide the overall performance of OCR systems. For more insight into which OCR stage is causing the most problems, a white-box evaluation, where each component of the system is accessible, is required.
- Performance evaluation of printed Arabic OCR suffers from the lack of public datasets. For objective performance evaluation, an accurate and free printed Arabic dataset is essential.
- Our investigation indicates the need for more research on Arabic OCR output correction with the use of Arabic contextual information.

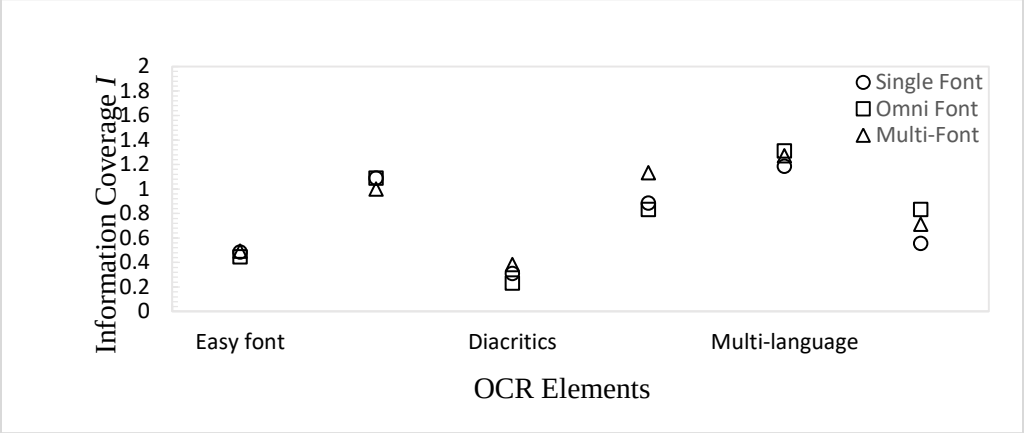



Fig. 9. The present status of printed Arabic OCR based on the number of publications for different OCR elements.

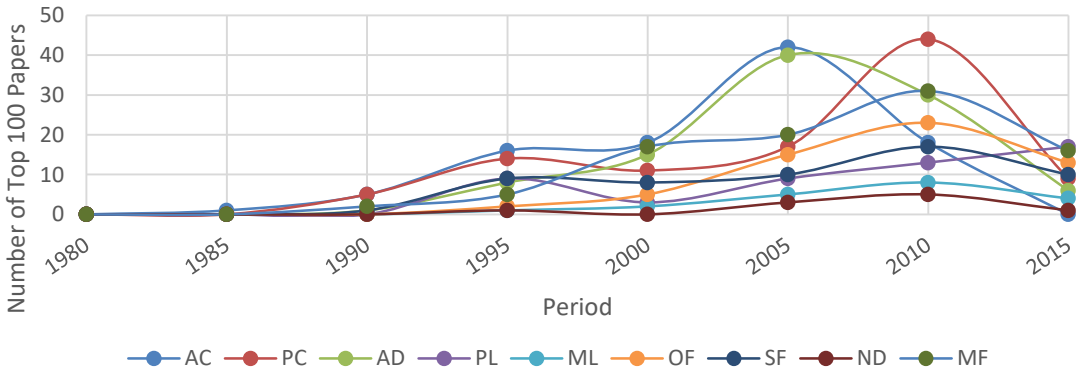



Fig. 10. Number of papers per 5-year period in the top 100 results returned by Google Scholar for different Arabic OCR related search phrases. AC = 'Arabic OCR'; PC = 'Arabic printed text'; AD = 'Arabic diacritics'; PL = 'Arabic OCR + page layout'; ML= 'Arabic OCR + multi-language'; OF = 'Arabic OCR + omni font'; SF = 'Arabic OCR + single font'; MF = 'Arabic OCR + multi font'; ND = 'Arabic OCR + noisy document'.

## VI. Conclusion

This paper has provided a comprehensive literature review of printed Arabic text recognition. At first, the specific characteristics of Arabic script that challenge the recognition process have been discussed. Then, the general methodology of printed Arabic OCR has been presented. This methodology was divided into five stages: preprocessing; segmentation; feature extraction; classification; and post-processing. Techniques applied at each stage of Arabic OCR have been discussed. Also, the issues related to the performance evaluation have been reviewed. Finally, we analyzed the remaining problems in the field of printed Arabic OCR and provide several direction for future research.

## References

[1] M. Alghamdi and W. Teahan, "Experimental evaluation of Arabic OCR systems," PSU Res. Rev., vol. 1, no. 3, pp. 229–241, 2017.

[2] B. Al-Badr and S. A. Mahmoud, "Survey and bibliography of Arabic optical text recognition," Signal Processing, vol. 41, no. 1, pp. 49–77, 1995.

[3] Y. M. Alginahi, "A survey on Arabic character segmentation," International Journal on Document Analysis and Recognition, vol. 16, no. 2. pp. 105–126, 2013.

[4] M. Saad and W. Ashour, "OSAC: Open Source Arabic Corpora," 6th Int. Conf. Electr. Comput. Syst. (EECS'10), Nov 25-26, 2010, Lefke, Cyprus., pp. 118–123, 2010.

[5] M. S. Khorsheed, "Off-line Arabic character recognition - A review," Pattern Analysis and Applications, vol. 5, no. 1. pp. 31–45, 2002.

[6] B. M. Al-Helali and S. A. Mahmoud, "Arabic Online Handwriting Recognition (AOHR): A Survey," ACM Comput. Surv., vol. 50, no. 3, p. 33, 2017.

[7] M. S. M. El-Mahallawy, "A Large Scale HMM-Based Omni Font-Written OCR System for Cursive Scripts," Faculty of Engineering, Cairo University Giza, Egypt, 2008.

[8] M. Khorsheed and W. Clocksin, "Multi-font Arabic word recognition using spectral features," Pattern Recognition, 2000. …, no. 2, pp. 543–546, 2000.

[9] B. H. Al-Badr, "A Segmentation-free Approach to Text Recognition with Application to Arabic Text," University of Washington, Seattle, WA, USA, 1995.

[10] R. J. Kannan and S. Subramanian, "An adaptive approach of Tamil character recognition using deep learning with big data-A survey," Adv. Intell. Syst. Comput., vol. 337, pp. 557–567, 2015.

[11] A. Lawgali, "A Survey on Arabic Character Recognition," vol. 8, no. 2, pp. 401–426, 2015.

[12] B. Al-Badr and R. M. Haralick, "Segmentation-free word recognition with application to Arabic," Proc. 3rd Int. Conf. Doc. Anal. Recognit., vol. 1, pp. 355–359, 1995.

[13] K. Jumari and M. Ali, "A survey and comparative evaluation of selected off-line arabic handwritten character recognition systems," J. Teknol., vol. 36, pp. 1–17, 2012.

[14] M. Sarfraz, S. N. Nawaz, and A. Al-Khuraidly, "Offline Arabic text recognition system," 2003 International Conference on Geometric Modeling and Graphics 2003 Proceedings. pp. 30–35, 2003.

[15] T. Pavlidis, "Recognition of printed text under realistic conditions," Pattern Recognit. Lett., vol. 14, no. 4, pp. 317–326, 1993.

[16] B. Al-Badr and R. M. Haralick, "A segmentation-free approach to text recognition with application to Arabic text," Int. J. Doc. Anal. Recognit., vol. 1, no. 3, pp. 147–166, 1998.

[17] A. M. AbdelRaouf, "Offline printed Arabic character recognition," 2012.

[18] M. Cheriet, N. Kharma, C. Liu, and C. Suen, Character recognition systems: a guide for students and practitioners. 2007.

[19] W. Cho, S. W. Lee, and J. H. Kim, "Modeling and recognition of cursive words with hidden Markov models," Pattern Recognit., vol. 28, no. 12, pp. 1941–1953, 1995.

[20] F. Drira and F. Lebourgeois, "Denoising textual images using local/non-local smoothing filters : A comparative study," Proc. - Int. Work. Front. Handwrit. Recognition, IWFHR, pp. 521–526, 2012.

[21] Z. Shi, S. Setlur, and V. Govindaraju, "Guide to OCR for Arabic Scripts," V. Märgner and H. El Abed, Eds. London: Springer London, 2012, pp. 79–102.

[22] A. Buades, B. Coll, and J. Morel, "A Review of Image Denoising Algorithms, with a New One," Multiscale Model. Simul., vol. 4, no. 2, pp. 490–530, 2005.

[23] R. Gonzalez and R. Woods, Digital image processing. 2002.

[24] H. Al-ani, N. Ban, and H. M. Abass, "Journal of Computing::Printed Arabic Character Recognition using Neural Network," vol. 5, no. 1, pp. 64–66, 2014.

[25] A. M. Al-Shatnawi, F. H. Al-Zawaideh, S. Al-Salameh, and K. Omar, "Offline Arabic Text Recognition -- An Overview," World Comput. Sci. Inf. Technol., vol. 1, no. 5, pp. 184–192, 2011.

[26] I. Ahmed, S. A. Mahmoud, and M. T. Parvez, "Guide to OCR for Arabic Scripts," V. Märgner and H. El Abed, Eds. London: Springer London, 2012, pp. 147–168.

[27] S. A. Mahmoud, "Arabic character recognition using fourier descriptors and character contour encoding," Pattern Recognit., vol. 27, no. 6, pp. 815–824, 1994.

[28] S. Taha, Y. Babiker, and M. Abbas, "Optical character recognition of arabic printed text," SCOReD 2012 - 2012 IEEE Student Conf. Res. Dev., pp. 235–240, 2012.

[29] A. M. Al-Shatnawi and K. Omar, "Skew Detection and Correction Technique for Arabic Document Images Based on Centre of Gravity Atallah Mahmoud Al-Shatnawi and Khairuddin Omar Department of System Science and Management , Faculty of Information Science and Technology," J. Comput. Sci., vol. 5, no. 5, pp. 363–368, 2009.

[30] I. S. I. Abuhaiba, "Skew Correction of Textural Documents," J. King Saud Univ. - Comput. Inf. Sci., vol. 15, pp. 73–93, 2003.

[31] C. Sun and D. Si, "Skew and slant correction for document images using gradient direction," in Proceedings of the Fourth International Conference on Document Analysis and Recognition, 1997, vol. 1, pp. 142–146.

[32] J. R. Parker, "Algorithms for Image Processing and Computer Vision," Vasa, p. 504, 2011.

[33] H. S. Baird, "The Skew Angle of Printed Documents," in Document Image Analysis, 1995, pp. 204–208.

[34] J. Hull, "Document image skew detection: Survey and annotated bibliography," Ser. Mach. Percept. Artif. …, pp. 40–64, 1998.

[35] J. X. Dong, P. Dominique, A. Krzyzak, and C. Y. Suen, "Cursive word skew/slant corrections based on Radon transform," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2005, pp. 478–482, 2005.

[36] A. Al-Shatnawi and K. Omar, "Methods of Arabic Language Baseline Detection – The State of Art," J. Comput. Sci., vol. 8, no. 10, pp. 137–143, 2008.

[37] Y. Alginahi, "Preprocessing Techniques in Character Recognition," pp. 1–20, 2010.

[38] F. Shafait, Adnan-ul-Hasan, D. Keysers, and T. M. Breuel, "Layout analysis of urdu document images," in 10th IEEE International Multitopic Conference 2006, INMIC, 2006, pp. 293–298.

[39] K. Bin Omar, R. Bin Mahmoud, M. N. Bin Sulaiman, and a. R. Bin Ramli, "The removal of secondaries of Jawi characters," 2000 TENCON Proceedings. Intell. Syst. Technol. New Millenn. (Cat. No.00CH37119), vol. 2, pp. 149–152, 2000.

[40] A. Amin, "Off-line Arabic character recognition," Pattern Recognit., vol. 31, no. 5, pp. 517–530, 1998.

[41] N. Arica and F. T. Yarman-Vural, "Optical character recognition for cursive handwriting," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 6, pp. 801–813, 2002.

[42] R. El-Hajj, L. Likforman-Sulem, and C. Mokbel, "Arabic handwriting recognition using baseline dependant features and hidden Markov modeling," in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2005, vol. 2005, pp. 893–897.

[43] M. Pechwitz and V. Margner, "Baseline estimation for Arabic handwritten words," in Proceedings - International Workshop on Frontiers in Handwriting Recognition, IWFHR, 2002, pp. 479–484.

[44] A. Zidouri, M. Sarfraz, S. N. Nawaz, and M. J. Ahmad, "PC based offline Arabic text recognition system," Seventh Int. Symp. Signal Process. Its Appl. 2003. Proceedings., vol. 2, 2003.

[45] H. Al-rashaideh, "Preprocessing phase for Arabic Word Handwritten Recognition," Inf. Transm. Comput. Networks J., vol. 6, no. 1, pp. 11–19, 2006.

[46] A. M. Zeki, "The segmentation problem in arabic character recognition the state of the art," in Proceedings of 1st International Conference on Information and Communication Technology, ICICT 2005, 2005, vol. 2005, pp. 11–26.

[47] S. S. Bukhari, F. Shafait, and T. M. Breuel, "High performance layout analysis of Arabic and Urdu document images," in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2011, pp. 1275–1279.

[48] G. Nagy, S. Seth, and M. Viswanathan, "A prototype document image analysis system for technical journals," Computer (Long. Beach. Calif)., vol. 25, no. 7, pp. 10–22, 1992.

[49] A. M. Al-shatnawi and K. Omar, "The Thinning Problem in Arabic Text Recognition - A Comprehensive Review," Int. J. Comput. Appl., vol. 103, no. 3, pp. 35–42, 2014.

[50] M. T. Parvez and S. a. Mahmoud, "Offline arabic handwritten text recognition: A Survey," ACM Comput. Surv., vol. 45, no. 2, p. 23:1–23:35, 2013.

[51] Z. Guo and R. W. Hall, "Fast fully parallel thinning algorithms," CVGIP Image Underst., vol. 55, no. 3, pp. 317–328, 1992.

[52] C. Hilditch, "Comparison of thinning algorithms on a parallel processor," Image Vis. Comput., vol. 1, no. 3, pp. 115–132, 1983.

[53] P. S. P. Wang and Y. Y. Zhang, "A Fast and Flexible Thinning Algorithm," IEEE Trans. Comput., vol. 38, no. 5, pp. 741–745, 1989.

[54] S. A. Mahmoud, I. AbuHaiba, and R. J. Green, "Skeletonization of Arabic characters using clustering based skeletonization algorithm (CBSA)," Pattern Recognit., vol. 24, no. 5, pp. 453–464, 1991.

[55] M. Melhi, S. S. Ipson, and W. Booth, "A novel triangulation procedure for thinning hand-written text," Pattern Recognit. Lett., vol. 22, no. 10, pp. 1059–1071, 2001.

[56] J. Cowell and F. Hussain, "Thinning Arabic characters for feature extraction," in Proceedings of the International Conference on Information Visualisation, 2001, vol. 2001–Janua, pp. 181–185.

[57] M. Altuwaijri and M. Bayoumi, "A new thinning algorithm for Arabic characters using self-organizing neural network," in Circuits and Systems, 1995. ISCAS'95., 1995 IEEE International Symposium on, 1995, vol. 3, pp. 1824–1827.

[58] M. A. Alghamdi and W. J. Teahan, "A New Thinning Algorithm for Arabic Script," Int. J. Comput. Sci. Inf. Secur. (IJCSIS), vol. 15, no. 1, pp. 204–211, 2017.

[59] L. M. Lorigo and V. Govindaraju, "Offline arabic handwriting recognition: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 5. pp. 712–724, 2006.

[60] S. Naz, K. Hayat, M. Imran Razzak, M. Waqas Anwar, S. A. Madani, and S. U. Khan, "The optical character recognition of Urdu-like cursive scripts," in Pattern Recognition, 2014, vol. 47, no. 3, pp. 1229–1248.

[61] R. Plamondon and S. N. Srihari, "On-line and off-line handwriting recognition: A comprehensive survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, pp. 63–84, 2000.

[62] M. Amara, K. Zidi, K. Ghedira, and S. Zidi, "New Rules to Enhance the Performances of Histogram Projection for Segmenting Small-Sized Arabic Words," in Hybrid Intelligent Systems, Springer, 2016, pp. 167–176.

[63] S. N. Nawaz, M. Sarfraz, A. Zidouri, and W. G. Al-Khatib, "An approach to offline Arabic character recognition using neural networks," in Electronics, Circuits and Systems, 2003. ICECS 2003. Proceedings of the 2003 10th IEEE International Conference on, 2003, vol. 3, p. 1328–1331 Vol.3.

[64] N. Arica and F. T. Yarman-Vural, "An overview of character recognition focused on off-line handwriting," IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, vol. 31, no. 2. pp. 216–233, 2001.

[65] J. Alkhateeb, "Word Based Off-line Handwritten Arabic Classification and Recognition," Ph. D. thesis, School of Computing, Informatics and Media, University of Bradford, 2010.

[66] L. O'Gorman, "The Document Spectrum for Page Layout Analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, no. 11, pp. 1162–1173, 1993.

[67] S. N. Srihari and G. Ball, "An assessment of Arabic handwriting recognition technology," in Guide to OCR for Arabic Scripts, Springer, 2012, pp. 3–34.

[68] A. Amin and G. Masini, "Machine Recognition of Multi Font Printed {Arabic} Texts," in Proceedings, Eighth International Conference on Pattern Recognition (Paris, France, October 27--31, 1986), 1986, pp. 392–395.

[69] J. F. Mari, "Machine Recognition and Correction of Printed Arabic Text," IEEE Trans. Syst. Man Cybern., vol. 19, no. 5, pp. 1300–1306, 1989.

[70] A. Ymin and Y. Aoki, "On the segmentation of multi-font printed Uygur scripts," in Pattern Recognition, 1996., Proceedings of the 13th International Conference on, 1996, vol. 3, pp. 215–219.

[71] A. Cheung, M. Bennamoun, and N. W. Bergmann, "Arabic optical character recognition system using recognition-based segmentation," Pattern Recognit., vol. 34, no. 2, pp. 215–233, 2001.

[72] M. S. Khorsheed, "Offline recognition of omnifont Arabic text using the HMM ToolKit (HTK)," Pattern Recognit. Lett., vol. 28, no. 12, pp. 1563–1571, 2007.

[73] A. Amin, "Off-line Arabic character recognition: the state of the art," Pattern Recognit., vol. 31, no. 5, pp. 517–530, 1998.

[74] H. Al-Yousefi and S. S. Udpa, "Recognition of arabic characters," IEEE Trans. Pattern Anal. Mach. Intell., vol. 14, no. 8, pp. 853–857, 1992.

[75] I. S. I. Abuhaiba, S. A. Mahmoud, and R. J. Green, "Recognition of handwritten cursive Arabic characters," IEEE Trans. Pattern Anal. Mach. Intell., vol. 16, no. 6, pp. 664–672, 1994.

[76] I. Supriana and A. Nasution, "Arabic Character Recognition System Development," Procedia Technol., vol. 11, no. Iceei, pp. 334–341, 2013.

[77] J. H. AlKhateeb, J. Jiang, J. Ren, and S. Ipson, "Component-based Segmentation of words from handwritten Arabic text," Int. J. Comput. Syst. Sci. Eng., vol. 5, no. 1, pp. 54–58, 2009.

[78] P. Xiu, L. Peng, X. Ding, and H. Wang, "Offline handwritten Arabic character segmentation with probabilistic model," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2006, vol. 3872 LNCS, pp. 402–412.

[79] E. J. Erlandson, J. M. Trenkle, and R. C. Vogt, "Word-level recognition of multifont Arabic text using a feature vector matching approach," in Document Recognition III, 1996, vol. 2660, pp. 63–71.

[80] A. Choudhary, "A review of various character segmentation techniques for cursive handwritten words recognition," Int. J. Inf. Comput. Technol, vol. 4, no. 6, pp. 559–564, 2014.

[81] S. Naz, A. I. Umar, S. H. Shirazi, S. B. Ahmed, M. I. Razzak, and I. Siddiqi, "Segmentation techniques for recognition of Arabic-like scripts: A comprehensive survey," Educ. Inf. Technol., pp. 1–17, 2015.

[82] B. A. Najoua and E. Noureddine, "A robust approach for Arabic printed character segmentation," in Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on, 1995, vol. 2, pp. 865–868.

[83] B. Parhami and M. Taraghi, "Automatic recognition of printed Farsi texts," Pattern Recognit., vol. 14, no. 1, pp. 395–403, 1981.

[84] L. Zheng, A. H. Hassin, and X. Tang, "A new algorithm for machine printed Arabic character segmentation," Pattern Recognit. Lett., vol. 25, no. 15, pp. 1723–1729, 2004.

[85] L. Lorigo and V. Govindaraju, "Segmentation and pre-recognition of arabic handwriting," in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 2005, vol. 2005, pp. 605–609.

[86] D. Koteswara Rao and A. Negi, "Advances in Signal Processing and Intelligent Recognition Systems: Proceedings of Second International Symposium on Signal Processing and Intelligent Recognition Systems (SIRS-2015) December 16-19, 2015, Trivandrum, India," M. S. Thampi, S. Bandyopadhyay, S. Krishnan, K.-C. Li, S. Mosin, and M. Ma, Eds. Cham: Springer International Publishing, 2016, pp. 633–644.

[87] A. M. Zeki, M. S. Zakaria, and C.-Y. Liong, "Segmentation of Arabic Characters: A Comprehensive Survey," Int. J. Technol. Diffus., vol. 2, no. 4, pp. 48–82, 2011.

[88] A. Cheung, M. Bennamoun, and N. W. Bergmann, "An Arabic optical character recognition system using recognition-based segmentation," Pattern Recognit., vol. 34, no. 2, pp. 215–233, 2001.

[89] C. H. Chen and J. L. DeCurtins, "Word recognition in a segmentation-free approach to OCR," Proc. 2nd Int. Conf. Doc. Anal. Recognit. ICDAR 93, pp. 573–576, 1993.

[90] B. A. Srinivas, A. Agarwal, and C. R. Rao, "An Overview of OCR Research in Indian Scripts," vol. 2, no. 2, 2008.

[91] Ø. Due Trier, A. K. Jain, and T. Taxt, "Feature extraction methods for character recognition-A survey," Pattern Recognit., vol. 29, no. 4, pp. 641–662, 1996.

[92] E. Moubtahij, A. H. Hicham, and K. SATORI, "Review of Feature Extraction Techniques for Offline Handwriting Arabic Text Recognition," Int. J. Adv. Eng. Technol., vol. 7, no. 1, pp. 50–58, 2014.

[93] D. V. Sharma, G. Saini, and M. Joshi, "Statistical Feature Extraction Methods for Isolated Handwritten Gurumukhi Script," vol. 2, no. 4, pp. 380–384, 2012.

[94] H. Al-Muhtaseb and R. Qahwaji, "Arabic Optical Character Recognition: Recent Trends and Future Directions," in Applied Signal and Image Processing: Multidisciplinary Advancements, IGI Global, 2011, pp. 324–346.

[95] M. Abd, S. Al Rubeaai, and G. Paschos, "Hybrid Features for an Arabic Word Recognition System," Davidpublishing.Com, vol. 3, no. 1, pp. 685–691, 2012.

[96] R. Saabni, "Efficient recognition of machine printed Arabic text using partial segmentation and Hausdorff distance," 6th Int. Conf. Soft Comput. Pattern Recognition, SoCPaR 2014, pp. 284–289, 2015.

[97] a. M. Elgammal and M. a. Ismail, "A graph-based segmentation and feature extraction framework for\nArabic text recognition," Proc. Sixth Int. Conf. Doc. Anal. Recognit., 2001.

[98] M. S. Khorsheed and W. F. Clocksin, "Structural Features of Cursive Arabic Script.," in BMVC, 1999, pp. 1–10.

[99] I. Ahmad, L. Rothacker, G. A. Fink, and S. A. Mahmoud, "Novel sub-character HMM models for arabic text recognition," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, pp. 658–662, 2013.

[100] D. Ghosh, T. Dube, and A. Shivaprasad, "Script Recognition-a review," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 12, pp. 2142–2161, 2010.

[101] H. Almohri, J. S. Gray, and H. Alnajjar, A real-time DSP-based optical character recognition system for isolated arabic characters using the TI TMS320C6416T. University of Hartford, 2007.

[102] M. Z. Hossain, "Rapid Feature Extraction for Optical Character Recognition," pp. 1–5, 2012.

[103] G. Kumar and P. K. Bhatia, "A detailed review of feature extraction in image processing systems," in International Conference on Advanced Computing and Communication Technologies, ACCT, 2014, pp. 5–12.

[104] H. Aboaisha, Z. Xu, and I. El-Feghi, "An investigation on efficient feature extraction approaches for Arabic letter recognition," 2012.

[105] M. Kef, L. Chergui, and S. Chikhi, "Comparative study of the use of geometrical moments for Arabic handwriting recognition," in Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012 6th International Conference on, 2012, pp. 303–308.

[106] M. A. Abd and G. Paschos, "Effective arabic character recognition using support vector machines," Innov. Adv. Tech. Comput. Inf. Sci. Eng., pp. 7–11, 2007.

[107] M. Oujaoura, R. El Ayachi, M. Fakir, B. Bouikhalene, and B. Minaoui, "Zernike moments and neural networks for recognition of isolated Arabic characters," Int. J. Comput. Eng. Sci., vol. 2, pp. 17–25, 2012.

[108] Z. Shaaban, "A new recognition scheme for machine-printed Arabic texts based on neural networks," in Proceedings of World Academy of Science, Engineering and Technology, 2008, vol. 31, pp. 25–27.

[109] I. A. Elrube, M. T. El Sonni, and S. S. Saleh, "Printed Arabic sub-word recognition using moments," World Acad. Sci. Eng. Technol., vol. 66, pp. 737–741, 2010.

[110] P. B. Pati and A. G. Ramakrishnan, "OCR in Indian scripts: A survey," IETE Tech. Rev., vol. 22, no. 3, pp. 217–227, 2005.

[111] D. Zhang and G. Lu, "A Comparative Study on Shape Retrieval Using Fourier Descriptors with Different Shape Signatures," Int. Conf. Intell. Multimed. Distance Educ., vol. 1, pp. 1–9, 2001.

[112] N. Ben Amor and N. E. Ben Amara, "Multifont Arabic Characters Recognition Using HoughTransform and HMM/ANN Classification," J. Multimed., vol. 1, no. 2, pp. 50–54, 2006.

[113] S. Touj, N. E. Ben Amara, and H. Amiri, "Generalized Hough Transform for Arabic Printed Optical Character Recognition.," Int. Arab J. Inf. Technol., vol. 2, no. 4, pp. 326–333, 2005.

[114] N. Ben Amor and N. E. Ben Amara, A Novel Method for Multifont Arabic Characters Features Extraction. INTECH Open Access Publisher, 2012.

[115] R. Mehran, H. Pirsiavash, and F. Razzazi, "A front-end OCR for omni-font Persian/Arabic cursive printed documents," in Digital Image Computing: Techniques and Applications, 2005. DICTA'05. Proceedings 2005, 2005, p. 56.

[116] M. Amara, K. Zidi, S. Zidi, and K. Ghedira, "Arabic Character Recognition Based M-SVM," in International Conference on Advanced Machine Learning Technologies and Applications, 2014, pp. 18–25.

[117] M. Shafii, "Optical Character Recognition of Printed Persian/Arabic Documents," 2014.

[118] S. Arora, D. Bhattacharjee, M. Nasipuri, L. Malik, M. Kundu, and D. K. Basu, "Performance Comparison of SVM and ANN for Handwritten Devnagari Character Recognition," Int. J. Comput. Sci. Issues, vol. 7, no. 3, pp. 1–10, 2010.

[119] M. S. Khorsheed, "Recognizing Cursive Typewritten Text Using Segmentation-Free System," vol. 2015, 2015.

[120] S. M. Awaida and M. S. Khorsheed, "Developing discrete density Hidden Markov Models for Arabic printed text recognition," Proceeding - 2012 IEEE Int. Conf. Comput. Intell. Cybern. Cybern. 2012, pp. 35–39, 2012.

[121] F. Slimane, S. Kanoun, J. Hennebert, A. M. Alimi, and R. Ingold, "A study on font-family and font-size recognition applied to Arabic word images at ultra-low resolution," Pattern Recognit. Lett., vol. 34, no. 2, pp. 209–218, 2013.

[122] H. A. Al-Muhtaseb, S. A. Mahmoud, and R. S. Qahwaji, "Recognition of off-line printed Arabic text using Hidden Markov Models," Signal Processing, vol. 88, no. 12, pp. 2902–2912, 2008.

[123] I. A. Doush and A. M. Al Trad, "Improving post-processing optical character recognition documents with Arabic language using spelling error detection and correction," Int. J. Reason. Intell. Syst., vol. 8, no. 3/4, p. 91, 2016.

[124] W. Magdy and K. Darwish, "Effect of OCR error correction on Arabic retrieval," Inf. Retr. Boston., vol. 11, no. 5, pp. 405–425, 2008.

[125] K. Taghva and E. Stofsky, "OCRSpell: An interactive spelling correction system for OCR errors in text," Int. J. Doc. Anal. Recognit., vol. 3, no. 3, pp. 125–137, 2001.

[126] A. H. Hassin, X.-L. Tang, J.-F. Liu, and W. Zhao, "Printed Arabic character recognition using HMM," J. Comput. Sci. Technol., vol. 19, no. 4, pp. 538–543, 2004.

[127] I. Aljarrah, O. Al-Khaleel, K. Mhaidat, M. Alrefai, A. Alzu'Bi, and M. Rabab'Ah, "Automated system for arabic optical character recognition with lookup dictionary," J. Emerg. Technol. Web Intell., vol. 4, no. 4, pp. 362–370, 2012.

[128] P. Damien, N. Wakim, and M. Eg??a, "Phoneme-viseme mapping for modern, classical arabic language," in 2009 International Conference on Advances in Computational Tools for Engineering Applications, ACTEA 2009, 2009, pp. 547–552.

[129] R. Prasad, S. Saleem, M. Kamali, R. Meermeier, and P. Natarajan, "Improvements in hidden Markov model based Arabic OCR," Proc. 19th Int. Conf. Pattern Recognit., pp. 1–4, 2008.

[130] P. Natarajan, S. Saleem, R. Prasad, E. MacRostie, and K. Subramanian, "Multi-lingual Offline Handwriting Recognition Using Hidden Markov Models: A Script-Independent Approach," in Arabic and Chinese Handwriting Recognition, 2008, pp. 231–250.

[131] Y. Bassil and M. Alwani, "OCR Post-Processing Error Correction Algorithm Using Google ' s Online Spelling Suggestion," J. Emerg. Trends Comput. Inf. Sci., vol. 3, no. 1, pp. 90–99, 2012.

[132] A. G. Krayem, "A high level approach to Arabic sentence recognition," Nottingham Trent University, 2013.

[133] S. V Rice, "Measuring the accuracy of page-reading systems," University of Nevada, Las Vegas, 1996.

[134] S. Mihov, K. U. Schulz, C. Ringlstetter, V. Dojchinova, V. Nakova, K. Kalpakchieva, O. Gerasimov, A. Gotscharek, and C. Gercke, "A corpus for comparative evaluation of OCR software and postcorrection techniques," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol. 2005, pp. 162–166, 2005.

[135] T. Kanungo, G. A. Marton, and O. Bulbul, "OmniPage vs. Sakhr: Paired model evaluation of two Arabic OCR products," in Electronic Imaging'99, 1999, pp. 109–120.

[136] T. Kanungo, G. a Marton, and O. Bulbul, "Performance evaluation of two Arabic OCR products," Proc. SPIE, pp. 76–83, 1999.

[137] K. Jung, K. I. Kim, and A. K. Jain, "Text information extraction in images and video: A survey," Pattern Recognit., vol. 37, no. 5, pp. 977–997, 2004.

[138] I. Ahmad, S. A. Mahmoud, and G. A. Fink, "Open-vocabulary recognition of machine-printed Arabic text using hidden Markov models," Pattern Recognit., vol. 51, pp. 97–111, 2016.

[139] M. A. Alghamdi, I. S. Alkhazi, and W. J. Teahan, "Arabic OCR evaluation tool," in 2016 7th International Conference on Computer Science and Information Technology (CSIT), 2016, pp. 1–6.

# Intelligent Irrigation Management System

Wafa Difallah[1]
Faculty of Technology,Laboratory of Energetic in arid zones, Tahri Mohammed University of Bechar
Bechar, Algeria

Fateh Bounnama[3], Belkacem Draoui[4]
Faculty of Technology,Laboratory of Energetic in arid zones, Tahri Mohammed University of Bechar
Bechar, Algeria

Khelifa Benahmed [2]
Faculty of Exact Sciences, Department of mathematics and computer science, Tahri Mohammed University of Bechar
Bechar, Algeria

Ahmed Saaidi [5]
Faculty of Exact Sciences, Department of mathematics and computer science, Tahri Mohammed University of Bechar
Bechar, Algeria

***Abstract*—It is widely known that water resources are decreasing around the world. Rapid urbanization, population growth, industries and the expansion of agriculture are increasing demand for freshwater. In most countries, including Algeria, irrigation is the largest consumer of water, with about 70% of all freshwater withdrawals being used for irrigation. Therefore, it can be said that solving the problem of water scarcity is based on the adjustment of irrigation. The aim of this paper is to shed light on the irrigation systems, how they can be applied, and what are their benefits. With the adoption of solar energy to feed the system; this energy source is strongly available in arid zones.**

***Keywords*—Internet of things; irrigation; soil; solar; water; wireless sensor network; intelligent**

## I. Introduction

Irrigation is strongly linked to food security. This sector is largely affected by water crisis. Algeria, like all African countries, is largely affected by the scarcity of water. In order to emerge from this crisis, further efforts should be made to develop the irrigation process. The integration of new technological developments such as wireless sensor networks and the Internet of Things are among the most important solutions to this crisis. Besides, solar power can be used to resolve energy constraints problem of WSN implementation.

In this paper an intelligent irrigation management system will be proposed.

The rest of the paper is organized as follows. Section 2 presents a literature review of smart irrigation system implementation. Then we will deal with irrigation strategies in section 3, followed by smart irrigation technologies in section 4. Intelligent irrigation management system architecture will be discussed in section 5; next we have results and discussion in section 6 before conclusion in section 7.

## II. Related Works

Following the critical and urgent nature of water resource problem, several studies have been proposed.

In [1] authors have found an important theory, which is "the optimal irrigation time exists when the rate of water loss is equal to the uniform inflow rate equivalent to the maximum infiltration requirement".

Sutar, S. et al. [2] developed an environment for drip irrigated agriculture and fertilizer tank control using WSN technologies and Ant Colony Algorithm to make the aggregation energy efficient.

Li-Ming, Z. et al. [3] proposed an auto water-saving irrigation based on wireless sensors networks and multi-source information fusion technology based on entropy.

A Microcontroller based controlled remote irrigation system for the agricultural plantation was developed by Kumbhar, S.R. and Ghatule, A.P.. [4]

To prove the importance of using WSN in the irrigation field, a comparison between automated drip irrigation system and non-automated drip irrigation system was done by Soorya, E. et al.. [5]

For the same purpose, Gao, L. et al. [6] combined wireless sensor network with fuzzy control system in an intelligent water-saving irrigation system to realize a remote on-line monitoring and controlling.

Mohammad, F.S. et al. [7] implemented an intelligent irrigation system under a drip irrigation system for the irrigation of tomato crops.

Tiny OS system was used in [8] for simple and cost effective smart irrigation system.

Another design of a low-cost remote monitoring system for agricultural ecosystems was implemented and validated by Kumar, J.P. et al. [9]

In [10], Shahidian, S. et al. validated a smart irrigation controller in a lettuce crop grown in a greenhouse using EC-5 Echo sensors.

In [11], a smart phone application was developed to quickly view soil moisture deficit, weather measurements, and the ability to input applied irrigation amounts into an online evapotranspiration-based irrigation scheduling tool.

To minimize power consumption and resolve the problem of energy constraint of wireless sensor network devices, solar

energy was used to feed irrigation management system in [12] and [13]. This energy resource was also used in [14] to power a soil moisture detector.

## III. Irrigation Methods

Irrigation takes many forms [15]. According to the way of transporting water and its application to the soil, we have: surface irrigation, drip irrigation and sprinkler irrigation.

There are many factors which must be taken into account in the choice of an irrigation method [16]. Table 1 below, represents a comparison between different irrigation strategies.

TABLE I. Comparison of Different Irrigation Strategies

| | **Surface irrigation** | **Drip irrigation** | **Sprinkler irrigation** |
|---|---|---|---|
| **Recommended crops** | The rice fields, the pastures (alfalfa, clover), arboriculture (citrus, banana), cereals. | line crops (vegetables, fruits), arboriculture, vine | line crops, field crops and tree crops |
| **Appropriate soil type** | Silty and clay soils | Clay and sandy soils | perfectly adapts to most types of soil |
| **Equipment** | Pump, Irrigation networks, sprinklers, Water cannons, Self-propelled sprinklers, drippers | The pumping unit control unit Main and secondary pipelines Ramps Drippers or distributors. | the pumping unit main and approach pipelines the ramps the sprinklers. |
| **Workforce** | Large workforce | Workforce economy | Workforce economy |
| **Cost** | low cost of energy and investment | low cost of energy and investment | High cost of energy and investment |

## IV. Smart Irrigation Technologies

Traditional irrigation strategies are not satisfactory for dealing with the shortage of irrigation water, this sector must benefit from modern technological advances. In this section the two main technologies that have rationalized the field of irrigation will be presented: wireless sensor networks and the Internet of Things.

### A. Wireless Sensor Networks (WSN)

Several technologies can be integrated in the field of irrigation such as sensor networks that constitute a set of sensor nodes dispersed in a geographical area to monitor a phenomenon. Each sensor node is composed of a: processing unit, detection unit, energy unit and transmission unit. In the field of irrigation several varieties of sensors can be used:

- Soil moisture (10SH, EC-5, …)
- Humidity (SHT75, HS220, …)
- Air temperature (TDR-3A, …)
- Soil temperature (200ss , …)
- Light intensity (On9558, …)
- Carbon dioxide (GMW120, …)
- Precipitation (Rain Bird RSD-CEx, …)
- Wind speed and direction (CM-100 compact, …)
- ... etc.

### B. Internet of Things (IoT)

The Internet of Things is also a technology that has rationalized several areas and not just the irrigation field. This diversity of application areas involves the diversity of objects connected to this network (sensor, drone, smartphone, etc.).

The number of these objects may exceed 50 billion in 2020. With the advent of the Internet of Things - the word that was first used by Kevin Ashton in 1990- the internet is acquiring a third dimension: in addition to the ability to connect anytime and anywhere, it is now possible to connect with any smart object (see Fig. 1).

There are different types of devices connected to the Internet of things network, or which connect other objects to the Internet such as: identification objects (bar codes, RFID markers,), sensors, drones, UAV (Unmanned Aerial Vehicle), smartphones and tablets (see Fig. 2.).

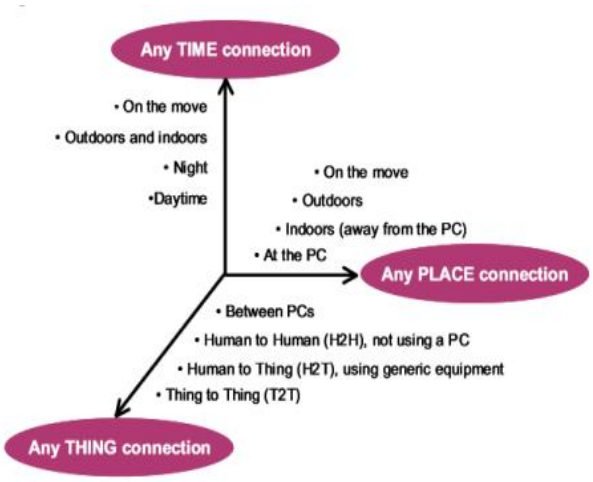



Fig. 1. IoT, A New Dimension [17].


Fig. 2. Iot Objects.

### C. Integration of WSN in an IoT Network

The determination of plants water needs is based on climatic parameters such as temperature, humidity, wind speed, etc. These parameters are detectable by specific sensors that are part of a WSN. These components are the basis for the development of any irrigation management system.

In an intelligent irrigation system, WSNs can be integrated into an IoT network, where the sensors will be hosts of the internet and not just data collection tools.

As it is mentioned in [18], this integration is feasible by 02 approaches:

*1) Proxy integration:* In this approach the base station is the only node connected to the internet, it acts as a proxy that represents an interface between the internet and the sensor nodes associated with it.

*2) Integration by adaptation of the TCP / IP standard:* In this integration architecture, the sensor network is fully open on the internet and the sensor nodes become real hosts of the Internet, addressable, and having the same concessions as an ordinary host. The projection of the TCP / IP standard as it is on the sensor networks is almost impossible, which necessitates adapting IP-based communication standards, and even developing new mechanisms that are inspired by them, and which would be alerted to the constraints of sensor networks in the new generation of the Internet (the Internet of Things).

## V. Architecture of Intelligent Irrigation Management System

Generally, an intelligent irrigation system takes place in three levels (see Fig. 5):

### A. Data Collection Level

Different Sensors can be used to collect environmental data. In this level, We have also the possibility to use solar energy powered sensors or solar powered sensor board such as "Waspmote" which can support 14 different sensors [19] (see Fig. 3 ).

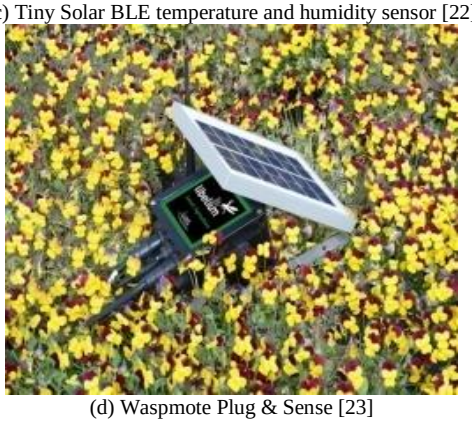
The gateway transmits the data collected by the sensor nodes to the controller to decide an irrigation action. For example, if the sensed temperature or humidity is below predefined thresholds, then activate the irrigation system. In this regard, Zigbee 802.15.4 communication model is used. It is the most suitable communication protocol for this type of application. It allows the transmission of small amount of data (environmental data) over long distances and with the minimal possible power consumption.

The user or manager of the irrigation system can control the irrigation process through smart phone, computer, or tablet.

(a) Edyn Soil Moisture Sensor [20]

(b) TMP102 Temperature Sensor [21]


(c) Tiny Solar BLE temperature and humidity sensor [22]



(d) Waspmote Plug & Sense [23]

Fig. 3. Solar Powered Sensors.

### B. Control Level

In this level, the microcontroller compares the sensed data with predefined threshold values. If the microcontroller finds that the sensed values are below a predefined threshold values then the auto irrigation system will be activated, otherwise the system stays inactive until the comparison gives the opposite result (see Fig. 4).

The choice of threshold values depends on many factors such as climatic conditions, soil texture, plant types, and irrigation method.

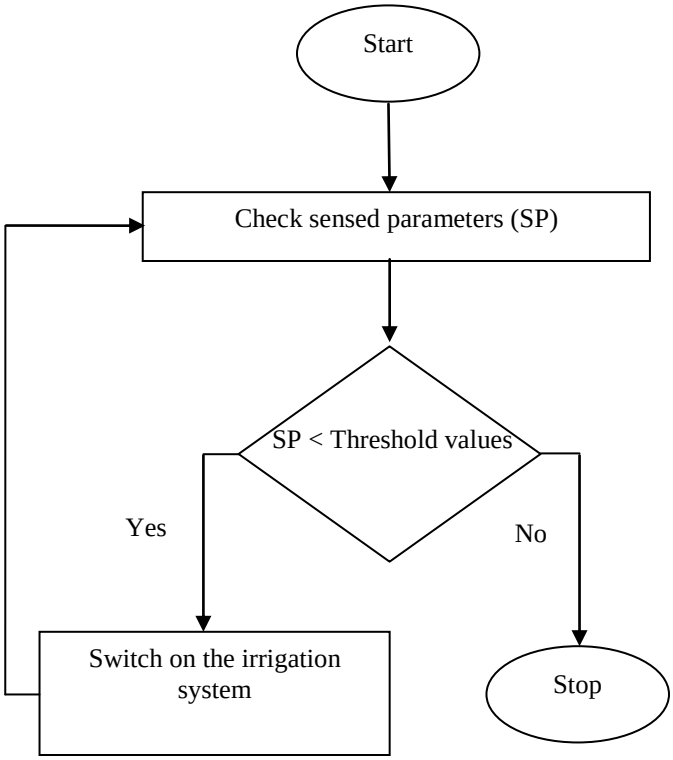



Fig. 4. Flowchart of the Microcontroller.

### C. Solar Powered Water Pumping Level

To pump water from the source to the storage tank, solar panels must be mounted near the pump; a regulator is used to charge a set of batteries to feed the pump at the source in order to pump the water into the storage tank.

In this step, Different devices are necessary: Source, pump, photovoltaic panels, motor pump-set, mounting structures and tracking system, controller (regulator), inverter (to change the direct current from the solar panels into alternating current for the pump), and water distribution system.

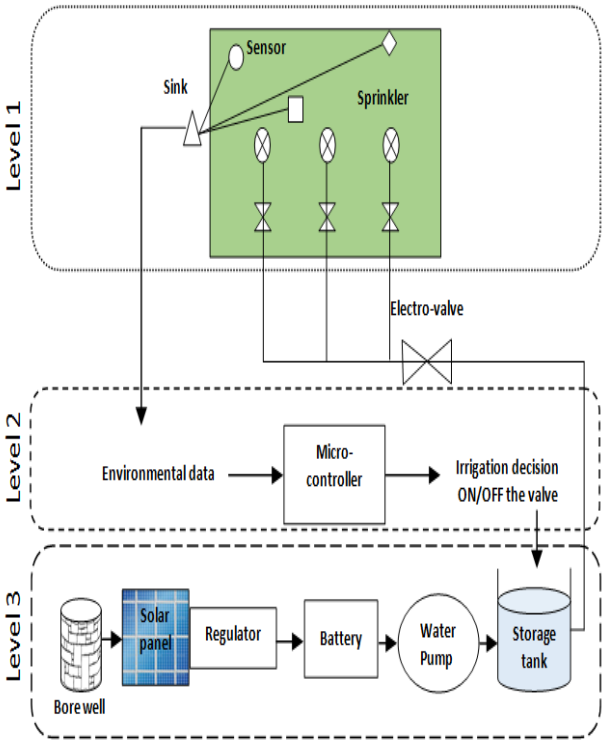



Fig. 5. Intelligent Irrigation Management System Architecture.

## VI. Architecture of Intelligent Irrigation Management System

The use of proper method of irrigation and its control is very important to reduce water consumption. Automation of irrigation system by monitoring soil moistures and other crop parameters using wireless sensor networks and the Internet of Things is one of the most promising applications that can provide optimal water use efficiency.

Implementation of smart systems in the irrigation field is based on the use of a combination of sensors such as temperature, wind speed and humidity.

Wireless communication and real time information provided by these systems play an important role in estimating the amount of water needed to meet the crop requirements, especially when it became part of the Internet of Things (IoT).

In this research axis, every day there are new technological developments, so it is difficult to develop a unified system adapted to all situations, but that does not prevent the establishment of some basic rules for the construction of a typical system:

- The system must be sensitive to climate change
- The system must be able to adapt with any type of plant and be effective during the different stages of growth.
- The system must be reliable with any irrigation strategy.
- The communication interface between the network and the end user should be as simple as possible.

## VII. CONCLUSION

Reduce water consumption through the use of intelligent technologies such as wireless sensor networks and the Internet of Things is one of the most promising applications to automate the irrigation process.

In this proposal, the Internet of Things technology is integrated as a communication tool, in addition to the use of solar energy not only to pump water from the source to the storage tank as it is already known, but also to feed the system and deal with the problem of energy constraints of sensor networks.

Since the system is a combination of three technologies (sensor networks, the Internet of Things and solar energy), therefore, it will benefit from the advantages of each technology, to mention but a few:

- The use of real time information to adjust irrigation strategies for any condition.
- Easy installation.
- Minimize labor and time needed to control the irrigation process
- The system guarantees higher productivity with efficient use of water.
- Automation of the irrigation control process by using the detected environmental parameters
- The sun is a stable and reliable source of energy. Fuel supply, transportation or rising operating costs are no longer a concern.
- Solar energy is a source of clean and inexhaustible energy.

## REFERENCES

[1] Dassanayake et al. Water saving through smarter irrigation in Australian dairy farming: Use of intelligent irrigation controller and wireless sensor network, paper presented at the 18th World IMACS / MODSIM Congress (2009) Cairns, Australia. From http://mssanz.org.au/modsim09.

[2] S. Sutar, et al. Irrigation and Fertilizer control for Precision Agriculture using WSN: Energy Efficient Approach. *International Journal of Advances in Computing and Information Researches*, 1(1) (2012).

[3] Z. Li-Ming, et al. The System of Water-Saving Irrigation based on WSN and MSIF. *International Journal of Computer Science Issues*, 9(6) (2012).

[4] S.R. Kumbhar and A.P. Ghatule, Microcontroller based Controlled Irrigation System for Plantation, *Proceedings of the International MultiConference of Engineers and Computer Scientist*, 2 (2013) Hong Kong.

[5] E. Soorya, et al. Smart drip irrigation system using sensor networks. *International Journal of Scientific & Engineering Research*, 4(2) (2013).

[6] L. Gao, et al. An Intelligent Irrigation System Based on Wireless Sensor Network and Fuzzy Control, *JOURNAL OF NETWORKS*, 8(5) (2013).

[7] F.S. Mohammad, et al. Adoption of an intelligent irrigation scheduling technique and its effect on water use efficiency for tomato crops in arid regions, Australian *Journal of Crop Science, AJCS*, 7(3) (2013) 305-313.

[8] P. Alagupandi, et al. *Smart Irrigation System for Outdoor Environment using Tiny OS*. Paper presented at the International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC), IEEE (2014).

[9] J.P. Kumar, et al. Implementing Intelligent Monitoring Techniques in Agriculture Using Wireless Sensor Networks, *International Journal of Computer Science and Information Technologies*, 5 (4) (2014) 5797-5800.

[10] S. Shahidian, et al. Smart irrigation control using inexpensive capacitance probes, *Proceedings International Conference of Agricultural Engineering* (2014) Zurich. From www.eurageng.eu.

[11] A.C. Bartlett, et al. A smart phone app to extend use of a cloud-based irrigation scheduling tool. *Computers and Electronics in Agriculture 111, Elsevier* (2015).

[12] J. Gutiérrez, et al. 2013. Automated Irrigation System Using a Wireless Sensor Network and GPRS Module. IEEE transactions on instrumentation and measurement.63, 1 (Jan.2014), 166-176. DOI= 10.1109/TIM.2013.2276487.

[13] W.Difallah et al. 2018. Design of solar powered smart irrigation system (SPSIS), International Conference on Software Engineering and Information Management. (ICSIM2018, January 4–6, 2018, Casablanca, Morocco ). ACM.

[14] M.H. Ariff and M.Z. Ibrahim, Solar Powered Soil Moisture Detector. Latest trends on communications and information technology,79-83.

[15] W. Difallah, et al. Implementing wireless sensor networks for smart irrigation. Taiwan Water Conservancy, 65 (3) (2017).

[16] W. R. Walker. SIRMOD III- Surface Irrigation Simulation, Evaluation and Design-Guide and Technical Documentation. (2003).

[17] ITU. Internet Reports 2005: The Internet of Things .2005. from http://www.itu.int/osg/spu/publications/internetofthings/InternetofThings_summary.pdf

[18] S.Sahraoui. Mécanismes de sécurité pour l'intégration des RCSFs à l'IoT (Internet of Things) "Security Mechanisms for Integrating WSNs into IoT (Internet of Things)". (2016).

[19] R.,Khan et al. Wireless Sensor Network Based Irrigation Management System for Container Grown Crops in Pakistan. World Applied Sciences Journal, 24(8) (2013).

[20] Edyn soil moisture sensor. From https://www.gardeners.com/buy/edyn-garden-soil-moisture-sensor-wifi-solar-powered/8592711.html

[21] TMP102 temperature sensor. From http://www.instructables.com/id/Solar-Powered-Remote-Temperature-Sensor/

[22] Tiny Solar BLE temperature and humidity sensor. From http://www.mouser.com/ds/2/100/CYALKIT-E03_Kit_Guide-968035.pdf

[23] Waspmote Plug & Sense. From http://www.libelium.com/smart-agriculture-monitoring-greenhouse-conditions-to-develop-new-products-in-the-food-industry/

# Securing and Monitoring of Bandwidth Usage in Multi-Agents Denial of Service Environment

Ogunleye G.O.[1]
Department of Computer Science, Federal University, Oye-Ekiti, Ekiti State, Nigeria.

Fashoto S.G[2]
Department of Computer Science, University of Swaziland, Kwaluseni, Swaziland

Mbunge Elliot[3]
Department of Computer Science, University of Swaziland, Kwaluseni, Swaziland

Arekete S.A[4]
Department of Computer Science, Redeemer's University, Ede, Osun State, Nigeria

Ojewumi T.O.[5]
Department of Computer Science, Redeemer's University, Ede, Osun State, Nigeria

***Abstract*—The primary purpose of Denial of Service attack (DoS) is to cripple resources so that the resources are made unavailable to the legitimate users. Due to the inadequate monitoring of activities on the network, it has resulted into huge financial losses. Bandwidth which is one of the resources being used on the network, if not properly monitored could result into misused and attack. This paper proposes a real time system for securing and monitoring the amount of bandwidth consumed on the network using the multi-agent framework technology. It also keeps a record of internet protocol (IP) addresses visiting the network and may be used as a starting point for the aspect of response in providing a comprehensive solution to DoS attacks. The bandwidth is pre-entered and an agent is assigned to monitor bandwidth consumption rate against the set threshold. If the bandwidth is consumed above the bandwidth limit and time set, then a DoS attack is suspected taking into considerations the DoS attack framework. This framework can be used as a replicate of what happen in the network scenario environment.**

***Keywords*—Bandwidth; mobile agent; multi-agents; DoS**

## I. Introduction

The internet is an interrelated computer networks on a universal system using protocol of the internet (TCP/IP) to connect devices all over the world. Over 4 billion users make use of the internet globally, with an increase of 20% internet users from Africa between 2017 and February 2018 [1,2]. The internet is used for different purposes at different levels to accomplish or support daily activities such as but not limited to research, downloading, electronic mail and group discussion. The over reliance of computer interconnectivity on the internet expose system resources and infrastructure to malicious users. It takes undue advantage of the situation to launch their denial of service (DoS) attacks by interrupting authorized users access which leads to unavailability of computer resources temporarily or indefinitely. The DoS attacks is a malicious attempt that can be achieved in different ways by a person or group of people through computer viruses, worm, Trojan horses, spyware, malware, phishing and so on. Recently, it was discovered that the DoS attacks growth is increasing exponentially on a yearly basis which is a cause of concerns especially on distributed enterprises and small-medium size businesses in 2017 in which 19 million malware attacks was identified and blocked [3,4]. One of the common cyber security threats is the distributed denial of service (DDoS) attack. A DDoS attack is a cyber-attack that takes place when multiple systems bombard the bandwidth or resources of a targeted system with so much traffic to render the services or the infrastructure of the web servers or websites unavailable or useless. The main focus of the attack is to render authorised user incapacitated from being able to perform normal transactions [5,6,7]. Conventional architecture of the internet is vulnerable to distributed denial of service (DDoS) attacks. According to National institute of standard and technology [8,9,10], several vulnerabilities preventive measures should be in place to mitigate malicious attacks but instead vulnerabilities are on the increase. In 2014, 7937 vulnerabilities were recorded as against 5186 in 2013 which shows an increase of 34.6 percent [11,12,13]. From the report of worldwide infrastructure security in 2010 that DoS attack is standardized and becomes scarier to the network operators which make them believe that there will be more problems with DDoS attack (Dobbins and Morales, 2010). The largest and the most impactful DDoS attacks were targeted at GitHub and it occurred on 28th of February 2018 but were reported in March 2018. Irrespective of the shortcomings of the DDoS defense detection mechanism to discover attack close to the target machine, describe the distinctive nature of it, and the inability to filter out legitimate packets from the attack packets, the defense detection accuracy mechanism is still very high. Recently, numerous DDoS defense detection strategies has been proposed such as SDN-based "moving target defense", to secure computer networks and operators from DDoS attacks by moving the computer networks and the operators from targeted virtual machines(VMs) to invulnerable environment [14,15,16].

Despite the high level of DDoS defense detection mechanism accuracy; there is still a need to develop a system capable of detecting and monitoring bandwidth to suspect DDoS attacks on the computer networks. In the fight against DoS attacks some problems have been identified.

The following are some of the attacks facing most networking environments.

- Difficulty in detecting highly distributed Denial of service especially close to its source IP address
- Difficulty in detecting DoS attack immediately without raising false alarm and
- Not delaying response in order to ensure that the attack is not a legal increase in user request on a website.

This paper seems to address the problems by proposing a secured system capable of monitoring the traffics on the network using the mobile agent technology.

## II. Literature Review

### A. Denial of Service

Denial of Service (DoS) attack occurs when the system denies legitimate users to have access to systems resources [17,18,19]. The DoS attack usually uses the single Internet connection and a single computer to flood a targeted system or resource [20,21,22]. The ultimate goal of DoS is to make services inaccessible [23,24,25] by either injecting computer viruses or flooding network traffic. Flooding network traffic can be achieved by incessant exploiting network vulnerable security loopholes, illegally access to network servers, and then brings down network services [26,27,28]. DoS attack utilises Transmission Control Protocol (TCP) connection buffers, exhausting bandwidth links, network router processing capacity [29], and application layer buffers that can lead to degradation of network performance [30], shun all network connections , and ultimately block the website [31,32]. The most active modern type of DoS is called Distributed Denial of Service (DDoS), which consists of multiple computers and Internet connections to target a single system. With the DDoS attack, incoming network traffic floods victim's system or resource indirectly using large-scale computer multi-agents connected via the internet [33]. DDoS attacks can happen in two ways over the internet. The first technique is to send the malicious network packets or codes to confuse either security application executing on victim's machine or Internet Protocol (IP). The second technique is to flood, interrupt and exhaust network connections by using the following three layers of OSI model; network layer, transport level and application. Once the computer is flooded with malicious codes, the multi-agent scans for another vulnerable computer entire the network. Multi-agents search and scan vulnerable computers over the network by employing random scanning, topological scanning, permutation scanning and local subnet scanning. According to Jelena and Peter [34], in random scanning, each compromised computer probes random addresses in either global or local Internet Protocol address space because attacking hosts duplicate themselves and execute uncoordinated attacks, thereby increasing the possibilities of packets collision and high traffic volume due to computers probing same address [35]. Topological scanning utilizes data stored in the victim's computer to discover new targets. With topological scanning, agents use valid URLs in the Web Servers in victim's computer to determine the next vulnerable computer in the network. The performance of this technique is almost similar to hit-list scanning [36]. Hit-list scanning occurs when attackers start by scanning the network and gather a list of potentially vulnerable computers before attaching. Once a list is created, attackers incessant rescanning the network to find the vulnerable computers, install malicious code and divide the list into half [37]. A newly infected computed is allocated half of the list, keep the remaining half, and scans the other residual list. When attackers find other vulnerable computers, they apply the same procedure decimating network performance due to the proliferation of infected computers in the network. Hit-list scanning makes sure that all vulnerable computers on the list are infected with malicious code. Local subnet scanning acts behind a firewall in an area that is considered to be infected by the malicious scanning program. The infected hosts scan and target vulnerable computers of its' own network using information in the local subnet addresses [38]. In permutation scanning, all machines share a common pseudorandom permutation list of Internet Protocol addresses. A block cipher of 32 bits is used to create a permutation list [38]. An infected host by either local subnet scanning or hit-list scanning starts scanning just after its point in the permutation list and scans through this list to find new targets. When it finds a host or a node that is already infected, it starts to scan again at a random point in the permutation list. According to [39], the process of scanning stops when the compromised host encounters sequentially a predefined number of already infected machines without finding new targets during that particular moment.

### B. Types of DDoS Attacks

DDoS attacks are typically be grouped as bandwidth and resource depletion attacks. Bandwidth depletion attacks are further sub-divided into two groups namely; flood attacks and amplification attacks. TCP SYN flood, UDP Flood, ICMP Flood and Smurf attacks are examples of bandwidth depletion attacks. With flood attacks, large volumes of malformed packets are sending continuously to the target computers to ensure that buffer overflow occurs and exploit vulnerable hosts. The ultimate goals of flood attacks are; to reduce processing capability, the memory of the victim's computer, and exhaust packet buffers and network bandwidth over the internet. Flood attacks can be mitigated by software patching. Resource depletion involves zombies sending messages to a broadcast IP address to cause all system in the subnet reached by the broadcast address to send a reply to the victim system [40]. Examples of resources of depletion attacks are; protocol exploits attacks and malformed packets attacks [41]. Figure 1 shows bandwidth depletion and resource depletion attacks. The list is not exhausted since attacks are dynamic and new on daily basis.

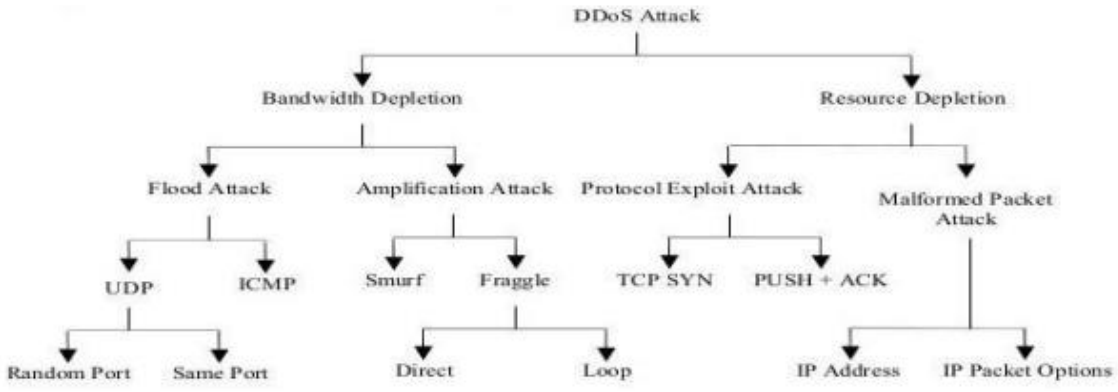



Fig. 1. DDoS Attacks Source (Keyur & Vivek, 2015)

### C. DDOS Defence and Detection Techniques

According to [42] a router-based packet filtering, ingress filtering, and source address validation are classified as DDoS prevention methods. Ingress filters depend on addresses that are on router that connects ISP's scope of control to fight against DDoS. Internal interface and external interface acts as a channel for internal addresses and external addresses respectively. These strategies work poorly for transit and backbone hosts but execute perfectly for leaf nodes. Router-based packet filtering and ingress filtering techniques does not affect attackers who show their IP addresses. Router-based packet filtering utilizes internet routing data to whether packets are arrived safely at its designated destination address without any modification [42]. Router-based packet filtering implements fast filtering to determine catch misconfigurations based on router's look-up table. Another effective DDoS prevention technique is called source-address validation. Past authors [42] states that source-address validation relies on building a second routing table of learned networks and matches subnet-to-subnet the 'allowed routes' and the filtered routes. The system changes gradually to topology modifications when the traffic is spoofed. This technique does not consider non-spoofing attacks.

DDoS detection techniques consist of artificial neural networks (ANN), data mining and fuzzy expert systems. Some authors used ANN to detect known and unknown DDoS attacks. ANN consists of the input layer, hidden layer, and output layers layer to make reasonable decisions. The input layer represents the typical patterns of attacks from genuine traffic. Input values in the input layer are feedforward to the preceding layer, hidden layer, which performs some computations by using activation function. The outputs of the hidden layer are used as inputs of the output layer. The learning algorithm and the nature of the problem determine the training function of ANN. Alan, et al.,[43] adopted supervised Backpropagation algorithm to train the network for known DDoS attacks. For unknown DDoS attacks, they trained ANN using unsupervised learning algorithm. Alan, et al, [43] successfully detected forged packets; the defense mechanism was activated to drop the packets while allowing genuine packets to pass through. Blocked packets are unblocked as soon as the system flags the traffic flagged to be normal. Some authors [43] applied Neural Network Classifier to detect a DDoS attack on DNS server. A Neural Network classifier used characteristics of the attacks as input values to classify whether the attacks exist or not. The performance and accuracy of neural network classifier depend largely on whether the selected features can really summarize the characteristics of DDoS attack. Abhilasha and Santosh [44] implemented Artificial Neural Network based on genetic algorithm and multivariate correlation analysis to detect DoS attacks. Support Vector Machine (SVM), data mining and fuzzy expert systems have been used successfully to detect both DoS and DDoS attacks on the internet. Fuzzy expert systems involve a set of rules written using conditional statements to specify attacks into categories and draw some conclusions from facts and rules [44]. All incoming network traffic instances are compared with rules in the system to check whether conditions are satisfied, if not then the system can detect some anomalies and inconsistent in traffic instances.

DoS and DDoS attacks can be detected by implementing Intrusion detection system (IDS) [44]. Intrusion detection system is the technique of analysing and monitoring network traffic to detect packets flow anomalies. According to [44] intrusion detection can also be implemented using Naïve bayes, Radial basis and rotation forest to discovered intruder network access pattern. Network security threats can be detected by using various types of IDS such as network-based IDS (NIDS), host intrusion detection system(HIDS), network behaviour anomaly detection. These Intrusion Detection Systems apply signature-based detection, anomaly-based detection and stateful protocol inspection to analyse and detect potentially unwanted traffic data and DoS threats. Mukkamala developed an IDS based on Multivariate Adaptive Regression Splines (MARS). It excels at finding optimal variable transformations and interactions, and the complex data structure that often hides in high-dimensional data on the internet. However, new emerging DoS and DDoS attacks require new prediction and detection techniques.

## III. Data Collection

The data that were required for the paper include the IP addresses sending request and the bandwidth of the network. The IP addresses were collected and stored in a database by an

IP_Collecting agent. This IP addresses were randomly generated for the purpose of this paper. The bandwidth was pre-entered and an agent was assigned to monitor bandwidth consumption rate against the set threshold. This constituted the network need and formed the basis for the action of the entire detection scheme.

### A. System Design

The information gathered by each component of the solution was stored in a relational database table. The relational database was designed using Microsoft Access. The solution consisted of three major parts handled by three agents. The first scheme in the detection mechanism is an agent that stores all Internet protocol (IP) addresses visiting the network. This part receives the addresses visiting a node and updates a record whenever it identifies a new address. The second part of the application is an agent that stores the total bandwidth capacity of the network. The total bandwidth capacity consumed will be monitored per time in relation to some allowable maximum of the total bandwidth. The allowable bandwidth is a fraction of the total bandwidth of the network that is paid for and received from the Internet Service Provider (ISP). The bandwidth threshold is a maximum above which the action agent blocks the network.

### B. IP_Collecting Agent

The IP_collecting agent is a specialized agent whose responsibility is to check all IP addresses visiting and/or sending request to the node as shown in Figure 2. It extracts all IP addresses, updates a record when it detects a new address. It will also display a history of all the IP addresses that have previously visited a node. It shows the time as well as the date they visited the network. In addition to monitoring the addresses required and displaying them, the collector agent also contains a method that would be activated in case of a system bridge or perceived attack.

### C. Timer Agent

Bandwidth consumption in the network was monitored by this agent as shown in Figure 3. It systematically observes all the in-flow traffic to the network. It monitors the amount of bandwidth consumed in relation to a threshold. The bandwidth threshold is preset as the maximum over which the overall usage should not exceed. This agent was called the threshold agent. It is responsible for checking the overall resource usage of the network. In this case, the resource is the bandwidth. Whenever, the threshold is exceeded, this agent alerts a third agent called the action agent.

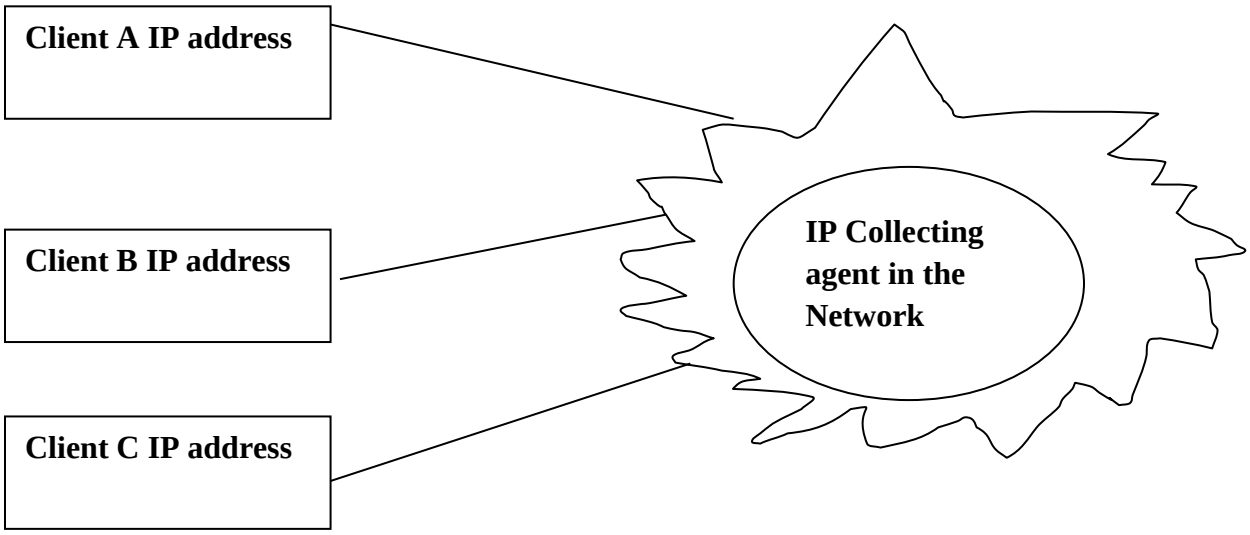



Fig. 2. IP_Collection Agent.

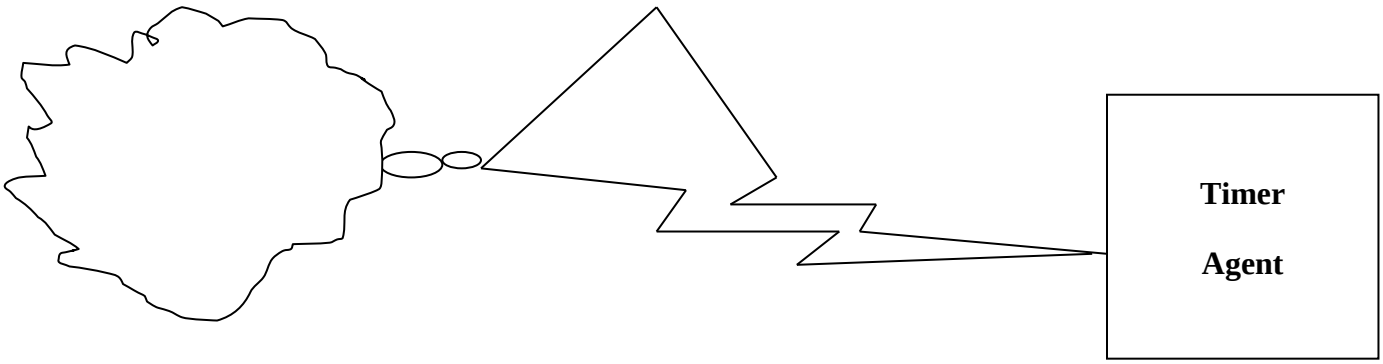



Fig. 3. Timer Agent Listening to the Incoming Traffic.

The threshold agent will broadcast or share this information if and only if the defined threshold is exceeded. This happens whether or not the rate is significantly exceeded or not. A threshold is set in order to reduce the level of false alerts that is, to a reasonable extent. To differentiate between a normal rise in usage and anomaly. The use of threshold also provides a baseline for monitoring this system. It is important to note that the agents' action is based on anomaly. In which case, it is the change in network usage. The agent monitors a particular node and ensures that the bandwidth could not be consumed more than a certain threshold. As soon as bandwidth consumption reaches the stipulated threshold, it alerts the action agent.

*D. Action Agent*

The third agent in this detection scheme is the action agent. The action agent is usually dormant most of the time until there is a security bridge. It is activated by the threshold agent. When activated, the action agent requests for the IP addresses that visited the network most recently and disconnects all such requests even if they have been granted. The action of the agent and its activities is dependent on the aforementioned agents. A monitoring agent is assigned to each node on the network.

*E. System Algorithm*

Let $X_p$ be the number of new IP addresses visiting the network at a particular time T. On the other hand, let $X_m$ be the average number of IP addresses that have visited the network over time. Under normal circumstances, the number of new IP addresses should be small for any time interval h. The monitoring agent calculates this value each time a new IP is detected. Whenever $X_p$ is greater than $X_m$, the Threshold is compared to the bandwidth that has been consumed. If the bandwidth usage is greater than the threshold, then the action agent would be activated. Another variable alert time $A_t$, is the time the monitoring or threshold agent has to wait to alert the action agent.

$A_t = X_p - X_m$ where $A_t$ is in seconds.

***Collecting agent listens to IP addresses {***

*If {new address is found, update database}*

*else {keep monitoring IP addresses}*

*}*

***Monitoring agent listens to incoming traffic {***

*If ( incoming traffic is greater than bandwidth threshold)*

*Then (Performs the calculation algorithm described above and alerts action agent at such time)*

*else (continue listening to traffic)*

*}*

***Action agent is activated by the monitoring agent {***

*Gets the most currently updated IP addresses and disconnects them from the network.*

*}*

The Unified Modeling Language diagram shows what part of the system may be activated and what is or should be expected as shown in figure 4.

**3.7. UML USE CASE Diagram**

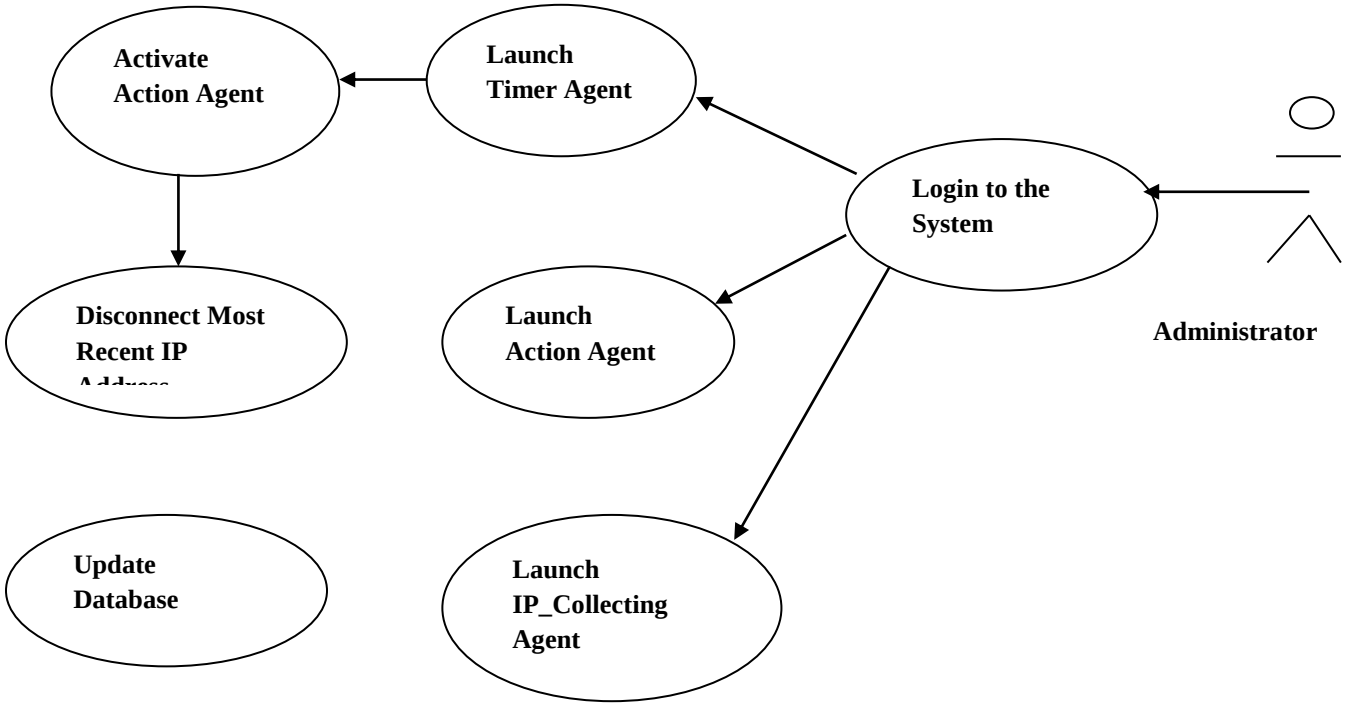



Fig. 4. Use Case Diagram for Accessing the System.

### 3.9. SYSTEM ARCHITECTURE

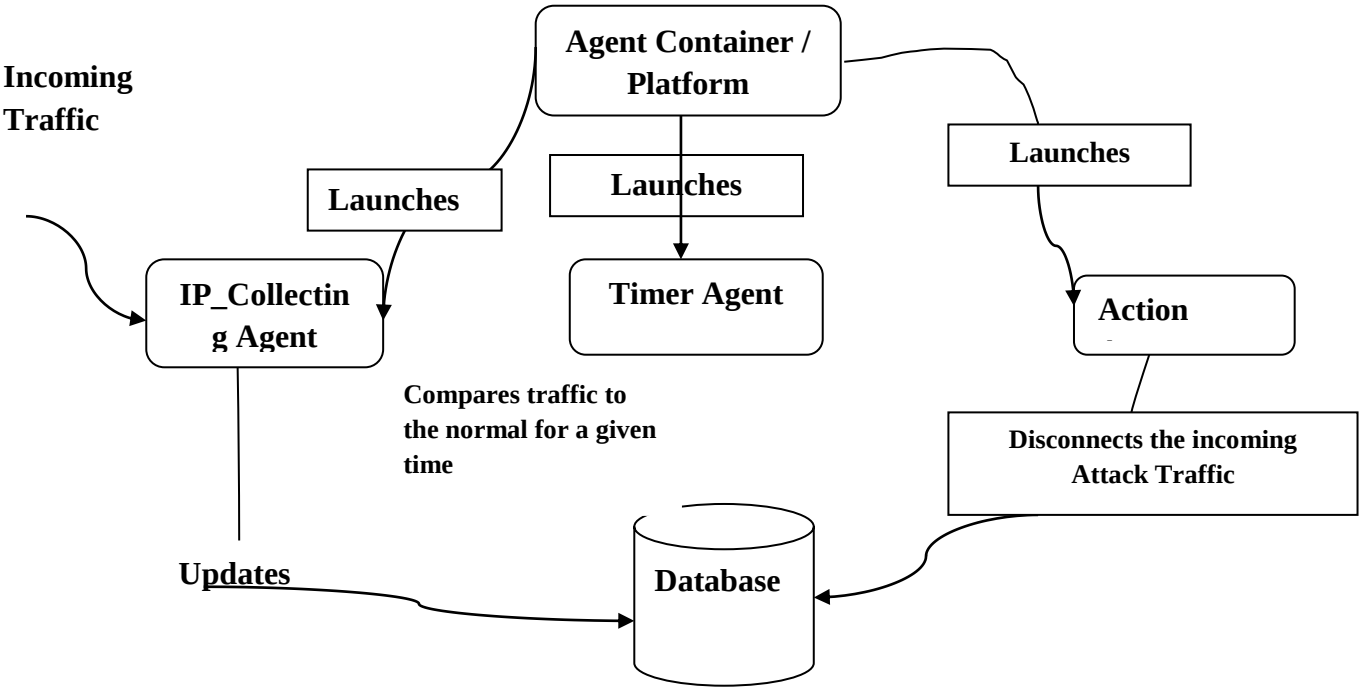



Fig. 5. System Architecture.

*F. Stages in Navigating through the Application*

*1)* Administrative login form where the network administrator logs in with password

*2)* Start the agent container and launch agents

*3)* The data monitoring agent is dispatched to monitor incoming request and store fresh IP addresses if not already in the database.

*4)* The bandwidth monitoring agent is launched to monitor incoming traffic.

*5)* The agents return their findings.

*6)* Decision is made based on the reports of the agents.

*7)* If an attack is suspected, the monitoring agent blocks all new IP addresses making request or disconnects those that have been honored.

The System architecture shows the interaction of various parts of the system and how they interact with one another as shown in figure 5. The agent platform houses all agents naturally. The agents, data collection agent, threshold agent and the action agent were all launched from here. More agents could also be created from the container. This is usually from the JADE platform. Immediately, the agents are launched, the monitoring and threshold agents are active, monitoring the visiting IP addresses and checking bandwidth consumption respectively. The monitoring agent observes the visiting IP addresses and updates the record whenever a new address is encountered. However, the Action agent is inactive for the time until an attack is detected. Whenever the threshold is exceeded, the monitoring agent activates the action agent. When activated, the action agent checks and disconnects all new IP addresses visiting the network for the first time from the updated database.

Normally, an anomaly based detection scheme encounters some challenges. Anomaly based DoS attack detection mechanisms analyse the normal behavior in a system and aim to detect attacks via identifying significant deviation from a normal behavior. Compared to signature based detection approaches, they can discover previously unseen attacks. The challenge therefore is in determining the threshold for anomalous behavior. A model that uses a tight threshold for legitimate behavior in the system may wrongly label normal behavior as malicious (false positive), whereas a loose threshold may lead to many attacks go undetected (false negative). This shortcoming is combated to an extent through the monitoring agent. It is assumed that when an attack is launched, the addresses used to launch the attack are new to network. In order to prevent total breakdown of system, recent requests from new IP addresses would be blocked instead.

*G. Network Interface*

The user interface is used to launch the network monitor that will display the activity screen. It also contains a button that will display the log details as shown in figure 6.

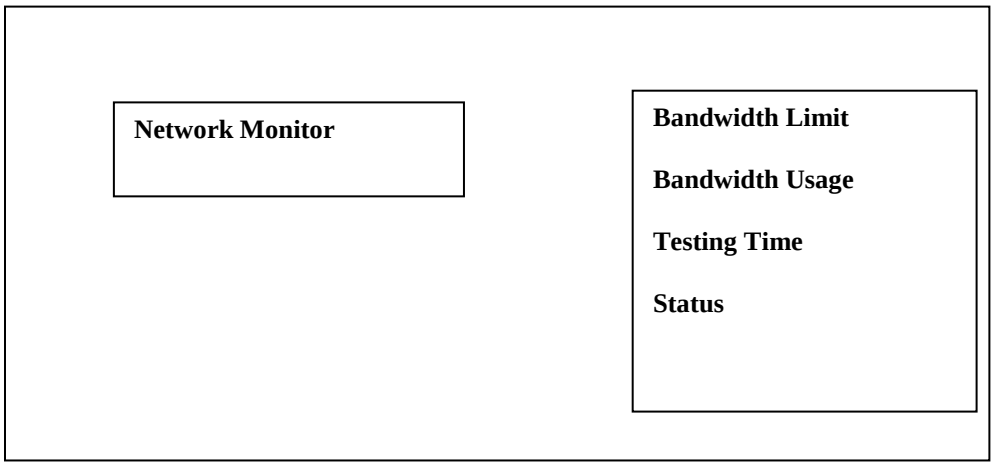



Fig. 6. Network Monitor Interface.

*H. Activity Screen*

The activity screen shows the IP addresses visiting the network as shown in figure 7. It displays the result of the monitoring agent at any instance. On the activity form, there is a start button to launch the agent. It also contains a start screen to stop an agent. On starting the form, an updated list is received. The activity screen also shows the bandwidth threshold as well as the bandwidth consumed at that time. The information displayed includes:

*1)* The IP address using or coming into the network.
*2)* The Time the address visited the network and
*3)* The date it was visited

*I. Log Details*

The event log could be viewed through a form interface. It is a report simply to display all the IP addresses as well as the time and date they visited a particular node. It is used by the network administrator to view the history of the IP addresses visiting the network and the particular node. It also stores registered IP addresses which would be used to decide on which IP address.

## IV. System Front End

This form provides restricted access to the system. It ensures the system's integrity and prevents unauthorized access. Providing the correct username and password allows the user to gain access into the system as shown in figure 8.

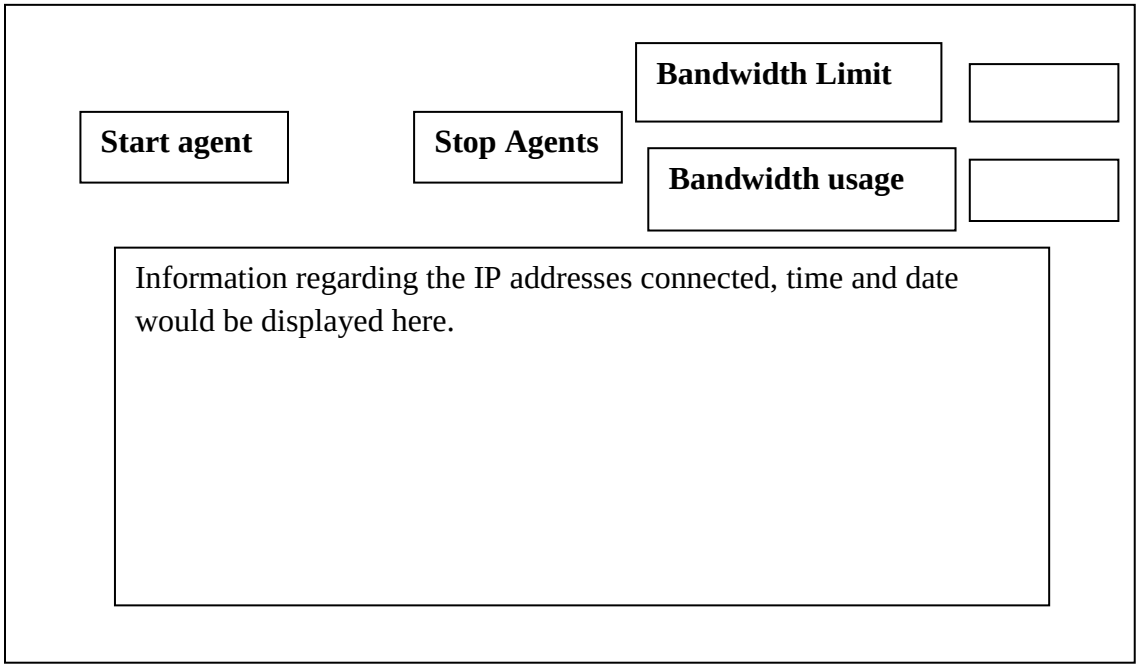



Fig. 7. Activity Form.

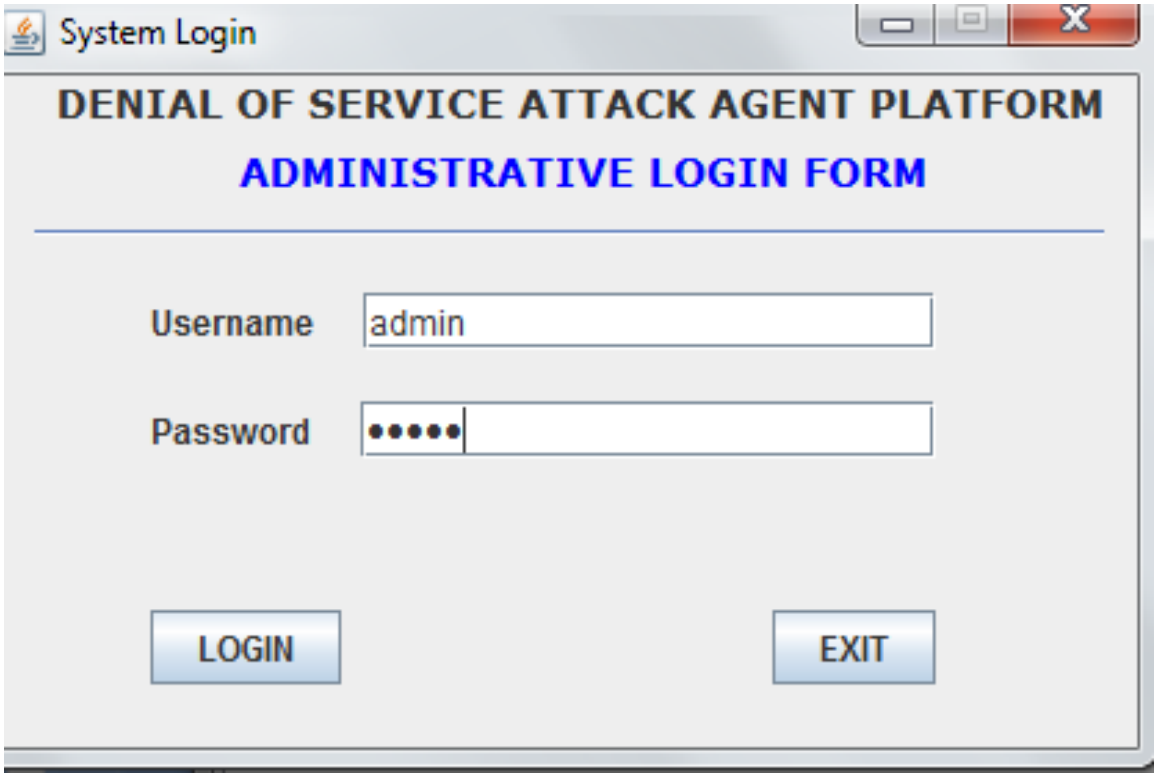



Fig. 8. Administrative Login Form.

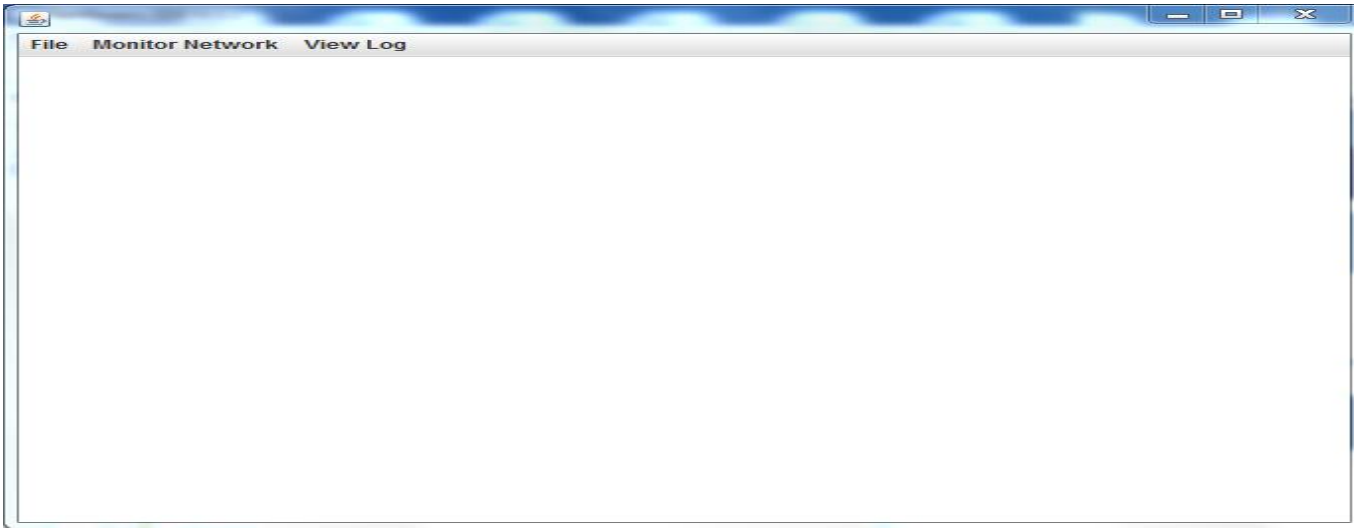



Fig. 9. The Network Monitor before the Agents are Initiated.

### A. System Interface

In order to launch or stop the system after access is obtained, figure 9 interface allows a user/administrator to either launch the network monitor which is already connected to the JADE environment from the agent platform itself. This interface was designed to provide a level of abstraction for the entire system. However, creating more agents and other maintenance tasks are done from the agent platform.

### B. Network Monitor

In this section, the IP_Collecting agent is launched. This also means the whole system is started. The "start agent" button initiates the IP_Agent and the Timer Agent as shown in figure 10.

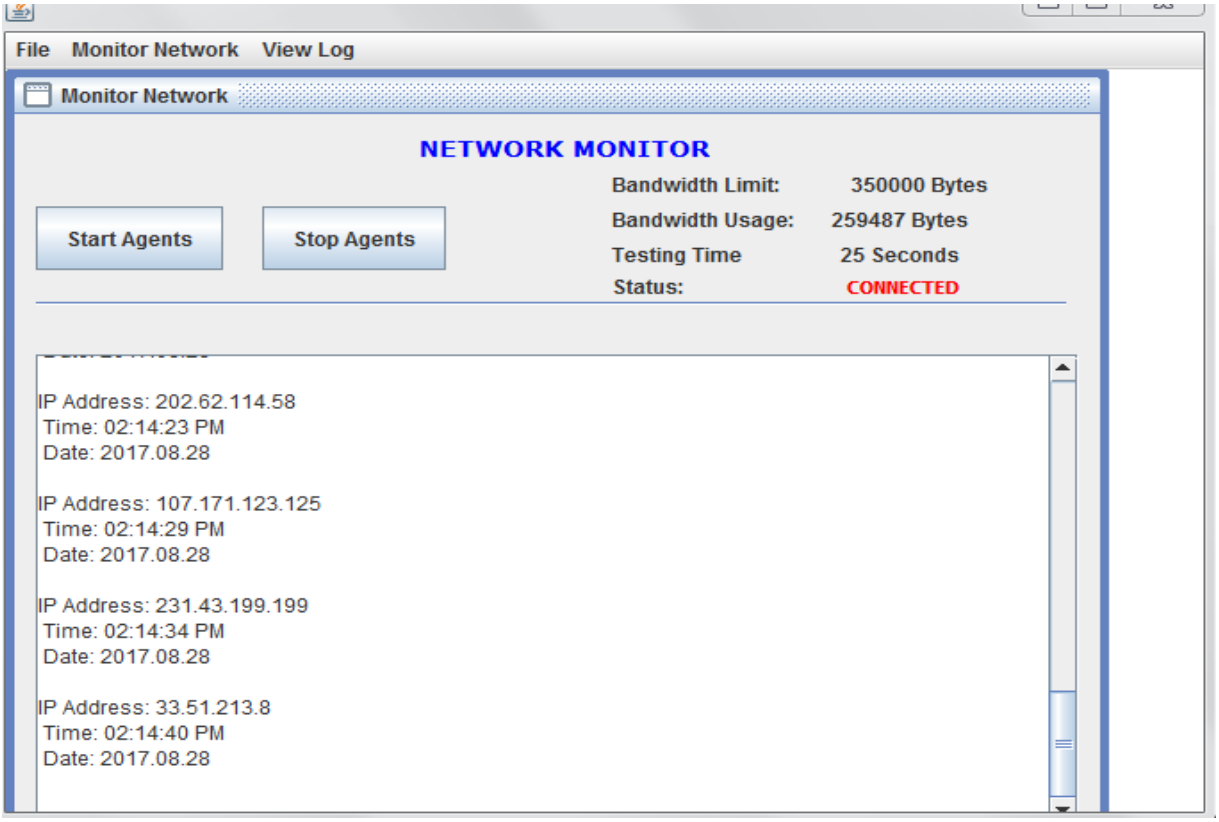



Fig. 10. Network Monitor before the Agents have been Launched.

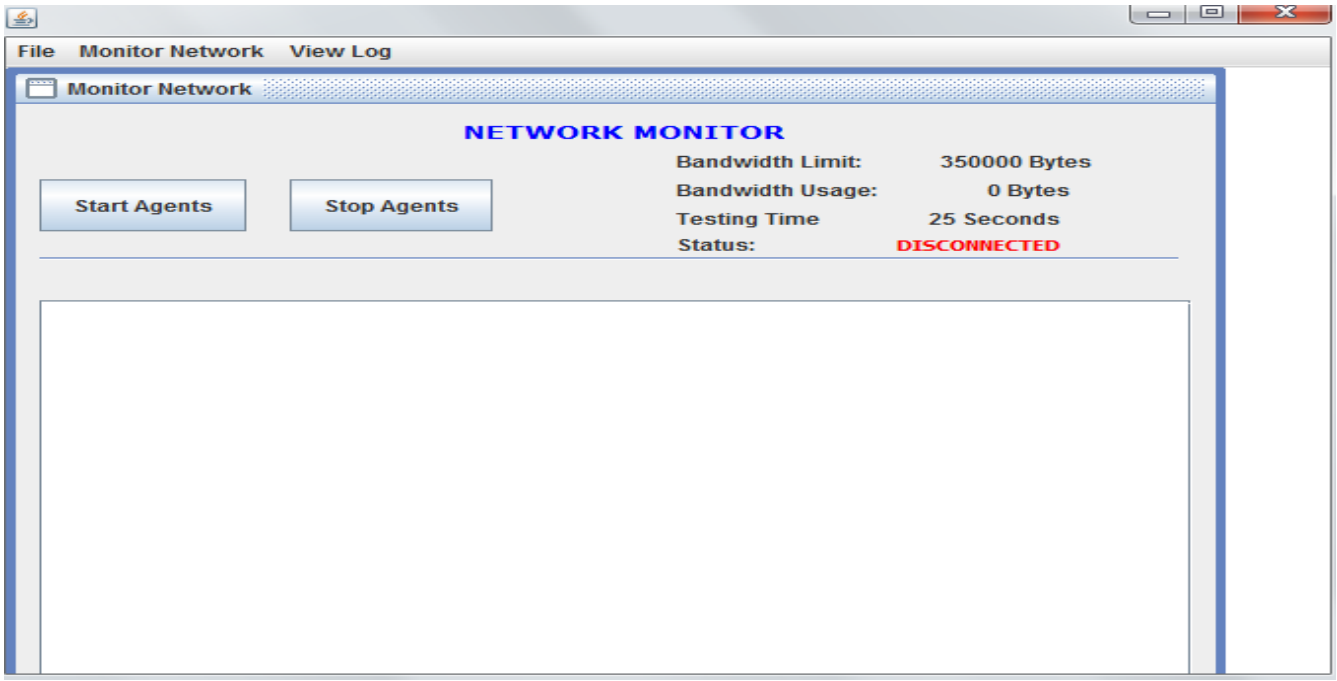



Fig. 11. The Network Monitor after the Agents are Initiated.

Figure 11 shows a simulated activity where the bandwidth is being consumed. It represents a normal traffic.

### C. *Network Monitor Showing a Simulated Denial of Service Attack*

Figure 12 shows a faster rate of bandwidth consumption within the observed period of time in which a Dos attack is suspected using the DoS attack detection rules.

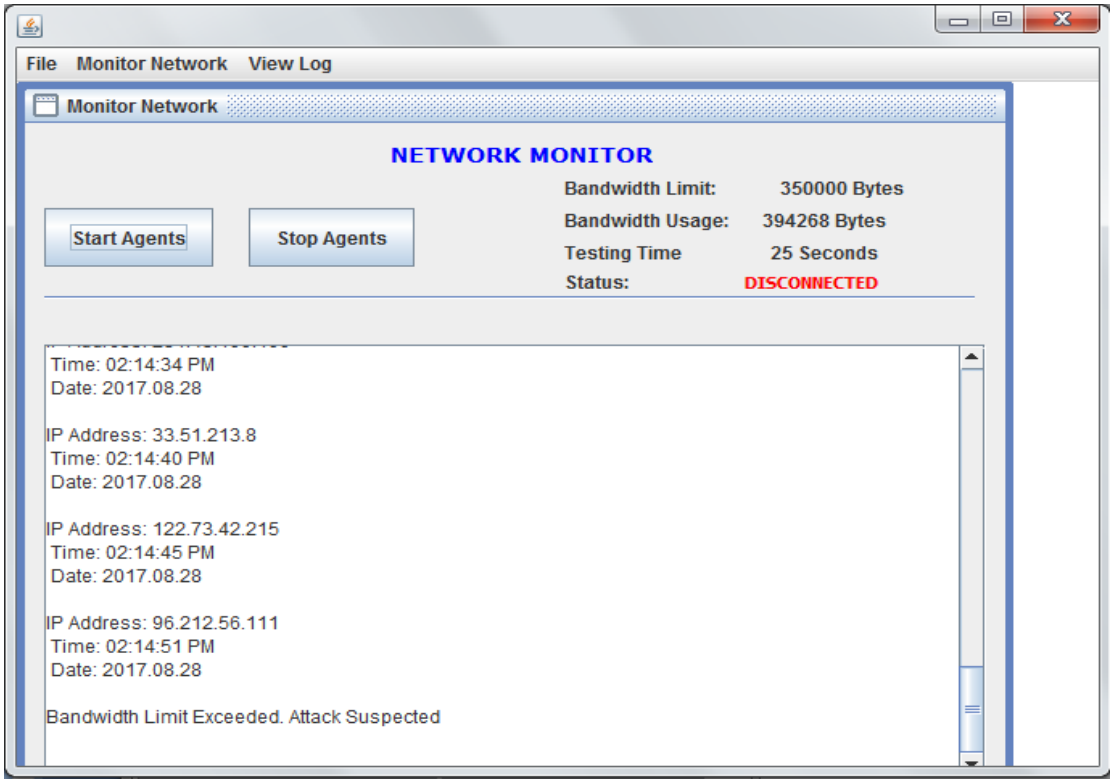



Fig. 12. Network Monitor Showing an Attack Traffic.

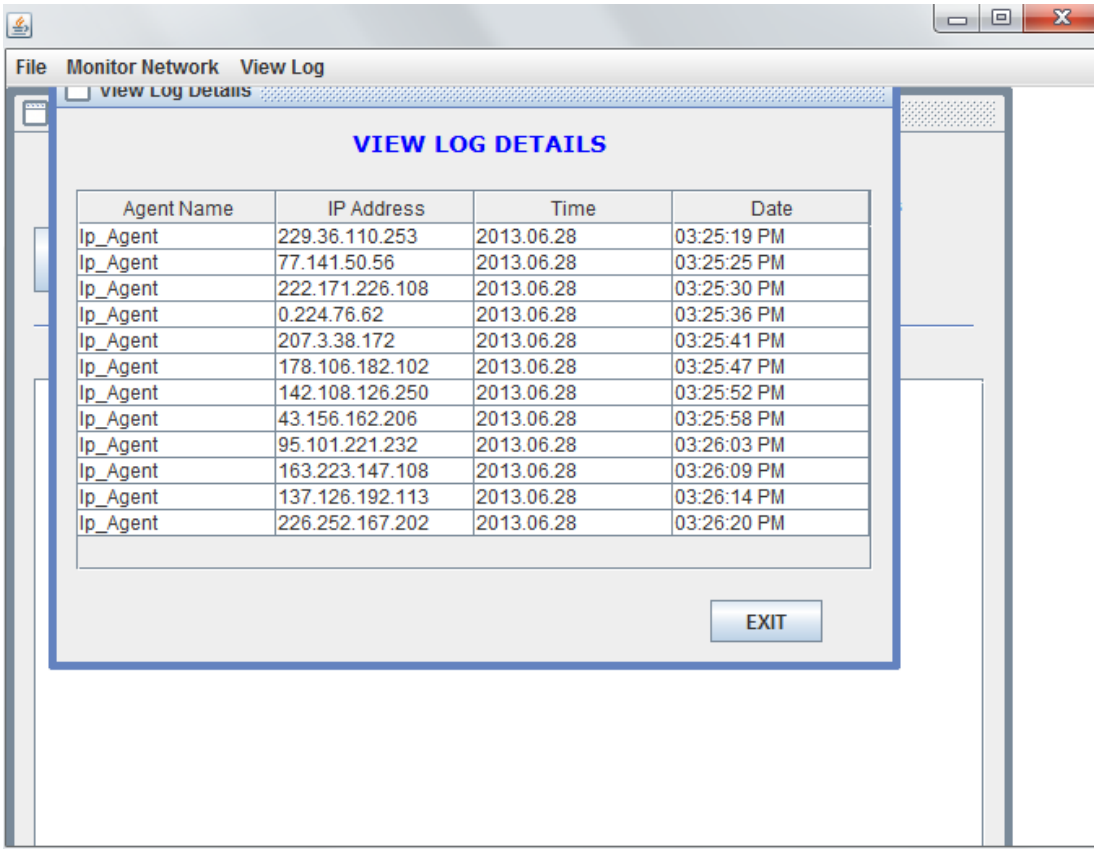



Fig. 13. IP Addresses Record.

### D. Log Details

The system also keeps a record of IP addresses that have visited the network which can be accessed from the view log details menu as shown in figure 13

### E. Results and Discussions

The result of this paper is a successful simulation of a Denial of service attack where a resource (in this case, a hypothetical bandwidth) was quickly consumed within an observed period of time. The result from the simulated attack traffic was compared against a control (i.e, an expected bandwidth consumption rate) in order to show what could happen in a DoS attack situation. In this paper, the amount of resource consumed is measured as well as the time it is consumed using the DoS detection rules. For the experiment, 350000 bytes was set as a threshold which may vary among organizations depending on the capacity of the network. If available resource is consumed within the test time i.e. 25 seconds then an attack is suspected and all incoming as well as honored requests are terminated. When a resource becomes unavailable, other users who may be legitimate users are prevented from accessing such resource. Scenarios like this can disrupt the proper functioning of the system and in some cases pave the way for more serious forms of attacks.

## V. CONCLUSION

Detection of DoS attack was demonstrated in this paper using mobile agent technology with focus on bandwidth consumption using hypothetical traffic. If the amount of resource consumed can be monitored and controlled by keeping track of the traffic in a network, then the devastation and loss caused by DoS attacks can be highly reduced if not totally eradicated. Hence, providing a threshold for the amount of resource that should be consumed would prevent computer networks from crashing totally before an action can be taken.

A drawback for this study is the temporal disconnection of requests for those that have been granted and the incoming ones. This may be solved by providing a better response method. A record can be maintained for visiting IP addresses without depending on internet service providers. The records obtained can now be further analysed to see the patterns of suspected IP packets and other methods for response can now be formulated by monitoring a particular network.

## REFERENCES

[1] Kessler, Gary C. 2000. Defenses Against Distributed Denial of Service Attacks. http://www.garykessler.net/library/ddos.html

[2] Gibson and Steve. 2001. The Strange Tale of the Denial of Service Attacks Against GRC.COM. http://www.crime-research.org/library/grcdos.pdf. Gibson Research Corporation.

[3] Andersen, D. and Balakrishnan, H. and Kaashoek, F. and Morris, R. 2001. Resilient Overlay Networks. ACM.

[4] Peng, T and Leckie, C and Ramamohanarao, K. (2007). Survey of Network-Based Defense Mechanisms Counteringthe DoS and DDoS Problems. http://doi.acm.org/10.1145/1216370.1216373. ACMJ258-03.

[5] Mukherjee B, Todd L, Levitt K(1994). "Network intrusion detection, IEEE Network, Vol. 8, No. 3, pp.26–41, 1994

[6] Teng H. S. Teng, K. Chen and S. C. Lu(1990). "Security Audit Trail Analysis Using Inductively Generated Predictive Rules". In Proceedings of the 11th National Conference on Artificial Intelligence Applications, pages 24-29, IEEE, IEEE Service Center, Piscataway, NJ, March 1990.

[7] Porras, 1992 "STAT: A State Transition Analysis Tool for Intrusion Detection". Master's Thesis, Computer Science Dept., University of California, Santa Barbara, July 1992.

[8] Dunigan T and Hinkel G (1999), "Intrusion detection and intrusion prevention on a large network: A case study", Proc. of workshop on intrusion detection and network monitoring, 1999.

[9] Ilgun, 1993, "USTAT: A Real-time Intrusion Detection System for UNIX," 16-28. Proceedings of the 1993 Computer Society Symposium on Research in Security and Privacy. Oakland, California, May 24-26, 1993. Los Alamitos, CA: IEEE Computer Society Press, 1993.

[10] Garvey T. D. and Lunt T. F (1991), "Model based intrusion detection". In Proceedings of the 14th National Computer Security Conference, pages 372-385, October 1991.

[11] S. Kumar and E. H. Spafford (1994). "An Application of Pattern Matching in Intrusion Detection". Technical Report CSD-TR-94-013, Purdue University, 1994.

[12] S. Kumar (1995). "Classification and Detection of Computer Intrusions". PhD Thesis, Department of Computer Science, Purdue University, August 1995.

[13] Mukkamala S,. Sung A.H and Abraham A. (2003), "Intrusion Detection Using Ensemble of Soft Computing Paradigms", Third International Conference on Intelligent Systems Design and Applications, Intelligent Systems Design and Applications, Advances in Soft Computing, Springer Verlag, Germany, pp. 239-248, 2003.

[14] Shah K., Dave N., Chavan S., Mukherjee S. (2004), A. Abraham and S. Sanyal, "Adaptive Neuro-Fuzzy Intrusion Detection System", IEEE International Conference on Information Technology: Coding and Computing (ITCC'04), USA, IEEE Computer Society, Volume 1, pp. 70-74, 2004.

[15] S Staniford-Chen, S, B Tung, and D Schnackenberg(1998), "The Common Intrusion Detection Framework (CIDF)". Proc. Information Survivability Workshop, Orlando FL, October 1998.

[16] Ning P, Jajodia S, Wang X(2002), "Design and implementation of a decentralized prototype system for detecting distributed attacks", Computer Communications, Vol. 25, pp. 1374-1391, 2002.

[17] Kim, Y. Cheong L., Chuah C. And Jonathan H. (2004) "Packetscore: Statistical-Based Overload Control Against Distributed Denial-Of-Service Attacks" IEEE INFOCOM 2004 ,The 23rd Annual Joint Conference of the IEEE Computer and Communications Societies, Hong Kong, China, March 7-11, 2004. IEEE.

[18] Dittrich, D. Dietrich, S. and Long, N.(2000). Analyzing distributed denial of service attack tools: The shaft case. In Proceedings of 14th Systems Administration Conference. New Orleans, Louisiana, USA, 329-339.

[19] Dobbins R. and Morales C.(2010), " Worldwide lnfrastructure Security Report, "Arbor Networks, Vol VII, Massachusetts, USA.

[20] Gavrilis, D. and Dermatas, E. (2005) Real-time detection of distributed denial-of-service attacks using RBF networks and statistical features. Computer Networks and ISDN Systems, 48, 235–245.

[21] Kampanakis P., Perros H., and Beyene T. (2014), Sdn-based solutions for moving target defense network protection. In WoWMoM, pages 1–6. IEEE Computer Society.

[22] Debroy S., Calyam P., Nguyen M., Stage A., and Georgiev V. (2016). Frequency-minimal moving target defense using software-defined networking. 2016 International Conference on Computing, Networking and Communications (ICNC), 00:1–6.

[23] Roshan L., Travis N., Nishant C (2018), Dolus: Cyber Defense using Pretense against DDoS Attacks in Cloud Platforms, 19th International Conference on Distributed Computing and Networking, January 4–7, 2018, Varanasi, India. ACM, New York, NY, USA,

[24] Abhilasha, P., & Santosh, W. (2017). Denial-Of-Service Attack Detection Using Artificial Neural Network Based On Genetic Algorithm and Multivariate Correlaton Analysis . *International Journal of Innovative Research in Science, Engineering and Technology*.

[25] Adrien, B., & Martine, B. (2017). A Survey of Denial-of-Service and Distributed Denial of Service Attacks and Defenses in Cloud Computing. *Future Internet*, 1-19.

[26] Alan, S., Richard, E., & Tomasz, R. (2015). Detection of known and unknown DDoS attacks using Artificial Neural Networks. *Science Direct*.

[27] Aleksandar, M. (2016, May 5). *Intrusion Detection Systems*. Retrieved from University of Würzburg: http://www.uni-wuerzburg.de/fileadmin/10030200/IDS.pdf

[28] Charalampos, P., Michalis, M., & Olga, Z. (2017, May 5). Distributed Denial of Service Attacks. *The Internet Protocol Journal*. Retrieved from CISCO: https://www.cisco.com/c/en/us/about/press/internet-protocol-journal/back-issues/table-contents-30/dos-attacks.html

[29] Elliot, M., Ralph, M., Nation, C., & Antony, C. (2017). Fraud Detection in E-Transactions using Deep Neural Networks-A Case of Financial Institutions in Zimbabwe. *International Journal of Science and Research* , 1036-1041.

[30] Felix, L., Stuart, R., Michael, S., & Ljiljana, T. (2000). Distributed Denial of Service Attacks. *IEEE Xplore*.

[31] Jelena, M., & Peter, R. (2002). A Taxonomy of DDoS Attack and DDoS Defense Mechanisms. *citeseerx*.

[32] Jelena, M., & Peter, R. (2014). A taxonomy of DDoS attack and DDoS defense mechanisms. *ACM SIGCOMM Computer Communications Review*, 39-53.

[33] Jun, W., Wang, X., & Lee, B. Y. (2010). Detecting DDoS Attack towards DNS Server Using a Neural Network Classifier. *Springer*, 118-123.

[34] Keyur, C., & Vivek, P. (2015). Distributed Denial of Service(DDoS) Attack Techniques and Prevention on Cloud Environment. *International Journal of Innovations & Advancement in Computer Science*, 1-6.

[35] Kihong, P., & Heejo, L. (2001 ). On the Effectiveness of Route-Based Packet Filtering for Distributed DoS Attack Prevention in Power-Law Internets. *citeseerx*.

[36] Munivara, P., Rama, M. ,., & Venugopal, R. (2014). DoS and DDoS Attacks: Defense, Detection and Traceback Mechanisms - A Survey. *Global Journal of Computer Science and Technology: ENetwork, Web & Security* .

[37] National CyberSecurity Center. (2018, May 05). *Understanding denial of service (DoS) attacks*. Retrieved from National CyberSecurity Center: https://www.ncsc.gov.uk/guidance/understanding-denial-service-dos-attacks

[38] Nicholas, W. (2002). Potential Strategies for High Speed Active Worms:A Worst Case Analysis. *citeseerx*.

[39] Ogunleye, Adewale, & Alese. (2013). DETECTING DDOS ATTACK IN AN AGENT-BASED VIRTUAL KNOWLEDGE COMMUNITY. *Seria Informatică*.

[40] Oyebode E.O., Fashoto S.G.,Ojesanmi O.A. and Makinde O.E.(2011). Intrusion Detection System for Computer Network Security. Australian Journal of Basic and Applied Sciences, 5(12): 1317-1320, ISSN 1991-8178,2011.

[41] Parneet, K., Manish, K., & Abhinav, B. (2017). A review of detection approaches for distributed denial of service attacks. *Systems Science & Control Engineering An Open Access Journal*, 301-320.

[42] Qijun, G., & Peng, L. (2018, May 05). *Denial of Service Attacks* . Retrieved from Pennsylvania State University : https://s2.ist.psu.edu/ist451/DDoS-Chap-Gu-June-07.pdf

[43] Santos, K., Chandra, S., Phani, R., Dawood, B., & Sudhakar. (2013). Intrusion Detection System- Types and Prevention. *International Journal of Computer Science and Information Technologies*, 77 - 82.

[44] Vangie, B. (2018, May 5). *DDoS attack - Distributed Denial of Service*. Retrieved from Webopedia: https://www.webopedia.com/TERM/D/DDoS_attack.html

# Internet of Things and Healthcare Analytics for Better Healthcare Solution: Applications and Challenges

Zuraida Abal Abas[1], Zaheera Zainal Abidin[2], Ahmad Fadzli Nizam Abdul Rahman[3], Hidayah Rahmalan[4], Gede Pramudya[5], Mohd Hakim Abdul Hamid[6]
Fakulti Teknologi Maklumat & Komunikasi,
Universiti Teknikal Malaysia Melaka,
76100 Durian Tunggal, Melaka, Malysia

***Abstract*—The total number of population in the world will keep on increasing. This will eventually pose challenges towards quality of life for example issues related to healthcare. Hence, a proper solution needs to be devised in order to face the challenges. Internet of Things (IoT), which is one of the digital technologies, that is becoming a trend now can offer promising solution. This paper serves as a short communication in introducing IoT and its application in healthcare domain as well as the analytics combined with the technology. Some examples are presented according to the categories of the application. It must be noted that the analytics play an important role in making the IoT healthcare as a comprehensive solution. At the end of the paper, challenges in making this digital as an accessible solution is discussed.**

***Keywords*—Internet of things; analytics; healthcare; applications; challenges**

## I. Introduction

Today, as recorded in world population clock, the total number of world population is about 7.6 billion in 2018 [1]. The total world population will keep on increasing as illustrated in Figure 1 [1]. The population growth is driven by several factors such as fertility rate, life expectancy (mortality) and migration [1]–[3]. Although the growth rate is decreasing or slower compared to previous time, the total number of world population will be projected to reach at about 9.8 billion in 2050, as projected by the United Nation [4]. This is actually a huge amount of number. Imagine the world is full with huge number of people; there must be several implications. Among the implications will definitely have a strong impact on quality of life.

Having the term 'quality of life', it is a normal setting to relate it with theory of Maslow's Hierarchy of Needs in mind as illustrated in Figure 2. The diagram of Maslow's Hierarchy of Needs is not an alien to most people, which has been seeing it from young age. In general, there are five levels of needs: physiological, safety, belonging, esteem and finally self-realization. The interest of this paper is to highlight the health component that belongs to the safety level. The health component in the safety level usually brings together unity among various stakeholders, practitioner, and industries across the globe in order to bring innovative health solution.

It is worth to highlight that the quality of life is one of the issues that need to be handled in efficient way while facing growing population phenomenon. Again, one of the domains that are very related to quality of life is healthcare service. It must be noted that the growing population will somehow create new and lasting challenges for healthcare worldwide [5]. In order to face these challenges, the emerging digital technology such as machine learning, cloud computing, Artificial Intelligence and Internet of Things (IoT) should be utilized. The world will benefit a lot more when utilizing digital technology especially in a healthcare setting.

This paper serves as a short communication in introducing IoT and its application in healthcare domain as well as the analytics combined with the technology. Section II presents the revolution of digital technology. Section III presents IoT for healthcare application while section IV presents the analytics and its role in IoT applications. Finally section V presents discussion and conclusion.

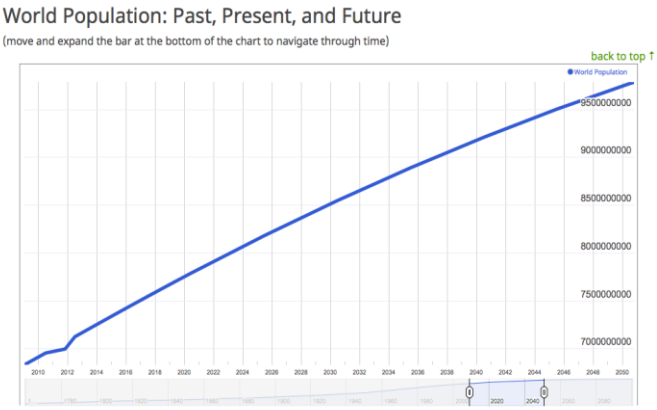



Fig. 1. The World Population taken from [1].


Fig. 2. The Maslow's Hierarchy of Needs taken from [6].

## II. Digital Technology: the Revolution

It can be seen obviously that the world today is surrounded with digital technology massively. The digital technology is advancing continuously and being embraced by many industries. Figure 3 [7] illustrates the digital waves, the revolution of digital technology that we are experiencing today.

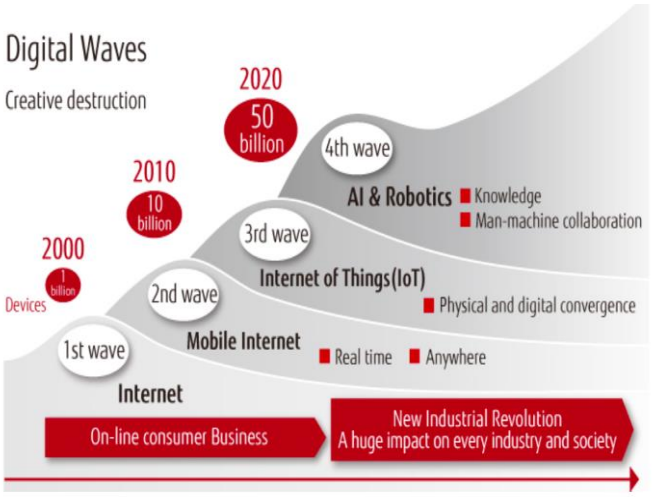



Fig. 3. The Digital Wave taken from [7].

The Internet was born during the first wave of digital technology, which give individual access and opportunity to wide variety of new and exciting borderless information and solution. Mobile Internet rose during the second wave of digital technology in which, there is an explosive growth of powerful smartphones and table devices. The world is now currently face the third wave of digital technology in which Internet of Things (IoT) emerged. It can be seen that the internet-enabled devices, expected to exceed 50 billion by 2020. IoT is about tiny embedded sensors and computers in equipment, machinery and devices, which can be link together and link to the cloud to generate new value for various industry and society. Robotics and Artificial Intelligence (AI) will be the main technology in the fourth wave of digital technology. It deals with the machine that is able to assist in knowledge creation process and decision making based on the combination of text and image data from the Internet as well as data from various sensors. The examples of modern digital technology are Cloud-based services, wearable sensors, biomarker detectors, mobile apps, the Internet of things, big data and AI [8].

As what can be seen now, IoT is one of the digital technology that surround today's daily live. IoT is an ecosystem of connected physical objects that are accessible through the Internet. These physical objects are connected using RFID, sensors/actuators and communication technologies[9].

## III. IoT For Healthcare Application

It is obvious that the initiative for embedding IoT in healthcare application is becoming a trend [10]. It is reported in [11], the IoT healthcare spending is expected to increase and total up to about $800 billion in 2018. In general, the current application of IoT healthcare can be divided into two categories. The first one is for improvement of care while the second one is on tracking, monitoring and asset maintenance.

With regards to the first categories of application, which is improvement of care, there are two examples that will be shared in this paper. The first one is the IoT solution developed for Diabetic retinopathy. Diabetic retinopathy is a critical disease that is a leading cause for blindness in adult. It is quite a surprise to discover that the lifetime cost for the treatment is about USD1 million. However the cost will be way cheaper if it is treated early, which amount only USD50. The key to prevent this disease is to have an early detection in screening. Unfortunately, there is only 40% of patient that turn up for screening. In order to promote early detection, the Intelligent Retinal Imaging System (IRIS), LLC (limited liability company) of Pensacola has developed a testing solution screening which is accessible in rural and city area by using automated camera [12]. The solution works by sending the retina image captured by the automated camera to the clouds. The clouds have the IRIS developed algorithm for detecting the disease. The diagnosis report is then send back within 24 hours.

The second example of IoT healthcare application is continuous Glucose monitoring system [13]. The glucose monitoring system works by inserting the sensor to the body. The glucose data will be send wirelessly through transmitter to various display devices. With this solution, alert and alarm can be set in order to track the condition if the glucose level is heading high or low.

'AutoBed' is an example of the second categories of IoT healthcare application, which is tracking, monitoring and asset maintenance. 'AutoBed' is the prototype of 'smart bed' developed by the GE healthcare team collaborated with Mount Sinai Hospital, New York [14], [15]. Mount Sinai Hospital has more than 1000 beds and there are more than 59000 in-patients in a year. Hence, it can be concluded that there is a massive daily operation that take place in the hospital, specifically in assigning the patient to the available bed. 'AutoBed' is an IoT healthcare application prototype that can help in optimizing the process of assigning patient to available bed. Once the patient is admitted, the admitting nurse will enter all the required information of the patient into the electronic medical record form. The information includes the gender and 'triage' (the urgency of the case). Once the form is completed, then the bed request is triggered. The 'AutoBed' algorithm works by matching the information obtained from the electronic medical record form with the available bed in real-time using location awareness devices such as RFID, infrared and computer vision. With this innovation, it was found that the waiting time is decreasing by one-hour for more than 50% of incoming emergency room patient. In other words, it took about 7 calls to place a patient, but on the other hand, it is only 1 call with 'AutoBed'.

## IV. Analytics and Its Role in IoT Applications

Having IoT as a platform to generate the data alone is a waste. There is a major obstacle in making this an innovative solution to the healthcare services. It is true that with the IoT technology, fresh data from the patient is coming out and delivered easily to the healthcare provider. However, data without meaningful insight is meaningless. Many have reported that they face challenges when trying to interpret and transform these data into actionable oriented insights [16]. One of the challenges is the business owner or the technology provider need to apply some of the latest techniques of analytics to transform the data into actionable insight [16]. Therefore, the latest appropriate analytics are needed when developing the innovative products or services that will positively affect patient outcomes as well as deliver commercial growth.

It must be noted that it is hard for IoT solution to working efficiently without Analytics. Analytics is a scientific process of transforming data into insights for making better decisions. It is a multidisciplinary field that uses mathematics, statistics, predictive modeling & machine learning techniques to discover meaningful pattern and knowledge. Analytics is a very significant field nowadays that surround today's daily life. Generally there are three types of analytics: Descriptive analytics, predictive analytics and prescriptive analytics [17].

Descriptive analytics is the most basic type. It serve to answer question like "what has happened?". Descriptive analytics analyze the real time incoming and historical data for having insight on how to approach the future. Basic statistics and mathematical techniques are among the most frequently used techniques for descriptive analytics. Figure 4 illustrates a screenshot of dashboard displaying heart rate and sleep data based on the resulted wearable sensors. It must be noted that the result displayed in the dashboard is obtained using descriptive analytics.

On the other hand, predictive analytics is a higher level of analytics that serve to answer question like "what could happen in the future based on previous trends and pattern?". It analyzes past data patterns and trends as well as tries to predict what could happen in the future. Among the frequent techniques that are used for predictive analytics are linear regression, multivariate regression, logistic regression, decision trees, random forest, naïve bayes and k-means. Figure 5 illustrates a screenshot of the Glucose monitoring system that used predictive analytics to predict the glucose level is heading low within 20 minutes.

Prescriptive analytics is an advanced analytics. It serves to answer question like "what should a business do?". Prescriptive analytics could advice on possible outcomes and results in action that is likely to maximize key business metrics. Optimization and simulation is among the famous techniques for prescriptive analytics.

IoT serves as a platform to obtain medical and health related data in healthcare application setting. Combining analytics with IoT will ensure all the relevant data is turned into actionable insight. As a matter of facts, IoT and analytics become a perfect combination for better healthcare solution.

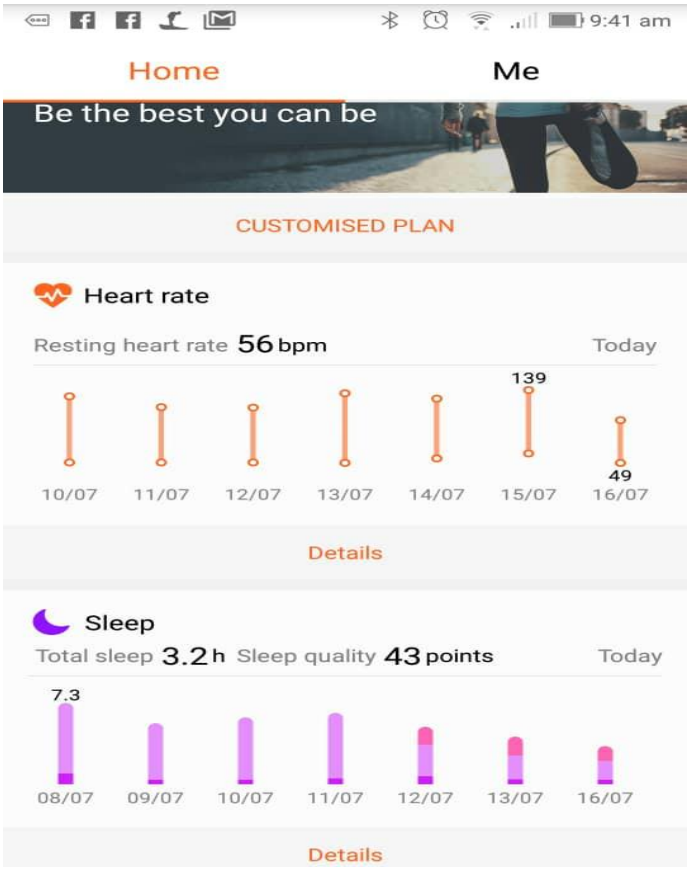



Fig. 4. An Example of Dashboard Displaying Result using Descriptive Analytics.


Fig. 5. An Example of Dashboard Displaying Result using Predictive, Analytics, taken from [13].

## V. Discussion and Conclusion

Huge amount of medical data are available through this third digital wave of technology, which is IoT. It should also be noted that the amount of data from the utilization of this technology is increasing exponentially. As a matter of facts, analytics offers better way to quickly provides relevant insights using this data.

From the healthcare perspective, the wide use of IoT devices such as wearable sensors as well as the mobile data analysis application will transform the patient from 'passive patients' to 'active patients'. Passive patients refer to patient that relies heavily on the doctors or the healthcare providers when it comes to addressing the health issues. On the other hand, active or activated patient refer to patient that take matters into their own hands. Take for example, a person or patient with wearable device such as fitbits, or those with at home monitor for glucose or blood pressure level and also those with mobile applications that transmit health related data. This example of patients will monitor and control their own health instead of just waiting for the regular check-up. Therefore, there is an active participation in the personalized healthcare, which leads to increase engagement in preventive care and also increase effectiveness of health practices.

One of the challenges that might be faced when adopting this technology is to ensure that all patients have equal opportunity to access this technology and get benefit from it. Therefore, it is important for all parties like government, health human resource and technology provider companies to work together for making this technology accessible to all. A solid policy in adopting this technology should be carefully formulated in order to ensure it is beneficial to all.

Although it is projected that the IoT healthcare analytics will soon be the technology that is widely adopted, it should be highlighted here that IoT healthcare analytics would never replace the health human resource. The IoT healthcare analytics technology could never be equal to human. Rather, it only complements the health human resource by providing data from sensors for better diagnoses and treatment plan. Furthermore, it is able to reduce inefficiencies and waste in healthcare system. In some cases, health human resource could focus in more significant responsibilities. Therefore, health human resource should never feel threaten by adopting the technology (like IoT, artificial intelligence, robots etc) as 'human touch' is still a top priority in delivering healthcare services.

## Acknowledgment

Special thanks to Ministry of Education Malaysia for granting a grant TRGS/1/2016/FKP-AMC/01/D00005 and Universiti Teknikal Malaysia Melaka.

## References

[1] "Current world population." [Online]. Available: http://www.worldometers.info/world-population/. [Accessed: 27-Jul-2018].

[2] "Impact of migration on population change," European Union, 2013. [Online]. Available: https://www.espon.eu/sites/default/files/attachments/demifer_PB_migr_impact.pdf.

[3] S. L. Colby and J. M. Ortman, "Projections of the size and composition

of the U.S. population: 2014 to 2060 Population estimates and projections," 2015.

[4] "World population projected to reach 9.8 billion in 2050, and 11.2 billion in 2100," United Nations. [Online]. Available: https://www.un.org/development/desa/en/news/population/world-population-prospects-2017.html. [Accessed: 27-Jul-2018].

[5] D. Ramanayake, V. Giovanni, B. Morgagni, and F. Fi, "Critical review of growing population and climate challenges: How to effects the future social structural changes and world conflict in 2100," Int. J. Dev. Res., vol. 08, pp. 20133–20139, 2018.

[6] "Maslow's Pyramid." [Online]. Available: https://www.vecteezy.com/vector-art/91591-maslow-s-pyramid.

[7] "How digital technology will transform the world: The digital transformation has arrived," Fujitsu Journal, 2016. .

[8] L. Li, "China ' s manufacturing locus in 2025 : With a comparison of " Made-in-China," Technol. Forecast. Soc. Chang., no. May, pp. 0–1, 2017.

[9] S. Madakam, R. Ramaswamy, and S. Tripathi, "Internet of Things ( IoT ): A Literature Review," J. Comput. Commun., vol. 3, no. May, pp. 164–173, 2015.

[10] H. H. Nguyen, F. Mirza, M. A. Naeem, and M. Nguyen, "A review on IoT healthcare monitoring applications and a vision for transforming sensor data into real-time clinical feedback," in 2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2017, pp. 257–262.

[11] V. Velamoor et al., "Unlocking digital health : Opportunities for the mobile value chain," 2015.

[12] "From the person, to the cloud, and back: How remote patient monitoring and intelligent feedback loops are driving the personalization of care," 2017.

[13] "Dexcom Continuous glucose monitoring." [Online]. Available: https://www.dexcom.com/continuous-glucose-monitoring.

[14] "How bed-tracking technology allowed Mt. Sinai Medical Center to reduce admission wait times," 2013. [Online]. Available: https://www.beckershospitalreview.com/healthcare-information-technology/how-bed-tracking-technology-allowed-mt-sinai-medical-center-to-reduce-admission-wait-times.html.

[15] B. G. Thomas et al., "Automated bed assignments in a complex and dynamic hospital environment," Interfaces (Providence)., vol. 43, 2013.

[16] I. Lee and K. Lee, "The Internet of Things ( IoT ): Applications , investments , and challenges for enterprises," Bus. Horiz., vol. 58, no. 4, pp. 431–440, 2015.

[17] S. Ransbotham, D. Kiron, and P. K. Prentice, "Minding the analytics gap," MIT Sloan Manag. Rev., 2015.

# Implementation of Forward Chaining and Certainty Factor Method on Android-Based Expert System of Tomato Diseases Identification

Kurnia Muludi[1]
Department of Computer Science
Lampung University
Lampung, Indonesia

Radix Suharjo[2]
Department of Plant Protection
Lampung University
Lampung, Indonesia

Admi Syarif[3], Fitria Ramadhani[4]
Department of Computer Science
Lampung University
Lampung, Indonesia

***Abstract*—Plant disease is one of the reasons that cause the destruction of plant. It affects plant productivity and quality. Most of the farmers made mistake in cope with this problem because of the lack of knowledge. Expert system is a solution that has been widely used for identifying disease. This paper presents an Android-based expert system to help identifying tomato diseases. Data used in this expert system consist of 16 data of tomato diseases, 53 data of symptoms, and 20 variety of rules. This paper implements forward chaining and certainty factor method. Forward chaining is used as a reasoning method to get the result of disease identification. Certainty factor is used as a calculation method to obtain accuracy degree of identification results. Testing has been done through two stages, internal and external. The result from internal testing shows that tomato expert system works properly and fit perfectly in various android devices. External testing is done by giving questionnaire to 44 respondents. The result of questionnaires shows that tomato expert system is categorized as "good" by them.**

***Keywords*—Expert system; forward chaining; certainty factor; tomato diseases; android**

## I. Introduction

Tomato is a commodity plant that is being cultivated by many farmers in Indonesia due to its benefit. Tomato can be categorized as vegetable and fruit. Tomato is also used as food ingredient and medicine [1]. Based on statistical data of vegetable productivity in Indonesia in 2016, tomato becomes big seven along with onion, garlic, and chili with 883.233 ton[a]. Tomato becomes one of horticulture commodities that has high economic value yet still needs more concern, especially in production and quality enhancement [2]. This matter is showed by statistical data of tomato in Lampung province between 2014 until 2016 in Table I[a]. Crop areas increase twice in 2016 yet is not followed by production result. So, the productivity decreases treble with only 4,54 Ton/Ha.

Plant disease is one of the most important reasons that lead to the destruction of plants and crops [3]. Detecting diseases at early stage enable people to overcome and treat them appropriately. This process requires an expert but the number of experts are still far from ideal condition. Moreover, it needs high cost to do consultation. Therefore, expert system helps a great deal in identifying diseases and describing methods of treatment to be carried out.

TABLE I. Table of Tomato Statistical Data in Lampung Province between 2012 - 2016

| Year | Crop area (Ha) | Production (Ton) | Productivity (Ton/Ha) |
|---|---|---|---|
| 2014 | 2.366 | 23.776 | 10,05 |
| 2015 | 2.143 | 24.490 | 11,43 |
| 2016 | 5.211 | 23.638 | 4,54 |

This paper develops an expert system for identifying tomato diseases implementing forward chaining and certainty factor method. These methods are used to get identification results with accuracy degree. This expert system is designed to be android based, so that it can be used by many people anywhere and anytime.

## II. Literature Review

### A. Artificial Intelligence

Artificial Intelligence (AI) is the area of computer science that focus on creating expert machines that can engage on behaviors that humans consider intelligent [4]. It has many areas in computer science. There are 12 fields in AI, such as Natural Language Processing (NLP), Speech (voice) Understanding, Robotics and Sensory Systems, and Expert Systems [5].

### B. Expert System

Expert system is a field in artificial intelligence intended to serve as consultant for decision making. It uses a collection of facts, rules of thumb, and other knowledge about a limited domain to help making inferences in the domain. It is called expert systems because it addresses problems normally thought to require human specialists for solution [6]. Expert systems can be viewed as having two environments: the development environment and the consultation environment. The development environment is used by an expert system builder to build the components and put knowledge into the knowledge base. The consultation environment is used by a non-expert to obtain expert knowledge and advice. The structure of expert system is showed in Fig 1 [5].

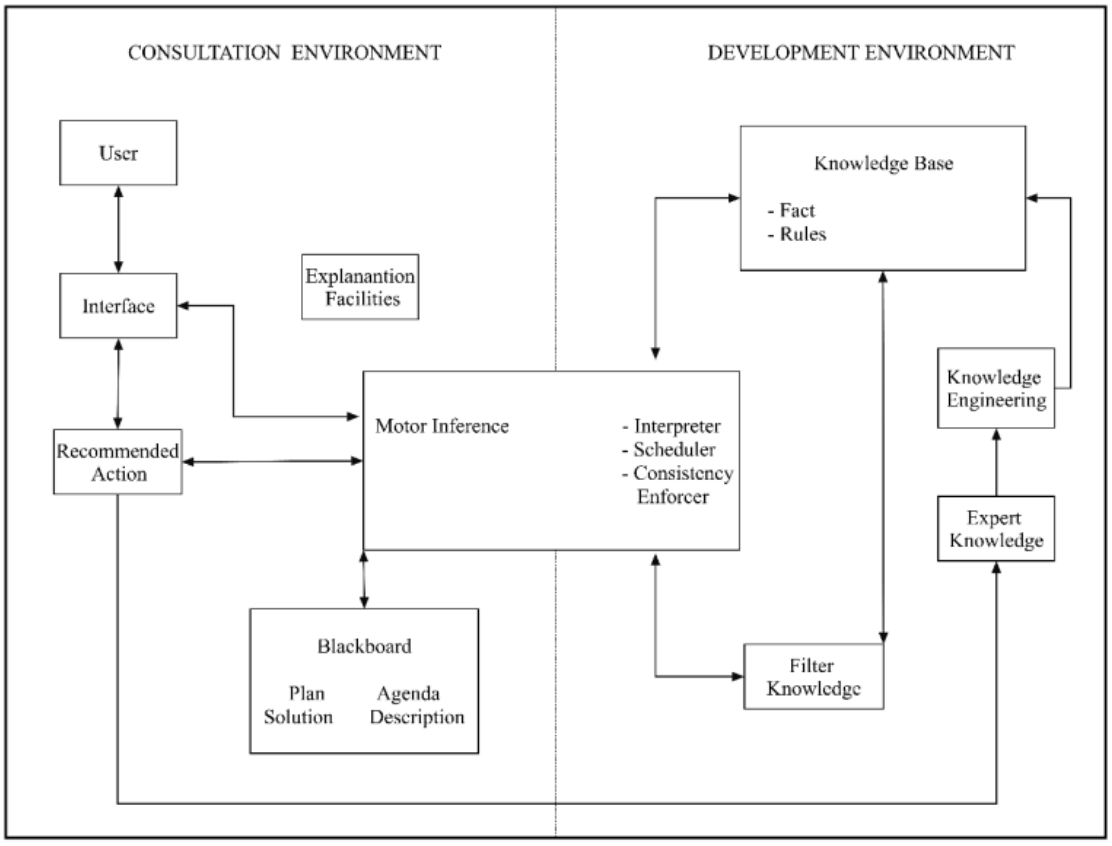



Fig. 1. The Structure of Expert System.

### C. Inference Engine

Expert knowledge often can not be represented in single rule. Instead, rules may be chained together dynamically to cover numerous conditions. The process of chaining multiple rules together is called inference. The component conducting inference in an expert system is called the inference engine. Inference engine provides a methodology for reasoning about information in the knowledge base and on the blackboard, and for formulating conclusions. This component provides directions about how to use the system's knowledge by developing the agenda that organizes and controls the steps taken to solve problems whenever consultation takes place. Therefore, inference engine is controller structure or rule interpreter, also is known as brain in expert system. There are two popular approaches for inferencing, such as forward chaining and backward chaining [5].

Forward-chaining is a bottom-up computational model. Forward chaining works with a set of known facts and applies rules to generate new facts whose premises match the known facts, and continues until it reaches a predetermined goal, or until no further facts can be derived whose premises match the known facts. Backward chaining is a reasoning model in expert system that using a goal-driven search. It is a top-down computational design and starts with a goal or hypothesis and then looks for rules to support the hypothesis. It attempts matching the variables that lead to valid facts in the data and indicates that the inference moves backward from the intended goal to determine facts that would satisfy that goal [7].

Forward chaining is appropriate for problems in which most of the data is given in the initial problem statement as how this model works. Forward chaining works forward to find conclusion from facts, starts with the initial facts. Meanwhile, backward chaining starts with hypothesis or goal to know whether the hypothesis accepted or rejected. Backward chaining is appropriate for problem in which goal or hypothesis is given in the problem statement [7]. In Conclusion, forward chaining is used as a reasoning because the system is identifying diseases based on symptoms or facts.

### D. Certainty factor

To develop an expert system calculation method is needed to get accuracy degree of the identification results. There are some calculation methods used in expert system, such as heuristic method, classical probability, bayes theorem, and certainty factor.

Heuristic is a method introduced by George Polya in 1945. He defines heuristic as rules for choosing branches in a state space that are most likely to lead to an acceptable problem solution. This method has weakness because it often works based on experience or intuition, so the result maybe inaccurate. Also, it uses limited information and seldom able to predict the exact behavior of the state space farther along in the search [8].

Theory of classical probability was proposed by Pascal and Fermat in 1654. Classical probability considers ideal system that do not become worn out like dice, card, and coin.

Classical probability only considers probability that will be searched, as showed by the following formula:

$$P = W / N \quad (1)$$

Where W is the number of wins and N is the number of equally possible events [9].

Bayes theorem relates cause and effect in such a way that by understanding the effect people can learn the probability of its cause. This theory was proposed by Thomas Bayes in 1763 in his paper *Essay towards Solving a Problem in the Doctrine of' Chances* [8]. Bayes theorem and classical probability requires the application of mathematical probability in uncertainty management. Bayes theorem is better than classical probability because it has prior value. Unfortunately, to obtain past data is becoming problem for expert. As a result, if prior data is lacking then it will affect the conclusion, even can give wrong conclusion [10].

One of the most known expert systems attempts to develop representations of uncertainty as an extension of deterministic rule-based expert systems is MYCIN. MYCIN was designed to assist physicians in the diagnosis and treatment of bacterial infections. Shorliffe Buchanan in 1975 introduced certainty factor (CF) method through MYCIN. When a CF-based system recommends a decision (for example, when MYCIN suggests treatment for a suspected infection), it makes use of the CFs assigned to the competing diseases to assess the amount of evidence for each [11]. Here is the formula of CF [12]:

$$CF(H,E) = MB(H,E) - MD(H,E) \quad (2)$$

Description :

*CF(H,E)* : Certainty factor from hypothesis H which is affected by symptom (evidence) E

*MB(H,E)* : Measurement of increased belief on hypothesis H affected by symptom E

*MD(H,E)* : Measurement of increased disbelief on hypothesis H affected by symptom E

*1) Parallel CF*: Parallel CF is a CF obtained by some premises from rule. The value of parallel CF is affected by CF User for each premise and operator used in the premise. Here is the formulas for parallel CF [12]:

$$CF\ (x\ AND\ y) = \mathbf{min}\ (CF\ (x), CF\ (y)) \quad (3)$$

$$CF\ (x\ OR\ y) = \mathbf{max}\ (CF\ (x), CF\ (y)) \quad (4)$$

$$CF\ (NOT\ x) = \neg CF(x) \quad (5)$$

*2) Sequential CF:* Sequential CF is obtained from calculation result of parallel CF of all premises in a rule with CF from expert. Sequential CF is defined as follows [12]:

$$CF\ (x,y) = CF(x) \times CF(y) \quad (6)$$

Description :

*CF (x,y)* : Sequential CF

*CF(x)* : Parallel CF from all premises in a rule

*CF(y)* : CF from expert

*3) Combined CF*: Combined CF is final CF from conclusion candidate. This CF is affected by all parallel CF from a rule that will produce conclusion. Combined CF is used when a conclusion obtained from several rules at once. Here is the formula for Combine CF [12]:

$$CF(x,y) = \begin{cases} CF(x)+CF(y)-(CF(x)\times CF(y)), \text{ where } CF(x)>0 \text{ and } CF(y)>0 & (7) \\ CF(x)+CF(y)/(1-(\mathbf{min}(|CF|,|CF(y)|))), \text{ where one of } (CF(x),CF(y))<0 & (8) \\ CF(x)+CF(y)\times(1+CF(x))), \text{ where } CF(x)<0 \text{ and } CF(y)<0 & (9) \end{cases}$$

## E. Expert Systems in Agriculture

Expert system has been implemented in many areas, such as medical, finance, automotive, and agriculture. During this five years, many expert systems in agriculture were developed, like tobacco expert system, sugarcane expert system, and red chili peppers expert system. CaneDES is a web-based expert system for disorder diagnosis in sugarcane. This expert system diagnoses disorders in sugarcane which is caused by biotic and abiotic and implements a new classified knowledge representation technique [13]. Another expert system that has been developed is an expert system for tobacco disease management that is applied in web. This tobacco expert system uses two methods for disease diagnosis, such as rule based inference and ontology [14]. Expert system for diagnosing diseases of red chili uses forward chaining and certainty factor method. Based on testing results, this expert system can diagnose red chili plant disease with high accuracy [15].

## F. Android

Nowadays web-based systems which can be accessed on the PC or laptop is slowly getting unpopular. People move to palm tools gadget like smartphone which makes activity more comfortable and efficient. So that, every business is developing applications for the mobile devices. One of the most popular operating systems used in mobile device is android [16]. Android is a software stack for mobile device that includes an operating system, middle-ware and key applications. Android is mainly based on Linux operating system which uses java-like languages for running applications [17]. The platform was officially announced in October 2008 [18].

## G. Diseases of Tomato

In this research, tomato diseases are divided into two categories, biotic and abiotic. Biotic disease is caused by organism such as fungus, virus, and bacteria. While, abiotic disease is caused by environmental factors like pesticide, less of calcium, and sunshine intensity. There are 12 diseases in biotic category, such as Late Blight (*Phythopthora infestans*), early blight (*Alternaria* solani), leaf mold (*Fulvia fulva*), Fusarium wilt (*Fusarium oxysporum f.sp. lycopersici*), Bacteria wilt (*Ralstonia solanacearum*), Tobacco Mosaic (Tobacco Mosaic Virus/TMV), Cucumber Mosaic (Cucumber

Mosaic Virus/CMV), crinkle (Tobacco Crinkle Virus), Tomato Yellow Leaf Curl (Tomato Yellow Leaf Curl Virus/TYCLV), gray mold (*Cercospora sp.*), powdery mildews (*Oidium sp.)* [19]. There are 4 diseases in abiotic category, such as pesticide poisoning, lack of calcium, high intensity of sunshine, and low intensity of sunshine [20].

## III. System Design

### A. Use Case Diagram

Use case diagram describes interaction between user and system, like what action user can do in this system. This diagram shows that tomato expert system has 5 functions as shown in Fig 2, such as do diagnose, see disease's detail information, access help menu, access about menu, and access instruction menu.

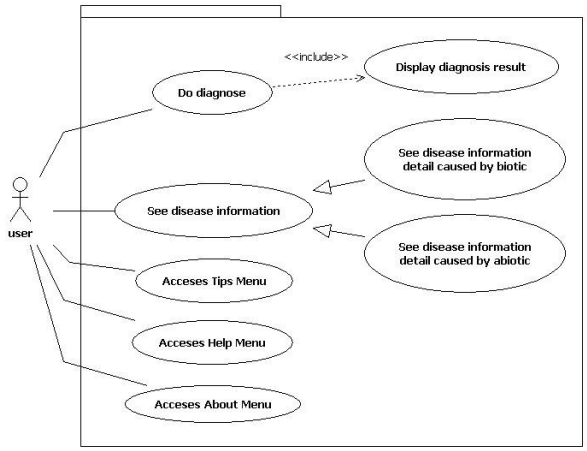



Fig. 2. Use Case Diagram of Tomato Expert System.

### B. Calculation of Certainty Factor Method

Calculation method implemented in this expert system is certainty factor. In developing expert system, rule-based is made into two types operator, OR and AND. Operator AND is used for main symptoms. Meanwhile, symptoms using operator OR is displayed in list form. For example, user done a consultation on the tomato leaf with the following symptoms.

- There's spot with diameter 0,5 cm (G22)
- There's gray-brownish ring on the spot (G23)
- There's brown-reddish mold on the spot (G24)

Expert gives CF value 0.79 for the selected symptoms. The rule: "**IF** (There's spot with diameter 0,5 cm (G22) **OR** There's gray-brownish ring on the spot (G23)) **AND** There's brown-reddish mold on the spot (G24) **THEN** Corynespora Spot Leaf (B12)".

Then the calculation is:

*Parallel CF = CF (G24) AND CF (G22 OR G23)*
*= **min** (G24, **max** (G22, G23))*
*= **min** (0.83, **max** (0.79, 0.81))*
*= **min** (0,83, 0.81)*
*= 0.81*

Sequential CF = *CF from Expert* × *Parallel CF*
*= 0.79 × 0.81 = 0.6399*

Percentage of accuracy = 63,99 %

Based on the calculation result, the conclusion for following symptoms chosen by user is Corynespora Spot Leaf with the accuracy 63,99%.

## IV. Testing and Results

The system was developed by using Android Studio, XAMPP on TOSHIBA Satellite L840 and android smartphone OPPO F1s A1601. Two stages of testing, internal and external, is done to see the effectiveness of the system. Internal testing is done by testing system's compatibilty and accuracy degree of identification results. To see whether the system is compatible or not, the system is installed in several android devices with different version and screen size. As a result, the system is compatible with Android 4.4 (Kitkat), 5.0 (Lollipop), 6.0 (Marsmallow), and 7.0 (Nougat) version and with android 4,5 inch, 5 inch, and 5,5 inch. The system's function works properly and the interface fits perfectly as shown in Fig 3.

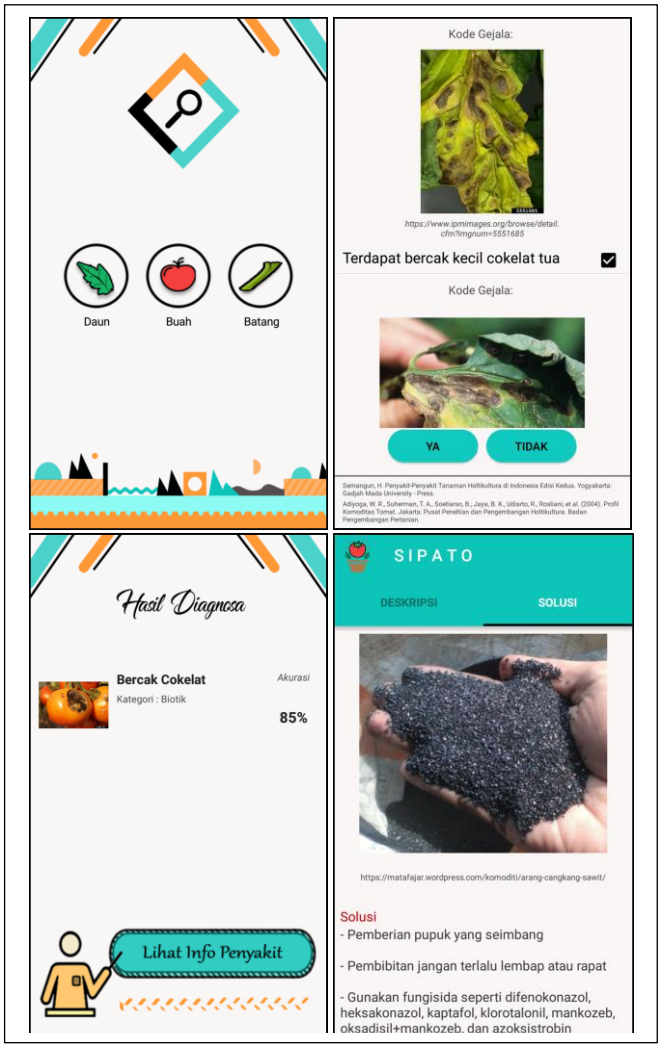



Fig. 3. Interface of Tomato Expert System.

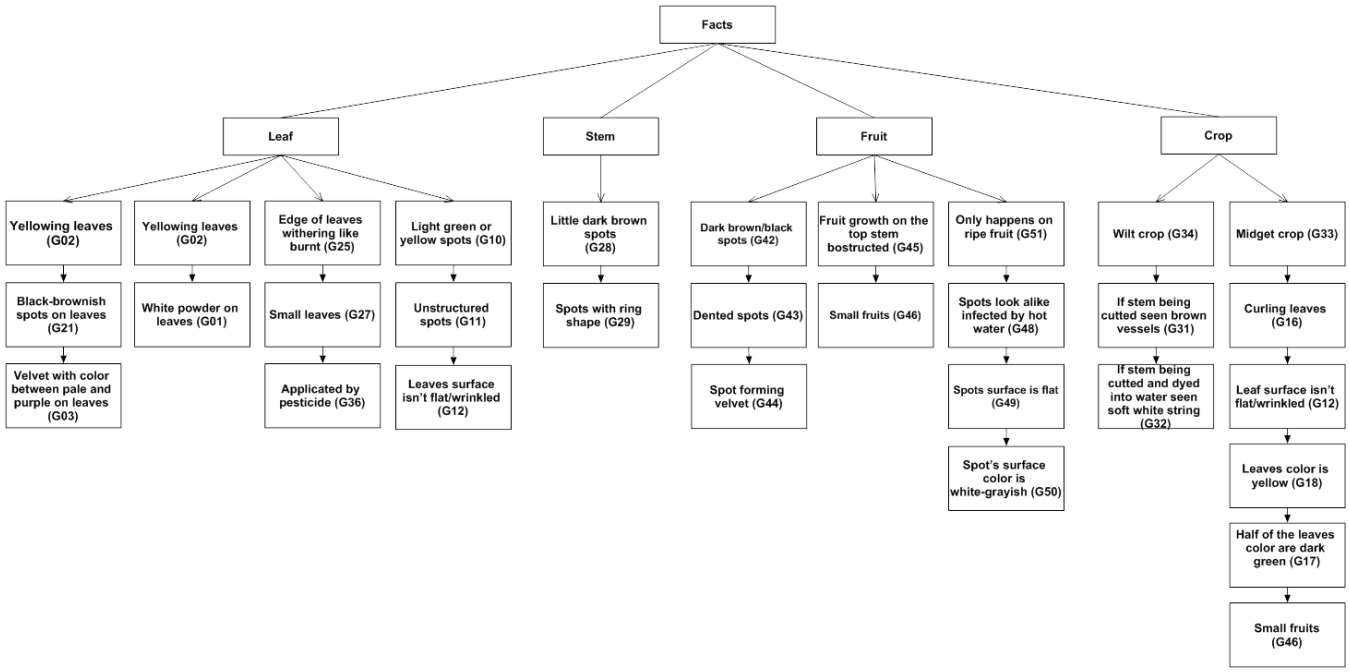



Fig. 4. Facts of Symptom.

This paper provides some facts and compares identification results between the system and expert. It use 33 facts divided into 4 symptom locations, such as leaf, stem, fruit, and crop as shown in Fig 4.

Testing result shows that system gives the accurate identification results. Also, there's degree of accuracy from implementation of certainty factor method. Comparison of identification results is showed in Table II.

TABLE II. TABLE OF COMPARISON IDENTIFICATION RESULTS BETWEEN TOMATO EXPERT SYSTEM AND EXPERT

| **Facts** | **Diagnosis** | |
|---|---|---|
| | ***System Calculation*** | ***Expert Judgement*** |
| G02, G21, G03 | Late Blight : 66,4% | Late Blight |
| G02, G01 | Powdery Mildews : 61,62% | Powdery Mildews |
| G25, G27, G36 | Pesticide Poisoning : 62,37% | Pesticide Poisoning |
| G10, G11, G12 | Tobacco Mosaic : 63,99% | Tobacco Mosaic |
| G28, G29 | Early Blight : 65,6% | Early Blight |
| G42, G43, G44 | Early Blight : 66,42% | Early Blight |
| G45, G46 | Cucumber Mosaic : 57,75% | Cucumber Mosaic |
| G51, G48, G49, G50 | High intensity of sunshine : 65,57% | High intensity of sunshine |
| G34, G31, G32 | Bacteria Wilt : 71,4% | Bacteria Wilt |
| G33, G16, G12, G18, G19, G46 | Yellow Leaf Curl : 65,57% | Yellow Leaf Curl |

External testing is done by giving questionnaire to several respondents. The respondent is divided into 3 groups. Group 1 is experts consist of 4 lecturers from Department of Plant Protection Lampung University. Group 2 is people who are related with this domain consist of 10 farmers and 20 students of Agriculture Faculty of University of Lampung. Group 3 is common people who do not understand about tomato diseases consist of 10 students of Computer Science of University of Lampung.

The respondents were asked to access the system directly and give their assessment by completing the questionnaire. Each statement has 5 answer points with different value, such as Completely Disagree (CD) = 1, Disagree (D) = 2, Slightly Agree (SA) = 3, Agree (A) = 4, and Completely Agree (GA) = 5. Then, those questionnaires being generated into percentage with this following formula [21]:

$$P = (((a \times A)/(nxN) \times 100\%) \quad (10)$$

$P$ = Percentage of statement

$a$ = Number of respondents choosing this point

$A$= Value of statement

$n$ = Number of all respondents

$N$ = Highest value of statement

Next step is deciding interval for each category using this following formula [21]:

$$I = 100\% / k \quad (11)$$

Description :

$I$ = Interval

$K$ = Number of categories

TABLE III. TABLE OF INTERVAL CATEGORY

| Percentage | Description |
|---|---|
| 0% $\geq P \leq$ 20% | Not Good |
| *20% > P ≤ 40%* | Less Good |
| *40% > P ≤ 60%* | Good Enough |
| *60% > P ≤ 80%* | Good |
| *80% > P ≤ 100%* | Very Good |

This research uses 5 categories, so interval for each category is 20%. The description for each category is showed in Table III.

Based on the questionnaires result, it shows that this expert system provides useful information for people especially tomato farmers with 87,33%. It helps in identifying diseases well with 79,44%. System works properly and appropriate with 84%. System is easy to understand and to use with 82,77%. Data in the system is complete and clear with 74,44. It makes people understand about expert system with 79,11%. System provides right solution with 76,11%. Graphic for external testing result is showed in Fig 5.

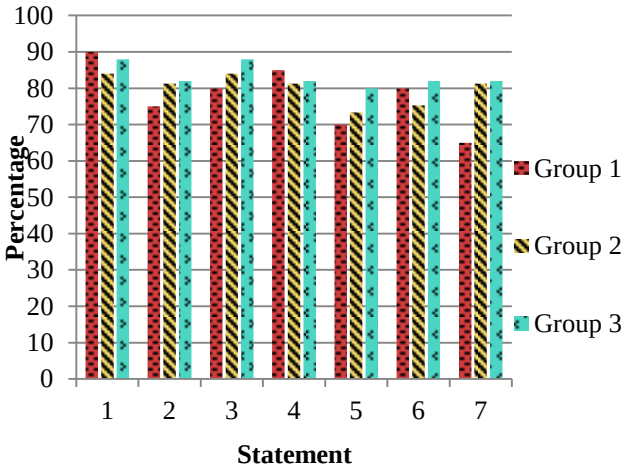



Fig. 5. Graphic of External Testing Result.

## V. CONCLUSIONS

An Android-based tomato expert system implements forward chaining and certainty factor method to identifying tomato diseases has been established. The system which has been installed on various android devices shows the accurate identification result as it is expert. Moreover, all functions work properly and the interface fits perfectly. The result from 44 respondents shows that this tomato expert system is categorized as “good”.

## REFERENCES

[1] Bambang Cahyono, *Technique of Cultivating Superior Tomato with Organic and Anorganic Method*. Depok: Pustaka Mina, 2016.

[2] Roydatul Zikria, *Tomato Commodity Outlook 2014*. Jakarta: Agriculture Data and Information System Center, 2014.

[3] S S Abu-Naser, K A Kashkash, and M Fayyad, "Developing an Expert System for Plant Disease Diagnosis," *Journal of Artificial Intelligence*, vol. 1, no. 2, pp. 78-85, 2008.

[4] P Santosh Kumar Patra, Dipti Prava Sahu, and Indrajit Mandal, "An Expert System for Diagnosis of Human Diseases," *International Journal of Computer Applications*, vol. 1, no. 13, pp. 71-73, 2010.

[5] Efraim Turban, Jay E Aronson, and Ting- Peng Liang, *Decision Support Systems and Intelligent Systems*. New Jersey: Prentice-Hall, Inc, 2005.

[6] Herbert E Rauch, "Probability Concepts For An Expert System Used For Data Fusion," in *Seventeenth Annual Asilomar Conference on Circuits, Systems and Computers*, California, 1983, pp. 55-60.

[7] Ajlan S Al-Ajlan, "The Comparison between Forward and Backward," *International Journal of Machine Learning and Computing*, vol. 5, no. 2, pp. 106-113, 2015.

[8] George F Luger, *Artificial Intelligence Structures and Strategies for Complex Problem Solving*. England: Pearson Education Limited, 2005.

[9] Joseph Giarratano and Gary Riley, *Expert Systems Principles and Programming*. U.S.A: PWS Publishing Company, 1998.

[10] Yeow Wei Liang and Rohana Mahmud, "A Comparison Model for Uncertain Information in Expert System," *International Conference on Uncertainty Reasoning and Knowledge Engineering*, pp. 127-130, 2012.

[11] Eric J Horvitz, John S Breese, and Max Henrion, "Decision Theory in Expert Systems and Artificial Intelligence," *International Journal of Approximate Reasoning*, vol. 2, pp. 247-302, 1988.

[12] Kusrini, *Expert System Application Deciding User Certainty Factor with Quantification Question Method*. Yogyakarta: ANDI, 2008.

[13] S S Hasan et al., "CaneDES: A Web-Based Expert System for Disorder Diagnosis in Sugarcane," *Sugar Tech: An International Journal of Sugar Crops and Related Industries*, vol. 17, no. 4, pp. 418-427, 2015.

[14] H Ravisankar, K Sivaraju, D Damodar Reddy, and N Srinivasa Rao, "Web Based Expert System for Tobacco Disease," *Journal of Entomology and Zoology Studies*, vol. 6, no. 1, pp. 5-11, 2018.

[15] Fahrul Agus, Hernandha Eka Wulandari, and Indah Fitri Astuti, "Expert System With Certainty Factor For Early Diagnosis Of Red Chili Peppers Diseases," *Journal of Applied Intelligent System*, vol. 2, no. 2, pp. 52-66, 2017.

[16] Chetan Bulla, Sachin Bakanetti, Jayant Bhosale, Kiran Patil, and Poornima Gujanal, "My Campus Android Application," *International Journal of Engineering Science and Computing*, vol. 7, no. 6, pp. 12631-12632, 2017.

[17] Mohamed Abd El-Latif Mowad, Ahmed Fathy, and Ahmed Hafez, "Smart Home Automated Control System Using Android Application and Microcontroller," *International Journal of Scientific & Engineering Research*, vol. V, no. 5, pp. 935-939, 2014.

[18] Suhas Holla and Mahima M Katti, "Android Based Mobile Application Development and Its Security," *International Journal of Computer Trends and Technology*, vol. 3, no. 3, pp. 486-490, 2012.

[19] H Semangun, *Horticulture plant diseases in Indonesia*, 2nd ed. Yogyakarta: Gadjah Mada University - Press, 2007.

[20] W Setiawati, I Sulatrini, and N Gunaeni, *Implementation of PHT Technology on Tomato Plant*. Bandung: Plant Vegetable Research Office, 2001.

[21] Djarwanto, *Knowing Some of Statistical Test in Research*. Yogyakarta: Liberty, 1996.

# Applying Floyd's Inductive Assertions Method for Verification of Generalized Net Models Without Temporal Components

Magdalina Todorova[1], Nora Angelova[2]
Faculty of Mathematics and Informatics
Sofia University "St. Kl. Ohridski"
Sofia, Bulgaria

***Abstract*—Generalized Nets are extensions of Petri Nets. They are a suitable tool for describing real sequential and parallel processes in different areas. The implementation of correct Generalized Nets models is a task of great importance for the creation of a number of applications such as transportation management, e-business, medical systems, telephone networks, etc. The cost of an error in the models of some of these applications can be very high. The implementation of models of similar applications has to use formal approaches to prove that the developed models are correct. A foundation stone of software verification, which is suitable for verification of Generalized Nets models with transitions without temporal component, is Floyd's inductive assertion method. This article presents a modification of Floyd's inductive assertion method for verification of flowcharts, which allows Generalized Nets without temporal component to be verified. Using an illustrative example, we show that the offered adaptation is appropriate for the purpose of training university students in the Informatics and Computer Sciences in formal methods of verification.**

***Keywords*—Floyd's inductive assertions method; generalized nets; verification; formal methods; education**

## I. Introduction

Generalized Nets (GNs) [1, 2] are a means of modeling sequential and parallel processes in a variety of areas, including medicine, industry, transport, software protection, etc. They were introduced in 1982 by Krassimir Atanassov as a further extension of the standard Petri Nets (PNs) and their modifications and extensions. GNs are defined in a way that is fundamentally different from the ways of defining Regular PNs, E-nets, Time PNs, Colored PNs, Self-modifying PNs, Stochastic PNs, Predicate-transition nets, and other PNs. In the 1980s, it was proved that the functioning and the results of the work of each of these types of nets can be described by a GN. Moreover, it was proved that for each of the classes of standard or extended PNs, there exists a GN that is universal for this class, i.e., it represents the functioning and the results of the work of each of the elements of the respective class of nets. In the following years, similar results were published for the Super nets, Numerical PNs, Fuzzy PNs, and others PNs.

Automatic tools for execution of GNs [3–10] have been developed and are currently in process of improvement.

In parallel with the scientific research related to GN, a great number of GN models have been developed, which simulate real-life processes. Designing such GN models is useful as researching their characteristics allows for focusing the attention of the real-life system developers on the most important (from the point of view of performance quality) elements, as well as to eliminate the unnecessary details when realizing the real-life systems. After developing a GN model, which presents the behavior of a real-life system, research should be conducted to discover how adequate is the developed model, as evaluated against the criteria for the respective real-life system.

Checking the model adequacy is done in two steps: model verification (if it meets the requirements), and model validation (if the requirements posed to the model are adequate to the real-life system). For some models, such a check is obligatory to be done.

This article presents a method of formal verification of GNs, which are without temporal components. The method is a modification of Floyd's inductive assertion method for verification of flowcharts.

The rest of the paper is organized as follows. Part II presents the definitions of a GN, of its main component – the transition, as well as of the GN loop. Part III is dedicated to an adaptation of Floyd's method of verification of flowcharts for verification of GNs with transitions without temporal components. As a result, methods for proving the partial correctness and the termination of such GNs are proposed. In Part IV, these methods are illustrated by a simple example. Applying Floyd's method for verification of practically applicable GN models is a subject of another paper due to volume constrains. Part V of the paper provides comments on applying the presented method for GN verification and ideas for further research in the field.

## II. Generalized Nets

GNs are defined in a way that [1, 2] is fundamentally different from the ways of defining the other types of PNs.

*Definition* 1 (*Transition*). Every transition is described by a seven- tuple:

$$Z = \langle L', L'', t_1, t_2, r, M, \Box \rangle,$$

where:

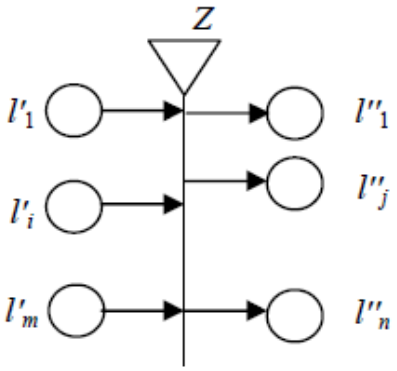



Fig. 1. GN Transition

- $L'$ and $L''$ are finite, non-empty sets of places (the transition's input and output places, respectively); for the transition $Z$ in Fig. 1 these are: $L' = \{l'_1, l'_2, ..., l'_m\}$ and $L'' = \{l''_1, l''_2, ..., l''_n\}$;
- $t_1$ is the current time-moment of the transition's firing;
- $t_2$ is the current value of the duration of its active state;
- $r$ is the transition's condition, determining which tokens will transfer from the transition's inputs to its output places. The parameter $r$ has the form of an Index Matrix (IM) [1, 11]:

| $r$ | $l''_1$ | ... | $l''_j$ | ... | $l''_n$ |
|---|---|---|---|---|---|
| $l'_1$ | | | | | |
| ⋮ | | | $r_{i,j}$ | | |
| $l'_m$ | | | ($r_{i,j}$ – predicate) | | |
| | | | ($1 \le i \le m, 1 \le j \le n$) | | |

where $r_{i,j}$ is the predicate which gives the condition for transfer from the $i$-th input place to the $j$-th output place. When $r_{i,j}$ has truth-value "*true*", then a token from the $i$-th input place can be transferred to the $j$-th output place; otherwise, this is impossible;

- $M$ is an IM of the capacities of transition's arcs:

$M =$

| | $l''_1$ | ... | $l''_j$ | ... | $l''_n$ |
|---|---|---|---|---|---|
| $l'_1$ | | | | | |
| ⋮ | | | $m_{i,j}$ | | |
| $l'_i$ | | | ($m_{i,j} \ge 0$ – natural number or $\infty$) | | |
| ⋮ | | | ($1 \le i \le m, 1 \le j \le n$) | | |
| $l'_m$ | | | | | |

- □ is called transition type. It is an object having a form similar to a Boolean expression. It may contain as variables the symbols that serve as labels for transition's input places. It is an expression consisting of variables and the Boolean connectives $\wedge$ and $\vee$ determining the following conditions:

$\wedge\left(l_{i_1}, l_{i_2}, ..., l_{i_u}\right)$ – every place $l_{i_1}, l_{i_2}, ..., l_{i_u}$ must contain at least one token,

$\vee\left(l_{i_1}, l_{i_2}, ..., l_{i_u}\right)$ – there must be at least one token in all places $l_{i_1}, l_{i_2}, ..., l_{i_u}$, where $\{l_{i_1}, l_{i_2}, ..., l_{i_u}\} \subset L'$. When the value of a type □ (calculated as a Boolean expression) is "*true*", the transition can become active, otherwise it cannot.

*Definition* 2 (*Generalized Net*) [1, 2]: The ordered four-tuple:

$$E = \langle\langle A, \pi_A, \pi_L, c, f, \theta_1, \theta_2\rangle, \langle K, \pi_K, \theta_K\rangle, \langle T, t^o, t^*\rangle, \langle X, \Phi, b\rangle\rangle$$

is called a Generalized Net if:

- $A$ is a set of transitions (Definition 1);
- $\pi_A$ is a function giving the priorities of the transitions, i.e., $\pi_A : A \to N$, where $N$ is the set of natural numbers;
- $\pi_L$ is a function giving the priorities of the places, i.e., $\pi_L : L \to N$, where L = $pr_1 A \bigcup pr_2 A$. Naturally, $L$ is the set of all GN-places;
- $c$ is a function giving the capacities of the places, i.e., $c : L \to N$;
- $f$ is a function that calculates the truth values of the predicates of the transition's conditions. For the (ordinary) GNs, described in this section, the function $f$ obtains values "false" or "true", or values from set {0, 1}. If $P$ is the set of the predicates used in a given model, then $f$ can be defined as $f : P \to \{0, 1\}$;
- $\theta_1$ is a function giving the next time-moment for which a given transition $Z$ can be activated, i.e., $\theta_1(t) = t'$, where $pr_3\, Z = t$, $t' \in [T, T + t^*]$ and $t \le t'$. The value of this function is calculated at the moment when the transition terminates its functioning;
- $\theta_2$ is a function giving the duration of the active state of a given transition $Z$, i.e., $\theta_2(t) = t'$, where $\mathrm{pr}_4 Z = t \in [T, T + t^*]$ and $t' \ge 0$. The value of this function is calculated at the moment when the transition starts functioning;
- $K$ is the set of the GN's tokens;
- $\pi_K$ is a function giving the priorities of the tokens, i.e., $\pi_K : K \to N$;

- $\theta_K$ is a function giving the time-moment when a given token can enter the net, i.e., $\theta_K(\alpha) = t$, where $\alpha \in K$ and $t \in [T, T + t^*]$;
- $T$ is the time-moment when the GN starts functioning. This moment is determined with respect to a fixed (global) time-scale;
- $t^o$ is an elementary time-step, related to the fixed (global) time-scale;
- $t^*$ is the duration of the GN functioning;
- $X$ is a function which assigns initial characteristics to each token when it enters input places of the net;
- $\Phi$ is the characteristic function which assigns new characteristics to each token when it transfers from an input to an output place of a given transition;
- $b$ is a function giving the maximum number of characteristics a given token can receive, i.e.,

$$b : K \rightarrow N.$$

It can be concluded that similarities between PNs and GNs exist, however, there are also differences. The GN transitions have a more complex structure of that of the PN ones. It must be noted that the GN transition contains: an index matrix with predicates that determine whether a token from $i$-th input place can go to the $j$-th output place; an index matrix with natural numbers that determine the capacities of the arc between $i$-th input and $j$-th output place; and a special condition, that determines whether the transition can be activated. The GN definition is also more complex than the definition of a PN. The GN-tokens enter the net with initial characteristics, determined by the characteristic function $X$. Upon entering a new place, the GN-tokens obtain new characteristics, defined by the characteristic function $\Phi$. In contrast to the Colored PNs and the Predicate-Transition Nets, the GN-tokens can keep all their characteristics and they can be used for evaluation of the truth-values of the transition condition predicates [1, 2].

*Definition* 3 (*GN loop*). A sequence of places of a GN, which a given token can go through sequentially and can reach the starting position, is called a *GN loop*.

## III. Verification of GNs, based on Floyd's Method of Verification of a Flowchart

The method presented here can be used for verification of GN models without temporal component (without the components: $\theta_1$, $\theta_2$, $\theta_K$, $T$, $t^o$ and $t^*$). Respectively, the GN transitions do not contain the temporal components $t_1$ and $t_2$. The reason for this is that this restriction and the existence of index matrices with predicates and characteristic functions cause each component of a GN to have a respective segment as in a flowchart. The GNs, which will be verified through a technique following Floyd's method for verification of flowcharts [12], consist of the components given in Fig. 2. In order to increase the readability, we will use the notations for Floyd's method for verification of flowcharts.

Three types of characteristics of the GN (grouped as three vectors) are distinguished:

- An *input* vector $\bar{x} = (x_1, x_2, ..., x_n)$, which sets the values of the initial characteristics $x_1, x_2, ..., x_n$ of the GN tokens.
- An *intermediate* vector $\bar{y} = (y_1, y_2, ..., y_m)$, which is used as temporary storage during the GN execution and describes the changing values of the characteristics $y_1, y_2, ..., y_m$ of the GN tokens.
- An *output* vector $\bar{z} = (z_1, z_2, ..., z_k)$, that yields the values of output characteristics $z_1, z_2, ..., z_k$ of tokens when the GN execution terminates.

These characteristics are specific for each GN model under verification.

Three types of non-empty domains are distinguished as well:

- an input domain $D_{\bar{x}} = D_{x1} \times D_{x2} \times \ldots \times D_{xn}$;
- an intermediate domain $D_{\bar{y}} = D_{y1} \times D_{y2} \times \ldots \times D_{ym}$ and
- an output domain $D_{\bar{z}} = D_{z1} \times D_{z2} \times \ldots \times D_{zk}$

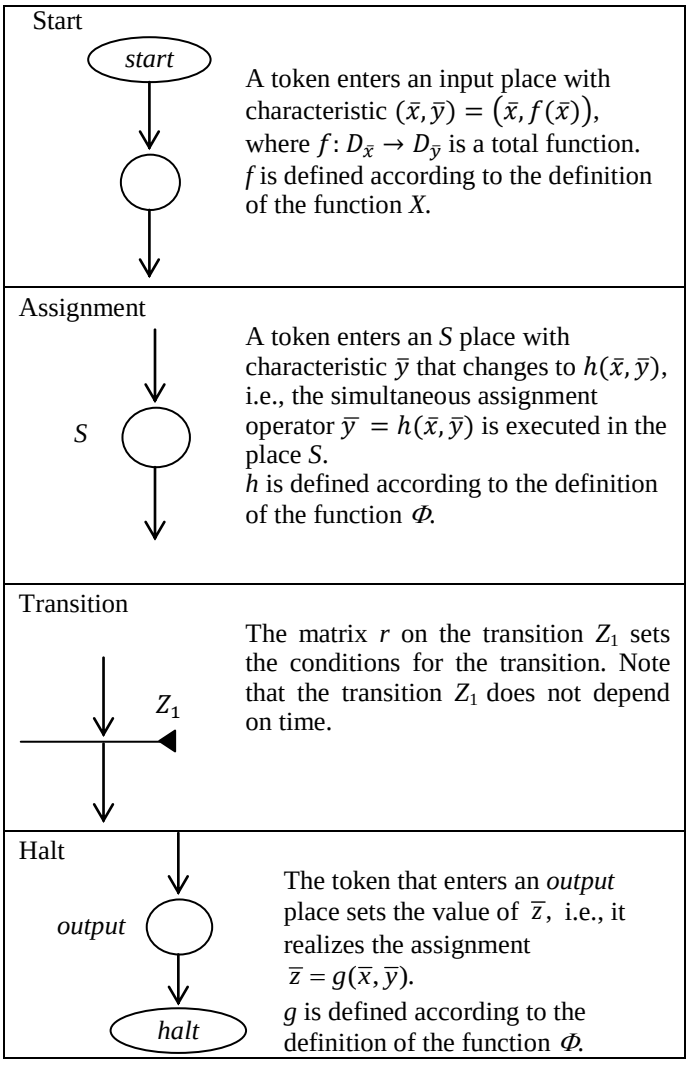



Fig. 2. Components of a GN.

As in the case of the flowcharts, the verification of a GN without temporal components depends on the following predicates:

- An *input* predicate, which will be denoted by $\varphi(\bar{x})$. It is a total predicate over $D_{\bar{x}}$, which describes those elements (data) that may be used as values of the initial characteristics $x_1, x_2, ..., x_n$ of the GN tokens.
- An *output* predicate, which will be denoted by $\psi(\bar{x}, \bar{z})$. It is a total predicate over $D_{\bar{x}} \times D_{\bar{z}}$, which describes the relationships that must be satisfied between the input and the output values of the characteristics of the GN tokens at the termination of the GN execution.

The predicates $\varphi(\bar{x})$ and $\psi(\bar{x}, \bar{z})$ set the input-output specification with respect to which the GN will be verified.

*Definition* 4. GN *P* without temporal components is *partially correct with respect to* $\varphi(\bar{x})$ *and* $\psi(\bar{x}, \bar{z})$ if for every input vector $\bar{\varepsilon}$, for which the input predicate $\varphi(\bar{\varepsilon})$ is *true* and the computation of *P* terminates, $\psi(\bar{\varepsilon}, P(\bar{\varepsilon}))$ is *true*.

*Definition* 5. GN *P* without temporal components *terminates over* $\varphi(\bar{x})$, if for every input vector $\bar{\varepsilon}$, for which $\varphi(\bar{\varepsilon})$ is *true*, the execution of *P* terminates (stops).

*Definition* 6. GN *P* of the type described above is *totally correct with respect to* $\varphi(\bar{x})$ *and* $\psi(\bar{x}, \bar{z})$, if for every vector $\bar{\varepsilon}$, for which the input predicate $\varphi(\bar{\varepsilon})$ is *true*, the execution of *P* terminates over $\varphi(\bar{x})$ and the output predicate $\psi(\bar{\varepsilon}, P(\bar{\varepsilon}))$ is *true*.

Verifying a GN without temporal components with respect to the input-output specification means proving its total correctness regarding this specification. What follows from the definitions above is that the GN *P* is totally correct with respect to $\varphi(\bar{x})$ and $\psi(\bar{x}, \bar{z})$, if *P* is partially correct with respect to $\varphi(\bar{x})$ and $\psi(\bar{x}, \bar{z})$ and *P* terminates over $\varphi(\bar{x})$.

### A. Partial Correctness of a GN

A technique for proving that a GN of the type described above is partially correct with respect to input predicate $\varphi(\bar{x})$ and output predicate $\psi(\bar{x}, \bar{z})$ will be presented. It is similar to this for flowcharts [12].

Let us execute the following three steps:

*Step* 1 (Cutpoints).

Each GN loop connects to a cut (see cutpoints $S_1$ and $S_2$ on Fig. 5). *Start* and *halt* cuts are added to this set of cutpoints (see Fig. 3). Only paths which start and end at cutpoints and which have no intermediate cutpoints are considered. For each path $\alpha$ from cutpoint *i* to cutpoint *j* there is a predicate $R_\alpha(\bar{x}, \bar{y})$ over $D_{\bar{x}} \times D_{\bar{y}}$ and a vector $r_\alpha(\bar{x}, \bar{y})$

$$r_\alpha : D_{\bar{x}} \times D_{\bar{y}} \to D_{\bar{y}}.$$

The predicate $R_\alpha(\bar{x}, \bar{y})$ indicates the condition for this path to be traversed, and the vector $r_\alpha(\bar{x}, \bar{y})$ describes the transformation of the values of $\bar{y}$ affected by path $\alpha$ traversion.

This function can be derived by means of the *backward-substitution technique* [12]. First, the values of the $R_\alpha(\bar{x}, \bar{y})$ and $r_\alpha(\bar{x}, \bar{y})$ (values in cut *j*), are set to *true* (we will denote *true* by *T*) and $\bar{y}$, respectively. Then, at each step, the old *R* and *r* are used to construct the new *R* and *r*, moving backwards toward the cutpoint *i*. The description of the new values in the components is shown in Fig. 3. The resulting *R* and *r* in the cut *i* are the desired $R_\alpha(\bar{x}, \bar{y})$ and $r_\alpha(\bar{x}, \bar{y})$.

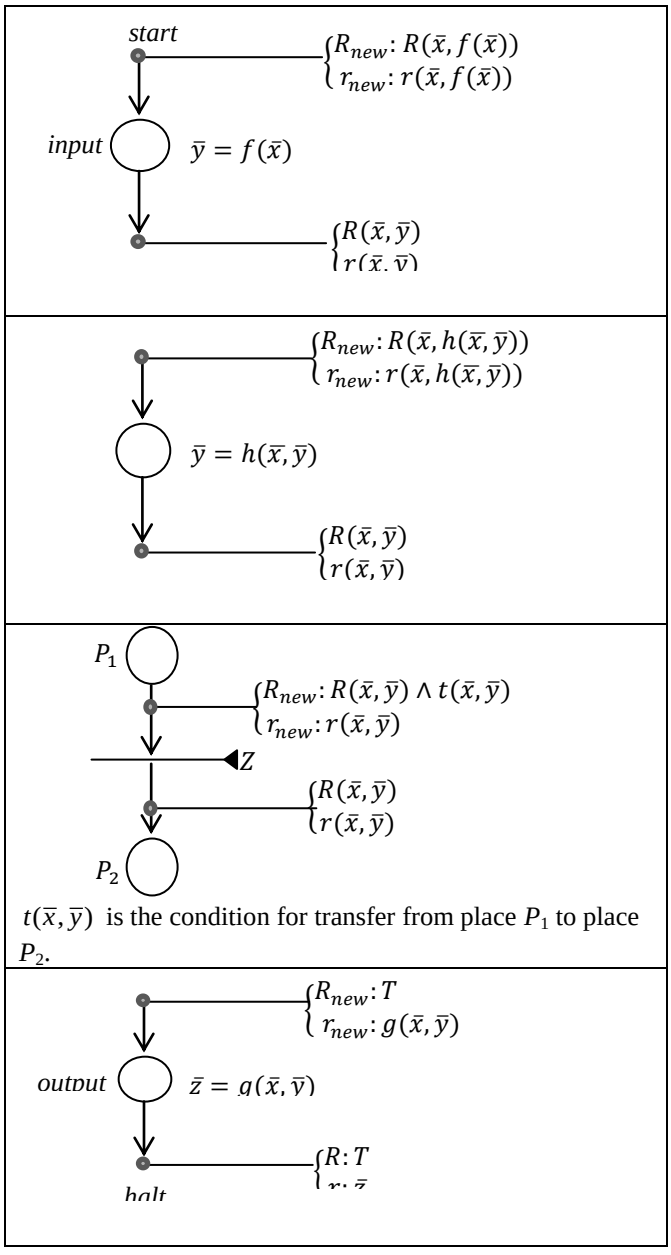



Fig. 3. Rules for constructing the new predicate $R(\bar{x}, \bar{y})$ and the new function $r(\bar{x}, \bar{y})$ for different types of components.

Fig. 4 presents an example for constructing $R_\alpha(\bar{x},\bar{y})$ and $r_\alpha(\bar{x},\bar{y})$ for path $\alpha$, where the condition for transfer from the input place $P_1$ to the output place $P_2$ is $t_1(\bar{x},\bar{y})$, and the condition for transfer from the input place $P_2$ to the output place $P_3$ is $t_2(\bar{x},\bar{y})$.

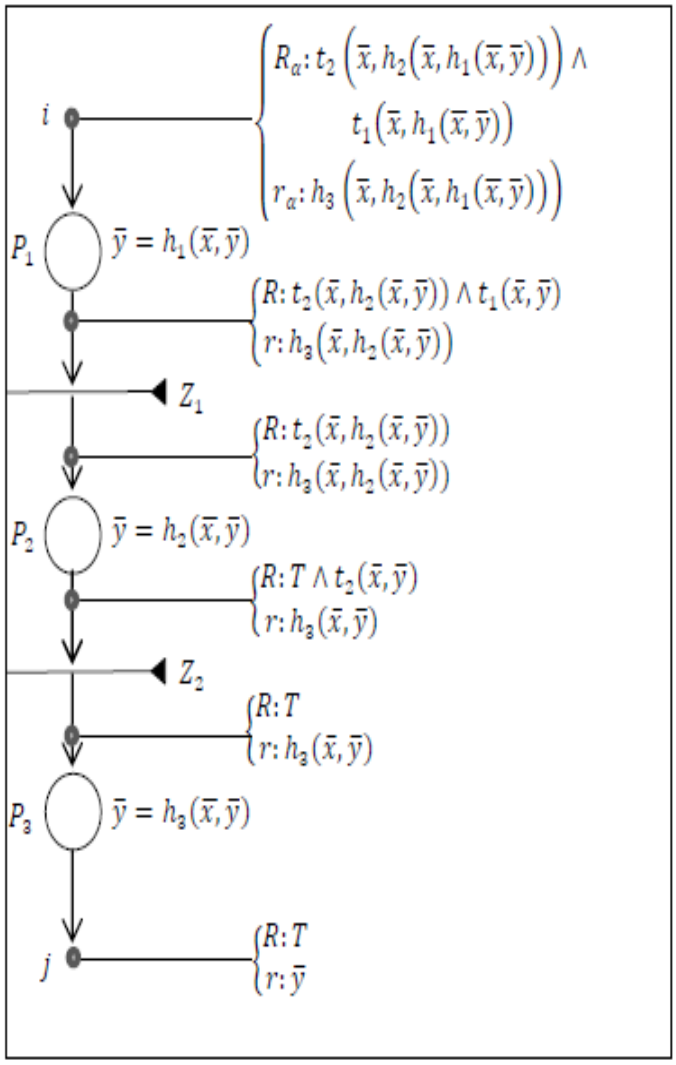



Fig. 4. Constructing the $R_\alpha$ and $r_\alpha$ functions for the path $\alpha$.

*Step* 2 (Inductive assertions).

With each cutpoint *i* of the GN, a predicate $p_i(\bar{x},\bar{y})$ is associated. This predicate is called *inductive assertion*. It characterizes the relation between the values of the characteristics $\bar{x}$ and $\bar{y}$ of the tokens at this point, i.e., $p_i(\bar{x},\bar{y})$ will have the property that, whenever the implementation reaches point *i*, $p_i(\bar{x},\bar{y})$, must be *true* for the current values of $\bar{x}$ and $\bar{y}$ at this point. The input predicate $\varphi(\bar{x})$ is attached to the *start* cutpoints, and the output predicate $\psi(\bar{x},\bar{z})$ is attached to the *halt* cutpoints.

*Step* 3 (Verification conditions).

The final step is to build the verification conditions for each path of the GN:

- For each path $\alpha$ for which *i* is cutpoint *start*

$$\forall\bar{x}\left[\varphi(\bar{x}) \wedge R_\alpha(\bar{x}) \Rightarrow p_j\left(\bar{x}, r_\alpha(\bar{x})\right)\right] \quad (1)$$

- For each path $\alpha$, from the *i* cutpoint to cutpoint *j*

$$\forall\bar{x}\forall\bar{y}\,[p_i(\bar{x},\bar{y}) \wedge R_\alpha(\bar{x},\bar{y}) \Rightarrow p_j(\bar{x}, r_\alpha(\bar{x},\bar{y}))] \quad (2)$$

- For each path $\alpha$ for which *j* is cutpoint *halt*

$$\forall\bar{x}\forall\bar{y}\,[p_i(\bar{x},\bar{y}) \wedge R_\alpha(\bar{x},\bar{y}) \Rightarrow \psi(\bar{x}, r_\alpha(\bar{x},\bar{y}))] \quad (3)$$

and to prove that all these conditions are *true*.

If the constructed verification conditions for all paths that cover the GN are satisfied, the GN is partially correct with respect to $\varphi(\bar{x})$ and $\psi(\bar{x},\bar{z})$. This leads to the follow theorem.

*Theorem* 1. The following steps are applied to a given GN *P* without temporal components, an input predicate $\varphi(\bar{x})$ and an output predicate $\psi(\bar{x},\bar{z})$:

- The loops of the GN are cut.
- An appropriate set of inductive assertions is found.
- The verification conditions (1), (2) and (3) are constructed.

If all the verification conditions are true, then *P* is partially correct with respect to $\varphi(\bar{x})$ and $\psi(\bar{x},\bar{z})$.

### *B. Termination of a GN*

The following is a description of a method of proving the termination of a GN without temporal components regarding an input predicate $\varphi(\bar{x})$. The method was proposed by Floyd for a flowchart [12]. Well-founded sets are used [12].

Note that the paths in steps 1 and 3 do not contain intermediate cutpoints.

Let's perform the following three steps:

*Step* 1 (Good assertions).

Select a set of cutpoints that cut the loops of the GN. Associate an assertion $q_i(\bar{x},\bar{y})$ with every cutpoint *i*, which is a good assertion [12], i.e.,

- For each path $\alpha$ from the *start* cutpoint to cutpoint *j*, the following is satisfied:

$$\forall\bar{x}\left[\varphi(\bar{x}) \wedge R_\alpha(\bar{x}) \Rightarrow q_j\left(\bar{x}, r_\alpha(\bar{x})\right)\right] \quad (4)$$

- For each path $\alpha$ from the *i* cutpoint to cupoint *j*, the following is satisfied:

$$\forall\bar{x}\forall\bar{y}\left[q_i(\bar{x},\bar{y}) \wedge R_\alpha(\bar{x},\bar{y}) \Rightarrow q_j\left(\bar{x}, r_\alpha(\bar{x},\bar{y})\right)\right] \quad (5)$$

*Step* 2 (Well-founded set).

Choose a well-founded set $(W, \prec)$ and with every cutpoint $i$ of the GN associate a partial function $u_i(\bar{x},\bar{y})$

$$u_i:\ D_{\bar{x}} \times D_{\bar{y}} \longrightarrow W$$

which is a good function [12], i.e., for every curpoint $i$, is satisfied:

$$\forall\bar{x}\forall\bar{y}[q_i(\bar{x},\bar{y}) \Rightarrow (u_i(\bar{x},\bar{y}) \in W)] \quad (6)$$

*Step* 3 (Termination conditions).

Show that the termination conditions hold. This means that for every path $\alpha$ from a cutpoint $i$ to a cutpoint $j$, which is a part of some GN loop, the following is satisfied:

$$\forall\bar{x}\forall\bar{y}\left[q_i(\bar{x},\bar{y}) \wedge R_\alpha(\bar{x},\bar{y}) \Rightarrow \left(u_i(\bar{x},\bar{y}) \succ u_j\left(\bar{x}, r_\alpha(\bar{x},\bar{y})\right)\right)\right] \quad (7)$$

This means that after each time a path, which is a part of a loop, is executed, the values of the functions $u_i$, that are associated with the cuts, strictly decrease. As $(W, \prec)$ is a well-founded set, i.e. there are no infinite decreasing sequences of elements of $W$, then the number of the path executions is limited. This leads to the follow theorem.

*Theorem* 2. The following steps are applied to a given GN $P$ of the type described above and an input predicate $\varphi(\bar{x})$:

- The loops are cut and "good" (satisfying (4) and (5)) inductive assertions are found.
- A well-founded set is selected and "good" (satisfying (6)) partial functions are found.
- The termination conditions (7) are checked.

If all the termination condition are true, then $P$ terminates over $\varphi$.

## IV. Illustrative Example

The generalized net in Fig. 5 implements the model of a sequential program, finding $z = x_1^{x_2}$ (where $0^0$ is considered to be equal to 1), where $x_1$ is an integer and $x_2$ is a nonnegative integer.

A start component, a halt component, 2 assignment components and 9 transition components are shown in the figure. The execution of the GN begins at entering the token with characteristic

$$(\bar{x},\bar{y}) = (x_1, x_2, y_1, y_2, y_3),$$

into an *input* place, where $y_1$ means the current value of the base, $y_2$ means the current value of the exponent and $y_3$ – the current value of the exponential result.

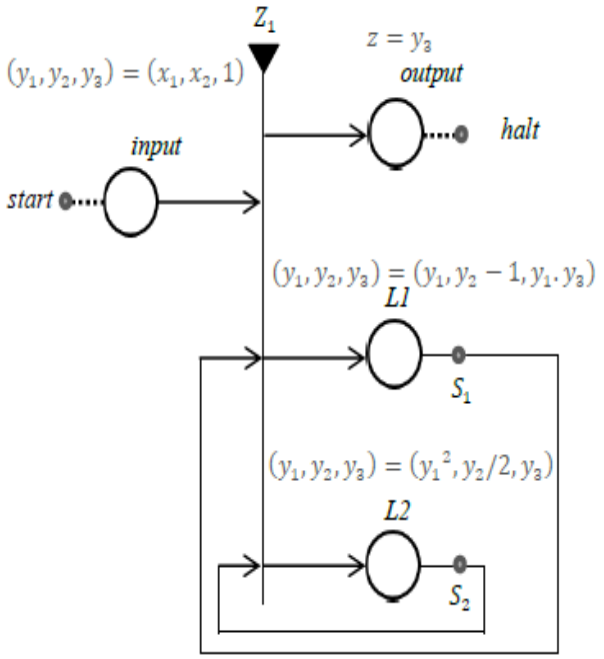



Fig. 5. GN which implements finding $z = x_1^{x_2}$.

The moment the token enters an *input* place with characteristic $(\bar{x},\bar{y}) = (x_1, x_2, y_1, y_2, y_3)$, it gets the value $(x_1, x_2, x_1, x_2, 1)$. The figure also shows how $\bar{y} = (y_1, y_2, y_3)$ of the characteristic of the token changes in places $L_1$ and $L_2$; and shows that, when it enters into an *output* place, it receives characteristic $z = y_3$.

The GN has one transition $Z_1$, with predicate matrix $r_1$ of transition:

| $r_1$ | *output* | $L1$ | $L2$ |
|---|---|---|---|
| *input* | $y_2 = 0$ | $y_2 \neq 0 \wedge odd(y_2)$ | $y_2 \neq 0 \wedge \neg odd(y_2)$ |
| $L1$ | $y_2 = 0$ | $y_2 \neq 0 \wedge odd(y_2)$ | $y_2 \neq 0 \wedge \neg odd(y_2)$ |
| $L2$ | $y_2 = 0$ | $y_2 \neq 0 \wedge odd(y_2)$ | $y_2 \neq 0 \wedge \neg odd(y_2)$ |

We will perform GN verification by the method described above over:

- the input predicate: $\varphi(\bar{x}) : x_2 \geq 0$ and
- the output predicate: $\psi(\bar{x},\bar{z}) : z = x_1^{x_2}$

where $\bar{x} = (x_1, x_2), \bar{y} = (y_1, y_2, y_3), \bar{z} = z$.

### A. Partial Correctness

Let us cut the two loops of the GN at points $S_1$ and $S_2$ (see Fig. 5), and attach to the cutpoint $S_1$ and cutpoint $S_2$ the assertion:

$$p(\bar{x},\bar{y}): x_2 \geq 0 \wedge y_2 \geq 0 \wedge y_3.y_1^{y_2} = x_1^{x_2}.$$

This GN is covered by the following paths:

| | | |
|---|---|---|
| 1. start → halt | 4. $S_1 \rightarrow S_1$ | 7. $S_2 \rightarrow S_1$ |
| 2. start → $S_1$ | 5. $S_2 \rightarrow S_2$ | 8. $S_1 \rightarrow$ halt |
| 3. start → $S_2$ | 6. $S_1 \rightarrow S_2$ | 9. $S_2 \rightarrow$ halt |

We will prove the verification conditions only for paths 1 and 7 below. The other verification conditions are proved likewise. The functions *R* and *r* for the paths 1, 2, …, 9 have the form:

1. *R*: $x_2 = 0$, *r*: 1
2. *R*: $x_2 \neq 0 \wedge odd(x_2)$, *r*: $(x_1, x_2-1, x_1)$
3. *R*: $x_2 \neq 0 \wedge \neg\, odd(x_2)$, *r*: $(x_1^2, \frac{x_2}{2}, 1)$
4. *R*: $y_2 \neq 0 \wedge odd(y_2)$, *r*: $(y_1, y_2-1, y_1.y_3)$
5. *R*: $y_2 \neq 0 \wedge \neg\, odd(y_2)$, *r*: $(y_1^2, \frac{y_2}{2}, y_3)$
6. *R*: $y_2 \neq 0 \wedge \neg\, odd(y_2)$, *r*: $(y_1^2, \frac{y_2}{2}, y_3)$
7. *R*: $y_2 \neq 0 \wedge odd(y_2)$, *r*: $(y_1, y_2-1, y_1.y_3)$
8. *R*: $y_2 = 0$, *r*: $y_3$
9. *R*: $y_2 = 0$, *r*: $y_3$

The verification condition for path 1 has the form:

$$\varphi(\bar{x}) \wedge (x_2 = 0) \Rightarrow \psi(\bar{x}, 1),$$

i.e.,

$$(x_2 \geq 0) \wedge (x_2 = 0) \Rightarrow 1 = x_1^{x_2}$$

is evidently satisfied.

The verification condition for path 7 has the form:

$$p(\bar{x},\bar{y}) \wedge y_2 \neq 0 \wedge odd(y_2) \Rightarrow p(\bar{x}, y_1, y_2 - 1, y_1.y_3)$$

i.e.,

$$x_2 \geq 0 \wedge y_2 \geq 0 \wedge y_3.y_1^{y_2} = x_1^{x_2} \wedge y_2 \neq 0 \wedge odd(y_2) \Rightarrow$$

$$x_2 \geq 0 \wedge (y_2 - 1) \geq 0 \wedge y_1 y_3.y_1^{y_2 - 1} = x_1^{x_2}$$

Since all verification conditions are true, it follows that the GN is partially correct.

*B. Termination*

Let choose the well-founded set $(N, <)$, that is, the set of natural numbers with the usual ordering <. We cut the two loops at points $S_1$ and $S_2$ (see Fig. 5.) Next, we choose $q_{S_1}(\bar{x},\bar{y})$ and $q_{S_2}(\bar{x},\bar{y})$ to be $y_2 \geq 0$. Let $u_{S_1}(\bar{x},\bar{y})$ and $u_{S_2}(\bar{x},\bar{y})$ are equal to $y_2$.

*Step* 1. $q_{S_1}(\bar{x},\bar{y})$ and $q_{S_2}(\bar{x},\bar{y})$ are good assertions.

Here we will prove the condition only for path 6. The assertions for other paths are proved similarly.

$$q_{S_1}(\bar{x}, y_1, y_2, y_3) \wedge y_2 \neq 0 \wedge \neg odd(y_2) \Rightarrow q_{S_2}(\bar{x}, y_1^2, \frac{y_2}{2}, y_3)$$

i.e.,

$$y_2 \geq 0 \wedge y_2 \neq 0 \wedge \neg odd(y_2) \Rightarrow \frac{y_2}{2} \geq 0$$

*Step* 2. $u_{S_1}(\bar{x},\bar{y})$ and $u_{S_2}(\bar{x},\bar{y})$ are good functions.

The condition for cutpoint $S_1$ is:

$$q_{S_1}(\bar{x},\bar{y}) \Rightarrow \left( u_{S_1}(\bar{x},\bar{y}) \in N \right)$$

that is, $y_2 \geq 0 \Rightarrow y_2 \geq 0$. The condition is evidently satisfied.

The condition for $S_2$ cutpoint is proved similarly.

*Step* 3. The termination condition holds.

Only paths 4, 5, 6 and 7 are considered as they are parts of some GN loop.

*Path* 4 $S_1 \rightarrow S_1$

$$q_{S_1}(\bar{x}, y_1, y_2, y_3) \wedge y_2 \neq 0 \wedge odd(y_2) \Rightarrow$$

$$(u_{S_1}(\bar{x}, y_1, y_2, y_3) > u_{S_1}(\bar{x}, y_1, y_2 - 1, y_1.y_3))$$

i.e.,

$$y_2 \geq 0 \wedge y_2 \neq 0 \wedge odd(y_2) \Rightarrow y_2 > y_2 - 1$$

*Path* 5 $S_2 \rightarrow S_2$

$$q_{S_2}(\bar{x}, y_1, y_2, y_3) \wedge y_2 \neq 0 \wedge \neg odd(y_2) \Rightarrow$$

$$(u_{S_2}(\bar{x}, y_1, y_2, y_3) > u_{S_2}(\bar{x}, y_1^2, \frac{y_2}{2}, y_3))$$

i.e.,

$$y_2 \geq 0 \wedge y_2 \neq 0 \wedge \neg odd(y_2) \Rightarrow y_2 > \frac{y_2}{2}$$

*Path* 6 $S_1 \rightarrow S_2$

$$q_{S_1}(\bar{x}, y_1, y_2, y_3) \wedge y_2 \neq 0 \wedge \neg odd(y_2) \Rightarrow$$

$$(u_{S_1}(\bar{x}, y_1, y_2, y_3) > u_{S_2}(\bar{x}, y_1^2, \frac{y_2}{2}, y_3))$$

i.e.,

$$y_2 \geq 0 \wedge y_2 \neq 0 \wedge \neg odd(y_2) \Rightarrow y_2 > \frac{y_2}{2}$$

*Path* 7 $S_2 \rightarrow S_1$

$$q_{S_2}(\bar{x}, y_1, y_2, y_3) \wedge y_2 \neq 0 \wedge odd(y_2) \Rightarrow$$

$$(u_{S_2}(\bar{x}, y_1, y_2, y_3) > u_{S_1}(\bar{x}, y_1, y_2 - 1, y_1.y_3))$$

i.e.,

$$y_2 \geq 0 \wedge y_2 \neq 0 \wedge odd(y_2) \Rightarrow y_2 > y_2 - 1$$

Since all conditions of the three steps are *true*, the GN terminates for every $x_2 \geq 0$ natural number.

## V. Applications and Ideas for Further Research

The necessity for developing methods and environments for formal verification of GN models is triggered by the implementation of real-life methods of software protection by using GNs [13]. Using Floyd's inductive assertion method for verification of flowcharts, adapted for GNs, GN models which realize sequential processes, and sequential programs in particular [14], can be verified. It can also be applied for verification of GNs, which model parallel processes. To this end, a transition's type component □ is applied.

Following our belief that training in applying formal methods for developing of correct software is the most efficient method of implementing these methods in the software industry, we intend on introducing the method described above in teaching students of specialty Informatics and Computer Sciences at Sofia University. In order to achieve this, we will develop and add tools for its application to the educational framework presented in [15]. The ideas presented in [16, 17, 18] are implemented in the educational framework. The resulting educational framework may be applied not only in teaching programming and data structures [19, 20], but also in creating GN models of applications, such as: classical transaction processing systems [21], mobile information applications for public access [22, 23], business process models [24, 25], software services models [26, 27], data models [28], etc. Thus, students will be stimulated to search for out of the box solutions of the tasks given in the programming courses [29] and in the courses in discrete mathematics (discrete structures) [30].

Furthermore, we consider designing and improving an educational framework with tools for verification of GN models with temporal components.

## VI. Related Work

The field of formal verification of GN models has not been studied by now. The most closely related work with the one presented here is the book by Zohar Manna [12]. This book represents an introduction in the mathematical theory of informatics, and is considered the main reference book in this area for many universities around the world, including the Sofia University. Chapter 3 of [12] explores the verification of computer programs. Sections 3-1 and 3-2 of this chapter contains definitions, related to the verification of programs presented in terms of flowcharts, and Floyd's inductive assertions method for program verification is provided. Since the methodology of development of Generalized Net models is difficult enough as such, in order to enhance the understanding of the herewith presented adaptation of Floyd's inductive assertions method for formal verification of GN models without temporal components, we opted to use as much as possible the denotations and theorems, formulated in [12].

## VII. Conclusion

The paper presents the authors' first attempt to achieve formal verification of GN models. Our research has been restricted to verification of GN models without temporal components. Since the proposed verification approach belongs to the set of formal verification methods, it bears all of their limitations, as well. Significant efforts are required to construct the input/output specification; to prove the conditions for partial correctness and for termination of the execution; the procedures are to be realized by the rare highly qualified experts in formal modelling. Using automated tools for formal theorem proving would reduce the mentioned difficulties. The active and impactful research on development of such tools has motivated us to continue the work in this direction by developing methods for formal verification of GN models featuring temporal components, as well as verification of the GN models of the applications proposed here in Part V.

## References

[1] K. Atanassov, Generalized Nets, World Scientific, Singapore, 1991.

[2] K. Atanassov, On Generalized Nets Theory, Prof. Marin Drinov Academic Publishing House, Sofia, 2007.

[3] T. Trifonov, and K. Georgiev, "GNTicker – A software tool for efficient interpretation of generalized net models," Issues in Intuitionistic Fuzzy Sets and Generalized Nets, vol. 3, Warsaw, 2005.

[4] T. Trifonov, K. Georgiev, and K. Atanassov, "Software for modelling with Generalised Nets," Issues in Intuitionistic Fuzzy Sets and Generalized Nets, vol. 6, pp. 36–42, 2008.

[5] V. Gochev, "An implementation of generalized nets using object-oriented programming in .NET framework, Management and education," vol. VI (4), pp. 227–231, 2010.

[6] K. Atanassov, D. Dimitrov, and V. Atanassova, "Algorithms for Tokens Transfer in the Different Types of Intuitionistic Fuzzy Generalized Nets," Cybernetics and Information Technologies, vol. 10, No. 4, pp. 22–35, 2010.

[7] D. G. Dimitrov, "Graphical Environment for Modeling and Simulation with Generalized Nets," Annual of "Informatics", Section Union of Scientists in Bulgaria, vol. 3, pp. 51–66, 2010.

[8] D. G. Dimitrov, "Software Products Implementing Generalized Nets," Annual of "Informatics", Section Union of Scientists in Bulgaria, vol. 3, pp. 37–50, 2010.

[9] D. G. Dimitrov, "Optimized Algorithm for Token Transfer in Generalized Nets," Recent Advances in Fuzzy Sets, Intuitionistic Fuzzy Sets, Generalized Nets and Related Topics, vol. 1, pp. 63–68, 2010.

[10] N. Angelova, M. Todorova, and K. Atanassov, "GN IDE: Implementation, Improvements and Algorithms," Comptes Rendus de L'Academie Bulgare des Sciences, Tome 69, vol. 4, pp. 411−420, 2016.

[11] K. Atanassov, Index Matrices: Towards an Augmented Matrix Calculus, Springer, Cham, 2014.

[12] Z. Manna, Mathematical theory of computation, New York, 1974.

[13] M. Todorova, and D. Orozova, "Software protection integrating registration-number and anti-debugging protections," 9th International Conference Information Systems & Grid Technologies, St. Kliment Ohridski University Press, 2015.

[14] M. Todorova, and D. Orozova, "Generalized Net Model of Sequential Programs," 20th International Symposium on Electrical Apparatus and Technologies (SIELA), 3 – 6 June, 2018, Bourgas, Bulgaria (in print), [Digests 20th International Symposium on Electrical Apparatus and Technologies, Bulgaria, pp. 265−266, 2018].

[15] M. Todorova and K. Kanev, "Educational framework for verification of object−oriented programs," The Joint International Conference on Human−Centered Computer Environments HCCE'2012, Hamamatsu, Japan, pp. 23−27, 2012.

[16] A. Shannon, D. Langova−Orozova, E. Sotirova, I. Petrounias, K. Atanassov, M. Krawczak, P. Melo−Pinto, and T. Kim, Generalized Net Modelling of University Processes, KvB Visual Concepts Pty Ltd, Monograph No. 7, Sydney, 2005.

[17] A. Shannon, K. Atanassov, D. Orozova, M. Krawczak, E. Sotirova, P. Melo−Pinto, I. Petrounias and T. Kim, Generalized nets and

information flow within a university, Warsaw School of Information Technology, Warsaw, 2007.

[18] D. Orozova, and K. Atanassov, "Generalized net model of the process of selection and usage of an intelligent e-learning system," Comptes Rendus de l'Academie bulgare des Sciences, tome 65, No. 5, pp. 591−598, 2012.

[19] I. Donchev, and E. Todorova, "Implementation of ADS Linked List Via Smart Pointers," International Journal of Advanced Computer Science and Applications, vol. 6, No. 2, pp. 196−203, 2015.

[20] I. Donchev, and E. Todorova, "Implementation of Binary Search Trees Via Smart Pointers," International Journal of Advanced Computer Science and Applications, vol. 6, No. 3, pp. 59−64, 2015.

[21] K. Kaloyanova, "Successful practices for learning information systems development," 7th International Technology, Education and Development Conference, pp. 4849−4855, 2013.

[22] I. Patias and V. Georgiev, "Mobile medical applications as instrument in supporting patients compliance," American Journal of Engineering Research, vol. 6, No. 8, pp. 96−102, 2017.

[23] I. Patias and V. Georgiev, "Modeling and implementation of bus rapid transit corridor based on isolated or coordinated traffic prioritization and automatic location," Journal of Emerging Research and Solutions in ICT, vol. 1, No. 2, pp. 17–24, 2016.

[24] E. Krastev and K. Shahinyan, "Computer assisted quality assessment of a set of business process models," Proceedings of the 9th IEEE European Modelling Symposium of Mathematical Modelling and Computer Simulation, IEEE Computer Society, pp. 180−186, 2015.

[25] V. Dimitrov, "Deriving semantics from WS-BPEL specifications of parallel business processes on an example," Computer Research and Modeling, vol. 7, No. 3, pp. 445−454, 2015.

[26] S. Hadzhikoleva and E. Hadzhikolev, "Model for automated integration of data from heterogeneous sources in the COMPASS-OK application for (self) evaluation and accreditation," International Journal of Applied Engineering Research, vol. 11, No. 12, pp 7648−7653, 2016.

[27] S. Hadzhikoleva, T. Rachovski and E. Hadzhikolev, "Generalized Net Model for Building Responsive Design of Web Pages," 20th International Symposium on Electrical Apparatus and Technologies SIELA 2018, 3 − 6 June 2018, Bourgas, Bulgaria (in print).

[28] V. Dimitrov, "Relationship Specified in Z-notation," Physics of Elementary Particles and Atomic Nuclei, Letters, vol. 8, No. 4 (167), pp. 655−663, 2011.

[29] K. Yordzhev, "The Bitwise Operations Related to a Fast Sorting Algorithm," International Journal of Advanced Computer Science and Applications, vol. 4, No. 9, pp. 103−107, 2013.

[30] K. Yordzhev, Sudoku, S-permutation matrices and bipartite graphs, LAP LAMBERT Academic Publishing, 2016.

# Location-aware Event Attendance System using QR Code and GPS Technology

Zakiah Ayop[1], Chan Yee Lin[2], Syarulnaziah Anawar[3], Erman Hamid[4], Muhammad Syahrul Azhar[5]
Center for Advanced Computing Technology (C-ACT),
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka,
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.

***Abstract*—Attendance process in a university's event is time consuming and tracking the attendance can be harder. In this paper, a smart event attendance system for a university using QR code and GPS technology is proposed with objective to speed up the process of taking students' attendance and tracking full attendance. The method of developing the system is based on two views; user view which is the mobile application used by the students, and admin view which is the web administration system used by the event organizer. From the evaluation, students' attendance can be traced from the GPS location combine with QR code. The results indicate that full attendance increases as the system validates attendance through users' identification, location and timestamp during user login and logout. The proposed system contributes to high satisfaction among the users that claim that the mobile application helps to speed up the event registration process.**

***Keywords*—Event attendance system; quick response (QR) code; global positioning system (GPS); android mobile application**

## I. INTRODUCTION

In this era, smartphones play a significant role in our daily lives. The emergence of mobile application, has been impacted by the convergent factors such as high-speed data network, relatively cheap devices, high-performing devices, easy-to-use market places for apps, and the need for simple, targeted applications while mobile [1].

Universiti Teknikal Malaysia Melaka (UTeM) is the 14th public university in Malaysia. This university consists of three campuses which are the main campus, technology campus, and city campus. The university organizes various events for the students from different campuses. Hence, there will be hundreds of students that will take part in the events, thus making the attendance taking process time consuming and may delay the start time of the event.

Therefore, the purpose of this study is two-fold: First, to investigate the requirements in event attendance for a university's event, and second, to develop a mobile application that utilizes the QR code and GPS location. A proof of concept for the proposed solution is developed. The system consists of admin view for event's organizer to create an encrypted QR code, and a user view for students to log in the university site by using unique matric number and password, scanning the QR code shown by the organizer and their current location which is tracked by the GPS as the attendance. The user view will then communicate the information collected to the admin view to confirm the attendance.

This paper is organized as follows. Section II discusses previous studies in attendance management system, QR code, and GPS solution. Section III describes the methodology used to develop the student attendance system. Section IV outlines the implementation of this study, the discussion on system evaluation is provided in Section V, followed by conclusion in Section VI.

## II. RELATED WORKS

Conventional attendance system is still used in most universities. However, this type of attendance system suffers problem like missing name, false attendance, missing attendance sheet, and tedious management. The advancement in attendance system has incorporate technological tools to improve the shortcomings in conventional system. In this section, various technologies used to support current work in the attendance system will be discussed.

An efficient web-based application for attendance management system is designed to track students' activity in the class by using the electronic methods [2]. Besides, the attendance records are stored in the database and this system is developed with the usage of Model, View and Controller (MVC) architecture with the assistance of power of Laravel Framework. The purpose of this system is to differentiate the hours of theoretical and practical lessons since the calculation method for the absence rate of students for these lessons are different.

On the other hand, biometric technologies such as face [3], fingerprint [4], and iris [5] recognition have been introduced as students' identification and reduce the false attendance problem. Although biometric identification prevents fake attendance and proxies, it requires some efficient recognition algorithms [6] and higher computation power on the mobile phone, thus increasing deployment cost.

The emergence of sensors has innovated the technology in smartphones and Student Identification (ID) card which facilitate the authentication process. Technology such as barcode [7], Bluetooth [8], RFID [9] and NFC [10] are used in attendance system to improve the weaknesses in biometric system. However, there is concern in substantial additional cost to the university, namely hardware reader to track the ID [11].

Hence, QR code based system which is a combination of mobile devices to display and scan the QR code is introduced. An online student attendance monitoring system (SAMS) based on QR code and mobile devices is developed in [11]. It

seems quantitatively easier to discern the students based on their diligence in attending classes and predict their performance due to the correlation between the attendance and academic performance. Besides, the main advantage of this system is to record and monitor student attendance in a more accurate and quicker way. There are two main components in this system, which are SAMS server and SAMS application. The system itself is available online and designed for access via mobile devices. A unique QR code is generated and sent for each student by email and it is used to record attendance. These QR code are presented by students to their lecturer either using their smart phone or with a print out, and later scanned by the lecturer using the SAMS application. However, we noticed that SAMS requires lecturers' intervention which can disturb the class delivery. Therefore, in order to avoid interruption, a proposed solution is offered by Masalha and Hirzallah [12], where students are required to scan the QR code by using a specific mobile application before or during the class. The QR code is pasted on each displayed lecture slide. The identity of the student is identified when he or she scans the generated QR code, and the attendance is taken and sent to the university's server [12]. Other solution in [12] also includes face recognition that is applied to perform identity verification. A location check will be performed to verify the users' location. However, lecturer need to design and develop a specific QR code for each student and this method is not suitable for the process attendance of event.

Nonetheless, the main weakness in current works in attendance system is the current location of students is not tracked when they take the attendance by using the student attendance system. This weakness can be seen in SAMS [11], the attendance checking system using QR code in University Sulaimaniyah in Iraq [13] and the smart attendance system in Institute of Hydropower Engineering and Technology [14]. Hence, cheating phenomena could have happened among students when they used these systems. Besides, the SAMS [11] is inappropriate for the process of students' attendance in event because a unique QR code is generated and delivered to each student by email, which is not suitable due to high overhead when the students coming from different faculties.

Consequently, we propose to develop an improved Event Attendance System which based on the features that has been discussed above. The event attendance system implemented in this project is a software application created using Android Studio to ensure only the authorized students can login into the system by using their unique matric number and password. Besides, the login and logout time of students and their current location which is tracked by GPS sensor will be recorded and stored in database as the attendance. In addition, the process of taking attendance can be speed up as the event organizer only needs to create an encrypted QR code with the event information provided.

## III. METHODOLOGY

### A. Requirement Analysis

Prior to this project, we performed requirement analysis in a meeting with the IT Operation officers from IT Centre (PPPK) in the UTeM. Based on the discussions, the functional requirements specification for the proposed system has been identified (refer Table I). Hardware and software requirement has been specified in Table II and Table III.

TABLE I. FUNCTIONAL REQUIREMENT

| **Event Attendance System** | |
|---|---|
| Purpose | Event Attendance |
| Platform | Android, Web |
| Student information | • Name<br>• Student Matric Number<br>• Faculty<br>• Course |
| Event Information | • Event name<br>• Event Location<br>• Event Date<br>• Event Start Time, End Time |

TABLE II. HARDWARE REQUIREMENT

| **Event Attendance System** | |
|---|---|
| Hardware | Purpose/Usage |
| Laptop with 2GHz CPU, 4GB RAM | Specification for system development |
| Mobile phone with Android 4.4 system | Mobile application settings for testing and deployment |

TABLE III. SOFTWARE REQUIREMENT

| **Event Attendance System** | |
|---|---|
| Software | Purpose/Usage |
| Android Studio 3.0 | Official IDE for system development |
| Android SDK v19 | Mobile application platform for testing |

To meet the requirements set by PPPK, the non-functional requirements will evaluate [15]:

*1) Performance:* The system is capable to scan the QR code based on various setting of lighting, angle and distance.

*2) User Acceptance:* To demonstrate how the design affects the usage of the application by the user, a preliminary study is conducted that use a quantitative methodology. A structured questionnaire is used to collect data survey from 20 students who have different levels of IT skill. The objective of this study is to measure the user satisfaction toward the application.

### B. System Flow

Figure 1 shows the flow chart of user view for Event Attendance System. The android application allows a student

to login into the system. After successfully login into the system, the student will select 'SCAN THE QR CODE' button to scan the QR code which is generated by the university's event organizer. After the scanning process, the information about the event which is included in the QR code, the student's location, and student identity will be sent to database server. This is to ensure that the student is within the event hall/location when he or she is registering their attendance. The attendance will only be saved in database when the student scans the QR code 15 minutes before the event starts, and needs to logout within 15 minutes after the event ended.

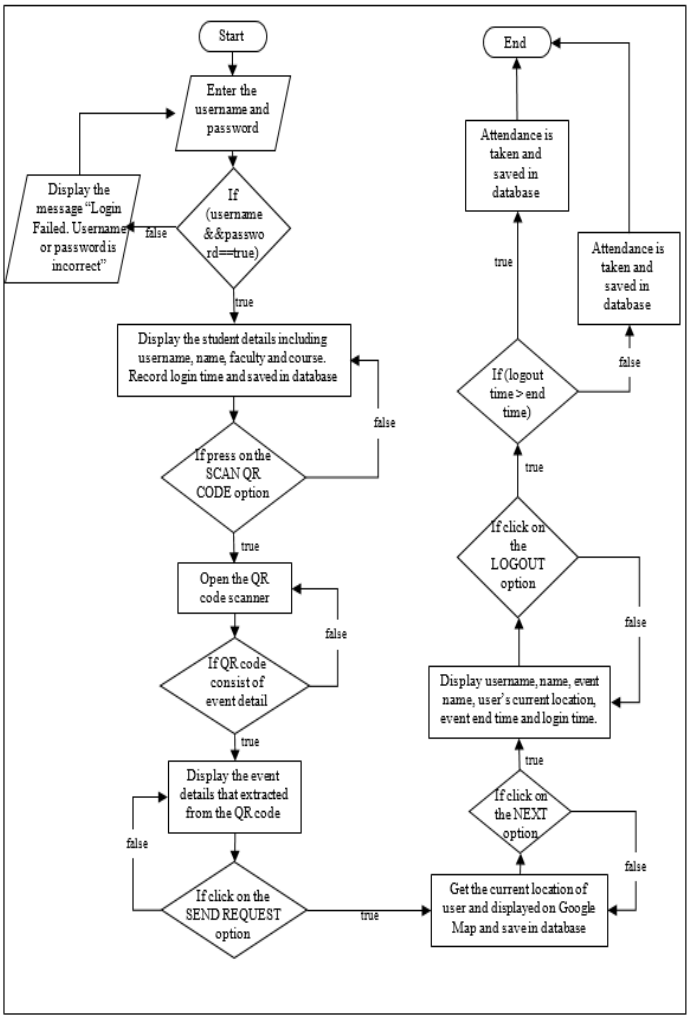



Fig. 1. Flow chart of user view

Figure 2 shows the flow chart of the admin view. The system administor is able to choose different options in the main menu page. If administrator select the Student Attendance option, he or she can search an event name for viewing all students' attendance who participated in the event. Besides, administrator is allowed to view the particular event details if he or she indicates the Event Details selection. In addition, administrator is able to add a new event details which will be saved in database and generate a QR code that consists of the event data in Add Event and QR code generator respectively. Lastly, administrator is capable to view the student details on Student Details page according to the searched matric number.

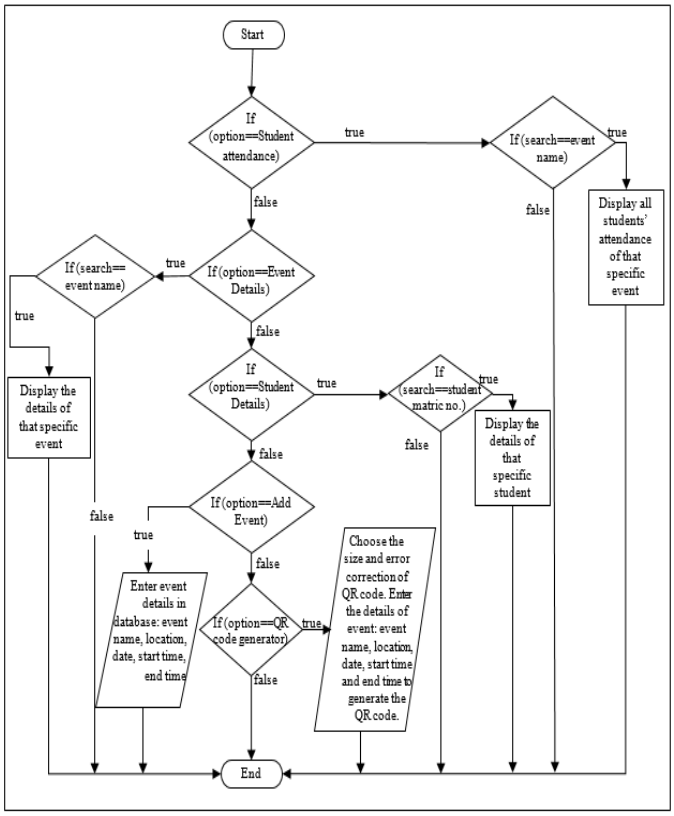



Fig. 2. Flow chart of admin view

## IV. IMPLEMENTATION

### A. System Architecture

Figure 3 shows the architecture of the proposed system. First, the user needs to log in the system by using their email address and password. After that, the student needs to scan the QR code that is provided by the organizer. When the student uses the android application to scan the QR code, the application will request the location of the student to ensure he or she is at the correct location to take the attendance.

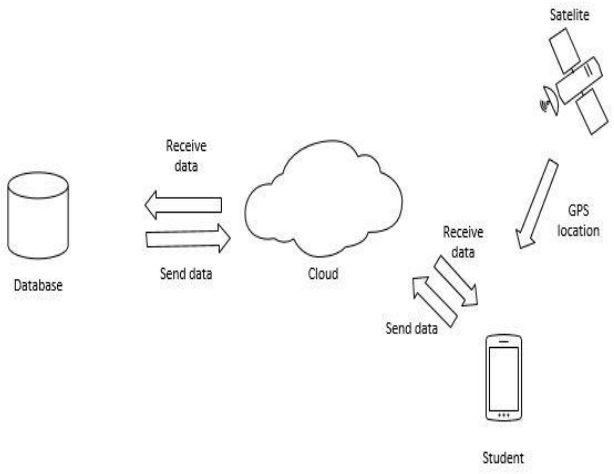



Fig. 3. System Architecture

### B. Experimental Setup

The system evaluation is conducted in Universiti Teknikal Malaysia Melaka (UTeM). 20 students who have different levels of IT skill from different faculties were selected to test the system. We created two events for the purpose of evaluation; Hacking Event and Workshop 2 Briefing, with different location for each event. We conducted three types of evaluation to meet the system requirements, namely system functionality, system performance and user acceptance.

The system consists of two views; user view which accessible through mobile application and admin view from web administration system. Further discussion on both views is provided in the Section C and D.

### C. User View: Mobile Application

Mobile application contains four modules: Login, QR code Scanner, GPS and Attendance modules. In Figure 4, the details of the student such as username, name, faculty, course and login time which are stored in database will be displayed if they successfully log into system. For attendance input, the student needs to press on the 'SCAN THE QR CODE' button for scanning the QR code.

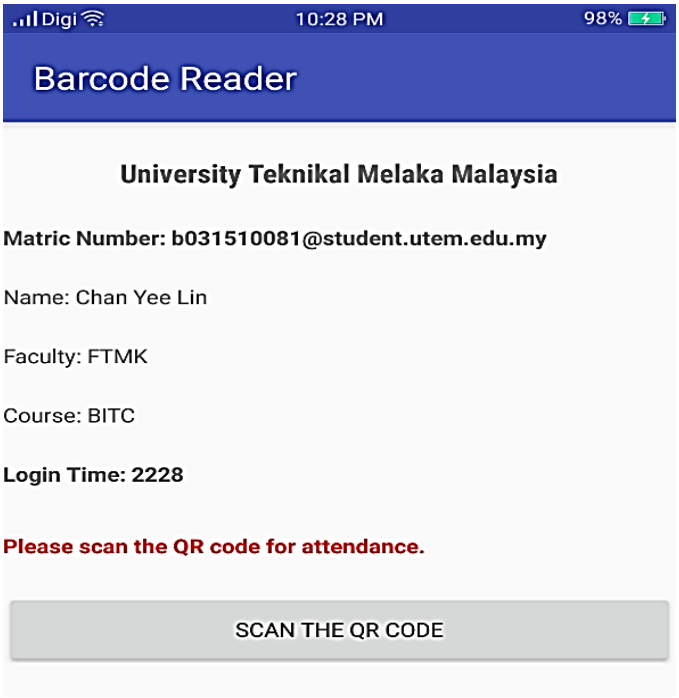



Fig. 4. Event Selection Interface

In Figure 5, the student needs to scan the QR code which contains the event's information by using the QR code scanner. The event's details that is extracted from the QR code will be displayed in this page and saved in database (refer Figure 6). Additionally, student can request their current GPS location by pressing Send Request button. The current location of the student will be tracked and later saved in database. The acceptance of student attendance is notified by displaying the username, student name and event's details on the mobile page.

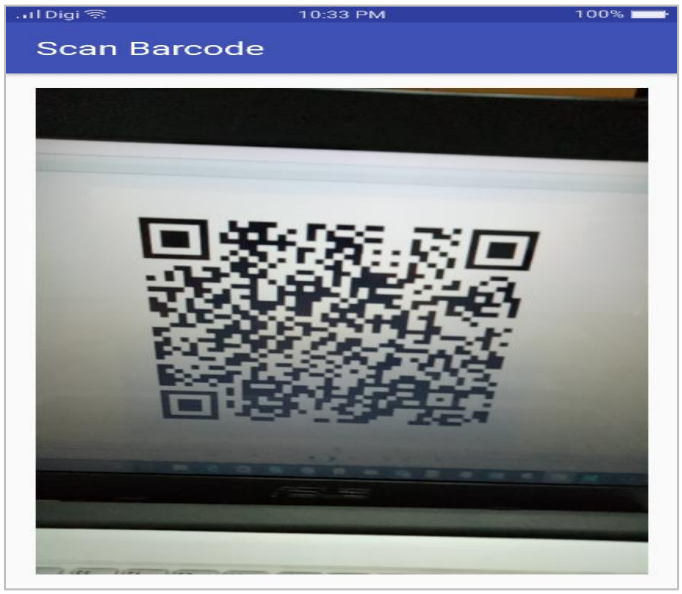



Fig. 5. Activity Scan Barcode Interface

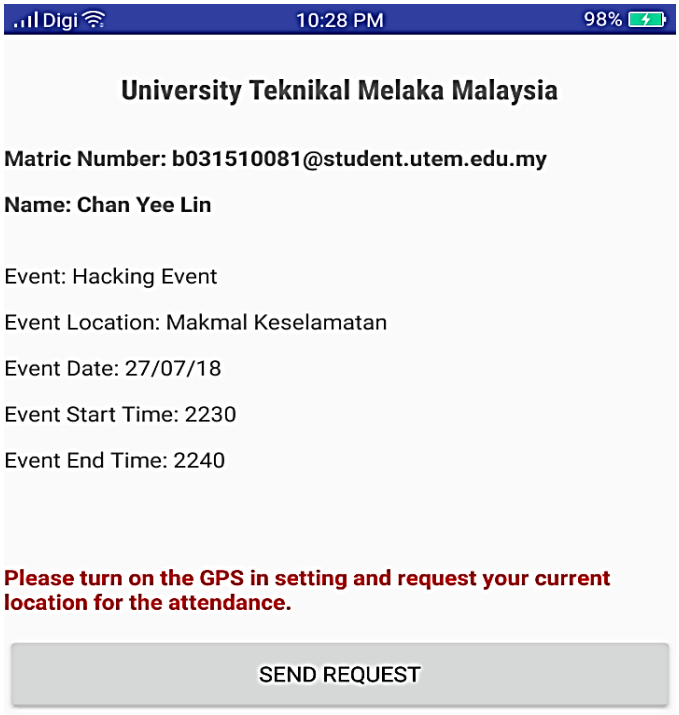



Fig. 6. Barcode Result and Request GPS Interface

Figure 7 shows the application tracking the current location of the student. In Figure 8, the event's details such as event name, current location and event end time will be displayed in this page. Besides, the username, student name, and login time of the student will be called from database and displayed on the page. To ensure the attendance is taken, Logout button should be pressed once the event ended for student to logout from the system. However, if student logout from the system before the event ended, the attendance will not be taken.

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

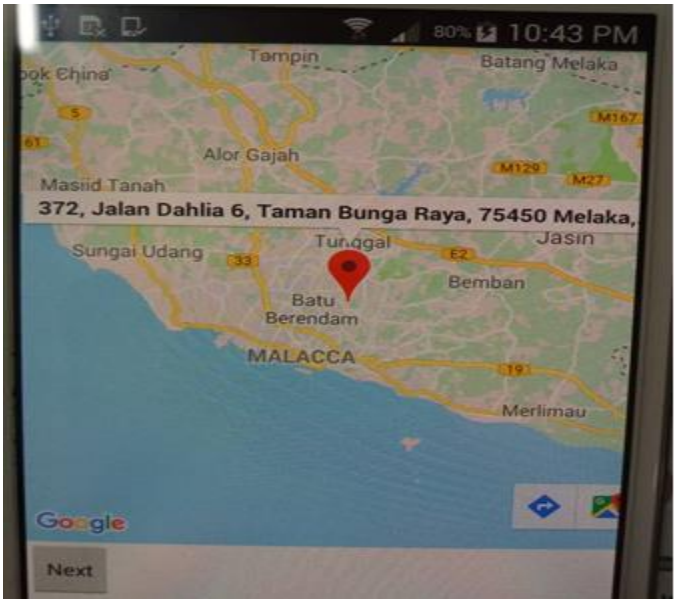



Fig. 7. Detect GPS Location Interface


Fig. 8. Logout page Interface

### *D. Admin View: Web Administration System*

On Web Administration main menu page (refer Figure 9), there are five modules that needs to be manage; Add Event (refer Figure 10), QR Code Generator (refer Figure 11), Event Details (refer Figure 12), Student Attendance (refer Figure 13), and Student Details (refer Figure 14).

To register new event, administrator can add the event details and save to the database as in Figure 10. Administrator will generate the event QR code on the page. Besides, administrator can choose the size and error correction of the QR code (refer Figure 11). Event details will be displayed on the page (refer Figure 12).


Fig. 9. Web Administration Main Menu

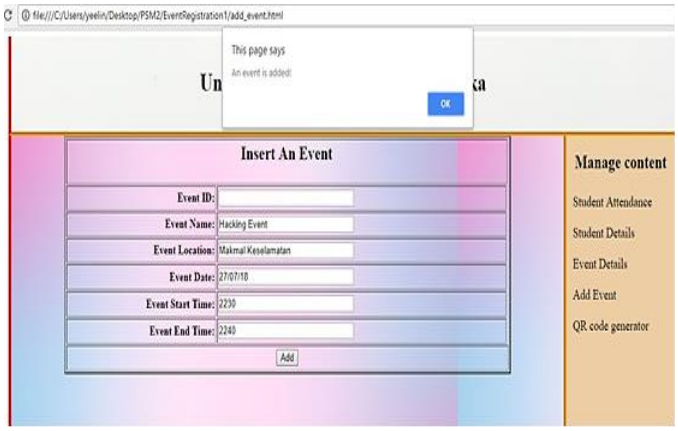



Fig. 10. Add Event

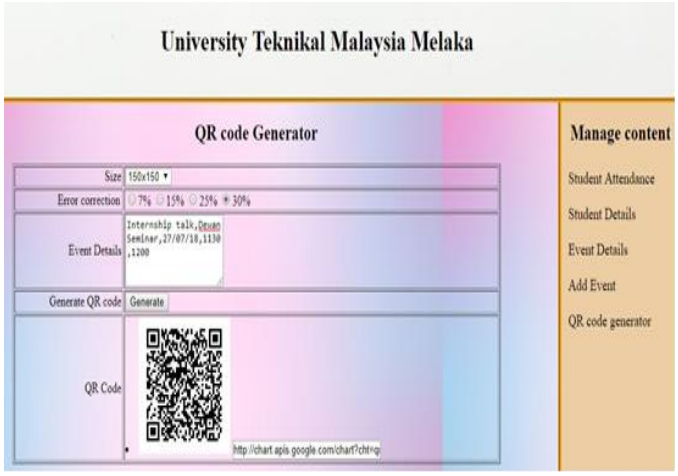



Fig. 11. QR Code Generator


Fig. 12. Event Details

In Figure 13, administrator can view the students' attendance according to a specific event. Students' attendance will be displayed on the page. Students' attendance list can be downloaded in Excel format.

In Figure 14, administrator can view the student's details according to a specific matric number and the student details will be displayed.

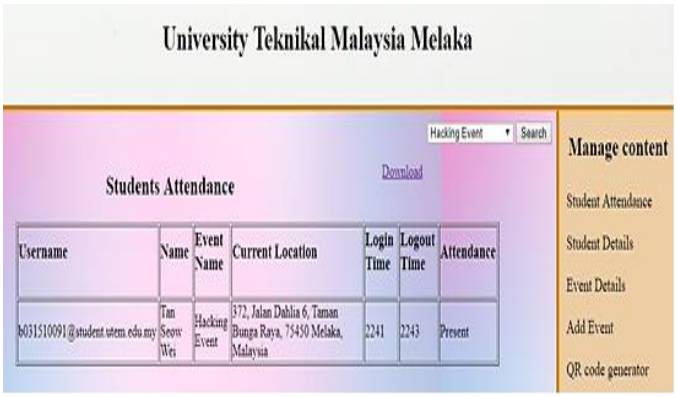



Fig. 13. Student Attendance Filtered By Events


Fig. 14. Student Details

## V. Evaluation And Discussion

System evaluation is conducted to verify system functionality, system effectiveness, and user satisfaction.

### A. System Functionality

User view and admin view are evaluated in integration test to examine the functionality of all the components together. Testing used wireless services provided by the university without increment of bandwidth from the IT Center and also mobile service provider. All modules in mobile application successfully connected to the database in the Web Administration system. This concludes that the user view and the admin view are functionally capable to operate.

Authentication for the system is unique as they used email which consists of student matric number. The application will check the identity based on the student database stored in the Firebase cloud. Registration of the event is design to make sure the student will participate in the event to the end. In this case, unless the student did not logout, their attendance will not be counted. In addition, there is condition to be met; they can only logout after the event ended.

### B. Performance Evaluation

Three parameters are used to evaluate the performance of the system:

*1)* Angle degree: the angle degree for handling the device when scanning the QR code.
*2)* Distance: the distance between the device and QR code
*3)* Brightness level: the level of brightness of the device

These performance parameters are selected to ensure the system can be performed in a high effectiveness [15].

Table IV, V and VI illustrate the performance of each parameter selected in this study. In Table IV, four angle degrees are evaluated namely 30, 45, 90 and 120 degrees. From the table, it can be seen that successful QR scanning applies only for 90 degrees' angle.

TABLE IV. Result for Angle Degree

| Angle for handling the device (degree) | Status (Pass/Fail) |
|---|---|
| 30 | Fail |
| 45 | Fail |
| 90 | Pass |
| 120 | Fail |

The results from Table IV shows that 90 degree is the preferred angle to handle the device when scanning the QR code. However, the status is Fail when user handle the device by using 30 degree, 45 degree, and 120 degree. This may be due to handling the device using these three angles will produce partial view of the QR code, thus making the system unable to detect the QR code. Besides, distance between the device and QR code plays an important role in this test.

In Table V, four distance evaluation is performed namely 3, 6, 9 and 12 cm. From the table it can be seen that successful QR scanning applies only on distance of 3 cm and 6 cm.

TABLE V. Performance Test of Distance between The Device and The QR Code

| Distance between the device and QR code (cm) | Status (Pass/Fail) |
|---|---|
| 3 | Pass |
| 6 | Pass |
| 9 | Fail |
| 12 | Fail |

From the results in Table V, 3cm and 6cm are the most suitable distance among these distances during the scanning of the QR code, whereas the status of 9cm and 12cm are Fail. These two distances of device are too far away from the QR code and the system is unable to detect the QR code.

In Table VI, four level of brightness evaluation is performed namely 10%, 30%, 60% and 100%. From the table, it can be seen that successful scanning applies on 60% and 100% brightness.

TABLE VI. PERFORMANCE TEST OF BRIGTNESS LEVEL OF THE DEVICE WHEN SCANNING THE QR CODE

| Level of brightness of device (%) | Status (Pass/Fail) |
|---|---|
| 10 | Fail |
| 30 | Fail |
| 60 | Pass |
| 100 | Pass |

From Table VI, 60% and 100% are the ideal brightness for the device to detect the QR code and extract the data inside it. Nevertheless, 10% and 30% of brightness is not suitable for device to detect the QR code because there is not enough light to decode the data that encoded in the QR code.

### C. User Acceptance Evaluation

Upon data collection, the level of acceptance was investigated by evaluating the application user interface quality, reliability, satisfaction and future use.

TABLE VII. USER ACCEPTANCE TEST

| Description | Mean | St. Dev. | Avg Mean |
|---|---|---|---|
| **Interface Quality** | | | |
| The way I interact with the application is pleasant. | 4.80 | 0.51 | 4.90 |
| The application is intuitive and user friendly and easy to understand. | 4.95 | 0.21 | |
| I think that I would not need the support of a technical person when using this application. | 5.00 | 0.00 | |
| I found that various functions in this system were well integrated. | 4.85 | 0.36 | |
| **Reliability** | | | |
| Time is saved and the purpose to avoid the event start lately is achieved. | 4.50 | 0.74 | 4.52 |
| The QR code scanner of the application is able to scan the QR code in high efficiency. | 4.90 | 0.30 | |
| User's current location is tracked correctly by using the application. | 4.15 | 0.85 | |
| **Satisfaction and Future Use** | | | |
| I feel convenient and confident when using the application for taking the attendance. | 4.75 | 0.62 | 4.80 |
| This application is an acceptable way for taking the students attendance. | 4.75 | 0.54 | |
| Overall, I am satisfied with the application for taking the attendance. | 4.90 | 0.30 | |

Table VII shows the feedback received from the users after they used the application. The highest average mean among three categories is the interface quality of the application which is 4.90. Besides, all users agree that they do not need technical support when using the application, hence the mean is 5.00. This application is accepted by most of the users because the system is user friendly and convenient to use, as the mean is 4.95. In terms of reliability, the average mean is 4.52. We conclude that this application is able to scan the QR code in high effectiveness due to its highest mean of 4.90. For tracking user's current location, this application score 4.15. Certain smartphones are slow to display GPS location; thus the score is low. The average mean of the satisfaction and future use category is 4.80. Therefore, it can be concluded that most of the users are satisfied with this application for taking the attendance as its mean is 4.90.

## VI. CONCLUSION AND FUTURE WORK

Location-aware Event Attendance System using QR code and GPS technology is implemented using android application and Firebase database in cloud to manage the attendance information. From the evaluation, the proposed system was capable to take the student attendance by scanning the QR code. The GPS location, time login and logout were tracked to ensure full attendance. We found positive feedback for the system in the user acceptance test. However, this system can only support android application which makes it inconvenient for iOS users. Furthermore, the proposed system is only capable of tracking the location without calculating the distance to the event venue. In addition, the application also needs strong Internet connection.

For future work, we plan to improve the application operability to support both android and iOS smart phone. To calculate the distance between the user and the venue, we propose to incorporate Google Maps Distance Matrix API in the application. To decrease false attendance and secure authentication, the authors also plan to apply factor-based authentication scheme with low cost method in the application [16]. This study can be extended to other areas such as recommender system.

## ACKNOWLEDGMENT

The authors gratefully acknowledge the help of the Knowledge and Communication Services Centre, UTeM in providing the requirement in this research. A high appreciation to Center of Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM) for supporting the work done in this paper.

## REFERENCES

[1] K.W. Tracy, "Mobile Application Development Experiences on Apple's iOS and Android OS," Ieee Potentials, vol. 31(4), pp. 30-4, July 2012.

[2] K. Jacksi, F. Ibrahim, and S. Ali, "Student Attendance Management System," Scholars Journal of Engineering and Technology, vol. 6(2), pp. 49-53, Feb 2018.

[3] S. Lukas, A.R. Mitra, R.I. Desanti, and D. Krisnadi, "Student attendance system in classroom using face recognition technique," In IEEE International Conference of Information and Communication Technology Convergence (ICTC), pp. 1032-1035, Oct 2016.

[4] M.M. Said, M.H. Misran, M.A Othman, M.M. Ismail, H.A. Sulaiman, A. Salleh, and N. Yusop, "Biometric attendance," In 2014 IEEE International Symposium Technology Management and Emerging Technologies (ISTMET), pp. 258-263, May 2014.

[5] S. Kadry, and M. Smaili, "Wireless attendance management system based on iris recognition," Scientific Research and essays, vol. 5(12), pp. 1428-35, Sep 2013.

[6] N. Mohamed Kutty and S. Mathai, "Face Recognition – A Tool for Automated Atttendance System," International Journals of Advanced Research in Computer Science and Software Engineering, vol. 7(6), pp. 334-336, June 2017.

[7] K.L. Sudha, S. Shinde, T. Thomas, and A. Abdugan, "Barcode based student attendance system," International Journal of Computer Applications, vol. 119(2), pp. 1-4, Jan 2015.

[8] S., Noguchi, M. Niibori, E. Zhou, and M. Kamada, "Student attendance management system with bluetooth low energy beacon and android devices," In 18th IEEE International Conference Network-Based Information Systems (NBiS), pp. 710-713, Sep 2015.

[9] M. Zhi, and M.M Singh, "RFID-enabled smart attendance management system," In Future Information Technology-II, pp. 213-231, Springer, Dordrecht, 2015.

[10] P. Subpratatsavee, T. Promjun, W. Siriprom, and W. Sriboon, "Notice of Violation of IEEE Publication Principles Attendance System Using NFC Technology and Embedded Camera Device on Mobile Phone," In IEEE 2014 International Conference Information Science and Applications (ICISA), pp. 1-4, May 2014.

[11] A. Rahni, N. Zainal, M.Z. Adna, N.E. Othman, and M.F. Bukhori, "Development Of The Online Student Attendance Monitoring System (Samstm) Based On QR-Codes And Mobile Devices," Journal of Engineering Science and Technology, vol. 1(10), pp. 28-40, May 2015.

[12] F. Masalha, and N. Hirzallah, "A students attendance system using QR code," International Journal of Advanced Computer Science and Applications, vol. 5(3), pp. 75-9, 2014.

[13] M.H.M Baban, "Attendance Checking System Using Quick Response Code for Students at the University of Sulaimaniyah," Journal of Mathematics And Computer Science, vol. 10, pp. 189-198, 2014.

[14] X. Wei, A. Manori, N. Devnath, N. Pasi, and V. Kumar, "QR Code Based Smart Attendance System," International Journal of Smart Business and Technology, vol. 5(1), pp 1-10, 2017.

[15] Y. Liu, J. Yang, and M. Liu, "Recognition of QR Code with mobile phones," In IEEE Control and Decision Conference, pp. 203-206, July 2008.

[16] M.A. Ferrag, L.A. Maglaras, A. Derhab and A.A. Kobra, "Taxonomy of Biometric-based Authentication Schemes for Mobile Devices," Unpublished.

# An Enhanced Malay Named Entity Recognition using Combination Approach for Crime Textual Data Analysis

Siti Azirah Asmai[1], Muhammad Sharilazlan Salleh[2], Halizah Basiron[3], Sabrina Ahmad[4]
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya
76100 Durian Tunggal, Melaka, Malaysia

***Abstract*—Named Entity Recognition (NER) is one of the tasks in the information extraction. NER is used for extracting and classifying words or entities that belong to the proper noun category in text data such as person's name, location, organization, date and others. As seen in today's generation, social media such as web pages, blogs, Facebook, Twitter, Instagram and online newspapers are among the major contributors to the generation of information. This paper presents an enhanced Malay Named Entity Recognition model using combination fuzzy c-means and K-Nearest Neighbours Algorithm method for crime analysis. The results showed that this combination method could improve the accuracy performance on entity recognition of crime data in Malay. The model is expected to provide a better method in the process of recognizing named entities for text analysis particularly in Malay.**

***Keywords*—Named entity recognition; information extraction; fuzzy c-means; k-nearest neighbors; malay language; crime data**

## I. Introduction

Information is one of the important sources in human life that is increasingly rising and technologically. At all times, various types of information have been generated on the internet and the amount of information is constantly increasing from time to time. Information consisting of various types such as text, images, audio, video, data, and so on are increasingly being generated on the internet which are largely unstructured. This growing number of information affects the daily lives of people in work, learning and lifestyle. Effective management and organization of information is a key strategy for addressing the problem of finding useful information. The appropriate techniques and methods are very necessary to process and extract the essential knowledge contained in this information.

Therefore, this paper presented the Malay named entity recognition using clustering and classification method. The rest of this paper is organized as follows. In Section 2, it discusses the related work for the named entity recognition task. Section 3 presents techniques and machine learning algorithms for NER. Then, Section 4 discusses the Malay NER and follow by its approach in Section 5. Next, the experiment result and discussion are elaborated in Section 6. Finally, Section 5 covers the conclusion.

## II. Related Work

Named Entity Recognition (NER) is important in analyzing the crime report to address the problem of crime due to the use of different languages in writing crime reports for each country. When a lot of information relates to crime occurrences are available on the web with many specific entities, many techniques can be used in NER for extracting useful information for better crime analysis and execution actions that explain by Hosseinkhani, Koochakzaei, and Keikhaee [1].

Shabat and Omar [2] have implemented NER tasks using an ensemble framework that focuses on designing models to extract specific criminal information from the Web. Their main goal is to integrate the set of features and classification algorithms in an orderly way to synthesize more precise classification procedures. Three base-classifiers specifically Naïve Bayes, Support Vector Machine and K-Nearest Neighbor classifiers are used for each of the feature sets and these three classifiers are combined using a weighted voting ensemble method.

Alkaff and Mohd [3] have analyzed online news, blogs and social networking sites on the internet using gazetteers and rule-based extraction for named entity recognition in identifying crime hot spots. Therefore, an accurate natural language processing technique is needed to be explored to capture and recognize named entity within open domain textual data effectively.

Execution processing recognition named entity analysis requires several steps to achieve the objectives of the research. The steps including the pre-processing stage, the annotation stage and evaluation or developing system stage. Based on Jurafsky and Martin [4] there some basic steps in the statistical sequence labelling approach to creating a named entity recognition system. The following Fig. 1 shows the steps illustration.

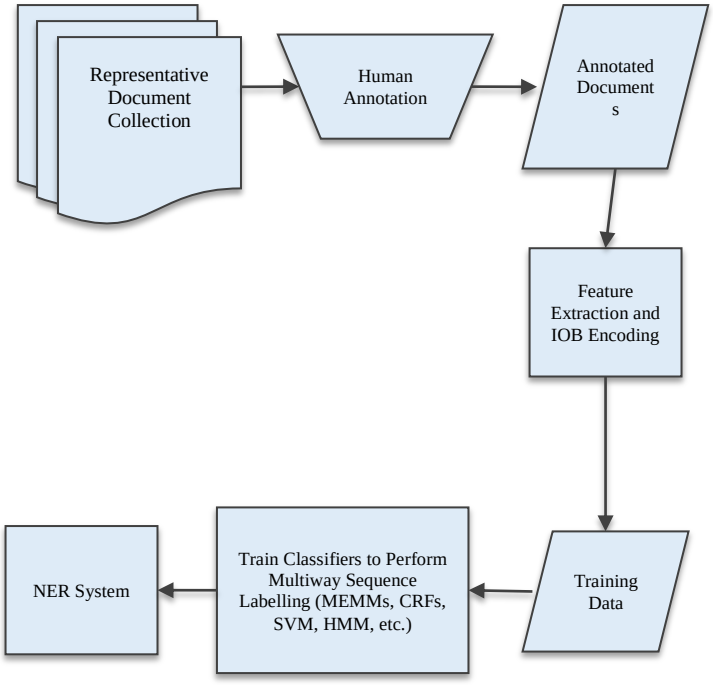



Fig. 1. Basic steps approach for NER

## III. Techniques and Machine Learning Algorithms For NER

Many method and techniques are being continuously developed which is it more focus on managing of information and knowledge. Previous knowledge management a strongly focuses on just keeping large amounts of data for data mining. Now the growing use of the Internet and the information burden placed a huge demand for managing intelligent information efficiently and effectively. This application of artificial intelligence methods and research in the growing area of human-machine interaction is ahead grounds for more investigations.

### A. Rule-based Approaches

In computer science, rule-based systems are used as a means of storing and manipulating knowledge to interpret information in a useful way. They are often used in artificial intelligence applications and research. Normally, the term Rule-Based System ('rules-based system') is used for systems involving a set of man-made rules or rules outlined. Today, these rules-based systems are widely being used and implemented for many kinds of problem and tasks. As developing the text analysis that focuses on NER task, the rule-based approach is used for the recognition of named entities by defining rules regarding the status of entity members' position in the phrase or sentence. The constraints in the implementation of this method lie in the capability of a pattern definition that is usually done by a linguist. Rule-based NER is also too dependent on the language used.

In general, the NER system using a rule-based approach has Part-of-Speech (POS) tagger, sentence or phrase syntax and orthographic, such as word capitalization pattern combined with the data dictionary. Eftimov et al. [4] state that the NER method using a rule-based approach uses a regular expression that combines information from the source terminology and interests of the feature entity. The main drawback of this method is the construction of manual rules, which are time-consuming and dependent on the domain. Eftimov et al. [4] combined the terminological-driven NER with rules-based NER as their proposed rule-based method called as DrNER extracting knowledge for evidence-based dietary recommendations. The basic structure of the rule-based expert system is shown in Fig. 2.

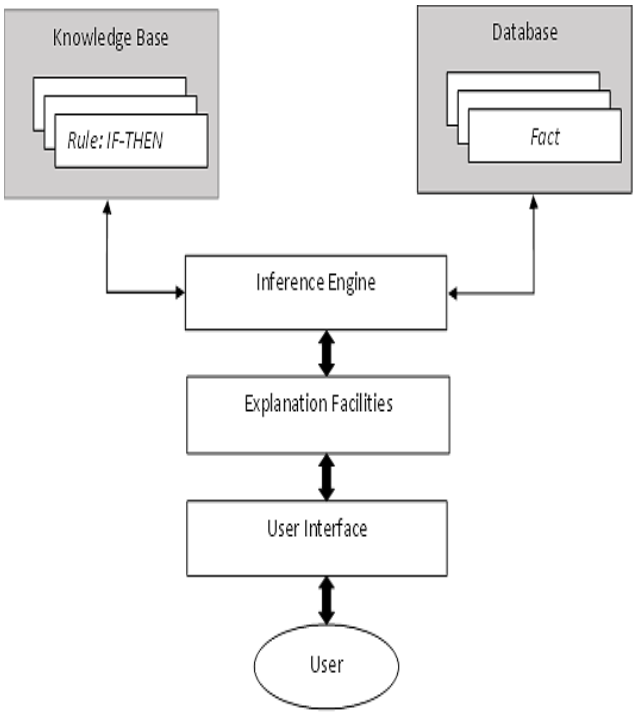



Fig. 2. Basic structure of rule-based expert system (Abraham, 2005)

### B. Learning-based Approaches

- Supervised Learning

The ability to learn unnamed entities is an essential part of the NER solution. Early studies were mostly based on the supervised learning (SL). The supervised learning algorithm is the process of forming a relationship model and dependence between predictive output and input characteristics so prediction of output values for new data can be predicted based on the relationships studied from previous datasets. Kotsiantis [5] stated that supervised machine learning is an algorithm that generates the general hypothesis based on externally supplied examples and hence is used in making predictions about future instances. In other meaning, the purpose of this learning is to build a brief model that distribute class labels based on predictor features.

Morwal [6], Chopra and Morwal [7] use Hidden Markov in named entity recognition. While Ahmed and Sathyaraj [8] applied maximum entropy to recognize entity sets from a given text such as name, location and organization. With the different variant of SL techniques, it offers tagging words of the test corpus from the define corpus that require a large set of heuristic rules and clusters.

- Unsupervised Learning

One of the learning based approaches for pattern recognition is unsupervised learning (USL). Unsupervised learning is an artificial intelligence algorithm (AI) that performs data isolation in a dataset using unlabelled or classified information where the isolation is based on the hidden features contained in the data. This algorithm acts on this information or data without guidance. The AI system used can arrange information based on similarities and differences in information although no category is provided among the data. The AI system algorithm also acts on data without prior training. Sathya and Abraham [9] stated that unsupervised learning model recognises information based on heuristic patterns and Reinforcement learning learns through trial and error interactions with their surroundings (rewards / penalties).

Unsupervised learning is also used in named entity recognition tasks. This learning-based is one of the approaches in solving the problems encountered in the task of named entity recognition. Li et al. [10] presented the unsupervised NER system without explicit human label efforts named TwiNER for targeted tweet streams in the Twitter application. The system not dependent on unreliable local linguistic features. Furthermore, S. Zhang and Elhadad [11] also proposed an unsupervised approach in the biomedical field for NER task by extracting named entities from biomedical text. This unsupervised approach for NER was conducted using three main step which are seed term collection, boundary detection and entity classification.

- Semi-supervised Learning

Semi-supervised learning is a technique that is a combination of supervised learning and unsupervised learning. A variety of semi-supervised learning method tries to generate high-quality training data automatically from the unlabelled corpus. By using the semi-supervised learning technique, it can produce considerable improvement in learning accuracy. This improvement in learning accuracy can help in the structured process of extracting named entities such as location, person, type of crime and other entities involved in the crime situation more accurately from any unstructured data like email messages, word processing documents and web blogs.

However, traditional semi-supervised learning methods remain to rely on the high quality of the labelled entity to learn the context of unlabelled data in textual data. Fuzzy semi-supervised clustering it offers a new opportunity to overcome classical methods and crisp semi-supervised hierarchical clustering. However, fuzzy semi-supervised clustering is still a new subject and not many studies have been done with fuzzy semi-supervised cluster related on named entity recognition in the literature. Diaz-Valenzuela, Vila, and Martin-Bautista [12] use fuzzy semi-supervised clustering approach to classifying scientific publications in digital web libraries. They use the concepts of fuzzy must-link and fuzzy cannot-link constraints for identifying optimum α-cut of a dendrogram.

Castellano, Fanelli, and Torsello [13] use a semi-supervised fuzzy clustering algorithm to group shapes into some clusters. Each cluster is represented by a prototype that is manually labelled and used to annotate shapes belonging to that cluster. To capture the evolution of the image set over time, the previously discovered prototypes are added as pre-labelled objects to the current shape set and semi-supervised clustering is applied again. Both of these recent studies improve the accuracy of the group clusters under the supervision of a limited number of labelled data.

## IV. MALAY NER

This research discusses the overview of Malay language based on some aspects related to this scope. The Malay language is also one of the language fields that get researchers interest to implement the named entity recognition task. It focuses on the identification of proper nouns in Malay. Like other languages, the Malay language also has its own characteristics in the presentation of information based on the order of sentences and the form of words that have certain meanings. The Discussions on the execution of named entity recognition in the Malay language include orthography, morphology, structure, and so on.

Alfred, Chin Leong, Kim On, and Anthony [14] explains that as one of the processes in Text Mining, a named entity recognition is very useful for information extraction by helping user for entities identification and detection like the person, location and organization. They also argue that different NER processes need to be applied to different languages due to morphological differences. So, a Rule-Based Named-Entity Recognition algorithm for Malay articles has been proposed based on a Malay part-of-speech (POS) tagging features and contextual features in dealing with Malay language articles. The use of a set of rules and manually-specified dictionary lists by the human is a method used in the Rule-Based NER algorithm in identifying named entities. Due to the lack of annotated corpus sources for the Malay language which can be used as training data, they have used rule-based methods rather than using machine learning method to identify person, organization and location as three named entities major types. The rule has been made based on the POS-tagging contexts. The F-Measure result's value during conducted the NER experimental was 89.47%.

Furthermore, another experiment was conducted by Sulaiman et al [15] to detect Malay named entity recognition. Stanford NER and Illinois NER tools are used to identify the Malay named entity using online news articles as a process of measuring the capabilities of this tool in the identification of Malay entities. Experimental comparisons have found that Stanford NER tends to yield higher results on F1 and Precision than Illinois NER. These two tools, Illinois NER and Stanford NER are developing based on machine learning method. They conclude that, for improvements in the named entity task in Malay, most NER Malays are used rule-based methods. After conducting experiments, they found that both NERs tools showed a low detection result for the Malay corpus because there were many errors when identifying entities. This is because of the morphological differences between Malay and English.

Besides that, Salleh, Asmai, Basiron, and Ahmad [16] was applied conditional random fields method in developed an automated Malay Named Entity Recognition (AMNER) conceptual model to recognize entities for the Malay language. Current approaches for Malay NER are more using a set of rules and list of dictionaries set by the human to identify entities. These rules work to extract the pattern of an entity such as location, organization and other entities based on their basic pattern. Due to limitation, the libraries or dictionaries used should always be updated for recognizing named entities. The Malay language features as the main factor on their development model as the guidance for the named entity recognition process. There are several structures in Malay language writing as follows.

### A. Orthography

In the execution of named entity recognition tasks, one of the things involved is the conventional spelling system of a language called orthography. The Malay language also has its own orthography in the spelling structure. Based on Cho [17], they explain that in the present time, the Latin alphabet has been used for orthography and spelling system for the Malay and Indonesian languages that have been made by Western linguists. Besides that, Zaidi, Rozan, and Mikami [18] stated that with the use of Malay language standard words using 26 letter alphabets known as Rumi in Malay, it is compatible with communication technology and has the potential to use only the text-based features for communicating in Malay. Orthography used in Malay includes spelling norms, hypotheses, emphasis, punctuation, capitalization, fractions of words.

### B. Morphology

Furthermore, morphology is also used in the research of named entity recognition. Morphology in linguistics is the study of the words inner structure and word formation that forms the essential part of today's linguistic study. It describes how the words are formed and their relationship to other word focus on the same language. By breaking the words down into smaller, meaningful part, this smallest meaningful part of a word is called a morpheme. Word structure and part of words analyzed by morphology include stems, prefixes, suffixes and root words. In addition, it also sees the part of speech, the way the context can change the word's pronunciation and meaning, as well as the intonation and pressure in one word.

## V. A Malay Named Entity Recognition Approach

The research is conducted through five phases represented in the form of research design. Each phase in the research design is intensively investigated and then used to facilitate the next phase of the research. The Phase One begins with data acquisition, data obtained in the form of web pages and unstructured. The Phase Two is pre-processing data and is followed by a Phase Three that focused on features extraction. Then, the development of the NER Malay model was carried out in Phase Four. Finally, an accuracy of the entity recognition is evaluated in Phase Five. Fig. 3 illustrates the design of the proposed Malay Named Entity Recognition (MNER) approach.

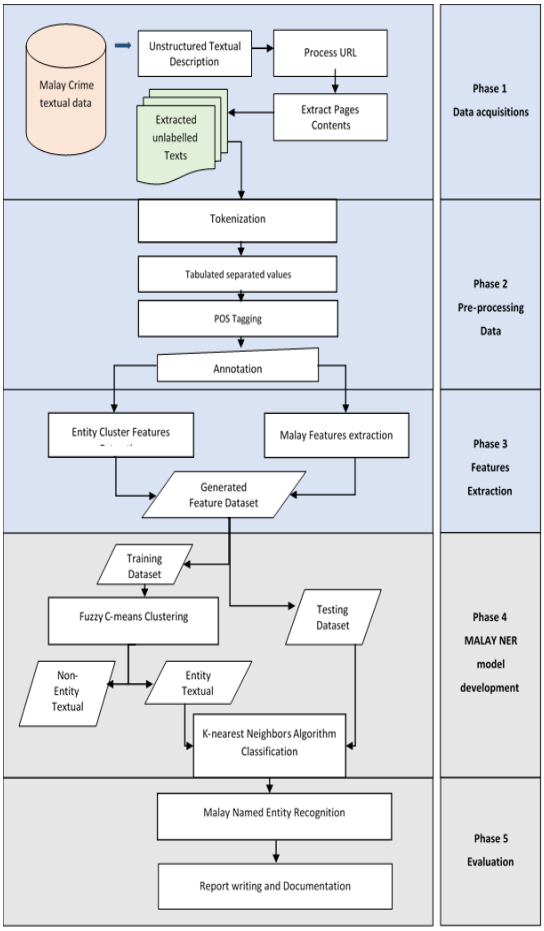



Fig. 3. The Proposed Malay Named Entity Recognition Design

### A. Data Acquisition

Based on research design in Fig. 3, data acquisition is conducted in Phase One. Data is obtained from the Malay Crime News PDRM Website in the form of web pages. These web pages contain some elements such as URL links, images, and texts that need to be processed as they are in unstructured form. The page contents are extracts to obtain the required information which as extracted unlabeled PDRM News Texts.

### B. Pre-processing Data

Pre-processing involved four tasks towards the data. As the process in Phase Two, the documents that contain many unstructured data need to delimit into meaningful units by performing tasks like tokenization, tabulation values, POS tagging and annotation. Then, after the annotation process was done, the data were divided into two parts: training data and testing data. The following Fig. 4 shows the process for pre-processing data.

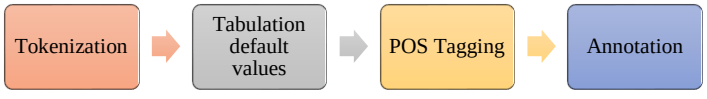



Fig. 4. Pre-processing Data

TABLE I. THE PENN TREEBANK PART-OF-SPEECH TAG SET

| Tag | Details |
|---|---|
| *CC* | conjunction, coordinating |
| *CD* | cardinal number |
| *DT* | determiner |
| *EX* | existential there |
| *FW* | foreign word |
| *IN* | conjunction, subordinating or preposition |
| *JJ* | adjective |
| *JJR* | adjective, comparative |
| *JJS* | adjective, superlative |
| *LS* | list item marker |
| *MD* | verb, modal auxillary |
| *NN* | noun, singular or mass |
| *NNS* | noun, plural |
| *NNP* | noun, proper singular |
| *NNPS* | noun, proper plural |
| *PDT* | predeterminer |
| *POS* | possessive ending |
| *PRP* | pronoun, personal |
| *PRP$* | pronoun, possessive |
| *RB* | adverb |
| *RBR* | adverb, comparative |
| *RBS* | adverb, superlative |
| *RP* | adverb, particle |
| *SYM* | symbol |
| *TO* | infinitival to |
| *UH* | interjection |
| *VB* | verb, base form |
| *VBZ* | verb, 3rd person singular present |
| *VBP* | verb, non-3rd person singular present |
| *VBD* | verb, past tense |
| *VBN* | verb, past participle |
| *VBG* | verb, gerund or present participle |
| *WDT* | *wh*-determiner |
| *WP* | *wh*-pronoun, personal |
| *WP$* | *wh*-pronoun, possessive |
| *WRB* | *wh*-adverb |
| . | punctuation mark, sentence closer |
| , | punctuation mark, comma |
| *:* | punctuation mark, colon |
| *(* | contextual separator, left paren |
| *)* | contextual separator, right paren |

- Tokenization

The text data file (.txt) that were presented in unstructured data consisted of sentences and paragraph which were tokenized as the process of separating a text into valuable elements, words, phrases, symbols or digits called tokens. The tokens were presented in a list as the input for further processing.

- Tabulation Values

Next, the token text file was processed to store data in a tabulator structure like spreadsheet data. The file was divided into three rows namely token data, part of speech tag (POS) and named entity tag. Before continuing to the annotation stage, entity tag column was set as default value "O" as outside or other.

- POS Tagging

Every token in the file was also annotated with POS tagging bands such as CC, CD, NN, VB and others. The description of The Penn Treebank POS tagset is based on Table 1.

- Annotation

Then, the file was annotated with entities types. There are five types of entities that are being worked out in this research. Those entities are person name, location, organization, date, and types of crime labelled as PERSON, LOCATION, ORGANIZATION, DATE and CRIME TYPE. For non-entity types, they are labelled as OTHER. The final preprocessing dataset produced is shown in both Fig. 6 and Fig. 7 respectively with their features extraction.

### *C. Features Extraction*

In Phase Three, the process of extracting features for the named entity recognition task has been performed. Feature extraction is divided into two parts. The first part, some features have been extracted for use in clustering process and in the second part; some other features have been extracted for use in the process of classification. The generated feature dataset is produced in this phase for further analysis. The features selected for both parts are as appropriate to carry out the task of recognizing named entities in the Malay language. The details process of extracting these features are discussed as shown in Fig. 5.

### *D. Malay NER Model Development*

Furthermore, in Phase Four, there are two types of learning used, namely clustering and classification. Fuzzy C-Means as clustering method is used to cluster the data either entity or non-entity. After that, the correct entities that have been clustered are labelled based on more detailed entity types which are person, location, organization, date, and type of crime. Then, these entities through the classification process by using K-nearest Neighbors Algorithm Classification.

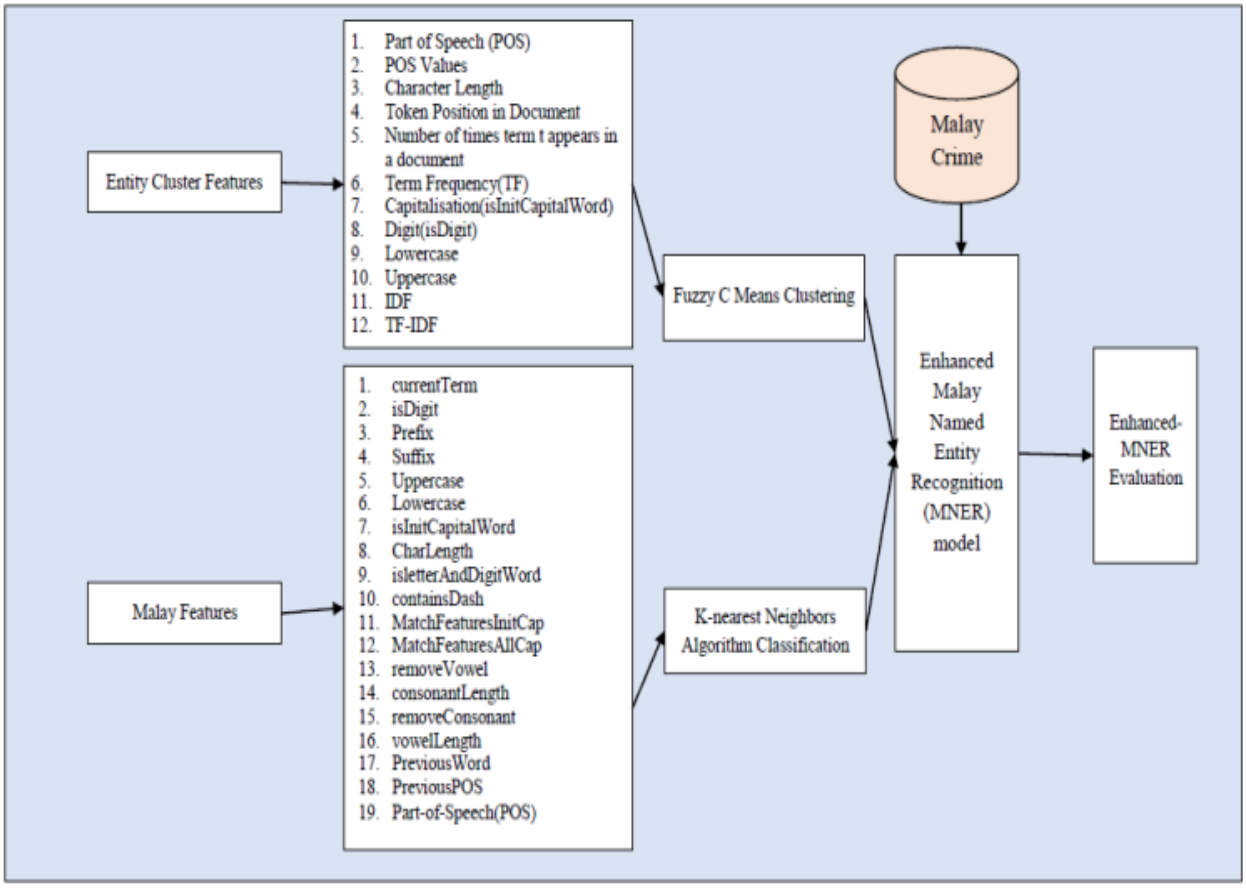



Fig. 5. MNER Features Extraction

| Row No. | Term | POS | POS_ValueNormalize | Character Length_normalize | Token Position in Document | noAppearNorm | Term Frequency(TF) | Lowercase | Uppercase | TFIDF | CLASS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | BELUKAR | NN | 0.06666666666666667 | 0.2692307692307692 | 1.0 | 0.07272727272727272 | 0.016597510373443983 | 0.0 | 1.0 | 0.028885687792435583 | NON_ENTITY |
| 2 | JADI | VB | 0.46666666666666667 | 0.15384615384615385 | 0.9958506224066339 | 0.03636363636363636 | 0.008298755186721992 | 0.0 | 1.0 | 0.014442843896217791 | NON_ENTITY |
| 3 | TEMPAT | NN | 0.06666666666666667 | 0.2307692307692307 8 | 0.9917012448132178 | 0.03636363636363636 | 0.008298755186721992 | 0.0 | 1.0 | 0.009446495420467065 | NON_ENTITY |
| 4 | JUAL | VB | 0.46666666666666667 | 0.15384615384615385 | 0.9875518672199171 | 0.05454545454545454 | 0.012448132780082987 | 0.0 | 1.0 | 0.017917004487751364 | NON_ENTITY |
| 5 | HEROIN | NN | 0.06666666666666667 | 0.2307692307692307 8 | 0.9834024896265556 | 0.05454545454545454 | 0.012448132780082987 | 0.0 | 1.0 | 0.017917004487751364 | NON_ENTITY |
| 6 | Jabatan | NNP | 0.13333333333333333 | 0.2692307692307692 | 0.9792531120331795 | 0.07272727272727272 | 0.016597510373443983 | 0.0 | 0.0 | 0.0 | ENTITY |
| 7 | Sumber | NNP | 0.13333333333333333 | 0.2307692307692307 8 | 0.9751037344839834 | 0.018181818181818182 | 0.004149377593360996 | 0.0 | 0.0 | 0.0 | ENTITY |
| 8 | Strategik | NNP | 0.13333333333333333 | 0.34615384615384615 | 0.9709543568464073 | 0.018181818181818182 | 0.004149377593360996 | 0.0 | 0.0 | 0.0 | ENTITY |
| 9 | Dan | NNP | 0.13333333333333333 | 0.11538461538461539 | 0.9668049792531112 | 0.12727272727272726 | 0.029045643153526972 | 0.0 | 0.0 | 0.0 | ENTITY |
| 10 | Teknologi | NNP | 0.13333333333333333 | 0.34615384615384615 | 0.9626556016597511 | 0.018181818181818182 | 0.004149377593360996 | 0.0 | 0.0 | 0.0 | ENTITY |
| 11 | Jabatan | NNP | 0.13333333333333333 | 0.2692307692307692 | 0.9585062240663901 | 0.07272727272727272 | 0.016597510373443983 | 0.0 | 0.0 | 0.0 | ENTITY |
| 12 | Integriti | NNP | 0.13333333333333333 | 0.34615384615384615 | 0.9543568464730291 | 0.018181818181818182 | 0.004149377593360996 | 0.0 | 0.0 | 0.0 | ENTITY |
| 13 | Dan | NNP | 0.13333333333333333 | 0.11538461538461539 | 0.9502074688796168 | 0.12727272727272726 | 0.029045643153526972 | 0.0 | 0.0 | 0.0 | ENTITY |
| 14 | Pematuhan | NNP | 0.13333333333333333 | 0.34615384615384615 | 0.9460580912863107 | 0.018181818181818182 | 0.004149377593360996 | 0.0 | 0.0 | 0.0 | ENTITY |
| 15 | Standard | NNP | 0.13333333333333333 | 0.3076923076923077 | 0.9419087136929146 | 0.018181818181818182 | 0.004149377593360996 | 0.0 | 0.0 | 0.0 | ENTITY |
| 16 | -LRB- | ( | 0.8 | 0.1923076923076923 2 | 0.9377593360995851 | 0.09090909090909091 | 0.020746887966804598 | 0.0 | 1.0 | 0.0 | NON_ENTITY |

Fig. 6. Sample of Feature Extraction for FCM

| Uppercase | Lowercase | isInitCapit | isDigit | isletterAn | MatchFea | isAllCapita | containsD | MatchFea | Term | POS | Class | id | Term-1 | Term-2 | Term-3 | POS-1 | POS-2 | POS-3 | CharLeng | Prefix | Suffix | removeVc | consonan | removeCo | vowelLeng | currentTe | AverageAt |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | .0 | PREMIS | NN | O | 1.0 | | | | | | | 6.0 | PREMI | REMIS | PRMS | 4.0 | EI | 2.0 | premis | 1.2 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | VCD/DVD | NN | O | 2.0 | PREMIS | | | NN | | | 7.0 | VCD/D | D/DVD | VCD/DVD | 7.0 | / | 1.0 | vcd/dvd | 1.3 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | .0 | HARAM | JJ | O | 3.0 | VCD/DVD | PREMIS | | NN | NN | | 5.0 | HARAM | HARAM | HRM | 3.0 | AA | 2.0 | haram | 1.0 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | .0 | DISERBU | VB | O | 4.0 | HARAM | VCD/DVD | PREMIS | JJ | NN | NN | 7.0 | DISER | SERBU | DSRB | 4.0 | IEU | 3.0 | diserbu | 1.3 |
| 1.0 | 1.0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | , | , | O | 5.0 | DISERBU | HARAM | VCD/DVD | VB | JJ | NN | 1.0 | , | , | , | 1.0 | , | 1.0 | , | .4 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | .0 | INDIVIDU | NN | O | 7.0 | 5 | , | DISERBU | CD | , | VB | 8.0 | INDIV | IVIDU | NDVD | 4.0 | IIIU | 4.0 | individu | 1.5 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | .0 | DITAHAN | VB | O | 8.0 | INDIVIDU | 5 | , | NN | CD | , | 7.0 | DITAH | TAHAN | DTHN | 4.0 | IAA | 3.0 | ditahan | 1.3 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Jabatan | NNP | ORGANIZA | 9.0 | DITAHAN | INDIVIDU | 5 | VB | NN | CD | 7.0 | Jabat | batan | Jbtn | 4.0 | aaa | 3.0 | jabatan | 1.3 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Sumber | NNP | ORGANIZA | 10.0 | Jabatan | DITAHAN | INDIVIDU | NNP | VB | NN | 6.0 | Sumbe | umber | Smbr | 4.0 | ue | 2.0 | sumber | 1.2 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Strategik | NNP | ORGANIZA | 11.0 | Sumber | Jabatan | DITAHAN | NNP | NNP | VB | 9.0 | Strat | tegik | Strtgk | 6.0 | aei | 3.0 | strategik | 1.7 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Dan | NNP | ORGANIZA | 12.0 | Strategik | Sumber | Jabatan | NNP | NNP | NNP | 3.0 | Dan | Dan | Dn | 2.0 | a | 1.0 | dan | .7 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Teknologi | NNP | ORGANIZA | 13.0 | Dan | Strategik | Sumber | NNP | NNP | NNP | 9.0 | Tekno | ologi | Tknlg | 5.0 | eooi | 4.0 | teknologi | 1.7 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Jabatan | NNP | ORGANIZA | 14.0 | Teknologi | Dan | Strategik | NNP | NNP | NNP | 7.0 | Jabat | batan | Jbtn | 4.0 | aaa | 3.0 | jabatan | 1.3 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Integriti | NNP | ORGANIZA | 15.0 | Jabatan | Teknologi | Dan | NNP | NNP | NNP | 9.0 | Integ | griti | ntgrt | 5.0 | leii | 4.0 | integriti | 1.7 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Dan | NNP | ORGANIZA | 16.0 | Integriti | Jabatan | Teknologi | NNP | NNP | NNP | 3.0 | Dan | Dan | Dn | 2.0 | a | 1.0 | dan | .7 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Pematuhai | NNP | ORGANIZA | 17.0 | Dan | Integriti | Jabatan | NNP | NNP | NNP | 9.0 | Pemat | tuhan | Pmthn | 5.0 | eaua | 4.0 | pematuhai | 1.7 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Standard | NNP | ORGANIZA | 18.0 | Pematuhai | Dan | Integriti | NNP | NNP | NNP | 8.0 | Stand | ndard | Stndrd | 6.0 | aa | 2.0 | standard | 1.5 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | -LRB- | ( | O | 19.0 | Standard | Pematuha | Dan | NNP | NNP | NNP | 5.0 | -LRB- | -LRB- | -LRB- | 5.0 | -- | 2.0 | -lrb- | 1.2 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | 1.0 | JIPS | NNP | ORGANIZA | 20.0 | -LRB- | Standard | Pematuha | ( | NNP | NNP | 4.0 | JIPS | JIPS | JPS | 3.0 | I | 1.0 | jips | .9 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | -RRB- | ) | O | 21.0 | JIPS | -LRB- | Standard | NNP | ( | NNP | 5.0 | -RRB- | -RRB- | -RRB- | 5.0 | -- | 2.0 | -rrb- | 1.2 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Jabatan | NNP | ORGANIZA | 22.0 | -RRB- | JIPS | -LRB- | ) | NNP | ( | 7.0 | Jabat | batan | Jbtn | 4.0 | aaa | 3.0 | jabatan | 1.3 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Pencegaha | NNP | ORGANIZA | 23.0 | Jabatan | -RRB- | JIPS | NNP | ) | NNP | 10.0 | Pence | gahan | Pncghn | 6.0 | eeaa | 4.0 | pencegaha | 1.8 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Jenayah | NNP | ORGANIZA | 24.0 | Pencegaha | Jabatan | -RRB- | NNP | NNP | ) | 7.0 | Jenay | nayah | Jnyh | 4.0 | eaa | 3.0 | jenayah | 1.3 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Dan | NNP | ORGANIZA | 25.0 | Jenayah | Pencegaha | Jabatan | NNP | NNP | NNP | 3.0 | Dan | Dan | Dn | 2.0 | a | 1.0 | dan | .7 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Keselamat | NNP | ORGANIZA | 26.0 | Dan | Jenayah | Pencegaha | NNP | NNP | NNP | 11.0 | Kesel | matan | Kslmtn | 6.0 | eeaaa | 5.0 | keselamat | 2.0 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Komuniti | NNP | ORGANIZA | 27.0 | Keselamat | Dan | Jenayah | NNP | NNP | NNP | 8.0 | Komun | uniti | Kmnt | 4.0 | ouii | 4.0 | komuniti | 1.5 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | -LRB- | ( | O | 28.0 | Komuniti | Keselamat | Dan | NNP | NNP | NNP | 5.0 | -LRB- | -LRB- | -LRB- | 5.0 | -- | 2.0 | -lrb- | 1.2 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | 1.0 | JPJKK | NNP | ORGANIZA | 29.0 | -LRB- | Komuniti | Keselamat | ( | NNP | NNP | 5.0 | JPJKK | JPJKK | JPJKK | 5.0 | | .0 | jpjkk | 1.1 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | 1.0 | -RRB- | ) | O | 30.0 | JPJKK | -LRB- | Komuniti | NNP | ( | NNP | 5.0 | -RRB- | -RRB- | -RRB- | 5.0 | -- | 2.0 | -rrb- | 1.2 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Jabatan | NNP | ORGANIZA | 31.0 | -RRB- | JPJKK | -LRB- | ) | NNP | ( | 7.0 | Jabat | batan | Jbtn | 4.0 | aaa | 3.0 | jabatan | 1.3 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Siasatan | NNP | ORGANIZA | 32.0 | Jabatan | -RRB- | JPJKK | NNP | ) | NNP | 8.0 | Siasa | satan | Sstn | 4.0 | iaaa | 4.0 | siasatan | 1.5 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Dan | NNP | ORGANIZA | 33.0 | Siasatan | Jabatan | -RRB- | NNP | NNP | ) | 3.0 | Dan | Dan | Dn | 2.0 | a | 1.0 | dan | .7 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Penguatku | NNP | ORGANIZA | 34.0 | Dan | Siasatan | Jabatan | NNP | NNP | NNP | 14.0 | Pengu | asaan | Pngtksn | 7.0 | euauaaa | 7.0 | penguatku | 2.5 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Trafik | NNP | ORGANIZA | 35.0 | Penguatku | Dan | Siasatan | NNP | NNP | NNP | 6.0 | Trafi | rafik | Trfk | 4.0 | ai | 2.0 | trafik | 1.2 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | -LRB- | ( | O | 36.0 | Trafik | Penguatku | Dan | NNP | NNP | NNP | 5.0 | -LRB- | -LRB- | -LRB- | 5.0 | -- | 2.0 | -lrb- | 1.2 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | 1.0 | JSPT | NNP | ORGANIZA | 37.0 | -LRB- | Trafik | Penguatku | ( | NNP | NNP | 4.0 | JSPT | JSPT | JSPT | 4.0 | | .0 | jspt | .9 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | -RRB- | ) | O | 38.0 | JSPT | -LRB- | Trafik | NNP | ( | NNP | 5.0 | -RRB- | -RRB- | -RRB- | 5.0 | -- | 2.0 | -rrb- | 1.2 |
| 1.0 | .0 | .0 | .0 | .0 | 1.0 | .0 | .0 | .0 | PREMIS | NN | O | 39.0 | -RRB- | JSPT | -LRB- | ) | NNP | ( | 6.0 | PREMI | REMIS | PRMS | 4.0 | EI | 2.0 | premis | 1.2 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | VCD/DVD | NN | O | 40.0 | PREMIS | -RRB- | JSPT | NN | ) | NNP | 7.0 | VCD/D | D/DVD | VCD/DVD | 7.0 | / | 1.0 | vcd/dvd | 1.3 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | .0 | HARAM | JJ | O | 41.0 | VCD/DVD | PREMIS | -RRB- | NN | NN | ) | 5.0 | HARAM | HARAM | HRM | 3.0 | AA | 2.0 | haram | 1.0 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | .0 | DISERBU | VB | O | 42.0 | HARAM | VCD/DVD | PREMIS | JJ | NN | NN | 7.0 | DISER | SERBU | DSRB | 4.0 | IEU | 3.0 | diserbu | 1.3 |
| 1.0 | 1.0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | , | , | O | 43.0 | DISERBU | HARAM | VCD/DVD | VB | JJ | NN | 1.0 | , | , | , | 1.0 | , | 1.0 | , | .4 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | .0 | INDIVIDU | NN | O | 45.0 | 5 | , | DISERBU | CD | , | VB | 8.0 | INDIV | IVIDU | NDVD | 4.0 | IIIU | 4.0 | individu | 1.5 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | .0 | DITAHAN | VB | O | 46.0 | INDIVIDU | 5 | , | NN | CD | , | 7.0 | DITAH | TAHAN | DTHN | 4.0 | IAA | 3.0 | ditahan | 1.3 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | .0 | PREMIS | NN | O | 47.0 | DITAHAN | INDIVIDU | 5 | VB | NN | CD | 6.0 | PREMI | REMIS | PRMS | 4.0 | EI | 2.0 | premis | 1.2 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | VCD/DVD | NN | O | 48.0 | PREMIS | DITAHAN | INDIVIDU | NN | VB | NN | 7.0 | VCD/D | D/DVD | VCD/DVD | 7.0 | / | 1.0 | vcd/dvd | 1.3 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | .0 | HARAM | JJ | O | 49.0 | VCD/DVD | PREMIS | DITAHAN | NN | NN | VB | 5.0 | HARAM | HARAM | HRM | 3.0 | AA | 2.0 | haram | 1.0 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | .0 | DISERBU | VB | O | 50.0 | HARAM | VCD/DVD | PREMIS | JJ | NN | NN | 7.0 | DISER | SERBU | DSRB | 4.0 | IEU | 3.0 | diserbu | 1.3 |
| 1.0 | 1.0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | , | , | O | 51.0 | DISERBU | HARAM | VCD/DVD | VB | JJ | NN | 1.0 | , | , | , | 1.0 | , | 1.0 | , | .4 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | .0 | INDIVIDU | NN | O | 53.0 | 5 | , | DISERBU | CD | , | VB | 8.0 | INDIV | IVIDU | NDVD | 4.0 | IIIU | 4.0 | individu | 1.5 |
| 1.0 | .0 | .0 | .0 | .0 | .0 | 1.0 | .0 | .0 | DITAHAN | VB | O | 54.0 | INDIVIDU | 5 | , | NN | CD | , | 7.0 | DITAH | TAHAN | DTHN | 4.0 | IAA | 3.0 | ditahan | 1.3 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Kuala | NNP | LOCATION | 55.0 | DITAHAN | INDIVIDU | 5 | VB | NN | CD | 5.0 | Kuala | Kuala | Kl | 2.0 | uaa | 3.0 | kuala | 1.0 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Lumpur | NNP | LOCATION | 56.0 | Kuala | DITAHAN | INDIVIDU | NNP | VB | NN | 6.0 | Lumpu | umpur | Lmpr | 4.0 | uu | 2.0 | lumpur | 1.2 |
| 1.0 | 1.0 | .0 | .0 | .0 | .0 | .0 | .0 | .0 | , | , | O | 57.0 | Lumpur | Kuala | DITAHAN | NNP | NNP | VB | 1.0 | , | , | , | 1.0 | , | 1.0 | , | .4 |
| 1.0 | 1.0 | .0 | 1.0 | 1.0 | .0 | .0 | .0 | .0 | 6 | CD | DATE | 58.0 | , | Lumpur | Kuala | , | NNP | NNP | 1.0 | 6 | 6 | 6 | 1.0 | 6 | 1.0 | 6 | .6 |
| .0 | .0 | 1.0 | .0 | .0 | 1.0 | .0 | .0 | .0 | Oktober | NNP | DATE | 59.0 | 6 | , | Lumpur | CD | , | NNP | 7.0 | Oktob | tober | ktbr | 4.0 | Ooe | 3.0 | oktober | 1.3 |
| 1.0 | 1.0 | .0 | 1.0 | 1.0 | .0 | .0 | .0 | .0 | 2017 | CD | DATE | 60.0 | Oktober | 6 | , | NNP | CD | , | 4.0 | 2017 | 2017 | 2017 | 4.0 | 2017 | 4.0 | 2017 | 1.3 |

Fig. 7. Sample of Feature Extraction k-NN Classification

- Fuzzy C-means Clustering Method

The research proposed the fuzzy c-means method that applies to Malay named entity recognition task. The experiment is conducted by analyses the data that have done the pre-processing stage. The data that consists with features set is processed by using clustering method called as fuzzy c-Means algorithm. Fuzzy clustering is categorized as an unsupervised learning method that influential for data analysis and model's construction. Sakinah [19] stated that the desired number of clusters and preliminary predictions for each grade of membership is the beginning of the FCM algorithm. Therefore, for each cluster, all data points have their respective membership grades. The goal algorithm is to guide the central cluster to the optimum location in the data space by gradually updating the membership grade along with prototype (cluster centers) of the data point.

Suganya and Shanthi [20] stated that fuzzy c-means use fuzzy division to allow the sharing of data by all groups with different grades of membership between 0 and 1. They explain that the fuzzy c-means algorithm works by providing membership to each data point equivalent to each cluster center. Membership value given was calculated based on the distance between the center of the cluster and data points. The membership value of each data increases according to the closeness of data to the specified cluster center. This fuzzy C-means clustering makes a performance to cluster data by iteratively searching for a set of fuzzy clusters and the associated cluster centers which represent the data structure. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. The following Fig. 8 is the algorithm for fuzzy C-Means clustering.

1. *Initialize* $U=[u_{ij}]$ *matrix,* $U^{(0)}$

2. *At k-step: calculate the centers vectors* $C^{(k)}=[c_j]$ *with* $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

3. *Update* $U^{(k)}$ , $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left( \frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

4. *If* $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$ *then STOP; otherwise return to step 2.*

Fig. 8. Fuzzy C-Means Clustering Algorithm

- K-Nearest Neighbors Algorithm

Classification is a machine learning technique in a supervised learning category that can be used to develop a model that describes the classification of important data. The development of the classifier is based on the class attributes involvement. Another method used in this experiment for classification is by using the K nearest neighbors algorithm. In pattern recognition, k-Nearest neighbors (k-NN) is one of the

algorithms that are very simple, efficient, effective and most widely used classification methods. KNN classifier is a straightforward classifier in classifying data where sample data is classified according to the nearest neighbor class.

The K number of the nearest neighbors used has been given first in achieving high precision in the classification and relies heavily on the data set used. As the most basic instance-based method, the data used in the KNN algorithm are represented in vector space. There are two steps that are used in simple K nearest neighbor algorithm, firstly is finding the K training example that is closest to the unknown example and the second step is to pick the most classify occur for these K examples. The following Fig. 9 is the pseudo code of k nearest neighbors algorithm.

***k-Nearest Neighbor***

1. Classify (X,Y,x) // *X:training data, Y:class labels of X, x:unknown* sample
2. Calculate "d (x, xi)" i =1, 2, ….., n; where d denotes the Euclidean distance between the points.
3. Arrange the calculated n Euclidean distances in non-decreasing order.
4. Let k be a +ve integer, take the first k distances from this sorted list.
5. Find those k-points corresponding to these k-distances.
6. Let ki denotes the number of points belonging to the ith class among k points i.e. $k \geq 0$
7. If ki >kj $\forall$ i $\neq$ j then put x in class i.

Note:
where $x_i$ is the training data point

Fig. 9. Pseudo code of k Nearest Neighbors algorithm

## VI. Result & Discussion

The collection of data is produced from PDRM news web pages in Malay languages cover on a few categories such as general topics, sports, crimes and others. Examples of the dataset before pre-processing are shown in Fig. 10 and after pre-processing in both Fig.6 and Fig. 7 respectively.


Fig. 10. Example of the dataset before pre-processing phase

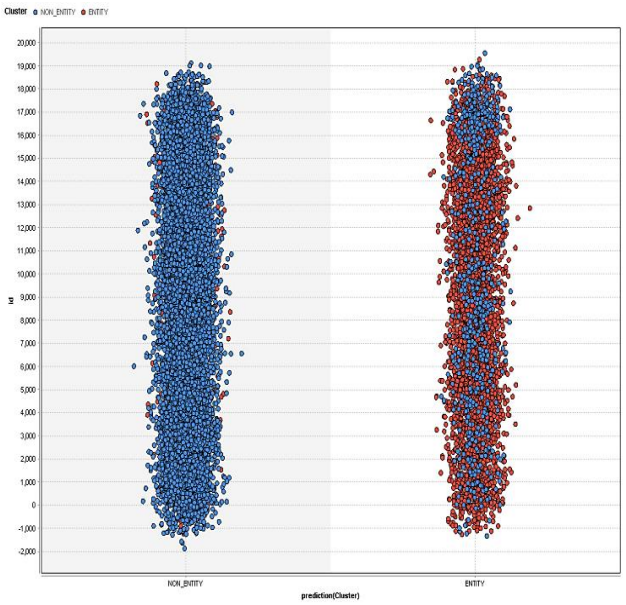



Fig. 11. Prediction FCM Clustering Chart

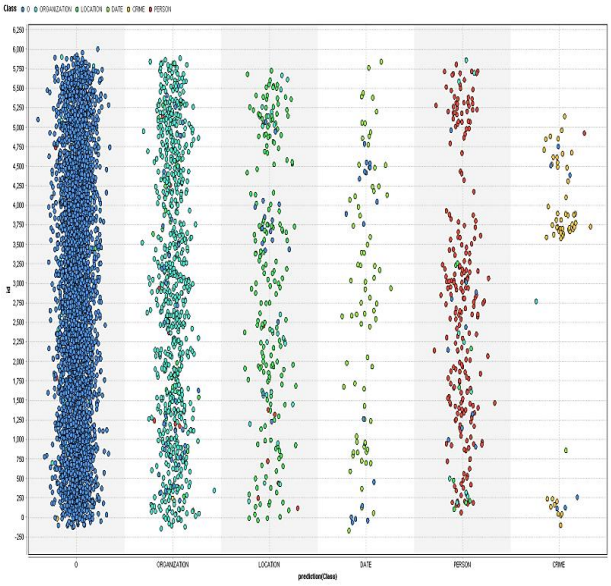



Fig. 12. k-NN Classification Chart

| | | True | | class precision | Accuracy |
|---|---|---|---|---|---|
| | | NON_ENTITY | ENTITY | | |
| Predicted | NON_ENTITY | 10451 | 1367 | 88.43% | 88.51% |
| | ENTITY | 646 | 5062 | 88.68% | |
| class recall | | 94.18% | 78.74% | | |

Fig. 13. FCM Clustering Result

| | | True | | | | | | class precision |
|---|---|---|---|---|---|---|---|---|
| | | OTHER | ORGANIZATION | LOCATION | DATE | CRIME | PERSON | |
| Predicted | OTHER | 4204 | 39 | 28 | 10 | 6 | 8 | 97.88% |
| | ORGANIZATION | 33 | 737 | 38 | 0 | 3 | 10 | 89.77% |
| | LOCATION | 16 | 23 | 194 | 0 | 0 | 7 | 80.83% |
| | DATE | 17 | 0 | 0 | 76 | 0 | 0 | 81.72% |
| | CRIME | 6 | 1 | 0 | 1 | 53 | 1 | 85.48% |
| | PERSON | 10 | 10 | 8 | 0 | 0 | 233 | 89.27% |
| class recall | | 98.09% | 90.99 % | 72.39 % | 87.36% | 85.48% | 89.96 % | Accuracy 95.24% |

Fig. 14. Result for Malay Named Entity Recognition

Based on prediction clustering chart of Fig. 11 and the cluster result in Fig. 13, the overall percentage accuracy had gave markedly good results based on clustering matching with 88.51% due to the calculation from all recall and precision results from all class entities. This accuracy was evaluated according to 17527 data samples, which have been pre-processed and undergone feature extraction. The precision result for NON_ENTITY class is 88.43% with 94.18% recall, whereas the precision for ENTITY class is 88.68% with 78.74% recall. Based on the analysis with other languages including English, NER has been implemented in the Malay language, which has the same characteristics as English in named entity recognitions such as capitalisation feature.

Then, for k-NN classification chart and result in the Fig.12 and Fig. 14 respectively, the prediction of classified entities consists of ORGANIZATION, LOCATION, DATE, CRIME, PERSON and OTHER is evaluated according to precision and recall. For ORGANIZATION entity, the precision is 89.77% and recall is 90.99%. For LOCATION entity, its precision is 80.83% and 72.39% recall. Next, the DATE entity produces 81.72% and 87.36% for both precision and recall respectively. For CRIME type entity, it produces both precision and recall as many as 85.48%. Then, for PERSON entity, it produces 89.27% for precision and 89.96% for recall. Lastly, for OTHER entity, the result for both precision and recall are 97.88% and 98.09% respectively.

## VII. Conclusions

As conclude, the overall accuracy produced for Malay NER analysis is 95.24% during k-NN classification. This accuracy that can be an overall perspective of the evaluation process can be improved by undergoing another experiment by increasing the training dataset for a better result. This is because the percentage of accuracy increment for recognizing Malay entities liable on the model trained and suitable features sets used. The generated model from the small amount of dataset during the training process affected the assessment of the test's results. Therefore, the bigger dataset is needed to develop the Malay model to increase the results. As significant, the produced NER model can help to extract text data by determining exact text or term in the Malay language as named entity for the further police investigation.

In addition, the selection of appropriate features need to be continuously focused as these features can affect the performance of the NER model especially for Malay language because the language has complex structure in sentences.

The proposed Malay NER model can be further improved by increasing the corpus references in Malay for solving the problem of ambiguities for recognizing named entity types in Malay texts.

## Acknowledgment

The publication is sponsored by Center of Advanced Computing Technology(C-ACT), Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka.

## References

[1] Hosseinkhani, M. Koochakzaei, S. Keikhaee, and J. H. Naniz, "Detecting Suspicion Information on the Web Using Crime Data Mining Techniques," Int. J. Adv. Comput. Sci. Inf. Technol., vol. 3, no. 1, pp. 32–41, 2014.

[2] H. Shabat and N. Omar, "Named Entity Recognition in Crime News Documents Using Classifiers Combination," vol. 23, no. 6, pp. 1215–1222, 2015.

[3] A. Alkaff and M. Mohd, "Extraction of nationality from crime news," J. Theor. Appl. Inf. Technol., vol. 54, no. 2, pp. 304–312, 2013.

[4] D. Jurafsky and J. H. Martin, "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition," Speech Lang. Process. An Introd. to Nat. Lang. Process. Comput. Linguist. Speech Recognit., vol. 21, pp. 0–934, 2009.

[5] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," Informatica, vol. 31, pp. 249–268, 2007.

[6] S. Morwal, "Named Entity Recognition Using Hidden Markov Model ( HMM ): An Experimental Result on Hindi , Urdu and Marathi Languages," vol. 3, no. 4, pp. 671–675, 2013.

[7] D. Chopra and S. Morwal, "Named entity recognition in english language using Hidden Markov Model," Int. J. Comput. Sci. Appications, vol. 3, no. 1, pp. 1–6, 2013.

[8] I. Ahmed and R. Sathyaraj, "Named Entity Recognition by Using Maximum Entropy," Int. J. Database Theory Appl., vol. 8, no. 2, pp. 43–50, 2015.

[9] R. Sathya and A. Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification," Int. J. Adv. Res. Artif. Intell., vol. 2, no. 2, 2013.

[10] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, and A. Sun, "Twiner: named entity recognition in targeted twitter stream," in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12, 2012, p. 721.

[11] S. Zhang and N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts," J. Biomed. Inform., vol. 46, no. 6, pp. 1088–1098, 2013.

[12] I. Diaz-Valenzuela, M. A. Vila, and M. J. Martin-Bautista, "On the Use of Fuzzy Constraints in Semisupervised Clustering," IEEE Trans. Fuzzy Syst., vol. 24, no. 4, pp. 992–999, 2016.

[13] G. Castellano, A. M. Fanelli, and M. A. Torsello, "Shape annotation by incremental semi-supervised fuzzy clustering," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial

Intelligence and Lecture Notes in Bioinformatics), 2013, vol. 8256 LNAI, pp. 193–200.

[14] R. Alfred, L. Chin Leong, C. Kim On, and P. Anthony, "Malay Named Entity Recognition Based on Rule-Based Approach," Int. J. Mach. Learn. Comput., vol. 4, pp. 300–306, 2014.

[15] S. Sulaiman, R. A. Wahid, S. Sarkawi, N. Omar, A. N. E. R. English, and G. Languages, "Using Stanford NER and Illinois NER to detect Malay Named Entity Recognition," vol. 9, no. 2, pp. 2–5, 2017.

[16] S. Salleh, A. Asmai, H. Basiron, and S. Ahmad, "A Malay Named Entity Recognition Using Conditional Random Fields," in 5th International Conference on Information and Communication Technology (ICoIC7), 2017, 2017, vol. 0, no. c, pp. 44–49.

[17] T. Cho, "Differences in the Romanized Spelling of Arabic Loanwords in Bahasa Melayu in Malaysia , and Bahasa Indonesia," MELAYU J. ANTARABANGSA DUNIA MELAYU JILID 9 BIL. 2 2016, 2016.

[18] M. Zaidi, A. Rozan, and Y. Mikami, "Orthographic Reforms of Standard Malay Online : Towards Better Pronunciation and Construction of a Cross-language Environment *," Education, no. March, pp. 129–159, 2007.

[19] S. Sakinah and S. Ahmad, "Fuzzy modeling through granular computing," Thesis, 2012.

[20] R. Suganya and R. Shanthi, "Fuzzy C-Means Algorithm-A Review," Int. J. Sci. Res. Publ., vol. 2, no. 11, pp. 2250–3153, 2012.

# Towards a Framework for Multilayer Computing of Survivability

Abolghasem Sadeghi[1], Mohammad Reza Valavi[2], Morteza Barari[3]
Department of Information and Computer Technology
Malek Ashtar University of Technology
Tehran, Iran

***Abstract*—The notion of survivability has an important position in today enterprise systems and critical functions. This notion has been defined in different ways. However, lacking a comprehensive and multilayer model for computing the survivability quantitatively, is the major gap happened in researches of this field; a model that is tally general and applicable in various applications. This research tries to design a comprehensive, multilayer as well as general model for modeling and computing the survivability. Considering that the Markov property is true in our proposed model, we used the Markov model. Using the proposed three layer architecture and designing a Markov structure, we could have been able to compute the survivability initially for each of infrastructure components separately and regardless of their functional dependency to each other. The computations were generalized to consider component dependencies as well as the upper layers entering dependencies in Markov model and could compute the survivability of each vital function for the highest architectural layer based on the underlying layers. Finally, a common and ordinary structure of crisis management has been studied and its results analyzed. We could examine the abilities of our model to compute the survivability of the whole crisis management system successfully.**

***Keywords*—Network survivability; survivability quantification; survivability computation; system survivability**

## I. Introduction

Today, all social, cultural, political and economic life aspects of societies and states are dependent on the information technology infrastructures and this dependency is ever increasing. Due to this dependency, important concerns on the functional quality and serving those infrastructures have been emerged. This issue is becoming more important day by day that whether these infrastructures can tolerate different challenges -including natural ones from flood and earthquake to human errors or adversary invasions- and can provide their major and essential services. Therefore we need to compute the resistance of infrastructures against such challenges for better planning, implementing and utilizing them. This will enable us to find appropriate solutions for improving the resistance property. This is explained by the survivability metric.

### A. Qualitative Definitions of Survivability

Like many other scientific subjects, there is no consensus and unanimous definition for the survivability. Table I summarizes definitions yet provided. The definitions have been ordered chronologically and their references have also been given. The definitions are dependent on the field within which are required and their own origins. There are multiple differences between these definitions, so there should be a deep understanding of intended problem to find the more suitable one. In our context, the dominant definition that many other researches have used, is the fourth definition. So we will use it too.

### B. Quantitative Definition of Survivability

All definitions contained in Table I have a qualitative approach. ANSI has provided a quantitative definition for survivability [27] that models the survivability concept parametrically. Fig. 1 shows this definition. In this definition, the measure of interest M has the value $m_0$ just before a failure occurs. The survivability of this system is represented by the following attributes:

- $m_a$ is the value of M immediately after the failure.
- $m_u$ is the maximum possible difference between $m_0$ and $m_a$ after failure.
- $m_r$ is the restired value of M after time $t_r$.
- $t_r$ is the time required for achieving the value of $m_0$ for M again or a reduced but acceptable value $m_0$ if $m_0$ is impossible to be fully restored.

The notion of survivability may seem similar or overlapping to certain notions of dependability field like reliability, availability, fault tolerance, maintainability, security and safety. These similarities and differences have been discussed in various important references such as [12,24,25,28-30]. Thus, we refer the reader to those references.

## II. Related Works

Various researches have been performed on the survivability implicitly and explicitly. By the explicit researches, we mean those that have been clearly focused on the survivability. However, implicit researches are those dealing with related concepts like system recovery or intrusion tolerance. Moreover, some of them have just provided a qualitative model in this field and have not indicated a way to inferring the level of survivability from such models. Others have attempted to make the issue quantitative and compute the survivability.

TABLE I. DIFFERENT DEFINITIONS PROVIDED FOR THE SURVIVABILITY

| No | Scientific area | Year | Definition | Reference |
|---|---|---|---|---|
| 1 | IT systems | 1988 | Survivability is the degree to which essential functions are still available even though some part of the system is down. | [2] |
| 2 | Telecommunication Systems | 1996 | Survivability is a property of a system, subsystem, equipment, process, or procedure that provides a defined degree of assurance that the named entity will continue to function during and after a natural or man-made disturbance. | [15] |
| 3 | Network Computing Systems | 1997 | Survivability is the ability of a network computing system to provide essential services in the presence of attacks and failures and recover full services in a timely manner. | [3] |
| 4 | Critical and Defense Systems | 1999 | Survivability is the capability of a system to fulfill its mission, in a timely manner, in the presence of attacks, failures or accidents. | [4-7] [9,10] |
| 5 | Critical and Defense Systems | 2000 | Survivability is the ability [of a system] to continue to provide service, possibly degraded or different, in a given operating environment when various events cause major damage to the system or its operating environment. | [1,11] |
| 6 | Networking & Telecommunication | 2005 | Survivability is the ability of system to deliver the minimum expected service when defined threats are faced and the system must retain those properties wanted by the users. | [19] |
| 7 | Telecommunication Systems | 2009 | Survivability is the system's ability to continuously deliver services in compliance with the given requirements in the presence of failures and other undesired events. | [16] |
| 8 | Aerospace & Military | 2009 | Survivability is the ability of a system to minimize the impact of finite-duration environmental disturbances on value delivery. | [20] |
| 9 | Networking & Telecommunication | 2015 | Survivability is a concept that describes the capability of a system to achieve timely recovery after the occurrence of undesired events | [26] |

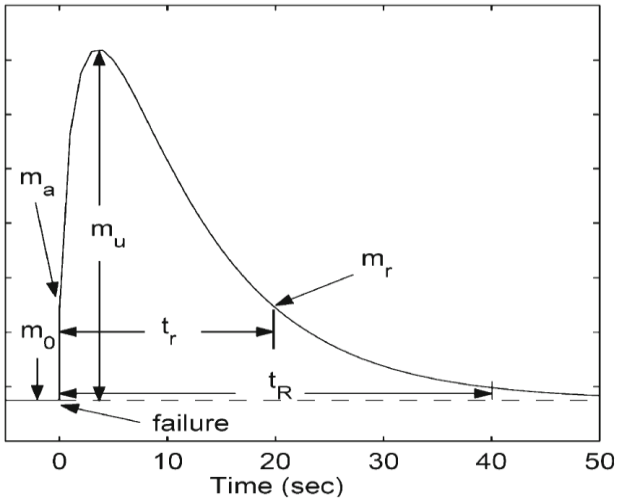



Fig. 1. ANSI Definition for Survivability.

SABER model in [13] dealt with providing an appropriate architecture for intrusion tolerance in the systems. This architecture has a conventional network security approach enabling it to continue the wanted services under an intrusion or attack using IDS sensors and a higher security level called SOS. This research includes only software attacks and has nothing covering other malicious and non-malicious undesired events.

ITDOS architecture [14] provided an intrusion-tolerant software structure for software systems using facilities based on CORBA firmware. Thereby, it is ensured that all CORBA-based softwares produced with proposed extensions are intrusion-tolerant. However, no review and implementation of mentioned architecture has been reported.

AWDRAT method [18] provides a self-adaptive method in software firmware to be able to detect the possibly compromised point comparing operating program behavior with the desired behavior. Then, a trust management system manages the restoration process changing the execution path from the previously compromised components to the unaffected ones to enable the system continue secure and trusted operation under attack conditions. Although this research had successful experimental results, it has focused merely on the software malicious threats and overlooked other threat aspects and is not fully comprehensive yet.

DPASA architecture [17] provides a model for system state recovery after a cyber attack. It uses a set of tools and methods for identifying, protecting and adaptive reactions. For assessing the model performance, it was entered in an applied example of JBI belonging to US Air Force laboratory and has modeled and represented accuracy of recovery and attack tolerance with several parameters. However, the parameters introduced in this research are not public and can not be generalized to other problems. Thereby, any new problem will need its own parameters to be extracted. This research emphasizes only on the cyber attacks and is also general and qualitative. Qualitative means it does not have any measureable parameters.

Willow architecture [8] claims creating survivability in wide critical and distributed systems. This architecture uses a combination of fault avoidance, fault elimination and fault tolerance. It disables vulnerable components under threatening conditions. Then replaces damaged components after a fault or intrusion. When an indispensable fault appears it will be able to strengthen the system against that using reconfiguration methods based on a control system feedback. This architecture reported an experimental study for the US air forces on the JBI basis whose results have not yet been published and has sufficed just the claim that the system has functioned successfully.

The previous works reviewed here were qualitative models that provide no accurate computations. Trivedi et al. in [16] and [26] have provided a model for computing survivability quantitatively that has been designed based on composition of availability and performability. They provided two Markov models for availability and performability in their research.

Then, they combined them and devised a complex Markov model for system survivability. For verification of the provided model, they modeled it on a given telecommunication system and computed the survivability supposing that the input variables of the model have been taken. This model has the advantage of being quantitative and computable. However, the research finally gives no accurate sense to the software service users on that how they can verify their service survivability. Moreover, the supposed problem in this research is a very well-known and already solved problem with clear solutions while imagination of such clarity for broader and more complex problems is generally difficult if not impossible. For this, we can say that the proposed method is slightly difficult to generalize to other problems in this field.

In Survivability Analysis of a Computer System under an Advanced Persistent Threat Attack, [31] has attempted to model and compute a software system survivability under APT attacks. For this purpose, they proposed an integrated model of process of an APT attack as well as different steps for defending it. Their final goal was to create a continuous time Markov model for this issue to compute the total survivability of the system under an APT attack. To create the model, different steps of the ATP attack have been modeled with Stochastic Reward Nets and its graph has been produced. Then, the reachability graph of this petri net is drawn as a continuous time Markov model for computing survivability. The graph introduces system recovery, system reachability, data confidentiality, and data accuracy as the four parameters of the survivability model to compute the survivability of this system. SRN net and Markov model are created here for computation of the aforementioned four parameters. Finally, to be able to verify the model, authors have obtained some of probable values required for the model from the valid references or supposed them and applying the values to the model. So, they computed survivability quadruplet probability measures. While an appropriate computation has been proposed in this model for survivability, the proposed model is allocated to APT cyber attacks and isn't suitable to be applied to other applications.

For computing the survivability generally in software layer, [21] attempted to provide measurable criteria for defining and assessing software survivability from the end user's viewpoint. Doing so, they have provided a framework for defining software survivability quantities and enabling the user to design and execute various policies for achieving survivability based on those quantities. A decision support model has also been proposed to realize the survivability quantities to ensure the minimum survivability for the software. The aforementioned quantities are classified into five groups as follow:

- Adaptability
- Recoverability
- Fault tolerance
- Reliability
- Performance degrading

Each group represents one of the characteristics of the survivability and each has several quantities that are survivability related quantities. Finally, survivability computations are classified into two groups: contribution-oriented and concern-oriented. The contribution-oriented functions compute those characteristics of the survivability that the user needs them essentially and must be met fully. In contrast, the concern-oriented functions deal computation of those characteristics and quantities of the survivability on which the user is concerned about but can tolerate violation of them up to a certain level.

The same author in [22] and [23] has attempted to use proof-carrying codes for survivability assessment. The general idea of this method is to enable the user to define his software survivability requirements and provide it to the software vendor. Then, the vendor will be able to provide the user with a system using proof-carrying code method that enables the user himself to assess his system survivability based on the initially proclaimed requirements. The main reference for introducing proof-carrying codes is [32] which is used in this method.

All works reported here from Dr. Zuo have computed and parameterized the survivability only based on outstanding characteristics of the software system itself. While the secure and correct execution of any software system is subjected to the security and correctness of infrastructure components performance that the system relies on them. Unfortunately, these researches have not discussed them and have not replied the ambiguity here.

## III. Basic Architecture

As mentioned in section 2 the general and widespread weakness in all works of this field was that the system user can not compute the overall system survivability based on his information about different layers. Some of works have dealt with computing the survivability of the infrastructure layer without enabling the user to use it for computing service/software survivability. Some of them have performed it in software layer without taking into account a logical and working dependency between the software layers with underlying layers. Naturally, these computations are not comprehensive and do not have enough accuracy and integrity. A suitable model is required for survivability computation that connects layers to remove this challenge. Fig. 2 shows this model.

The provided model is a set of various and heterogeneous agents and components that are set up beside each other randomly and unpredictably and each component can be connected to others and there is no predefined limitation for services that is provided to other components. Of course it is clear that we do not mean the practical limits like memory, connection link capacity, etc. Each of those components participates in one or more application belonging to the software layer. The total system is depicted as a set of functions or services in the top layer namely operation level. In this layer we deal with organizational processes as functional components of that layer. Functions are executed using several applications. In other words, each function need some applications for operating. In the given model of Fig. 2, the system is supposed to have X functions that use n applications for fulfilling their functions and services. Application systems are executed on the basis of k components.

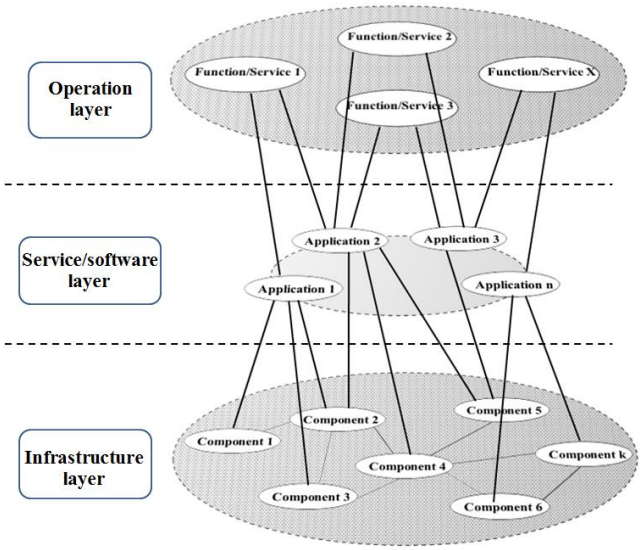



Fig. 2. Basic Model of Layers for Survivability Quantification.

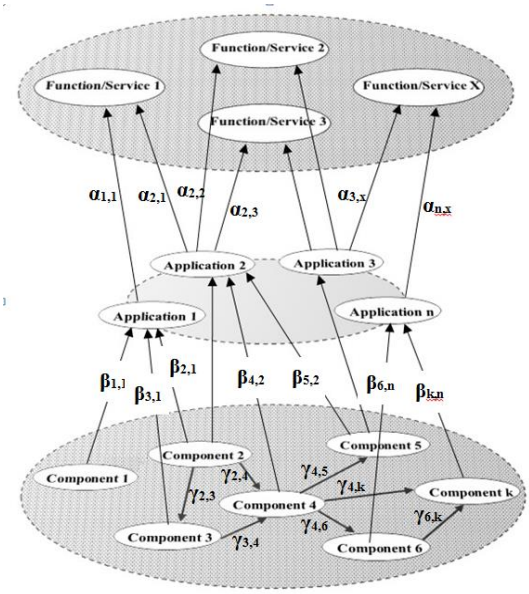



Fig. 3. Dependency Metrics in Survivability Model.

### C. Relations in the Model

The relationship between components must be understood and analyzed accurately to make the model efficient and practically useable. The components relationships to each other in the infrastructure layer is transversal while the relationship of infrastructure layer components with software layer applications and between software layer applications with functions of the operation level is longitudinal. In a real environment the components can serve each other. Therefore, it is required that to suppose the relationship between components a directional relationship for demonstrating that which component is client and which one is service provider.

Given the directionality of the graph, it must be cleared that does the graph have a loop or would be a DAG? Although it is acceptable to suppose that this graph can involve a loop, it may be a DAG. In this regard, what is important here is our attitude resolution and granularity. For instance, if a smart building management system is taken as a component then this component can provide the inputs needed for other systems like ventilation, cooling, electricity, etc. where the supposed component has low granularity in this situation. If the BMS system is separated to its basic facilities and modules and each module is considered as a component, then that components will be single task that lead us achieving a loop free graph. Whether or not, we supposed the graph of components as DAG and provide our algorithm based on it. Although this decomposition process helps to achieve a DAG, it is obvious that appropriate algorithms can be developed in future works considering the graph a cyclic one.

In the upper layer, it is possible to consider no direct dependency between them because when system A serves system B it means that some of components in system A serve some of components in system B. Indeed, this concept is considered in relationship between components. Therefore, there is no explicit transversal dependency between applications and independent set of applications form a function in operation layer. Fig. 3 depicts this notion.

Now, we must analyze dependency of components to each other, dependency of applications to components and dependency of functions to applications separately and quantify them. Doing so, the model edges are named according to fig. 2. Moreover, for the sake of facilitation in representing topics, the applications are symbolized with $AP_i$, functions with $FS_i$ and components with $CMP_i$. In this model, $\alpha_{i,j}$ represents the total dependency of $FS_i$ to $AP_i$. Further, $\beta_{i,j}$ represents dependency of $AP_j$ to $CMP_i$. $\gamma_{x,y}$ represents the dependency of component y to component x. α, β and γ coefficients are real numbers between 0 and 1. Now, we discuss properties of these coefficients in the graph of Fig. 3.

$$\forall j, \sum_i \alpha_{i,j} = 1 \tag{1}$$

Because, each function $FS_j$ is consisted of its applications and regardless of user mistakes, the full execution of applications means that the function $FS_j$ will be executed completely.

$$\forall j, \sum_i \beta_{i,j} = 1 \tag{2}$$

Meaning that the full operation of any application is subjected to the fact that all concerning components fulfill their tasks completely, because each system only is consisted of its components functioning well and no other components intervening correct application execution.

$$\forall cmp_j, \sum_i \gamma_{i,j} < 1 \tag{3}$$

Meaning that each component would be partially –and not fully- dependent on other components functionally. In fact, each component definitely has its own special and independent functionality that cause the above summation should be less than 1. If the required inputs for a client aren't provided from one of the service provider components, the function of client

component will be damaged proportionate to coefficient of dependency to the service providing component.

## IV. Survivability basic Model

In this section we propose our basic conceptual model for the survivability of any system generally. As we saw in section 1, the survivability aims at enabling the system to continue its vital and essential services and operations under crisis until recovery of failed subsystems. Thus, for modeling the survivability of any system it is required to consider three basic states. The first state is where the system operates normally and naturally. Under such state, the crisis has no degrading effects on the system and operates normally that is called Healthy state. The other is loss of the important and critical subsystems that results in the total failure and break down and is called Fail state. However, the third state is one that some of non critical subsystems are failed but the system can continue its fundamental operation until the problem is removed. This state is called Survive state. Tri-state Markov model in Fig. 4 represents these definitions. In this model, μ and ρ parameters show the transmission rate between various states of Markov model.

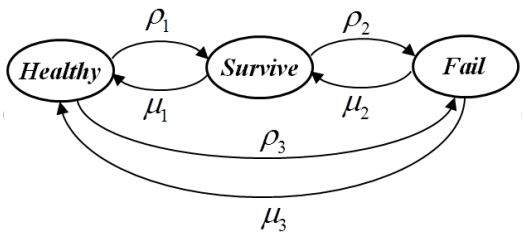



Fig. 4. Survivability basic Model.

## V. Computing Survivability

For multilayer computing the survivability across the layers of Fig. 2, it is required to start from the lowest layer and compute it separately for each infrastructure component regardless of its dependency to other components. Then, the infrastructure layer components survivability is computed taking into account their dependency. In the next step, the applications survivability in software layer is computed given their dependency to the infrastructure layer components and computations in that layer. Finally, the functions survivability in the operation layer will be computed based on computations of the software level.

### A. Computing the Survivability of a Single Infrastructure Component

The model depicted in Fig. 5 is Markov model for survivability of a single component of infrastructure layer. $attr_i$ is an attribute or subsystem of the component and αi is the probability of failing any $attr_i$. Some attributes or subsystems are critical for basic functioning of the component while others are not. The component could not tolerate failure of critical attributes and the component will enter the fail state. In the case of failing non critical attributes or subsystems, the component can continue its essential functions while entering the survive state. We show critical attributes with * mark in Fig. 5.

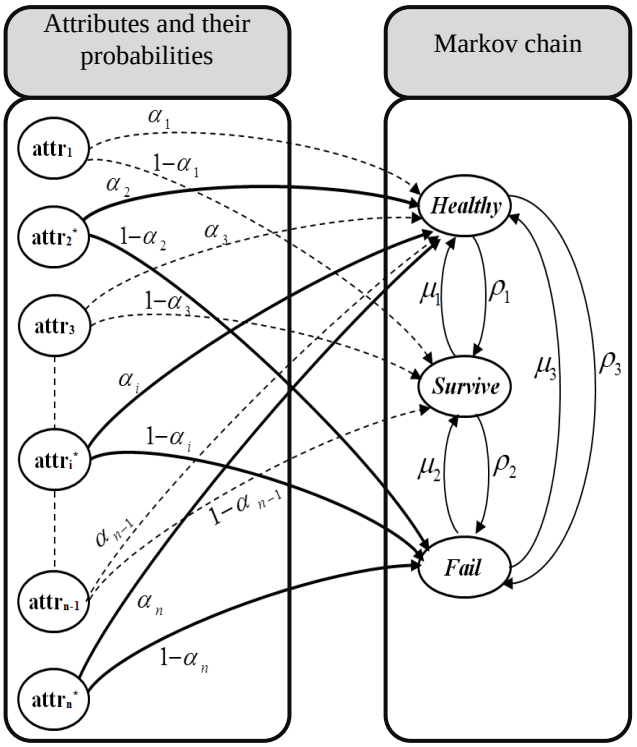



Fig. 5. Combined Markov Model of Survivability Quantification of a Single Component.

In the model shown in Fig. 5, each property of $attr_i$ has a bi-state Markov model as represented in Fig. 6.

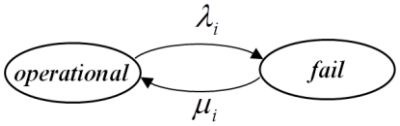



Fig. 6. Markov Model for Each Property.

In this model, $\lambda_i$ and $\mu_i$ are failure and recovery rate of the property i. In transient state, the probability of healthy and fail states in Markov model is computed as follow:

$$\left.\begin{aligned} \frac{d\,\pi_i^{op}}{dt} &= -\pi_i^{op}.\lambda_i + \pi_i^{f}.\mu_i \\ \pi_i^{op} + \pi_i^{f} &= 1 \end{aligned}\right\} \Rightarrow \begin{cases} \pi_i^{op} = \dfrac{\mu_i}{\mu_i+\lambda_i} + c.e^{-(\lambda_i+\mu_i)t} \\ \pi_i^{f} = \dfrac{\lambda_i}{\mu_i+\lambda_i} - c.e^{-(\lambda_i+\mu_i)t} \end{cases} \quad (4)$$

In Eq. (4), $\pi_i^{op}$ and $\pi_i^{f}$ mean the probability of healthy and failure states for the property i that are symbolized as $\alpha_i$ and 1-$\alpha_i$ in Fig. 5 for the sake of facilitation in reading and writing. The number c is an arbitrary constant. Thus, we have:

$$\alpha_i = \frac{\mu_i}{\mu_i+\lambda_i} + c.e^{-(\lambda_i+\mu_i)t} \quad (5)$$

In the steady state, the probability of healthy and fail states is as follow:

$$\left.\begin{aligned}\pi_i^{op}.\lambda_i=\pi_i^{f}.\mu_i\\ \pi_i^{op}+\pi_i^{f}=1\end{aligned}\right\}\Rightarrow\begin{cases}\pi_i^{op}=\dfrac{\mu_i}{\mu_i+\lambda_i}\\ \pi_i^{f}=\dfrac{\lambda_i}{\mu_i+\lambda_i}\end{cases} \quad (6)$$

In model shown in Fig. 5, values assigned to $\alpha_i$ are probability type while values of ρ and μ are rate. On the other hand, values of $\alpha_i$ are given and known already. Therefore, probabilities of Markov model tri-states must be obtained first to compute rates of $\rho$ and $\mu$. Then, $\rho$ and μ are computed based on the probabilities of three states. For this purpose, three sets are introduced for using in Eq. (7). *S* is a set including all properties of this component. The set *IC* is a subset of critical characteristics of *S* and the set *INC* includes non-criticals. Following section shows formulae for the survivability computation.

$$S=\{i\in \mathrm{N}\,|\,attr_i \text{ is valid}\}$$
$$IC=\{i\in S\,|\,attr_i \text{ is critical attribute}\}$$
$$INC=\{i\in S\,|\,attr_i \text{ is not critical attribute}\}$$
$$IC\cap INC=\phi \;\&\&\; IC\cup INC=S$$

$$\pi_{Healthy}=\prod_{i\in S}\alpha_i \quad (a)$$

$$\pi_{Survive}=\prod_{i\in IC}\alpha_i\left(\sum_{\substack{A\subset INC\\ A\neq INC}}\left(\prod_{j\in A}\alpha_j\prod_{k\in INC-A}(1-\alpha_k)\right)\right) \quad (b) \qquad (7)$$

$$\pi_{Fail}=\sum_{\substack{A\subset IC\\ A\neq IC}}\left(\prod_{i\in A}\alpha_i\prod_{j\in IC-A}(1-\alpha_j)\right)\sum_{B\subseteq INC}\left(\prod_{k\in B}\alpha_k\prod_{l\in INC-B}(1-\alpha_l)\right) \quad (c)$$

Given the practical conditions in this model, it is possible to consider all properties independent. Even with some of properties depending on each other practically, the desired independency can be obtained through changing the system design. According to this assumption, the probability of a fully healthy state equals to multiplication of all properties healthy states probabilities that is shown in Eq. (7)(a). for computing the probability of survive state, the failure probability of non-critical properties are considered with their different permutations and multiply it by the probability of critical properties healthy probability. This is shown in Eq. (7)(b). However, the probability for the fail state equals the state within which some of critical properties are failed regardless of whether non-critical properties are healthy or not that is shown in Eq. (7)(c).

The probabilities related to states of Markov model of Fig. 5 have been computed in Eq. (7). Now, we should prove that summation of these three states equals 1 according to Markov model conditions. In other word, following equation must be true.

$$\pi_{Healthy}+\pi_{Survive}+\pi_{Fail}=1$$

**Theorem**: prove that the following equation is true in Markov model of Fig. 5:

$$\pi_{Healthy}+\pi_{Survive}+\pi_{Fail}=1$$

**Proof**: first, for simplification of notations we define:

$$\left.\begin{aligned}X=\sum_{\substack{M\subset INC\\ M\neq INC}}\left(\prod_{i\in M}\alpha_j\prod_{j\in INC-M}(1-\alpha_k)\right)\\ Y=\sum_{\substack{N\subset IC\\ N\neq IC}}\left(\prod_{i\in N}\alpha_i\prod_{j\in IC-N}(1-\alpha_j)\right)\end{aligned}\right\}\Rightarrow\begin{cases}\pi_{Healthy}=\prod_{i\in S}\alpha_i\\ \pi_{Survive}=\prod_{i\in IC}\alpha_i.X\\ \pi_{Fail}=Y.\left(X+\prod_{i\in INC}\alpha_i\right)\end{cases}$$

Now, given the sections (a) to (c) of Eq. (7):

$$\begin{aligned}&\pi_{Healthy}+\pi_{Survive}+\pi_{Fail}=\\ &\prod_{i\in S}\alpha_i+\prod_{i\in IC}\alpha_i.X+Y.\left(X+\prod_{i\in INC}\alpha_i\right)=\\ &\prod_{i\in S}\alpha_i+\prod_{i\in IC}\alpha_i.X+Y.X+Y.\prod_{i\in INC}\alpha_i=\\ &\prod_{i\in IC}\alpha_i.\prod_{i\in INC}\alpha_i+\prod_{i\in IC}\alpha_i.X+Y.X+Y.\prod_{i\in INC}\alpha_i=\\ &X.\left(\prod_{i\in IC}\alpha_i+Y\right)+\prod_{i\in INC}\alpha_i.\left(\prod_{i\in IC}\alpha_i+Y\right)=\\ &\left(X+\prod_{i\in INC}\alpha_i\right)\left(Y+\prod_{i\in IC}\alpha_i\right)=\\ &\sum_{A\subseteq INC}\left(\prod_{j\in A}\alpha_j\prod_{k\in INC-A}(1-\alpha_k)\right).\sum_{A\subseteq IC}\left(\prod_{j\in A}\alpha_j\prod_{k\in IC-A}(1-\alpha_k)\right)=\\ &\sum_{A\in P(S)}\left(\prod_{i\in A}\alpha_i\prod_{j\in S-A}(1-\alpha_j)\right)\end{aligned} \quad (8)$$

The *P(S)* in the final result of Eq. (8) is the power set of *S*. Indeed, the final result in Eq. (8) contains all possible permutations of failure or healthy state probability for each of properties through a linear polynomial. Now, it must be proved that the last sentence of Eq. (8) equals 1. To do so, the mathematical induction method is applied. For base case S must have two members. We know that sum of two elements of *S* is 1. So:

$$\begin{aligned}&S=\{\alpha_1,\alpha_2\}\;,\;\alpha_1+\alpha_2=1\\ &\sum_{A\in P(S)}\left(\prod_{i\in A}\alpha_i\prod_{j\in S-A}(1-\alpha_j)\right)=\alpha_1\alpha_2+\alpha_1(1-\alpha_2)+\alpha_2(1-\alpha_1)+(1-\alpha_1)(1-\alpha_2)=\\ &\alpha_1\alpha_2+\alpha_1-\alpha_1\alpha_2+\alpha_2-\alpha_1\alpha_2+1-\alpha_1-\alpha_2+\alpha_1\alpha_2=1\end{aligned} \quad (9)$$

Therefore the theorem for S with two members is true. Now, suppose that for S with n members the desired sentence equals 1. It must be proved that the relation is also true for *S* with n+1 members. $S^n$ represents the set S has n members. Thus, we have:

$$\sum_{A\in P(S^{n+1})}\left(\prod_{i\in A}\alpha_i\prod_{j\in S^{n+1}-A}(1-\alpha_j)\right)= \tag{10}$$

$$\alpha_{n+1}\cdot\sum_{A\in P(S^n)}\left(\prod_{i\in A}\alpha_i\prod_{j\in S^n-A}(1-\alpha_j)\right)+(1-\alpha_{n+1})\sum_{A\in P(S^n)}\left(\prod_{i\in A}\alpha_i\prod_{j\in S^n-A}(1-\alpha_j)\right)=$$

$$\sum_{A\in P(S^n)}\left(\prod_{i\in A}\alpha_i\prod_{j\in S^n-A}(1-\alpha_j)\right)=1$$

Therefore, the sum of probability of three states will be always equal to 1.

Now, combining formulae in Eq. (4) and Eq. (7) for computing probabilities related to the model of Fig. 5 in transient state, we have:

$$\pi_{Healthy}=\prod_{i\in S}\alpha_i=\prod_{i\in S}\left(\frac{\mu_i}{\mu_i+\lambda_i}+e^{-(\lambda_i+\mu_i)t}\right) \quad (a)$$

$$\pi_{Survive}=\prod_{i\in IC}\alpha_i\left(\sum_{\substack{A\subset INC\\A\neq INC}}\left(\prod_{j\in A}\alpha_j\prod_{k\in INC-A}(1-\alpha_k)\right)\right)= \tag{11}$$

$$=\prod_{i\in IC}\left(\frac{\mu_i}{\mu_i+\lambda_i}+e^{-(\lambda_i+\mu_i)t}\right)\left(\sum_{\substack{A\subset INC\\A\neq INC}}\left(\prod_{j\in A}\left(\frac{\mu_j}{\mu_j+\lambda_j}+e^{-(\lambda_j+\mu_j)t}\right)\prod_{k\in INC-A}\left(\frac{\lambda_k}{\mu_k+\lambda_k}-e^{-(\lambda_k+\mu_k)t}\right)\right)\right) \quad (b)$$

$$\pi_{Fail}=\sum_{\substack{A\subset IC\\A\neq IC}}\left(\prod_{i\in A}\alpha_i\prod_{j\in IC-A}(1-\alpha_j)\right)\sum_{B\subseteq INC}\left(\prod_{k\in B}\alpha_k\prod_{l\in INC-B}(1-\alpha_l)\right)=$$

$$=\sum_{\substack{A\subset IC\\A\neq IC}}\left(\prod_{i\in A}\left(\frac{\mu_i}{\mu_i+\lambda_i}+e^{-(\lambda_i+\mu_i)t}\right)\prod_{j\in IC-A}\left(\frac{\lambda_j}{\mu_j+\lambda_j}-e^{-(\lambda_j+\mu_j)t}\right)\right)\sum_{B\subseteq INC}\left(\prod_{k\in B}\left(\frac{\mu_k}{\mu_k+\lambda_k}+e^{-(\lambda_k+\mu_k)t}\right)\prod_{n\in INC-B}\left(\frac{\lambda_n}{\mu_n+\lambda_n}-e^{-(\lambda_n+\mu_n)t}\right)\right) \quad (c)$$

Probabilities of the steady state in Fig. 5 are as follows:

$$\pi_{Healthy}=\prod_{i\in S}\alpha_i=\prod_{i\in S}\left(\frac{\mu_i}{\mu_i+\lambda_i}\right) \quad (a)$$

$$\pi_{Survive}=\prod_{i\in IC}\alpha_i\left(\sum_{\substack{A\subset INC\\A\neq INC}}\left(\prod_{j\in A}\alpha_j\prod_{k\in INC-A}(1-\alpha_k)\right)\right)=\prod_{i\in IC}\left(\frac{\mu_i}{\mu_i+\lambda_i}\right)\left(\sum_{\substack{A\subset INC\\A\neq INC}}\left(\prod_{j\in A}\left(\frac{\mu_j}{\mu_j+\lambda_j}\right)\prod_{k\in INC-A}\left(\frac{\lambda_k}{\mu_k+\lambda_k}\right)\right)\right) \quad (b)$$

$$\pi_{Fail}=\sum_{\substack{A\subset IC\\A\neq IC}}\left(\prod_{i\in A}\alpha_i\prod_{j\in IC-A}(1-\alpha_j)\right)\sum_{B\subseteq INC}\left(\prod_{k\in B}\alpha_k\prod_{l\in INC-B}(1-\alpha_l)\right)=\sum_{\substack{A\subset IC\\A\neq IC}}\left(\prod_{i\in A}\left(\frac{\mu_i}{\mu_i+\lambda_i}\right)\prod_{j\in IC-A}\left(\frac{\lambda_j}{\mu_j+\lambda_j}\right)\right)\sum_{B\subseteq INC}\left(\prod_{k\in B}\left(\frac{\mu_k}{\mu_k+\lambda_k}\right)\prod_{n\in INC-B}\left(\frac{\lambda_n}{\mu_n+\lambda_n}\right)\right) \quad (c) \tag{12}$$

adad

In this model, the number of failures of the system over time t is obtained from Eq. (13):

$$N_{Failure}(t)=\int_0^t(\rho_2+\rho_3)\,\pi_{Fail}(x)dx=2\int_0^t\rho_2\,\pi_{Fail}(x)\,dx \tag{13}$$

We assume $\rho_2=\rho_3$ *in* Eq. (13) because they implicitly describe an equivalent rate. $\rho_2$ is the rate of transmission from survive state to failure state, but $\rho_3$ is the rate of transmission from healthy state to failure state. Actually, both $\rho_2$ and $\rho_3$ describe the rate of failure of critical subsystems of Fig. (5). So, assuming them to be equal can be correct.

### *B. Survivability Propagation Model of Dependent Components in Infrastructure Layer*

At this step, we suppose that a technical component $CMP_i$ is functionally dependent on components $C_1$ to $C_n$. Thus, while the $CMP_i$ has its own independent survivability, its final survivability also depends on survivability of $C_1$ to $C_n$ with coefficients $\delta$ and $\overline{\delta}$. So, we must try to compute survivability of $CMP_i$ based on $C_1$ to $C_n$ survivability along with its own independent survivability. This process is called propagation in our notation. This is represented in Fig. 7.

Coefficients $\delta$ and $\overline{\delta}$ in Fig. 7 are obtained by the Eq. (14). $\gamma_{x,i}$ used in this formulae shows the dependency coefficient of $CMP_i$ to $C_x$ and has been taken from Fig. 3. In Eq. (14), $\pi_H^{C_x}$ shows the healthy state probability of infrastructure component $C_x$ that $CMP_i$ is dependent to.

$$\delta_x=\left(\pi_H^{C_x}+\pi_S^{C_x}\right)\gamma_{x,i}$$
$$\overline{\delta_x}=\pi_F^{C_x}\cdot\gamma_{x,i} \tag{14}$$

Now we define following sets for computing dependent component $CMP_i$ survivability.

$$S=\{l\in N\mid CMP_i\ is\ depended\ to\ CMP_l\}$$
$$IC=\{l\in S\mid CMP_l\ is\ critical\ \ for\ CMP_i\}$$
$$INC=\{l\in S\mid CMP_l\ is\ not\ critical\ \ for\ CMP_i\}$$

Based on coefficients $\delta$ and $\overline{\delta}$ we compute final survivability of $CMP_i$ through Eq. (15). In Eq. (15), $\pi_H^{abstract}$ means the probability of healthy state before including dependencies. The probability after including dependencies represented by $\pi_H^{final}$.

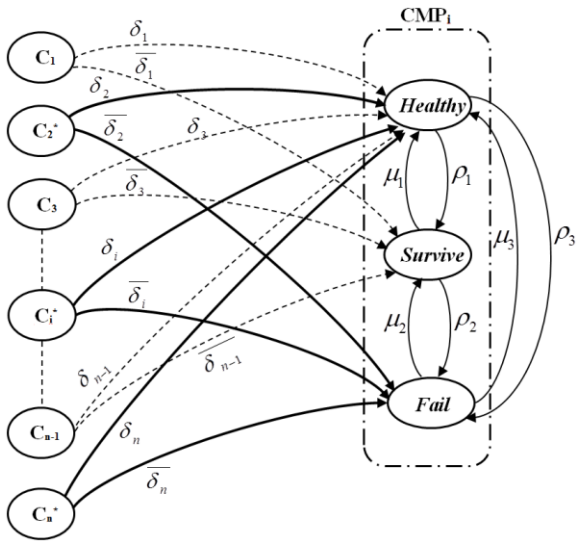



Fig. 7. Survivability Propagation in Infrastructure Layer of Model Among Dependent Components.

$$\pi_H^{final}(CMP_i)=\prod_{x\in S}\delta_x+\left(1-\sum_{x\in S}\gamma_{x,i}\right)\pi_H^{abstract}$$

$$\pi_S^{final}(CMP_i)=\prod_{x\in IC}\delta_x\left(\sum_{\substack{A\subset INC\\ A\neq INC}}\left(\prod_{j\in A}\delta_j\prod_{k\in INC-A}\overline{\delta_k}\right)\right)+\left(1-\sum_{x\in S}\gamma_{x,i}\right)\pi_S^{abstract} \quad (15)$$

$$\pi_F^{final}(CMP_i)=\sum_{\substack{A\subset IC\\ A\neq IC}}\left(\prod_{j\in A}\delta_j\prod_{k\in IC-A}\overline{\delta_k}\right)\sum_{B\subseteq INC}\left(\prod_{j\in B}\delta_j\prod_{k\in INC-B}\overline{\delta_k}\right)+\left(1-\sum_{x\in S}\gamma_{x,i}\right)\pi_F^{abstract}$$

$$\pi_H^{final}(CMP_i)+\pi_S^{final}(CMP_i)+\pi_F^{final}(CMP_i)=1$$

### C. Comprehensive Model for Multilayer Survivability Computation

Now we are completely ready for developing our model toward the multilayer computing of survivability. To do so, as we mentioned at the beginning of section 5, we shoud compute the survivability of applications of software layer based on finalized survivability of components. Then we compute the survivability of operation layer functions based on applications survivability of each function. In other words, we must propagate the survivability of infrastructure layer components to software layer applications. Then propagate the survivability of applications to operation layer functions. This process exactly follows the propagation method provided in section 5.2. Fig. 8 shows the process.

In Fig. 8, we compute the survivability of applications with respect to survivability of its underlying components that depends on. This process is similar to previous one for calculating survivability of dependant component $CMP_i$. When computation of survivability of all applications is done, then we take them into account for computing survivability of operation layer functions in a similar way. One can say, we propagate from software layer to operation layer.

One important point in Fig. 8 is that for healthy operating of any application, it is enough that each underlying component performing its essential functions only. So we can merge the healthy and survive state of components and name it as operational state as illustrated in right portion of Fig. 8.

## VI. Survivability of a Crisis Management System

For investigating about the proposed model, we imagined a crisis management system and tried to model it as well. Then we applied the model to the crisis management system for verification of our approach. Based on our studies, we extracted the general model of Fig. 9 for a common crisis management system.

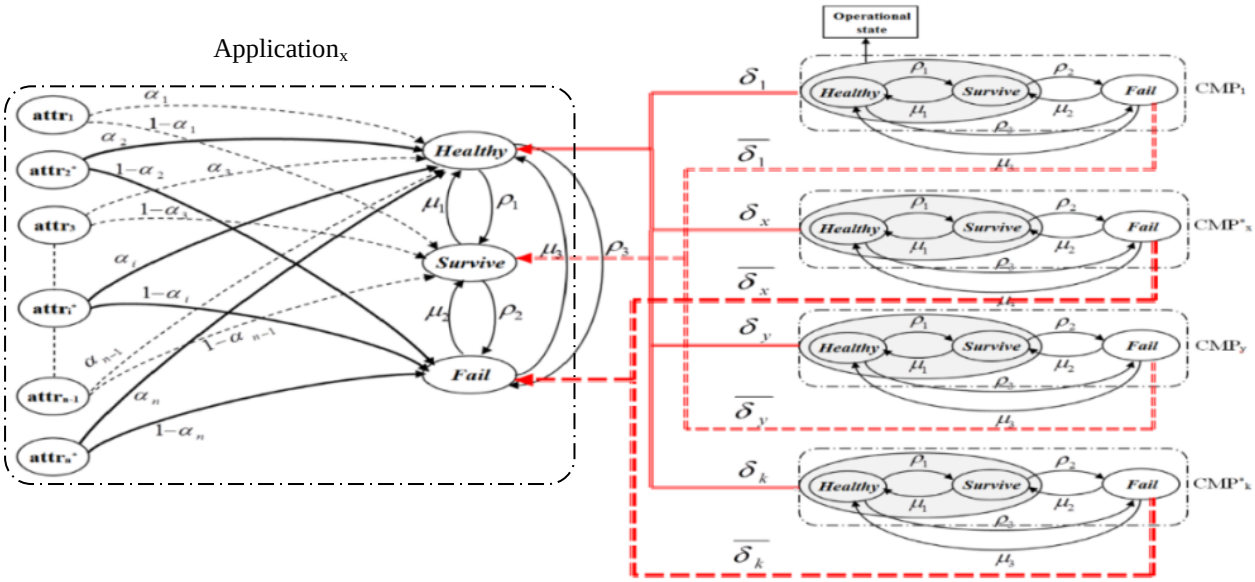



Fig. 8. Comprehensive Model of Survivability Computation.

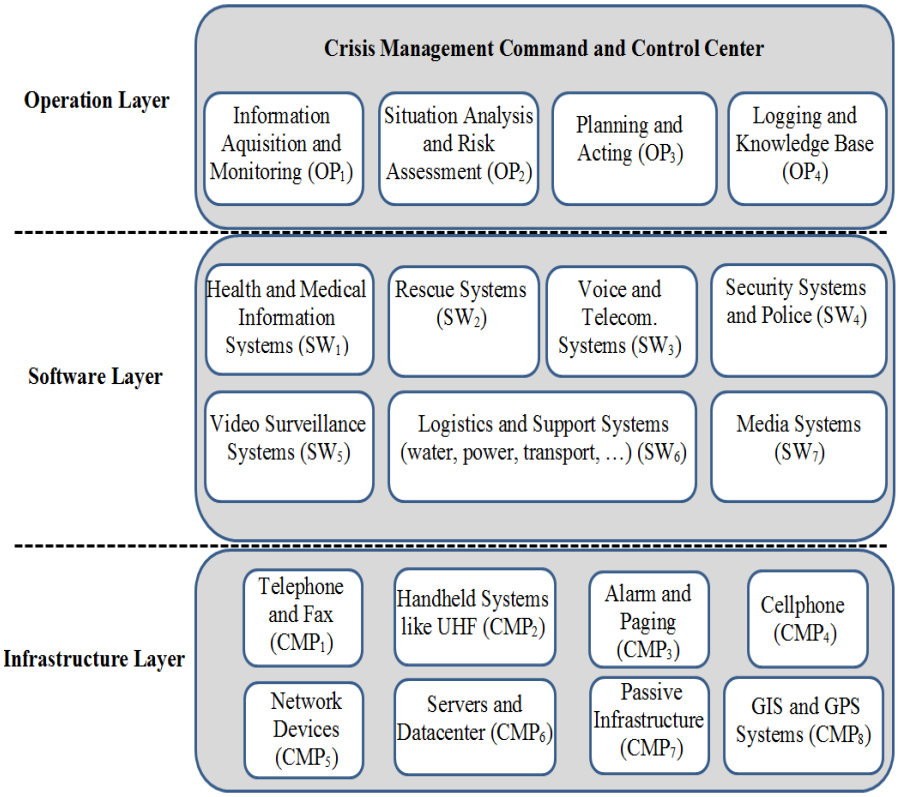



Fig. 9. Model of the Given Crisis Management System.

TABLE II. COMPUTATION OF CRITICAL OPERATION LAYER SURVIVABILITY

| Process | πH | πS | πF |
|---|---|---|---|
| OP1 | 9.631358E-01 | 3.874875E-02 | 9.092752E-03 |
| OP2 | 9.659744E-01 | 6.218598E-02 | 2.288903E-02 |

We have done a noticeable amount of calculations about all layers and components of the system for calculating survivability, but due to page number limitations we are not able to present all of them. Each interested reader can achieve them by email. Only for representing the achieved results at final stage, we present the calculated survivability of two critical processes OP1 and OP2 in Table II.

As seen from Table II, the essential disorder probabilities of OP1 and OP2 processes that are critical for the crisis management are 9.092752E-03 and 2.288903E-02 respectively, that are called $\pi_F^{OP_1}$ and $\pi_F^{OP_2}$.

Now, we are ready for computing the survivability probabilities of the total crisis management system overally. For this purpose, the total health probability of the crisis management structure is symbolized $\pi_H^{Overall}$ and probability of operation continuation in the failure conditions of non-critical process as $\pi_S^{Overall}$ and the probability of failure of total crisis management system as $\pi_F^{Overall}$. It is supposed that the computation processes performed for OP1 and OP2 are similarly performed for OP3 and OP4.

Since we want to find acceptable states operationally for processes, $\pi_S^{Overall}$ is also considered among those states. Thus, instead of direct computation of $\pi_S^{Overall}$ and $\pi_H^{Overall}$ values, the following quantity that is derived from the basic relation $\pi_H^{Overall}+\pi_S^{Overall}+\pi_F^{Overall}=1$ are introduced as the total system acceptable function probability.

$$\pi_F^{Overall} = 1-\left(1-\pi_F^{OP_1}\right)\left(1-\pi_F^{OP_2}\right) = 3.177366\text{E-}02$$
$$\pi_S^{Overall} = \left(1-\pi_F^{OP_1}\right)\left(1-\pi_F^{OP_2}\right)\left(1-\left(\pi_H^{OP_3}+\pi_S^{OP_3}\right)\left(\pi_H^{OP_4}+\pi_S^{OP_4}\right)\right)$$
$$\pi_H^{Overall} = \left(1-\pi_F^{OP_1}\right)\left(1-\pi_F^{OP_2}\right)\left(\pi_H^{OP_3}+\pi_S^{OP_3}\right)\left(\pi_H^{OP_4}+\pi_S^{OP_4}\right)$$

system acceptable operation probability is equal to:

$$\pi_H^{Overall}+\pi_S^{Overall}=1-\pi_F^{Overall}=96.822634E-02$$

## VII. CONCLUSION

This paper provides a general multilayer structure for systems survivability computation that is extendable to all common organization systems and operations. We designed a three layer model that connects the operational processes to application systems and application systems to the infrastructure layer. Then, the dependencies among these layers have been studied vertically (interlayer) and horizontally (intralayer). On the other hand, a new conceptual model was provided based on the Markov model characteristics for

survivability. Then, this model was used in a three stage structure for achieving our goal. In the first stage, the survivability of an infrastructure layer component was computed regardless of any dependencies and independently. Then, the horizontal dependency between the infrastructure layer components was entered in the computations and the survivability was computed applying those dependencies. In the final stage, the survivability computation model was provided taking into account the vertical dependencies for upper layers. The survivability of application systems and finally system operational processes have been computed including these dependencies. Finally, applying the total model in an important and frequently used problem such as the crisis management system, we could compute the real value of survivability for such system in the level of the crisis management critical and major processes and presented the abilities of our model. Utilizing this model will result in enabling the managers and planners to detect system weak points that make the highest loss in the survivability and efficiently protecting and retaining the system critical functions in crisis condition.

## REFERENCES

[1] Knight JC, Sullivan KJ. On the definition of survivability. University of Virginia, Department of Computer Science, Technical Report CS-TR-33-00. 2000.

[2] Deutsch MS, Willis RR. Software quality engineering: a total technical and management approach. Prentice-Hall Inc., 1988.

[3] Ellison RJ, Fisher DA, Linger RC, Lipson HF, Longstaff T. Survivable network systems: An emerging discipline. Carnegie-mellon Univ Pittsburgh PA Software Engineering Institute, 1997.

[4] Ellison RJ, Fisher DA, Linger RC, Lipson HF, Longstaff TA, Mead NR. An approach to survivable systems. In: NATO IST Symposium on Protecting Information Systems in the 21st Century, 1999, 754-759.

[5] Byon, Imju. Survivability of the US electric power industry. PhD Dissertation. Carnegie Mellon University, 2000.

[6] Caldera, Jose. Survivability requirements for the US health care industry. PhD Dissertation. Carnegie Mellon University, 2000.

[7] Ellison RJ, Fisher DA, Linger RC, Lipson HF, Longstaff TA, Mead NR. Survivability: Protecting your critical systems. IEEE Internet Computing, 1999, 3(6):55-63.

[8] Knight J, Heimbigner D, Wolf AL, Carzaniga A, Hill J, Devanbu P, Gertz M. The Willow architecture: comprehensive survivability for large-scale distributed applications. COLORADO UNIV AT BOULDER DEPT OF COMPUTER SCIENCE, 2001.

[9] Neumann PG. Practical Architectures for Survivable Systems and Networks: Phase-One Final Report. SRI INTERNATIONAL MENLO PARK CA COMPUTER SCIENCE LAB, 1999.

[10] Mead NR, Ellison RJ, Linger RC, Longstaff T, McHugh J. Survivable network analysis method. CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST, 2000.

[11] Ravindran B. Engineering dynamic real-time distributed systems: Architecture, system description language, and middleware. IEEE Transactions on Software Engineering, 2002, 8(1):30-57.

[12] Avizienis A, Laprie JC, Randell B. Fundamental concepts of dependability. University of Newcastle upon Tyne, Computing Science, 2001.

[13] Keromytis AD, Parekh J, Gross PN, Kaiser G, Misra V, Nieh J, Rubenstein D, Stolfo S. A holistic approach to service survivability. In: Proceedings of the 2003 ACM workshop on Survivable and self-regenerative systems in association with 10th ACM Conference on Computer and Communications Security, 2003, 11-22.

[14] Sames D, et al. Developing a heterogeneous intrusion tolerant CORBA system. In: Proceedings of IEEE International Conference on Dependable Systems and Networks, 2002, 239-248.

[15] Federal Standard l037C, U.S. Department of Commerce, National Telecommunications and Information Administration, Institute for Telecommunications Services, 1996.

[16] Heegaard PE, Trivedi KS. Network survivability modeling. Computer Networks, 2009, 53(8):1215-1234.

[17] Chong J, et al. Survivability architecture of a mission critical system: The DPASA example. In: 21st IEEE Annual Computer Security Applications Conference, 2005.

[18] Shrobe H, et al. AWDRAT: a cognitive middleware system for information survivability. AI Magazine. 2007, 28(3):73-80.

[19] M. Keshtgari, A. H. Jahangir,A General Framework For Network Survivability Performance Evaluation, PhD Dissertation, Sharif University of Technology, 2005.

[20] Richards MG, et al. Design for survivability: concept generation and evaluation in dynamic tradespace exploration. In: 2nd International Symposium on Engineering Systems, 2009.

[21] Zuo Y. A framework of survivability requirement specification for critical information systems. In: 43rd Hawaii International Conference on System Sciences (HICSS), 2010, 1-10.

[22] Zuo Y, Lande S. A logical framework of proof-carrying survivability. In: IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 472-481.

[23] Zuo Y. Incorporating Constraints to Software System Survivability Specification and Proof. In: Sixth International Symposium on Theoretical Aspects of Software Engineering (TASE), 2012, 67-74.

[24] Knight JC, Strunk EA. Achieving critical system survivability through software architectures. In: Architecting Dependable Systems II, 2004, 51-78.

[25] Avizienis A, Laprie JC, Randell B, Landwehr C. Basic concepts and taxonomy of dependable and secure computing. IEEE transactions on dependable and secure computing, 2004, 1(1):11-33.

[26] Trivedi KS, Xia R. Quantification of system survivability. Telecommunication Systems, 2015, 60(4):451-470.

[27] T1A1.2 Working Group on Network Survivability Performance, Technical Report No. 68, ATIS-T1.TR.68, 2001.

[28] Westmark VR. A definition for information system survivability. In: Proceedings of the 37th IEEE Annual Hawaii International Conference on System Sciences, 2004.

[29] Wang XG. Network System Survivability: A Survey. In: IEEE International Conference on Industrial and Information Systems, 2009, 232-235.

[30] Sterbenz JP, et al. Resilience and survivability in communication networks: Strategies, principles, and survey of disciplines. Computer Networks, 2010, 54(8):1245-1265.

[31] Rodríguez RJ, Chang X, Li X, Trivedi KS. Survivability Analysis of a Computer System Under an Advanced Persistent Threat Attack. In: International Workshop on Graphical Models for Security, 2016, 134-149.

[32] Necula GC. Proof-carrying code. In: Proceedings of the 24th ACM SIGPLAN-SIGACT symposium on Principles of programming languages, 1997,106-119.

# Developing A Model to Predict the Occurrence of the Cardio-Cerebrovascular Disease for the Korean Elderly using the Random Forests Algorithm

Haewon Byeon
Department of Speech Language Pathology
Honam University
Gwangju, Republic of Korea

***Abstract*—This study aimed to develop a model for predicting the cardio-cerebrovascular disease of the South Korean elderly using the random forests technique. This study analyzed 2,111 respondents (879 males and 1,232 females), who were age 60 or older, out of total 7,761 respondents, who completed the Seoul Welfare Panel Study. The result variable was defined as the cardio-cerebrovascular disease (e.g., hypertension, cerebral infarction, hyperlipidemia, cardiac infarction, and angina). As a result of developing a random forest-based model, the major determinants of the cardio-cerebrovascular diseases of the South Korean elderly were mean monthly household income, the highest level of education, subjective health condition, subjective friendship, subjective family relationship, smoking, regular exercise, age, marital status, gender, depression experience, economic activity, and high-risk drinking. Among them, mean monthly household income was the most important predictor of the cardio-cerebrovascular disease. Based on the developed prediction model, it is needed to develop a systematic program for preventing the cardio-cerebrovascular disease of the Korean elderly.**

***Keywords*—Prediction model; data mining; random forest; risk factors; cardio-cerebrovascular disease; stroke**

## I. Introduction

The cardio-cerebrovascular diseases include cerebrovascular diseases (e.g., cerebral hemorrhage and cerebral infarction), cardiac disorders (e.g., cardiac insufficiency, angina, and cardiac infarction), and vascular abnormalities (e.g., hypertension, diabetes, hyperlipidemia, and arteriosclerosis). As of 2013, the mortality due to cardio-cerebrovascular diseases accounts for more than 25% of the national mortality. The annual death toll is 50.3 people for cardiovascular diseases and 50.2 people for cerebrovascular per 100,000 population [1]. The cardio-cerebrovascular disease is the second leading cause of death in South Korea [1]. The cardio-cerebrovascular disease has increased by 1.35 times over the past decade and has become a critical health problem in South Korea [1].

Particularly, the cardio-cerebrovascular disease is a representative chronic disease of the elderly. It is known that the mortality rate increases rapidly with age. Especially, previous studies reported that it increased abruptly in the elderly over 70 years old [1]. Additionally, the cardio-cerebrovascular disease of the elderly is often accompanied by severe disability even if surgical treatment is successful. Therefore, they tend to have a hard time to return to the society even after recovery [2]. Consequently, it is essential to identify factors associated with the cardio-cerebrovascular disease and prevent them for achieving the successful aging.

As more people die from cardio-cerebrovascular diseases, there is a growing interest in managing and preventing the diseases. In the past 20 years, a number of studies have attempted to evaluate various risk factors for cardio-cerebrovascular diseases such as sociodemographic factors, lifestyle, and family history [3-6]. The results of these studies have identified various risk factors encompassing those that cannot be controlled (e.g., age and gender) and those that can be controlled (e.g., eat habits and physical activities) [3-6].

However, it has been pointed out that these individual risk factors have limitations in explaining the onset of a cardio-cerebrovascular disease [7]. Moreover, studies have indicated different factors as the most important risk factor. Additionally, although the cardio-cerebrovascular disease is known as a complex disease due to the interactions of multiple factors including sociodemographic factors (e.g., age and gender), environmental factors (e.g., lifestyle), and causative disease factors (e.g., hypertension and hyperlipidemia) [8-9], recent studies reported that psychological factors such as depression were major risk factors as well [10-11].

Moreover, the occurrence patterns and the risk factors of the cardio-cerebrovascular disease vary greatly among different ethnic groups. Therefore, it is difficult to establish a prevention and management strategy based on the results of previous studies conducted for different ethnic groups. Additionally, the lifestyle, an important factor in deciding the health, is determined by cultural influences as well as personal characteristics. Therefore, it is necessary to develop a model for predicting the cardio-cerebrovascular disease with reflecting the characteristics of the elderly living in the local communities in South Korea using big data.

The random forests technique has been used more frequently as a data mining algorithms for predicting the risk factors of target variables such as a disease or a disability [12-14]. The random forests technique is a method of combining multiple decision trees based on the ensemble technique in order to minimize the over-fitting, which is a shortfall of the

decision tree. The technique shows a good prediction ability, which is an advantage of this technique. This study aimed to develop a model for predicting the cardio-cerebrovascular disease of the South Korean elderly using the random forests technique.

Construction of this study is as follows. chapter II explains data source and materials and chapter III defines random forests and explains the procedure of final model development. Chapter IV compares the results of developed final prediction model. Lastly, chapter V presents discussion and direction for future studies.

## II. Materials and Methods

### A. Data Source

This study analyzed a portion of the raw data of the Seoul Welfare Panel Study, which was conducted by Seoul Welfare Foundation to survey Seoul citizens from Jun 1 to August 31, 2010. Seoul Welfare Panel Study was approved (#20113) by Statistics Korea in 2009 and it has been conducted to identify the welfare level of households residing in Seoul, understand the status of the welfare vulnerable class, and estimate the demand for welfare services [15]. This study targeted households in Seoul as of "the 2005 Population and Housing Census" and sampled using the stratified cluster sampling method for 25 districts in Seoul. The main survey items were income, economic level, health, living conditions, and the demand for welfare services. The survey was conducted by using the computer-assisted personal interviewing method: the interviewer visited the surveyed households and inputted responses according to a structured questionnaire into a portable computer. This study analyzed 2,111 respondents (879 males and 1,232 females), who were age 60 or older, out of total 7,761 respondents, who completed the survey.

### B. Measurements and Definitions of Variables

The result variable was defined as the cardio-cerebrovascular disease (e.g., hypertension, cerebral infarction, hyperlipidemia, cardiac infarction, and angina). The explanatory variables included age (60years or older and younger than 70years or 70years or older), gender (male or female), the highest level of education (below elementary school, junior high school, high school, and college graduation and over), economic activity (yes or no), mean monthly household income (less than 2 million KRW, 2-4 million KRW, and more than 4 million KRW), marital status (living with a spouse, married but not living with a spouse, or single), High-risk drinking (yes or no), Smoking (non-smoker, past smoker, current smoker), subjective health condition (good, normal, or poor), subjective family relationship (good, average, or bad), subjective friendship (good, average, or poor), regular exercise (no or yes), and the depression symptom in the past one month (no or yes).

## III. Analysis Methods

### A. Exploring Potential Factors of the Cardio-Cerebrovascular Disease in Old Age

The prevalence of the cardio-cerebrovascular disease between groups was analyzed by using the chi-square test. When the significance level of an explanatory variable was 0.1 or below, it was considered as a potential factor of the cardio-cerebrovascular disease and it was included in the random forest model.

### B. Random Forests Algorithm

The random forests technique [16] is an algorithm that creates various sample datasets using bootstrap. This method has an advantage of increasing the diversity of the decision tree because it repeats the process of randomly selecting several variables [17]. Unlike the decision trees, which present each node with the partition showing the most optimum results by using all variables, the random forests select explanatory variables randomly and use the method showing the most optimum results among the selected explanatory variable groups (Figure 1). The process of random forests is shown in Eq. (1).

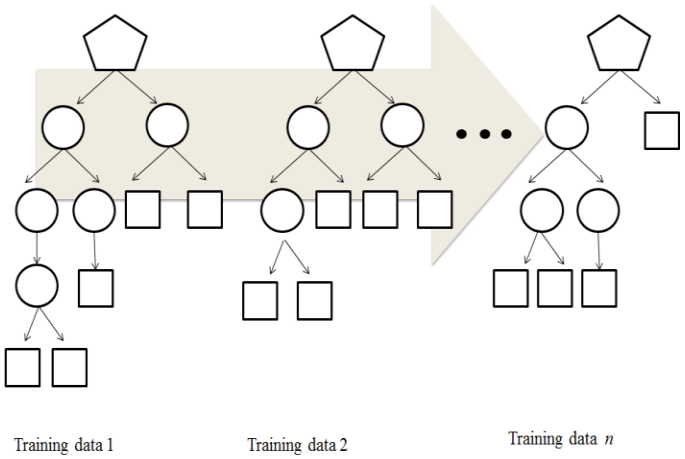



Fig. 1. Random forest classifier: source is Byeon [12]

1: Generate $B$ bootstrap samples $L_1, ..., L_B$ from the original training data set L
2: Grow a random forest tree using a random feature selection from bootstrapped data.
:randomly select $\sqrt{p}$ predictors at each node and split the data using the best predictors.
3: Construct train classifiers $C_b(\mathbf{x}), b = 1, ...B$ from each of $L_b$ samples
4: Aggregate the $B$ train classifiers.
Let $N_j$ be the number of times that classified j
$N_j = \sum_{b=1}^{B} I[C_b\mathbf{x} = j]$, for $j = 1, ...J$
5: final classification: $arg\,max_j N_j$

(1)

Another advantage of the random forests is to reduce the variance compared to the bagging method because it decreases the correlation between trees. Moreover, it presents more accurate results than other algorithms and it is useful to find an important variable in big data because it utilizes thousands of independent variables without eliminating variables [18]. Especially, when there are many input variables, it often shows similar or better prediction power than bagging or boosting. The input source of the R program for performing random forests analysis is shown in Fig 2.

```
X_train, X_test, y_train, y_test = train_test_split(

cvd.data, cvd.target, random_state=0)

forest = RandomForestClassifier(n_estimators=500, random_state=0)

forest.fit(X_train, y_train)



print("train: {:.3f}".format(forest.score(X_train, y_train)))

print("test: {:.3f}".format(forest.score(X_test, y_test)))
```

Fig. 2. Input source of the R program for performing random forests

In this study, the number of trees in the model was set to 500. The analysis was conducted using R version 3.4.2 and Waikato Environment for Knowledge Analysis (WEKA) version 3.6.0 [19].

## IV. Results

### A. General Characteristics of Subjects

The characteristics the data (n=2,111) were analyzed and the results showed that the 53.5% of study subjects were between 60 and 69 years old and the 58.4% of them were women. The majority of the subjects lived with their spouses (67.2%), were elementary school graduation or below (43.3%), had the mean monthly income less than 2 million KRW (64.7%), were not economically active (83.2%), did not exercise regularly (55.4%), had poor subjective health (39.4%), had good subjective family relationship (59.1%), had average subjective friendship, and did not experience a depression symptom in the past one month (74.4%). The prevalence of the cardio-cerebrovascular disease was 42.3 %.

### B. Potential Factors of Cardio-Cerebrovascular Disease in Old Age

Table 1 shows the general characteristics and potential factors of subjects according to the prevalence of cardio-cerebrovascular diseases. The prevalence of cardio-cerebrovascular diseases, which indicated the proportion of subjects suffering hypertension, cerebral infarction, hyperlipidemia, cardiac infarction, and angina, was 42.3% (n=894). The results of chi-square test showed that the elderly with cardio-cerebrovascular diseases and those without cardio-cerebrovascular diseases were there were significant ($p<0.05$) different in age, marital status, economic activity, smoking, the depression symptom in the past one month, subjective health condition, and subjective family relationship. The prevalence of cardio-cerebrovascular diseases was significantly higher for the elderly equal to or older than 70 years (50.8%), not living with a spouse (47.8%), not economically active (44.1%), former smoker (43.8%), depression symptom experience in the past one month (49.7%), poor subjective health (51.1%), and average family relationship (46.4%).

TABLE I. General Characteristics Of The Subjects By Cardio-Cerebrovascular Disease, N (%)

| Characteristics | cardio-cerebrovascular disease | | p |
|---|---|---|---|
| | *Yes (n=894)* | No (n=1,217) | |
| Age | | | <0.001 |
| 60-69 | 396 (35.0) | 734 (65.0) | |
| 70+ | 498 (50.8) | 483 (49.2) | |
| Gender | | | 0.315 |
| Male | 316 (41.1) | 518 (58.9) | |
| Female | 533 (43.3) | 699 (56.7) | |
| Marital Status | | | 0.002 |
| Living with a spouse | 563 (39.7) | 856 (60.3) | |
| Married but not living with a spouse | 20 (47.6) | 22 (52.4) | |
| Single | 311 (47.8) | 339 (52.2) | |
| The highest level of education | | | 0.071 |
| Below elementary school | 406 (44.4) | 508 (55.6) | |
| Junior high school | 160 (42.9) | 213 (57.1) | |
| High school | 185 (37.3) | 311 (62.7) | |
| College graduation and over | 143 (43.6) | 185 (56.4) | |
| Mean monthly household income | | | 0.056 |
| Less Than 2 million KRW | 601 (44.0) | 765 (56.0) | |
| 2-4 million KRW | 194 (39.9) | 292 (60.1) | |
| More than 4 million KRW | 31 (33.3) | 62 (66.7) | |
| Economic activity | | | <0.001 |
| Yes | 120 (33.9) | 234 (66.1) | |
| No | 774 (44.1) | 983 (55.9) | |
| Smoking | | | 0.035 |
| Non-smoker | 619 (43.1) | 817 (56.9) | |
| Past smoker | 202 (43.8) | 259 (56.2) | |
| Current smoker | 73 (34.1) | 141 (65.9) | |
| High-risk drinking | | | 0.299 |
| No | 122 (38.4) | 196 (61.6) | |
| Yes | 61 (33.7) | 120 (66.3) | |
| Regular exercise | | | 0.689 |
| No | 500 (42.7) | 670 (57.3) | |
| Yes | 394 (41.9) | 547 (58.1) | |
| Depression symptom in the past one month | | | <0.001 |
| No | 625 (39.8) | 945 (60.2) | |
| Yes | 269 (49.7) | 272 (50.3) | |
| Subjective health condition | | | <0.001 |
| Good | 158 (27.1) | 426 (72.9) | |
| Normal | 312 (44.7) | 386 (55.3) | |
| Poor | 424 (51.1) | 405 (48.9) | |
| Subjective family relationship | | | 0.014 |
| Good | 479 (39.5) | 735 (60.5) | |
| Average | 307 (46.4) | 355 (53.6) | |
| Bad | 78 (43.6) | 101 (56.4) | |
| Subjective friendship | | | 0.146 |
| Good | 277 (39.4) | 426 (60.6) | |
| Average | 481 (44.0) | 611 (56.0) | |
| Bad | 136 (43.0) | 180 (57.0) | |

### C. Predict Occurrence of Cardio-Cerebrovascular Disease for Korean elderly

The importance of variables (the decrement of node impurity) based on random forests is shown in Table 2 and Figure 3. The results showed that the major determinants of the

cardio-cerebrovascular diseases of the South Korean elderly were mean monthly household income, the highest level of education, subjective health condition, subjective friendship, subjective family relationship, smoking, regular exercise, age, marital status, gender, depression experience, economic activity, and high-risk drinking. Among them, mean monthly household income was the most important predictor of the cardio-cerebrovascular disease.

TABLE II. IMPORTANCE OF VARIABLES: THE DECREMENT OF NODE IMPURITY

| Importance of variables | Decrement of node impurity |
|---|---|
| Household income | 53.351 |
| Highest level of education | 39.502 |
| Ubjective health condition | 33.180 |
| Subjective friendship | 27.991 |
| Subjective family relationship | 25.104 |
| Smoking | 23.234 |
| Regular exercise | 16.904 |
| Age | 16.635 |
| Marital status | 16.501 |
| Gender | 14.559 |
| Depression experience | 12.522 |
| Economic activity | 11.815 |
| High-risk drinking | 8.330 |

Figure 4 shows the error rate graphs for each prediction model for each of the extracted 500 bootstrap samples. The error rate of the developed random forests was 0.24 and the prediction rate was 76.5%.

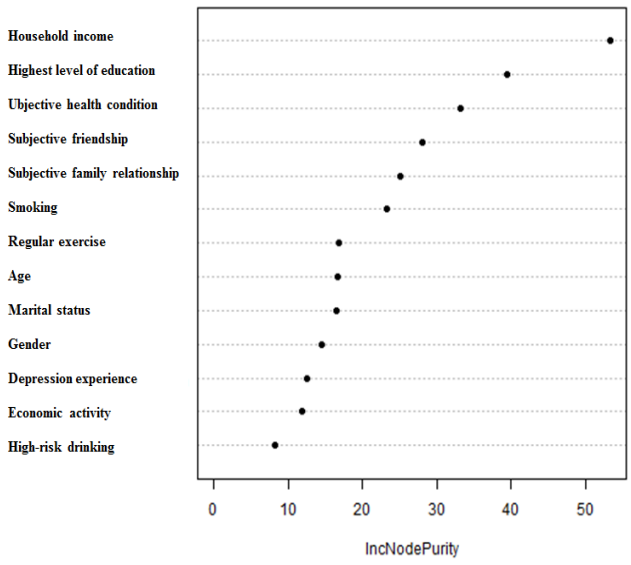



Fig. 3. Variable importance plot

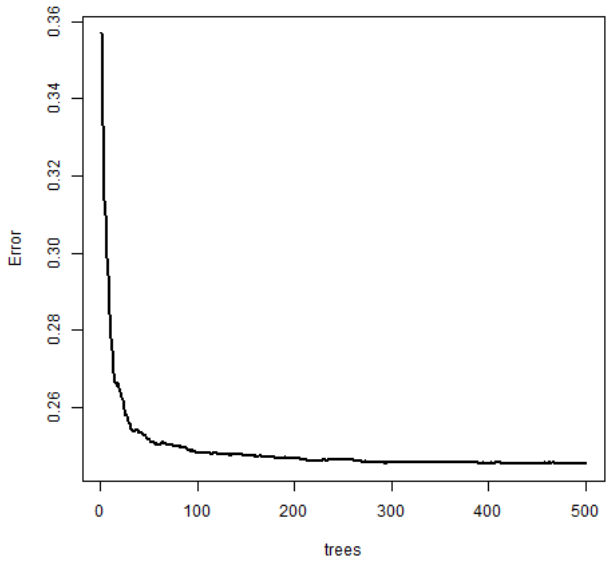



Fig. 4. Error rate graphs for each prediction model

## V. DISCUSSION

This study developed a model for predicting the cardio-cerebrovascular disease of the elderly living in the community using the random forests technique, which is a data minding algorithm based on the classification learning. This study constructed a prediction model of cardio-cerebrovascular disease considering multiple risk factors. The results of the constructed model showed that household income was the most important factor followed by the education level, which indicated that socio-economic factors were major risk factors.

Many previous studies have reported that the prevalence of the cardio-cerebrovascular disease is affected by the socioeconomic levels [20-22]. The mechanism of the socioeconomic factors can be explained by the changes in hemodynamics due to the increase of stress and the lack of health life practice, HDL-cholesterol, insulin resistance, and blood coagulation-related factors [20]. Therefore, the low-income and poorly educated groups should be sufficiently considered when establishing prevention programs.

The results of this study confirmed that depression was a major predictor of the cardio-cerebrovascular disease. Previous studies reported that the depression and the cardio-cerebrovascular disease were highly associated. Particularly, the cardio-cerebrovascular disorder was identified as a risk factor of depression [6, 23], and those with depression were at risk for the cardiovascular disease and their mortality risk was twice than others without it [24]. Morris et al. (1993) also found that cerebrovascular disorders were frequently accompanied by depression and people with depression had an 8-fold higher risk of death from cerebral infarction than others without it [25]. Byeon (2015) identified the cardio-cerebrovascular disease risk groups using the QUEST algorithm and also predicted that the elderly who experienced depression would have a higher risk of the cardio-cerebrovascular disease [6]. The results of this study indicated that the depression in the old age was a risk factor of the cardio-cerebrovascular disease. Therefore, it will be necessary

to develop the cardio-cerebrovascular disease prevention program for the elderly with depression for preventing the cardio-cerebrovascular disease of the elderly in the local community.

This study developed a depression prediction model for children from multicultural families by using CHAID algorithm and found that the experience of social discrimination is the most critical factor affecting depression. Although it is hard to compare the results of this study directly, the previous studies evaluating the relationship between social discrimination and mental health reported that the economic discrimination and the discrimination against a specific group (e.g., the elderly group) were significant predictor variables negatively influencing mental health [18]. Therefore, it is necessary to establish a legal system and pay social level interests to overcome the discrimination and prejudice against adolescents from multicultural families based on the results of this study.

The results of this study showed that the cardio-cerebrovascular disease prediction model based on the random forests technique had stronger prediction power than the previously developed cardio-cerebrovascular disease prediction model based on QUEST algorithm [6]. The random forests showed superior prediction performance than the decision tree and it produced more stable results because it made decisions by integrating the prediction results of multiple decision trees using the bootstrap sample [26]. Therefore, it was believed that using the random forests model would be more effective than using the decision tree model when estimating the importance of variables in the development of disease prediction models. It will be necessary to compare the predictive performance of the logistic regression model, the decision tree model, and the random forests in the future.

## VI. Conclusion

Based on the developed prediction model, it is needed to develop a systematic program for preventing the cardio-cerebrovascular disease of the Korean elderly.

## Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07041091).

## References

[1] Statistical Office, Cause of death statistics, National Statistical Office, Daejeon, 2015.

[2] P. B. Gorelick, R. L. Sacco, D. B. Smith, M. Alberts, L. M. Alexander, D. Rader, J. L. Ross, E. Raps, M. N. Ozer, L. M. Brass, M. E. Malone, S. Goldberg, J. Booss, D. F. Hanley, J. F. Toole, N. L. Greengold and D. C. Rhew, Prevention of a first stroke-a review of guidelines and a multidisciplinary consensus statement fromt The National Stroke Association. Journal of the American Medical Association, vol. 281, no. 12, pp. 1112-1120, 1999.

[3] N. Y. Han, E. A. Ko and S. Y. Hwang, Knowledge of stroke symptoms and risk factors among older adults. Korean Journal of Adult Nursing, vol. 21, no. 3, pp. 314-323, 2009.

[4] D. Ettehad, C. A. Emdin, A. Kiran, S. G. Anderson, T. Callender, J. Emberson, J. Chalmers, A. Rodgers and K. Rahimi, Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. The Lancet, vol. 387, no. 10022, pp. 957-967, 2016.

[5] S. Mottillo, K. B. Filion, J. Genest, L. Joseph, L. Pilote, P. Poirier, S. Rinfret, E. L. Schiffrin and M. J. Eisenberg, The metabolic syndrome and cardiovascular risk: a systematic review and meta-analysis. Journal of the American College of Cardiology, vol. 56, no. 14, pp. 1113-1132, 2010.

[6] S. Cho and H. Byeon, Predictive modeling of hypertension in Korean old adults using QUEST algorithm. Asia Life Sciences, pp. 21-30, 2015.

[7] T. W. Smith and J. M. Ruiz, Psychosocial influences on the development and course of coronary heart disease: current status and implications for research and practice. Journal of Consulting and Clinical Psychology, vol. 70, pp.5 48-568, 2002.

[8] R. M. Conroy, K. Pyörälä, A. P. Fitzgerald, S. Sans, A. Menotti, G. De Backer, P. De Bacquer, P. Ducimetière, P. U. Jousilahti, I. Keil, R. G. Njølstad, T. Oganov, H. Thomsen, A. Tunstall-Pedoe, H. Tverdal, P. Wedel and I. M. Wilhelmsen, Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. European Heart Journal, vol. 24, no. 11, pp. 987-1003, 2003.

[9] J. D. Berry, A. Dyer, X. Cai, D. B. Garside, H. Ning, A, Thomas, A. Thomas, P. Greenland, L. V. Horn, R. P. Tracy and D. M. Lloyd-Jones, Lifetime risks of cardiovascular disease. New England Journal of Medicine, vol. 366, no. 4, pp. 321-329, 2012.

[10] A. Rozanski, J. A. Blumenthal and J. Kaplan, Impact of psychological factors on the pathogenesis of cardiovascular disease and implications for therapy. Circulation, vol. 99, pp. 2192-2217, 1999.

[11] B. W. Penninx, Depression and cardiovascular disease: epidemiological evidence on their linking mechanisms. Neuroscience & Biobehavioral Reviews, vol. 74, pp. 277-286, 2017.

[12] H. Byeon, A prediction model for mild cognitive impairment using random forests. International Journal of Advanced Computer Science and Applications, vol. 6, no. 12, pp. 8-12, 2015.

[13] K. R. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, D. Rueckert and Alzheimer's Disease Neuroimaging Initiative. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. NeuroImage, vol. 65, pp. 167-175, 2013.

[14] H. Byeon, H. Jin, and S. Cho, Development of parkinson's disease dementia prediction model based on verbal memory, visuospatial memory, and executive function. Journal of Medical Imaging and Health Informatics, vol. 7, no. 7, pp. 1517-1521, 2017.

[15] Seoul Welfare Foundation, Seoul Welfare Panel Study 2010, Seoul Welfare Foundation, Seoul, 2010.

[16] L. Breiman, Random forests. Machine learning, vol. 45 no. 1, pp. 5-32, 2010.

[17] A. Singh, M. N. Halgamuge and R. Lakshmiganthan, Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbors algorithms. International Journal of Advanced Computer Science and Applications, vol. 8, no. 12, pp. 1-10, 2017.

[18] H. Byeon, D. Lee, and S. Cho, Assessment for the model predicting of the cognitive and language ability in the mild dementia by the method of data-mining technique. International Journal of Advanced Computer Science and Applications, vol. 7, no. 6, 2016.

[19] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, Burlington, 2016.

[20] J. F. Meschia, C. Bushnell, B. Boden-Albala, L. T. Braun, D. M. Bravata, S. Chaturvedi, and L. B. Goldstein, Guidelines for the primary prevention of stroke: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke, vol. 45, no. 12, pp. 3754-3832, 2014.

[21] K. Backholer, S. A. Peters, S. H. Bots, A. Peeters, R. R. Huxley, R. R. and M. Woodward, M, Sex differences in the relationship between socioeconomic status and cardiovascular disease: a systematic review and meta-analysis. Journal of Epidemiology and Community Health, jech-2016, 2016.

[22] C. T. January, L. Samuel Wann, J. S. Alpert, H. Calkins, J. E. Cigarroa,, J. C. Cleveland Jr, J. B. Conti, P. T. Ellinor, M. D. Ezekowitz, M. E. Field, K. T. Murray, R. L. Sacco, W. G. Stevenson, P. J. Tchou, C. M Tracy and C. W. Yancy, 2014 AHA/ACC/HRS guideline for the

management of patients with atrial fibrillation: executive summary: A report of the American College of Cardiology/American Heart Association Task Force on practice guidelines and the Heart Rhythm Society. Journal of the American College of Cardiology, 64(21), 2246-2280.

[23] B. E. Cohen, D. Edmondson and I. M. Kronish, State of the art review: depression, stress, anxiety, and cardiovascular disease. American Journal of Hypertension, vol. 28 no. 11, pp. 1295-1302, 2015.

[24] J. C. Barefoot and M. Schroll, Symptoms of depression, acute myocardial infarction, and total mortality in a community sample. Circulation. vol. 93, no. 11, pp. 1976-1980, 1996.

[25] P. L. Morris, R. G. Robinson and J. Samuels, Depression, introversion and mortality following stroke. Australian & New Zealand Journal of Psychiatry, vol. 27, no.3, pp. 443-449, 1993.

[26] J. Maroco, D. Silva, A. Rodrigues, M. Guerreiro, I. Santana and A. de Mendonça, Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC research notes, vol. 4, no. 1, 299, 2011.

# Proposal for a Feature Automation Solution for an IMS-KMS-IoT Platform based on SDN

Samba DIOUF[1], Kéba GUEYE, Samuel OUYA[2], Gervais MENDY[3]
Laboratory LIRT, Higher Polytechnic School, University Cheikh Anta Diop
Dakar, Sénégal

***Abstract*—The concept of the Internet of Things is a paradigm that is gaining more and more ground and soon the number of connected objects will be counted in billions. This will transform our lives and pose new challenges. To meet the challenges of objects several platforms have been proposed by previous work. Some are based on an IMS-IoT platform and others integrate IMS, IoT and SDN technologies. Our article proposes an architecture that integrates SDN into an IMS-IoT platform and the automation of the control layer (IMS-IoT platform) and the transport layer of the functional architecture in order to meet future requirements related to the configuration of these many objects, mobility and diversity of user terminals (smartphone, tablet, computers, etc.). The proposed architecture makes it possible to benefit from all the simplicity and efficiency provided by an SDN network but also the services offered by the IMS-IoT platform.**

***Keywords—IMS; IoT; SDN; automation; KMS***

## I. INTRODUCTION

The concept of the Internet of Things (IoT) incorporates the vision of the ubiquitous virtual connectivity of billions of physical objects or things across a global network infrastructure with interoperable, self-configurable and scalable capabilities. Objects can be part of various application domains and are therefore represented by different types of devices that have heterogeneous technical parameters and communication capabilities. This imposes various technical challenges in terms of adaptation, context awareness, device management discoveries, scalability, large data management, privacy and security.

In addition, the main asset of the IoT idea is its considerable impact on many aspects of everyday life and the behaviour of potential users. The most obvious effects of the Internet of Things, observable by a private user, would be visible in both the home and the workplace. In the first case, some examples of possible application scenarios in which the new paradigm, the Internet of Things, will play a prominent role in the near future are home automation, online health, life support autonomous and improved learning [1], [2]. In the second case, business users have observed similar consequences that can be detected in certain areas such as logistics, intelligent transport of people and goods, automation and industrial manufacturing, as well as business process management.

The "Building the Environment of Things as a Service" (BETaaS) paradigm is the result of a successful research initiative [3] that takes advantage of computer fog and introduces the innovative concept of Things as a Service (TaaS) into the medium M2M applications. Computer fog [4] was introduced as a concept that reduces the amount of data to be transported to the cloud and reduces the burden of processing cloud systems by implementing data processing, storage, analysis and the edge of the network. BETaaS has extended the computer fog paradigm by introducing a local cloud of M2M gateways that aggregate and analyse data from different heterogeneous IoT devices and expose unified service-oriented interfaces that can be used to access objects connected to the local cloud, thereby introducing the concept of TaaS. This approach offers distinct advantages such as reducing network complexity, reducing transported data, generated by things, and seamless service-oriented interfaces that provide access to different types of devices with technical parameters and capabilities. heterogeneous communication. However, to fully exploit the benefits of exposed IoT services, an effective mechanism for deploying, managing, and delivering end-user services is required.

This article discusses the main challenges of integrating Software Defined Networking (SDN) into an IMS-IoT platform and automating control and transport layers of data collected by objects. IP Multimedia Subsystem (IMS) is a technology that manages session establishment, management, and termination through Session Initiation Protocol (SIP) to enable the deployment of services across all IP networks. IMS provides a common IP interface that simplifies signalling and application development and facilitates the processing of multiple sessions that can be used by different applications. In addition, it enables access control and billing management, which enables consistent management of different applications on the network [5, 6]. The benefits of introducing IMS into the IoT domain have already been examined by several research teams. A cloud-based implementation of the generic framework is described later in [7], where the core IMS components are virtualized in a resource pool and provided as Infrastructure as a Service (IaaS). The framework concept is further extended in [8], where an object management and naming mechanism based on IMS uniform resource identifiers (URIs) is proposed. The article presented a name distribution agent that used SIP messages "MESSAGE" to assign specific URIs to "things". In [9], the integration of IoT devices into a SIPIMS platform is considered in the context of WLAN-3GPP networks. Different SIP message flows are designed to provide different IoT functions, such as recording, streaming, and notifications, based on request / report messages. Finally, in [10], an operator-class IoT service-oriented architecture that integrates IMS with wireless sensor networks (WSN) is

prototyped. The architecture uses the Sensor Web Enablement (SWE) framework to solve the heterogeneity of IoT devices by introducing two main components: a facilitator who introduces IoT services on the telecom operating platform and a gateway that interconnects the IoT device, activator and the WSN. As proof of concept, the proposed design was implemented in an application scenario with 20 WSN nodes and the network load was evaluated, showing that the introduction of the M2M gateway effectively reduces the data traffic generated.

This document proposes an architecture for the provision of IoT services such as e-health, e-learning, etc. on a network of all-IP telecommunications service providers, while facilitating access to services and terminal diversification. We are taking a step closer to integrating SDN into an IoT-IMS platform by introducing the automation of core elements to address the challenges of increased traffic and network complexity. SDN is defined by the Open Network Foundation as a technology where "the control of the network is decoupled from the transfer and is directly programmable".

Previously related works had proposed an implementation of various software modules that can have a dynamic control over network functions, to support network expansion and innovation and for more efficient use of network resources [11]. One of the key benefits of SDN is thought to be its ability to efficiently route traffic flows, making it easier to support Quality of Service (QoS) differentiation. In [12], a converged SDN and IMS framework is presented, which incorporates an Open Flow controller that communicates with the IMS application servers via RESTful programmable interfaces (APIs). This allows for application-based QoS classification and data flow routing. The integration of Open Flow into IMS is also proposed in [13], where an SDN IMS research benchmark is designed.

This paper is structured as follows: Section II discusses the state of the art IMS (OpenIMSCore), SDN and IoT technologies. Section III discusses the composition and characteristics of the proposed architecture across its different layers. Section IV shows the results obtained with the automation of the control and transport layers and an analysis of these results. Finally, Section V conclusion and future work.

## II. State of the Art

### A. OpenIMSCore

OpenIMSCore is the implementation of an IMS core network based on the open source solution SIP Express Router (SER). The main elements of the layer are: Home Subscriber Servers (HSS) and different Call Control Session Control (CSCF) servers. The control layer connects to the application servers (AS) in the service layer and to the user equipment (UE) via the transport layer. The CSCF-Proxy (P-CSCF) acts as the point of contact with the UE for access to the main network service; Examination-CSCF (I-CSCF) [12] [13]. See Figure 1.

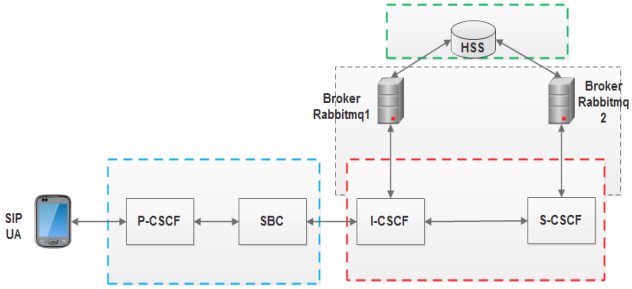



Fig. 1. IMS Architecture with Rabbitmq Servers at HSS.

- The P-CSCF is the point of entry into the IMS network for all signalling of a user's session. It is the point of contact between the user and the central server.
- The I-CSCF constitutes, for a session of a user, the entry point into the network of the operator to which the user has subscribed a service contract (SLA). It acts as an intermediary between the proxy server and the authentication server. It redirects a user's requests to his original operator. The I-CSCF entity assigns an S-CSCF entity to the UE and transfers the request to the selected S-CSCF entity. To perform this function, a DIAMETER exchange with the HSS entity is necessary. It also generates the information needed to generate tax tickets.
- The S-CSCF entity is responsible for authenticating of the user and providing the operational services. It provides routing information, billing information, and maintains session status by controlling a timer (time), querying the HSS to verify user rights to a service. At the same time, the S-CSCF stores the subscriber's position in the HSS in the network and tells the HSS its address so that an entity seeking to reach the subscriber knows which S-CSCF to contact.
- The HSS entity is a database that stores the data specific to each user. The main recorded data include among others
    a. SVP Table (SerVice Profile): Contains the service profile that is associated with each user.
    b. The public and private identities of the user (IMSU, IMPU and IMPI) along with the other authentication and authorization information.
    c. Charging information table (Chginfo): Contains pricing information.
    d. Application Server Table (APSVR): Contains the address of the AS to contact if the trigger points are compliant.
    e. Initial Filter Criteria table (IFC); which contains the initial filtering information, characterizing a certain service for the user. This table maps a point trigger that associates a point service set with an application server.

- The RabbitMQ: is a database server that interacts with the I-CSCF, S-CSCF and HSS entities to complete the authentication process before a user is allowed to log on.

### B. Software Defined Networking (SDN)

Software Defined Networking (SDN) is a new network paradigm that decouples the data plane and control plane. The SDN is based on the principle of providing a programming interface (API) for data packet transfer devices such as switches and routers, allowing the programming of network elements and the entire network, creating an architecture dynamic, flexible, automated and manageable [14] [15]. The main idea of the SDN is to deport the intelligence of the network in a fundamental element called controller. This controller manages elements of the data plane through APIs [16]. There are several types of SDN controllers. The most used in research and education are POX, OpenDaylight, FloodLight and Ryu [16] [17]. The SDN architecture is structured in three layers that interact with each other via APIs. We have the infrastructure layer or transmission layer. Its main role is to transmit data, monitor local information and collect statistics. We also have the control layer. This layer contains the controller (s) that uses Southern APIs to interact with the infrastructure layer. It uses the north APIs to communicate with the last layer called application layer. The different controllers use East / West APIs to interact with each other [18]. The SDN architecture is as shown in Figure 2:

SDN use the Application Interface (API) for communication between different entities. The development of the APIs allowed the programming of the SDN networks. Among these APIs there is Openflow. Openflow is an open protocol that allows the separation of control plans and data. It acts between the applications and the network. Thus, it allows to have a direct access to the data plane. It is also responsible for managing network flows such as routing, marking, filtering, etc. The OpenFlow protocol is an important, even fundamental element in the design of SDN solutions. It allows equipment and controllers to interact [11]. This approach as presented by the NFB involves deploying routers and switches compatible with OpenFlow protocol and SDN controllers to manage network flows on these devices.

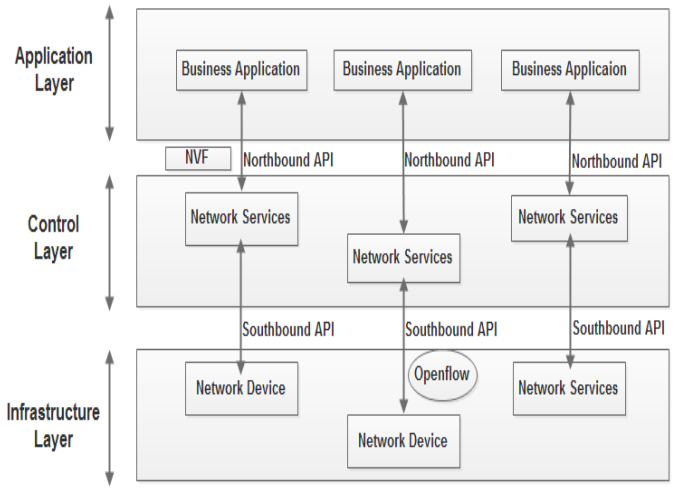



Fig. 2. SDN Architecture.

An SDN network makes it possible to reduce CAPEX and OPEX [10] and can be emulated thanks to a tool like mininet.

For academic purposes, the simplest approach to most features is to use Mininet. It can be easily distributed and deployed due to its packaging in a virtual machine. It also supports multiple controllers and others can be easily added. In addition, it does not have excessive material requirements [16] [18]. For these reasons, we have chosen it as a main tool for our integration of SDN into other platforms. This is what justifies the choice of this simulator.

### C. Kurento Media Server (KMS)

Traditional WebRTC applications are standardized so that browsers can communicate directly without the mediation of third-party infrastructures. This is sufficient to provide basic multimedia services, but features such as group communications, stream recording, streaming, or transcoding are difficult to implement. For this reason, the most interesting applications require the use of a multimedia server.

Kurento is an open source WebRTC multimedia server. It makes possible the creation of media processing applications based on the concept of pipelines. Media pipelines are created by interconnect modules called Media Elements. Each Media Element provides a specific feature. KMS contains Media Elements capable of recording and mixing streams, computer vision, etc. Kurento Media Server offers the capabilities of creating media pipelines through a simple JSON-RPC-based network protocol. However, to further simplify developer work, a client API that implements this protocol and directly leverages Media Elements and pipelines is provided. Currently, the Java and JavaScript client API is ready for developers [19]. Taking into account the integrated modules, the Kurento Toolkit is detailed in Figure 3.

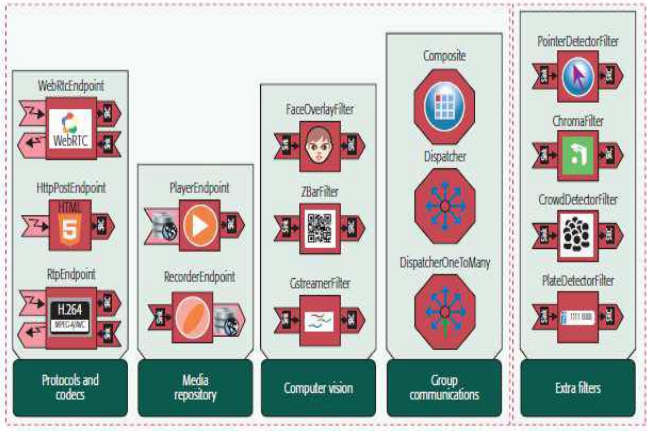



Fig. 3. Kurento Media Elements Toolbox.

### D. Web of Things( WoT)

The Web of Things (WoT) is a specialization of te Internet of Things (IoT). On the one hand, it provides an abstraction of the connectivity of smart objects. On the other hand, WoT adds a standard web standards-based application layer to simplify the creation of IoT applications. In IoT, the communication protocols are multiple (MQTT, CoAP, ARMQP ...), which creates groups of users.

The main interests of using WoT instead of IoT are the simplicity of development using APIs, standardization and simple coupling. The idea is that all intelligent objects can communicate using a Web language through an API. This API can be present in the intelligent object itself or in an intermediary that can act on behalf of the intelligent object [20].

In the field of research, each of these technologies is currently gaining momentum by proposing innovative solutions. Thus, each technology has developed independently from the others but thanks to IP convergence several technologies are integrating in order to diversify offers and improve the quality of service. Today, many researchers are taking advantage of IP to offer solutions that combine two of these technologies. It is in this case that the work of [21] deals with issues and proposes possible solutions for the Internet of the future through an IoT-IMS communication platform. By coupling SDN and IMS, the authors of [22] set up architecture for a QoS-compliant application in the virtualized IMS (V-IMS) environment. In the same vein [23] is putting in place an integrated ICT service creation platform, the IMS-based ICT service infrastructure is being built and some prototype services are being implemented. Added to this is a converged IMS and SDN ICT service infrastructure that is also presented. The possible use cases of the SDN on the creation of new services and the control of IMS service are also proposed. For some time, the IMS-IoT-SDN integration has been proposed as architecture to give access to information to several types of terminals. In this sense, the authors of [11] have set up a BETaaS M2M development and execution platform and an SDN-based IMS architecture to provide an open environment for the development and delivery of IoT services. All these works have the merit of proposing innovative solutions even revolutionary and constitute an important contribution in the field. However, these solutions can be improved by the automation techniques provided by the SDN. Our article, inspired by this work, proposes the integration of an SDN network into an IMS / IoT platform in order to facilitate access to users and to diversify user terminals. In addition, we also offer the automation of the control layer (IMS / KMS platform) and the transport layer of the functional architecture.

## III. Proposed Architecture

The functional architecture of the proposed solution is the integration of SDN technology into an IMS-IoT platform. This integration makes it easier for users to access the IMS-IoT network but also to diversify terminals (smartphone, tablet, computer, etc.). Figure 4 represents the architecture that we propose:

### A. Access Layer

The access layer consists essentially of access network, terminals, gateway and IoT devices. IMS provides a standardized framework for creating and deploying services across all IP networks, regardless of access technology, supporting mobile and wireless telecommunications environments. In the IoT device, we have an IoT gateway to interact with the IMS / KMS platform. This gateway is the equipment responsible for retrieving the data collected by the IoT devices and transmitting them to the access network so that the terminals can access them. The gateway plays an important role in the integration of SDN and IoT technologies. It can support several types of devices through a pilot architecture [11]. It can also be characterized by its representation of resources in different formats to facilitate its integration with other web applications [12].

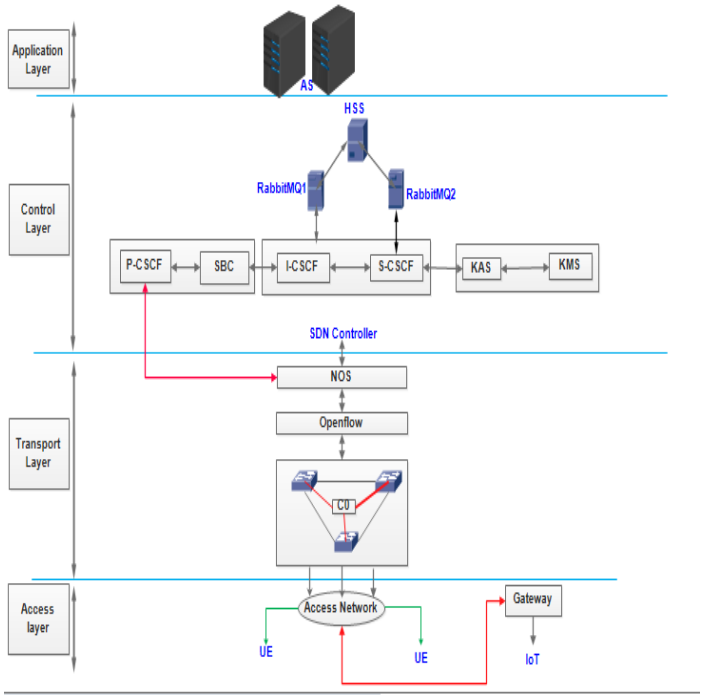



Fig. 4. Proposed Functional Architecture.

### B. Transport Layer

The transport layer consists of the SDN part. It is composed of NOS which represents the operating system, different Switch, a controller responsible for controlling the equipment. The SDN concept decouples the control plane and data plane by concentrating data transfer decisions into a central controller (software). A common standard protocol that defines the communication between the control and the data transfer plan is Openflow. In the current architecture, the transport layer corresponds directly to the data transport plan of the SDN network. This plan consists of Openflow switches. Each Openflow switch contains one or more flow tables and an abstraction layer that communicates with the controller.

### C. Control Layer

The control layer consists of a condensed IMS and KMS entity. The IMS part consists of P-CSCF, SBC, I-CSCF, S-CSCF, rabbitmq servers and HSS. In the KMS part we have the KMS and KAS entities. These two parts allow us to manage the IMS and WebRTC services for each one. In the SDN scenario, the data transmission decisions and the higher level control functions are executed by a logically centralized POX software controller. The network configuration is performed on a simplified network abstraction and the data flow configurations are communicated to the data routing devices via the Openflow protocol. The control plan is implemented using the SDN (NOS) network operating system

and a set of software modules that perform the various network control functions. The NOS exposes open APIs (North APIs) to enable the development of various applications that can interact and provide better network management. In such a scenario, the basic IMS / KMS elements can be implemented as modules for communication with the NOS via the northern APIs.

### D. Application Layer

The IMS / KMS control layer uses multiple interfaces to communicate with service layer that provide the capability to deliver applications implemented by Application Servers (AS). The latter communicate with the S-CSCF using the SIP protocol to perform the session control service. They also communicate with the HSS via the DIAMETER protocol to retrieve or update the subscriber's profile information. AS are the most common type of application servers used in IMS scenarios. They support SIP applications and respond to requests with SIP messages to enable subsequent processing of user requests to S-CSCF. SIP applications can include presence service or sessions involving multimedia elements such as voice, video, chat, games, and so on.

## IV. Results and Discussion

To illustrate the functionality of the proposed architecture and the automation of the control layer and the transport layer we can refer to the following figures 5, 6 and 7. These figures show all the tests we have done. To do these tests we used a mininet simulator. Mininet is an open source network emulator that can run virtual hosts, switches, the controller (s) and virtual machine, or deployed on real hardware. For our case, mininet has enabled us to emulate the different entities of the IMS / KMS platform thanks to the automation of the two intermediate layers (control and transport). Thus, Figures 5 and 6 show the launch of the program via the python automation.py command. This program supports the automation of the control layer and the transport layer. The execution of such a program is as follows: the terminal creates the network and adds the controller whose role is to control the various equipments network Then there is the creation of hosts, switches and links between equipments. It is only after this step that the system starts the network devices (controller, switches, CLI, etc.) just after the computers are set up. If the process proceeds correctly the terminal displays mininet.

Figure 7 highlights the communication tests of the automation of the IMS / KMS platform. By doing a global pingall test we notice an exchange of flows between the different entities.

```
root@sambademo-VirtualBox:~/mininet# python automatisation.py
*** Creating network
*** Adding controller
Unable to contact the remote controller at 127.0.0.1:6633
*** Adding hosts:
HSS ICSCF KAS KMS PCSCF Rabit1 Rabit2 SBC SCSCF
*** Adding switches:
s1 s2 s3
*** Adding links:
(10.00Mbit 10ms delay 0% loss) (10.00Mbit 10ms delay 0% loss)
```

Fig. 5. First Part of the Program Launch.

```
*** Starting 3 switches
s1 s2 s3 ...(10.00Mbit 10ms delay 0% loss) (10.00Mbit 10ms dela
y 0% loss) (10.00Mbit 10ms delay 0% loss) (10.00Mbit 10ms delay
 0% loss) (10.00Mbit 10ms delay 0% loss) (10.00Mbit 10ms delay
0% loss) (10.00Mbit 10ms delay 0% loss)
la connexion des hotes
HSS HSS-eth0:Rabit1-eth1 HSS-eth1:Rabit2-eth1
ICSCF ICSCF-eth0:SBC-eth1 ICSCF-eth1:SCSCF-eth0 ICSCF-eth2:Rabi
t1-eth0
KAS KAS-eth0:SCSCF-eth1 KAS-eth1:KMS-eth0
KMS KMS-eth0:KAS-eth1
PCSCF PCSCF-eth0:s1-eth3 PCSCF-eth1:SBC-eth0
Rabit1 Rabit1-eth0:ICSCF-eth2 Rabit1-eth1:HSS-eth0
Rabit2 Rabit2-eth0:SCSCF-eth2 Rabit2-eth1:HSS-eth1
SBC SBC-eth0:PCSCF-eth1 SBC-eth1:ICSCF-eth0
SCSCF SCSCF-eth0:ICSCF-eth1 SCSCF-eth1:KAS-eth0 SCSCF-eth2:Rabi
t2-eth0
*** Starting CLI:
mininet>
```

Fig. 6. Launch of the Program Continuation and End.

```
mininet> pingall
*** Ping: testing ping reachability
HSS -> ICSCF KAS KMS PCSCF Rabit1 Rabit2 SBC SCSCF
ICSCF -> HSS KAS KMS PCSCF Rabit1 Rabit2 SBC SCSCF
KAS -> HSS ICSCF KMS PCSCF Rabit1 Rabit2 SBC SCSCF
KMS -> HSS ICSCF KAS PCSCF Rabit1 Rabit2 SBC SCSCF
PCSCF -> HSS ICSCF KAS KMS Rabit1 Rabit2 SBC SCSCF
Rabit1 -> HSS ICSCF KAS KMS PCSCF Rabit2 SBC SCSCF
Rabit2 -> HSS ICSCF KAS KMS PCSCF Rabit1 SBC SCSCF
SBC -> HSS ICSCF KAS KMS PCSCF Rabit1 Rabit2 SCSCF
SCSCF -> HSS ICSCF KAS KMS PCSCF Rabit1 Rabit2 SBC
*** Results: 0% dropped (72/72 received)
mininet>
```

Fig. 7. IMS / KMS Platform Automation Communication Test.

Based on these tests we can say that this proposal is an important contribution to the management of data collected by IoT devices. In that it offers users a lot of ease and diversifies the terminals. It also automates the IMS-KMS platform. This automation is more than important. Indeed, when a component relies on manual processes then it can become a bottleneck. Utilizing administrative tools and data center switches that enable high levels of automation allows IT departments to eliminate such bottlenecks. At the same time, network automation helps to increase agility and reliability, accelerate deployments, and improve the performance of mission-critical business applications.

## V. Conclusion

Our paper proposes a functional architecture for an IMS-IoT platform, the integration of the SDN and the automation of the transport layer and the control layer of this architecture in order to facilitate the management, the access to the users, to diversify the user terminals etc. The technologies used to do this work are described as well as the essential components of the architecture without forgetting the different challenges posed by these technologies.

This architecture is for us the center of future work that can be focused on the study and load balancing of the various entities and the virtualization of some network functions.

## Acknowledgment

The authors would like to thank the LIRT Laboratory Manager for the help he provided during the research.

## References

[1] A. Hakiri, P. Berthou, A. Gokhale, and S. Abdellatif, “Publish/subcribe-enabled software defined networking for efficient and scalable IoT communications”, IEEE Journal on Communications Magazine, vol. 53, n°9, 2015, pp. 48-54.

[2] N. Omnes, M. Bouillon, G. Fromentaux, and O. Le Grand, “ A programmable and virtualized network & it infrastructure for the internet of things: How can nfv sdn help for facing the upcoming challenges”, Proc. IEEE ICIN, 2015, pp. 64-69

[3] C. Vallati, E. Mingozzi, G. Tanganelli, N. Buonaccorsi, N. Valdambrini, N. Zonidis, B. Martnez, A. Mamelli, D. Sommacampagna, B.

Anggorojati, S. Kyriazakos, N. Prasad, F. Nieto, and O. Rodriguez, "Betaas: A platform for development and execution of machine-tomachine applications in the internet of things," WPC, pp. 1–21, 2015.

[4] I. Stojmenovic and S. Wen, "The Fog computing paradigm: Scenarios and security issues," *Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2014, pp. 1–8.

[5] "IMS Tutorial | IP Multimedia Subsystem | Radio-Electronics.Com." [Online]. Available: http://www.radioelectronics.com/info/telecommunications_networks/ims-ip-multimediasubsystem/tutorial-basics.php. [Accessed: 07-Jul-2015].

[6] Katov, Anton Nikolaev; Anggorojati, Bayu Kyriazakos, Sofoklis; Mihovska, Albena Dimitrova; Prasad, Neeli R. 18th International Symposium on Wireless Personal Multimedia Communications, 2015 (WPMC)

[7] H.Liao, K. Chang, C.Chen, J. Chen, H.Chao, T.Wu, "IoTIMS Communication Platform for Future Internet Implementation," The 11th Conference on Information Technology and Applications in Outlying Islands, Taiwan, May 25-26, 2012.

[8] McKeown, N .; Anderson, T .; Balakrishnan, H .; Parulkar, G .; Peterson, L .; Rexford, J .; Shenker, S .; Turner, J. OpenFlow: Favoriser l'innovation dans les réseaux de campus. *ACM SIGCOMM Comput. Commun. Rev.* 2008, *38*, 69-74. [ Google Scholar ] [ CrossRef ]

[9] OpenFlow Switch Specification v1.1.0. Disponible en ligne: https://www.opennetworking.org/wp-content/uploads/2013/04/openflow-spec- v1.0.0.pdf (consulté le 2 avril 2018).

[10] CAPEX OPEX Available: http://onosproject.org/wp-content/uploads/2014/11/Whitepaper-ONOS-final.pdf

[11] C. Tranoris, S. Denazis, N. Mouratidis, P. Dowling, and J. Tynan, "Integrating OpenFlow in IMS Networks and Enabling for Future Internet Researchand Experimentation," in *The Future Internet*, A. Galis and A. Gavras, Eds. Springer Berlin Heidelberg, 2013, pp. 77–88.

[12] Imène ELLOUM I, 20 1 2. "Management of mobility between access networks and Quality of Service in an NGN / IMS approach. " Thèse doctora t : Unive rsity of Ca rtha ge (Tu nisie )

[13] N. KAG-TEUBE, Y. P. K. DODOAGNEN, S. OUYA, K. GUEYE, '' Proposed solution for improving the reliability of HSS data by integrating a queue manager'', international conference on advanced communicationtechnology (ICACT), february 11-14 2018

[14] Faris Keti, "Computer networks virtualization cost effective and distance learning achievement", Journal of University of Duhok, Vol. 20,No.1(Pure and Eng. Sciences), Pp 90-98, 2017 eISSN: 2521-4861 & pISSN: 1812-7568

[15] Guohui Wang, T. S. Eugene Ng, Anees Shaikh, "Programming your network at run-time for big dataapplications", HotSDN'12, Helsinki, Finland, August 13, 2012.

[16] Filip Holik, Josef Horalek, '' Implementing SDN into Computer Network Lessons'', Journal of Telecommunication, Electronic and Computer Engineering, Vol. 9 No. 1-3

[17] Enns E et al (2011). "Network Configuration Protocol (NETCONF), RFC 6241". Retrieved from: https://tools.ietf.org/html/rfc6241, on 8 Jan 2016

[18] Rogério Leão Santos de Oliveira, Using Mininet for Emulation and Prototyping Software-Defined Networks, 978-1-4799-4340-1/14/$31.00 ©2014 IEEE

[19] B. Garcia, L. Lopez-Fernandez, M. Gallago and F. Gortazar, « Kurento : The Swiss Army Knife of WebRTC Media Servers », in IEEE Communications Standards Magazine, vol. 1, no. 2, pp. 44-51, 2017

[20] D. Guinard and V. Trifa, Building the web of Things. Manning Publications Co, 2016

[21] C.-Y. Chen, H.-C. Chao, T.-Y. Wu, C.-I. Fan, J.-L. Chen, Y.-S. Chen and J.-M. Hsu, "IoT-IMS Communication Platform for Future Internet:," International Journal of Adaptive, Resilient and Autonomic Systems, vol. 2, no. 4, pp. 74–94, 34 2011.

[22] HM Yoon, Z Liu, WS Yang, JH Kim, JO Lee, '' Architecture for QoS-enabled Application Service Deployment in Virtualized Environment'', Advanced Science and Technology Letters Vol.142 (SIT 2016)

[23] Chung-Shih Tang, Chin-Ywu Twu, Jen-Hong Ju, and Ying-Dian Tsou, '' Collaboration of IMS and SDN to Enable New ICT Service Creation'', Asia-Pacific Network Operation and Management Symposium (APNOMS) 2014

# A Novel Approach for Background Subtraction using Generalized Rayleigh Distribution

Pavan Kumar Tadiparthi[1]
Associate Professor
Department of Information Technology,
MVGR College of Engineering, India

Srinivas Yarramalle[2]
Professor
Department of Information Technology,
GIT, GITAM University, India

Nagesh Vadaparthi[3]
Professor
Department of Information Technology,
MVGR College of Engineering, India

***Abstract*—Identification of the foreground objects in dynamic scenario video images is an exigent task, when compared to static scenes. In contrast to motionless images, video sequences offer more information concerning how items and circumstances change over time. Pixel based comparisons are carried out to categorize the foreground and the background based on frame difference methodology. In order to have more precise object identification, the threshold value is made static during both the cases, to improve the recognition accuracy, adaptive threshold values are estimated for both the methods. The current article also highlights a methodology using Generalized Rayleigh Distribution (GRD). Experimentation is conducted using benchmark video images and the derived outputs are evaluated using a quantitate approach.**

***Keywords*—Background subtraction; segmentation; generalized rayleigh distribution (GRD); quantitative evaluation; image analysis***

## I. Introduction

The most imperative characteristic of an intelligent vision based inspection system is background subtraction, which is considered to be a primitive step for object recognition and tracking. Typically, pixel by pixel comparison is practiced for either detection or tracking with a predefined object dataset. However, this procedure of searching and comparing against each pixel requires a huge computational time and as an improvement to this approach, background subtraction methods are coined for the optimization of both search and computational time. In many of the Human computer interactive systems, background subtraction is considered during the pre-processing procedure to optimize the cost. As such, background subtraction has become a significant method and has deeply penetrated with strong roots in the area of computer vision. Since background modelling considerably influences the performance on the whole vision system, it is imperative to make use of an excellent background subtraction methodology. However, most of the background modelling techniques need to combat the challenges due to dynamic or non-static backgrounds, unexpected or steady lighting changes; motion in the object and shade, Background modelling methods should intelligently overcome such issues. To overcome these challenges, many models are presented in the literature [1-11], [13], [15], [16], [17], [18].

## II. Literature Review

Haiying et al [1] has proposed a modified Gaussian mixture model based on the optimization of the GMM and Combining the spatial information. H.Zhou et al [2] have proposed a foreground detection methodology in which the authors have tried to improvise the codebook. Viswanth et al [3] suggested and modeled an approach using non parametric background modeling. In this approach, a single Spatio Temporal Gaussian is used for modeling the back ground pixels. However, this methodology fails as the adequate features are not obtainable from the section. Yuhan.L et al [4] proposed a robust back ground subtraction methodology based on the adaptive dictionary strategy and penalized splitting approach. Lu yang et al [5] considered a pixel for modeling the background information in case of complex scenes. The consideration pixel manipulates the distance between the pixels and it was used for updating the back ground model as a substitute of local descriptors. Chen et al [6] proposed a model using varying learning rate and also adaptively selecting the number of Gaussians. This model performs better in particular cases of extraction of dynamic background information and sudden illumination variations. However, this model cannot handle strong dynamic back ground and also fails in case of capturing the paused objects. Chien et al [7] proposed a foreground object detection method by using a threshold value. In this article, the authors have assumed that the camera considered for capturing the videos are tolerant to noise and posses a zero –mean Gaussian distribution. But this assumption has affected the selection of the threshold. Lui et al [8] proposed an approach based on the binary descriptors. In this article, the authors have generated the back ground instances using binary descriptors. The developed model has proved to be robust against lighting changes and dynamic back ground and is tested against the environmental changes.

Haung et al [10] proposed a method for back ground modeling based on binary descriptors. However, this method can reduce the effect of noise and capable for the extraction of rough shaped images from the foreground objects. Hedayathi et al [11] proposed a statistical frame work for back ground subtraction, which acquired better performance in terms of segmentation.

Stephan Kopf's et al [15] has proposed a model for automatic scaling and cropping. The main limitation with reference to this approach is that, this methodology needs pre-identification of certain parameters which helps during the cropping of the selected regions. In the motion detection, the objects can be recognized only when the object is in moving condition. Therefore, appropriate reorganization of objects can

be well planned, only if the back ground information from these motion images is subtracted. This methodology is very much useful for recognization of images acquired from surveillance cameras. Stauffer and Grim Son [18] have presented an approach using pixel wise operation for the identification of the back ground images. The main restraint of this model is that, extracting the back ground pixel information from the static camera is relatively difficult and therefore, this methodology leaves an unsolved issue about the problem with respect to the images acquired from the static cameras. Tao Mei and Xian et al [16] have presented a model for background subtraction, where image mosiacing is considered. The limitation of this approach is that, mosaicing of background information is highly impossible. AL-Najdawi et al [13] have utilized the kalman filter for the purpose of seeking optimal estimation in tracking. D.Farin, P.de et al [17] have proposed a model for the extraction of individual frames. The limitation of this model is with respect to identification of back ground pixels from the frames from the static cameras. D. Hari Hara Santosh et al [9] have utilized the Gaussian Mixture models for the effective identification of the foreground information. In this article, the authors have addressed the concepts of object tracking using Blob Analysis. However, this methodology has its limitations while dealing with sudden, drastic lighting changes.

To overcome these challenges many techniques were therefore planned, in particular by considering mostly the statistical frame works, with the very criteria that, the efficiency of statistical models will be relatively high and helps towards better identification of the object pixels more appropriately. Based on this approach, many models have been further developed using the mixture models, such as Gaussian Mixture models and models based on the Rayleigh distribution.

The Rayleigh distribution cannot model the natural images having specals, impulsive behaviour leading to heavy tails. Therefore in this article, an attempt is made to overcome the limitation highlighted and proposes a methodology by considering a Generalized Rayleigh distribution (GRD). The main advantage of using the GRD is that, it is more accurate in the reverberation regions.

The rest of the paper is organized as in Section III where the Generalized Rayleigh distribution is considered. Section IV presents the details about data sets and Section V emphasizes the methodology. Section VI the experimentation carried out is highlighted. The performance evaluation and the results derived are highlighted in Section VII. The Section VIII concludes the paper and describes the future scope.

## III. Generalized Rayleigh Distribution

The early roots of considering Rayleigh Distribution for enhancing the Background Subtraction is considered by Michael Unger et al [19]. In this article, the authors have considered the model for Background Subtraction for unconstrained images acquired from cameras that are in motion, as they have low resolutions and less distortion.

However, with the availability of sophisticated technological cameras, capturing the images with different resolutions is highly difficult. Hence, to overcome this disadvantage, the work presented by Michael Unger et al [19] has been extended by considering Generalized Rayleigh Distribution. The main advantage of proposed method is that, the generalization process allows estimating the back ground images in particular situations where the information is suppressed. Beyond the above mentioned advantage, the model also includes,

*1)* It has the advantage over the other distribution, where the degree of freedom can be easily obtained using the maximum value which is generally unique.

*2)* These distributions are maximum when it approaches towards Y-axis.

*3)* The back ground information can easily be interpreted using the Rayleigh distribution with the maximum value.

Another limitation with respect to the Rayleigh distribution is that it can't handle specals having heavy tails. Therefore, the generalizations of the models help to overcome these limitations. Hence, as a contributing factor in this article, we propose a model based on generalized distribution to overcome the limitations of Rayleigh distribution. This distribution has an advantage to handle both high tail and low tail impulsive noises.

A continuous random variable $x$ is said to follow a Rayleigh distribution if its probability density function (pdf) is given by

$$f(x) = bxe^{-\frac{bx^2}{2}}, x \geq 0 \quad (1)$$

$= 0$ , otherwise

The Cumulative distribution function (cdf) is given by the formula

$$F(x) = 1 - e^{-\frac{bx^2}{2}}, x \geq 0 \quad (2)$$

$= 0$ , otherwise

Where $b = \frac{1}{\sigma 2}$, where $\sigma$ , is the scale parameter.

## IV. Data Set

To exhibit the proposed work, a Bench mark data set of Video images from www.changedetection.net [12] has been considered for experimentation. The dataset consists of 6 different video categories with a total of 31 videos comprising of 90,000 frames. These videos are mainly based on Baseline, Camera Jitter, Dynamic Background, Intermittent Object Motion, Shadow and Thermal.

## V. Methodology

*1) Post processing:* In order to extract the background image, each of the inputs has to be first preprocessed by considering the pixels having least deviation. The least deviation pixels have to be considered as background pixels. However, the main limitation in choosing the background pixels is that, in particular situation, the foreground and background shall share similar information with respect to color, size and orientation. Therefore, whenever we need to

estimate the background images, lighting conditions play a vital role. This lighting condition, if interpreted exactly helps to model the background images. In this article, the usage of Generalized Rayleigh Distribution helps to overcome this disadvantage because of its ability to handle low illumination images.

*2) Back ground subtraction*: In background subtraction technique, motion objects were identified by deducting the present image from the background image. The initial frame of the video progression was taken as reference image for background frame. The present frame will be deducted from the considered background frame. The background pixel is decided on the basis of the resultant difference, i.e. if the output of the subtraction reference pixel value is greater than the reference pixel value, then it is considered to be a background pixel, else it is considered as a foreground pixel.

*3) Frame difference method*: Here, we estimate the difference in values between two consecutive frames, 't ' and 't-1'. If the resultant value is better, then the value is taken to be the threshold value, and the pixel will be treated as background pixel.

The estimated threshold values, from both the cases are considered and are given as input to the model Generalized Rayleigh Distribution proposed in section III of the article.

The probability density functions (pdf) against each of the intensity values are given as input to the model and the respective values are estimated. These values which are below the threshold value are considered as background information else they are considered as foreground information.

*4) Fusion technique:* In this article, two methods for estimating the background pixel, viz., background subtraction method and frame difference method are highlighted. However, each of these methods have their own limitations, i.e., if we consider the background subtraction method, the boundaries and contour will be intact, however the output result may be affected due to the noise parameter, in contrary, in frame difference method, result will have minimal impact due to noise, but in this case, the complete information regarding the boundaries and contours may be a bit influenced. To overcome these limitations, in this article, we have considered the fusion concept using "AND" operation.

*5) Adaptive background subtraction*: The best possible threshold value can be estimated using the adaptive threshold technique and it is estimated using the formula given by

$$F(x,y)=C(x,y)-R(x,y) \quad (3)$$

$$F(x, y) = 1, \text{ if } F(x, y) \geq T$$

(where T is the Threshold value, ( using the methodology proposed by N. Otsu[21] ) and zero otherwise. Here C(x, y) denotes the current frame, and R(x, y) represents the considered reference background image, F(x, y) denotes, the deviation between the present frame and the reference frame.

*6) Frame differencing method and Adaptive frame difference fused methods*: Here the optimal threshold values are estimated in line with the heuristics given by W. Jun-Qin[22]. Adaptive background subtraction the difference between video frame at time t and the frame at time t-1. The optimal threshold is thereby estimated. In case of Adaptive frame difference fused method, the choice of the adaptive threshold value is based on the difference value obtained by subtracting the background reference frame from the current frame and then these values are fused to get a unique threshold value.

## VI. Experimentation

Blob analysis is considered for the effective identification of the background and foreground regions. Each pixel value is extracted based on threshold values obtained from background subtraction method and frame difference methods and the corresponding pixels are categorized into either background or foreground. In general, pixels with minimum threshold values will be mostly considered as background pixels. The pixels with high threshold values are given as inputs to the Generalized Rayleigh Distribution (GRD) presented in section III. Basing on the log likelihood estimation of the pixels, each pixel is categorized either as a back ground pixel or a foreground pixel. The experimentation is carried out in matlab environment and the results obtained are shown below. In this article, we have experimented with numerous ways of estimating the background pixel; namely Fusion method, adaptive Background subtraction, Adaptive frame difference, Adaptive frame difference fused with Adaptive background subtraction. The significance of each of these methods are presented in section V. The results were also compared with the model based on Gaussian Mixture Model.

## VII. Performance Evaluation and Experimental Results

In order to validate the model, we have considered the performance for quantitative analysis are metrics Precision, Recall, Accuracy, F- Score, MSE, RMSE, FNR, FPR, PSNR[20].In order to validate the model, we have performed the experimentation with different frames 257,863,1005,1954 respectively. The formulas for the identification of Precision, Recall, Accuracy, F-Score. Recall is expressed in terms of the number of allocated foreground pixels to that of actual foreground pixels; and the evaluation outcome of this metric showcase, the exact number of true foreground pixels that are classified as foreground pixels. Precision is defined in terms of the number of exact foreground pixels against the allocated foreground pixels; it signifies the exactness of the pixels that were classified as true foreground pixels against the allocated foreground pixels. The performance of the model can be justified by the value of calculated precision, if it is high, it signifies high performance. On the other hand, if method allocates the majority of the pixels to background, the output precision value may be high, but proportionally, the value of recall declines. To identify the trade-off between recall and precision, F-measure is also considered. The other performance metrics considered include; Mean Squared Error (MSE).Root Mean Square Error (RMSE), False Negative Rate (FNR) False Positive Rate (FPR) and Peak Signal to Noise Ratio (PSNR) [20]

The formulas for the calculation of the above metrics are given by

$$\text{Precision} = TP / (TP+FP) \quad (4)$$

$$\text{Recall} = TP / (TP + FN) \quad (5)$$

$$\text{Accuracy} = TP+ TN / (TP+TN+FP+FN) \quad (6)$$

$$\text{F-Score} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{recall}) \quad (7)$$

$$MSE = FP+FN / M *N \quad (8)$$

$$RMSE = \sqrt{(MSE)} \quad (9)$$

$$FNR=FN/(TP+FN) \quad (10)$$

$$FPR=FP/(FP+TN) \quad (11)$$

$$PSNR=10\log_{10} (R^2/MSE) \quad (12)$$

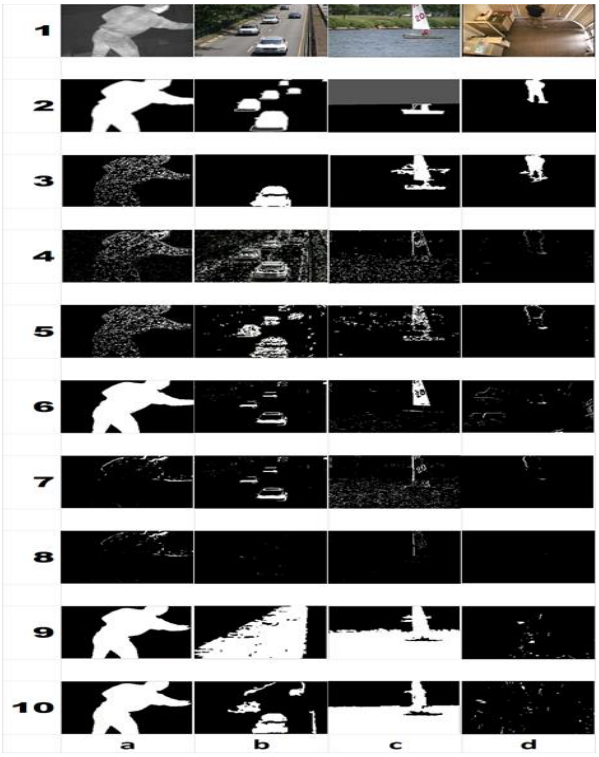



Fig. 1. Foreground Detection Results of Thermal, Baseline, Dynamic Background, Shadow from the CDnet2014 Dataset.

Where, TP-the number of Foreground pixels classified as foreground, FN-the number of Foreground pixels classified as background, FP- the number of pixels of background pixels classified as foreground, TN- the number of back ground pixels classified as background.

Experimentation is performed with the developed model, by considering the data set presented in the section-IV. The results derived are presented in the following Fiig-1

1. Original frame. 2. Ground truth .3.Back ground subtraction.4.frame difference. 5 .Back ground subtraction and frame difference .6.Adaptive back subtraction. 7. Adaptive frame difference. 8. Adaptive back ground subtraction and adaptive frame difference. 9. GMM. 10. GRD.

We evaluated the different background modelling methods discussed in sectionV. The scenarios used to evaluate different methods thermal, baseline, dynamic background, shadow. There are many videos for each scenario. We selected one typical frame work from each video. Fig 1 a-d are selected from four categories in the CDNet 2014 dataset. Fig .1 (1) show the original frame of the video and Fig 1. (2) are the results of the ground truth data .Fig 1.(3)-(10)are the foreground detection results of the state of the art background modelling methods.Tabel I-IV presents nine performance evaluation metrics of the eight back ground modeling methods in the CDNet 2014 dataset.

The performance of the different methods can be confirmed by the recall, precision,F-Score and other metrics. For each evaluation metric, we give the results of the back ground modeling methods in different Scenes via Figs.2 to 37.

**Therma**l: As shown on Fig.1 (a), their results are closer to the ground truth data. Fig .5 indicates the F-Score are greater than 50%

**Base line**: These videos contain a noise free static back ground Fig.1 (b) shows foreground detection results of every method. The proposed method (GRD) successfully detected the foreground object. It can also observed that the F-Score of each method in Fig.14 is very high, greater than 63%.

**Dynamic background**: As shown in Fig 1 (c), the proposed method (GRD) is more effective than the other methods when dealing with dynamic backgrounds. Fig .21 shows that their Recall is very high when compared with other methods.

**Shadows**: The methods differ in the capability of classifying shadow pixels as back grounds. As shown in Fig 1. (d) shows foreground detection results of every method. Fig .32 shows that their F -score is very high when compared with other methods.

TABLE I. EVALUATION METRICS OF DIFFERENT METHODS ON THREMAL VIDEO FROM CD NET DATASET

| Metrics\ Methods | BACKGROUND SUBTRACTION | FRAME DIFFERENCE | BACKGROUND SUBTRACTION AND FRAME DIFFERENCE | ADAPTIVE BACKGROUND SUBTRACTION | ADAPTIVE FRAME DIFFERENCE | ADAPTIVE BACKGROUND SUBTRACTION AND ADAPTIVE FRAME DIFFERENCE | GMM | GRD |
|---|---|---|---|---|---|---|---|---|
| PRECISION | 0.0625 | 0.0226 | 0.1225 | 0.0707 | 0.105 | 0.0473 | 0.0511 | 0.0362 |
| RECALL | 0.041 | 0.1287 | 0.0829 | 0.0923 | 0.076 | 0.0615 | 0.0643 | 0.1327 |
| ACCURACY | 0.97 | 0.8865 | 0.9719 | 0.9628 | 0.9724 | 0.9605 | 0.9607 | 0.9248 |
| F-SCORE | 0.0495 | 0.0385 | 0.0989 | 0.0801 | 0.0882 | 0.0535 | 0.057 | 0.0569 |
| MSE | 0.0166 | 0.0529 | 0.0158 | 0.0205 | 0.0153 | 0.0221 | 0.0213 | 0.0387 |
| RMSE | 0.129 | 0.23 | 0.1257 | 0.1431 | 0.1239 | 0.1488 | 0.1458 | 0.1967 |
| FPR | 0.0119 | 0.0999 | 0.0113 | 0.0217 | 0.0116 | 0.0229 | 0.0225 | 0.0615 |
| FNR | 0.959 | 0.8713 | 0.9171 | 0.9077 | 0.924 | 0.9385 | 0.9357 | 0.8673 |
| PSNR | 65.955 | 60.9316 | 66.1809 | 65.0508 | 66.3061 | 64.7143 | 64.8912 | 62.2877 |

TABLE II. EVALUATION METRICS OF DIFFERENT METHODS ON BASELINE VIDEO FROM CD NET DATASET

| Metrics\ Methods | BACKGROUND SUBTRACTION | FRAME DIFFERENCE | BACKGROUND SUBTRACTION AND FRAME DIFFERENCE | ADAPTIVE BACKGROUND SUBTRACTION | ADAPTIVE FRAME DIFFERENCE | ADAPTIVE BACKGROUND SUBTRACTION AND ADAPTIVE FRAME DIFFERENCE | GMM | GRD |
|---|---|---|---|---|---|---|---|---|
| PRECISION | 0.0339 | 0.0171 | 0.0366 | 0.04 | 0.0411 | 0.0234 | 0.0227 | 0.0609 |
| RECALL | 0.0568 | 0.2112 | 0.2123 | 0.1059 | 0.1066 | 0.0253 | 0.094 | 0.0681 |
| ACCURACY | 0.9585 | 0.7764 | 0.907 | 0.9441 | 0.9482 | 0.9662 | 0.9745 | 0.9662 |
| F-SCORE | 0.0425 | 0.0317 | 0.0625 | 0.0581 | 0.0594 | 0.0243 | 0.0366 | 0.0643 |
| MSE | 0.0323 | 0.0617 | 0.0672 | 0.0425 | 0.0407 | 0.0267 | 0.011 | 0.0255 |
| RMSE | 0.1798 | 0.2483 | 0.2593 | 0.206 | 0.2017 | 0.1635 | 0.1048 | 0.1596 |
| FPR | 0.0267 | 0.2136 | 0.0827 | 0.0421 | 0.0387 | 0.0177 | 0.0209 | 0.0182 |
| FNR | 0.9432 | 0.7888 | 0.7877 | 0.8941 | 0.8934 | 0.9747 | 0.906 | 0.9319 |
| PSNR | 63.0694 | 60.2637 | 59.8884 | 61.886 | 62.0724 | 63.8943 | 67.7608 | 64.1051 |

TABLE III. EVALUATION METRICS OF DIFFERENT METHODS ON DYNAMIC BACKGROUND VIDEO FROM CD NET DATASET

| Metrics\ Methods | BACKGROUND SUBTRACTION | FRAME DIFFERENCE | BACKGROUND SUBTRACTION AND FRAME DIFFERENCE | ADAPTIVE BACKGROUND SUBTRACTION | ADAPTIVE FRAME DIFFERENCE | ADAPTIVE BACKGROUND SUBTRACTION AND ADAPTIVE FRAME DIFFFERENCE | GMM | GRD |
|---|---|---|---|---|---|---|---|---|
| PRECISION | 0.23 | 0.0111 | 0.038 | 0.0104 | 0.0078 | 0.0635 | 0.013 | 0.0238 |
| RECALL | 0.0922 | 0.1207 | 0.0421 | 0.0765 | 0.1159 | 0.0153 | 0.013 | 0.3333 |
| ACCURACY | 0.9896 | 0.8929 | 0.9848 | 0.9224 | 0.8574 | 0.9891 | 0.7324 | 0.7972 |
| F-SCORE | 0.1316 | 0.0204 | 0.0399 | 0.0183 | 0.0146 | 0.0246 | 0.033 | 0.0444 |
| MSE | 0.0051 | 0.0337 | 0.0072 | 0.0341 | 0.0485 | 0.0053 | 0.004 | 0.005 |
| RMSE | 0.0711 | 0.1836 | 0.085 | 0.1846 | 0.2203 | 0.0727 | 0.025 | 0.0268 |
| FPR | 0.0027 | 0.0999 | 0.008 | 0.0696 | 0.1358 | 0.002 | 0.2676 | 0.1962 |
| FNR | 0.9078 | 0.8793 | 0.9579 | 0.9235 | 0.8841 | 0.9847 | 0.566 | 0.6667 |
| PSNR | 71.1247 | 62.8896 | 69.5814 | 62.8415 | 61.3044 | 70.9357 | 80.1485 | 79.6116 |

TABLE IV. EVALUATION METRICS OF DIFFERENT METHODS ON SHADOW VIDEO FROM CD NET DATASET

| Metrics\ Methods | BACKGROUND SUBTRACTION | FRAME DIFFERENCE | BACKGROUND SUBTRACTION AND FRAME DIFFERENCE | ADAPTIVE BACKGROUND SUBTRACTION | ADAPTIVE FRAME DIFFERENCE | ADAPTIVE BACKGROUND SUBTRACTION AND ADAPTIVE FRAME DIFFFERENCE | GMM | GRD |
|---|---|---|---|---|---|---|---|---|
| PRECISION | 0.1182 | 0.0517 | 0.0438 | 0.0117 | 0.0641 | 0 | 0.0081 | 0.022 |
| RECALL | 0.2734 | 0.159 | 0.2796 | 0.0686 | 0.0994 | 0 | 0.0342 | 0.0897 |
| ACCURACY | 0.9893 | 0.9832 | 0.9709 | 0.9663 | 0.9889 | 0.993 | 0.9737 | 0.9768 |
| F-SCORE | 0.165 | 0.0781 | 0.0757 | 0.0199 | 0.0779 | 0.078 | 0.0131 | 0.0353 |
| MSE | 0.0099 | 0.0156 | 0.0265 | 0.0303 | 0.0104 | 0.0066 | 0.0239 | 0.0214 |
| RMSE | 0.0993 | 0.1247 | 0.1626 | 0.1741 | 0.1022 | 0.0815 | 0.1546 | 0.1463 |
| FPR | 0.0079 | 0.0131 | 0.0261 | 0.0292 | 0.0069 | 0.002 | 0.0215 | 0.019 |
| FNR | 0.7266 | 0.841 | 0.7204 | 0.9314 | 0.9006 | 1 | 0.9658 | 0.9103 |
| PSNR | 68.2231 | 66.2475 | 63.9405 | 63.348 | 67.9736 | 69.942 | 64.3793 | 74.849 |

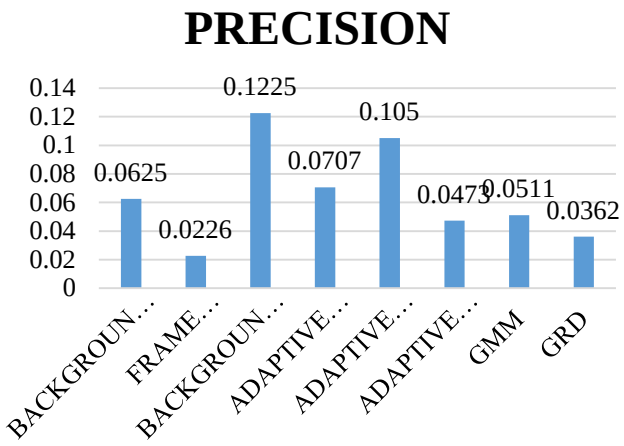



Fig. 2. Precision of Different Methods on Thermal Data Set.

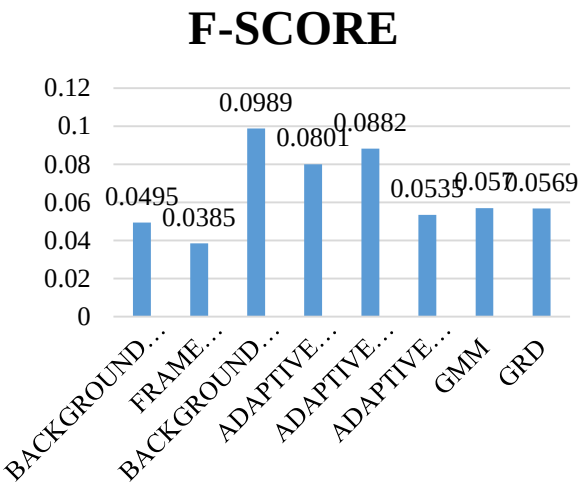



Fig. 5. F-Score of Different Methods on Thermal Data Set.

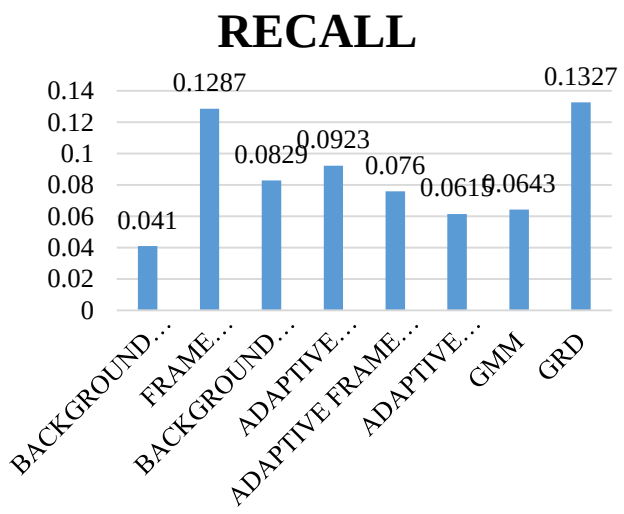



Fig. 3. Recall of Different Methods on Thermal Data Set.

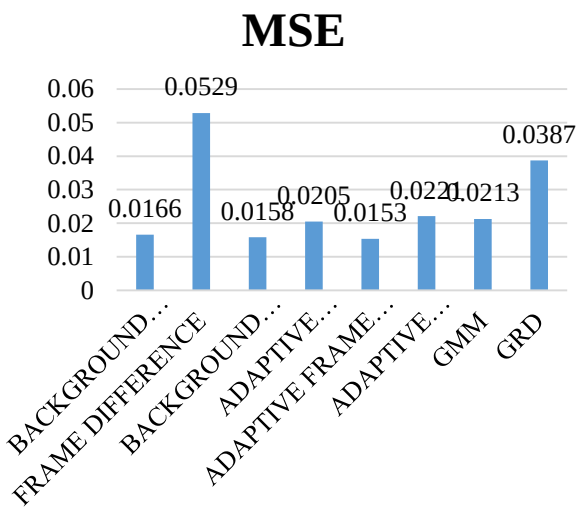



Fig. 6. MSE of Different Methods on Thermal Data Set.

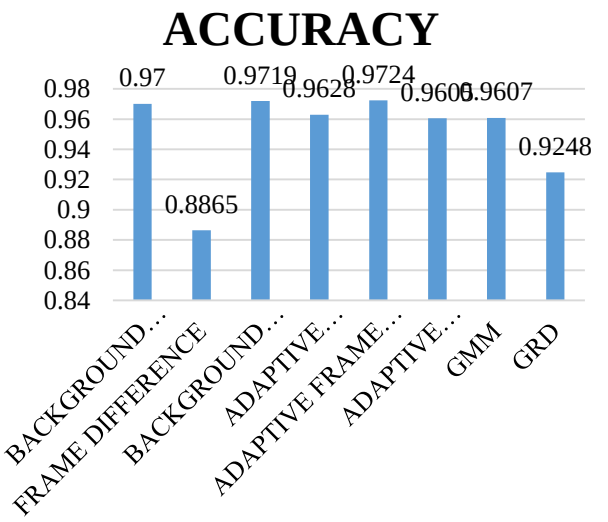



Fig. 4. Accuracy of Different Methods on Thermal Data Set.

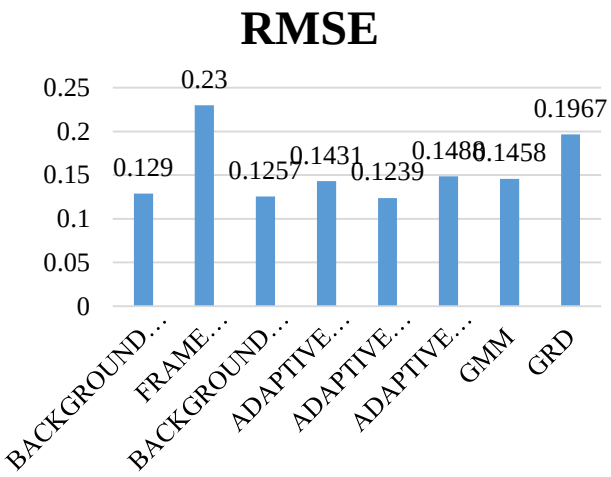



Fig. 7. RMSE of Different Methods on Thermal Data Set.

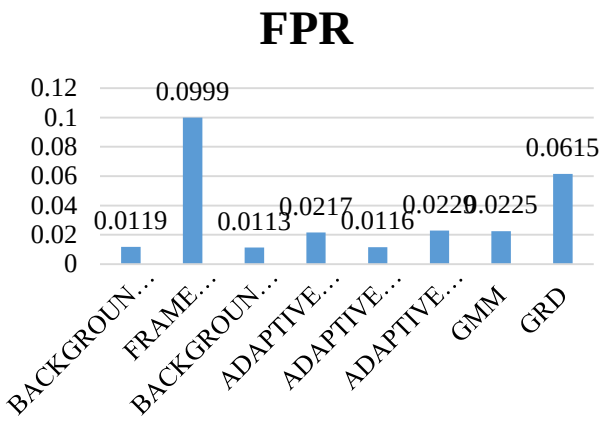



Fig. 8. FPR of Different Methods on Thermal Data Set.

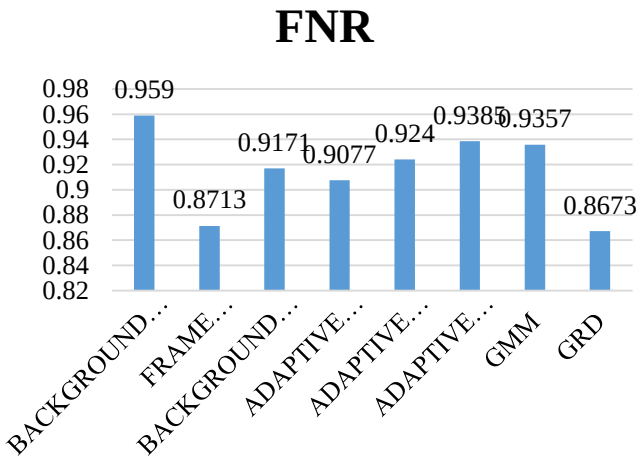



Fig. 9. FNR of Different Methods on Thermal Data Set.

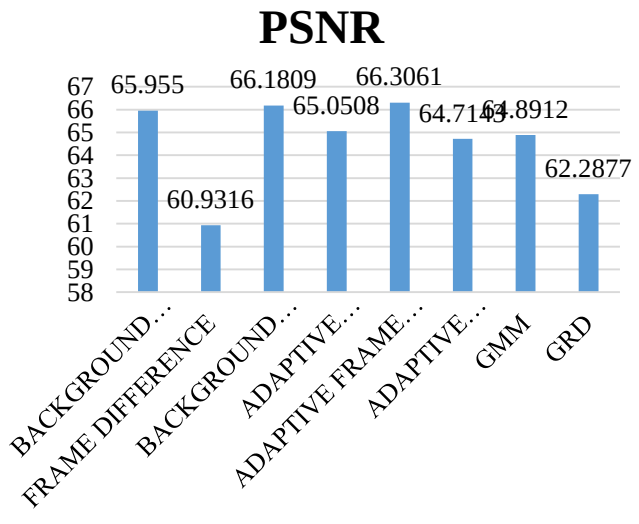



Fig. 10. PSNR of Different Methods on Thermal Data Set.

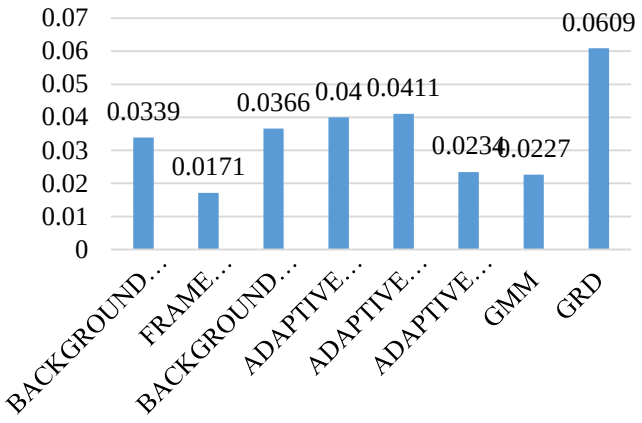



Fig. 11. Precision of Different Methods on Base Line Video Frames.

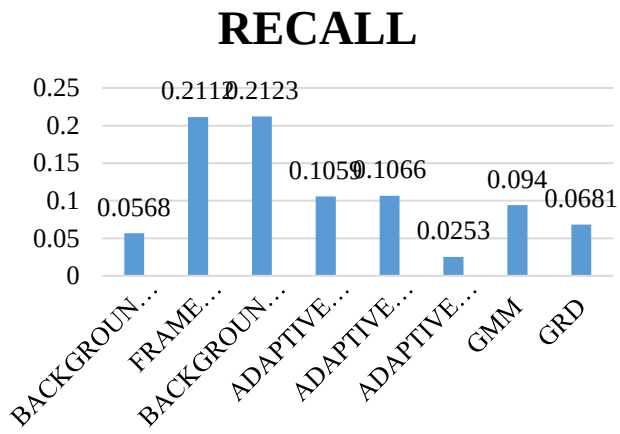



Fig. 12. Recall of Different Methods on Base Line Video Frames.

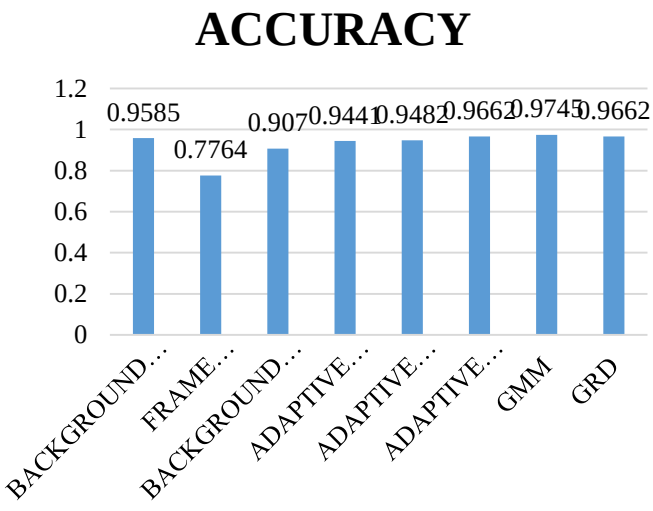



Fig. 13. Accuracy of Different Methods on Base Line Video Frames.

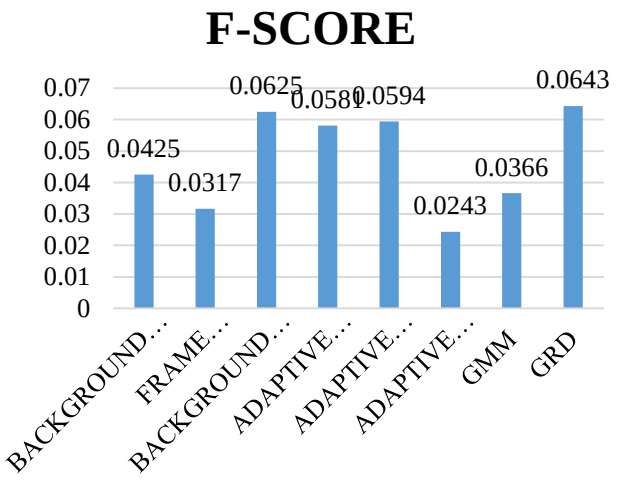



Fig. 14. F-Score of Different Methods on Base Line Video Frames.

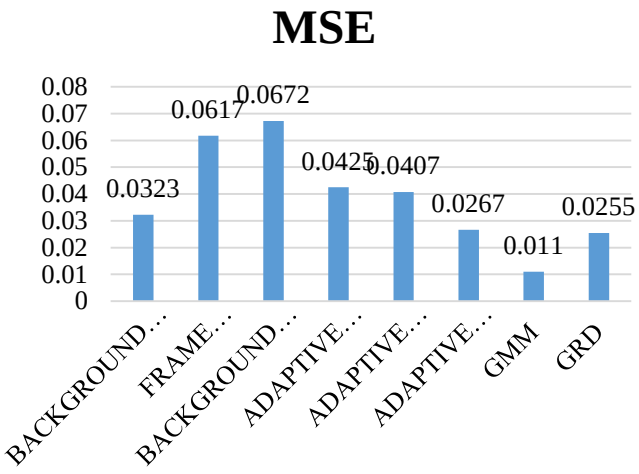



Fig. 15. MSE of Different Methods on Base Line Video Frames.

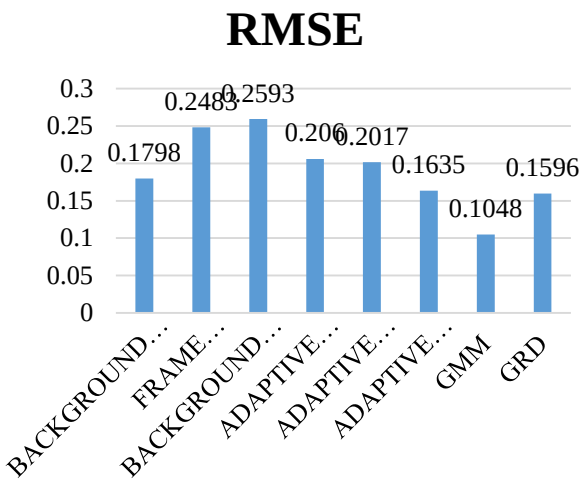



Fig. 16. RMSE of Different Methods on Base Line Video Frames.

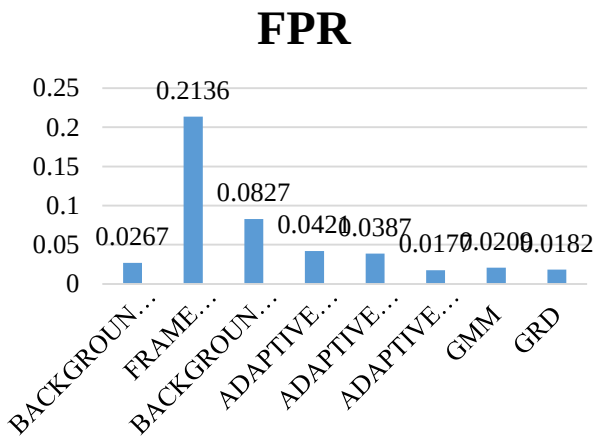



Fig. 17. FPR of Different Methods on Base Line Video Frames.

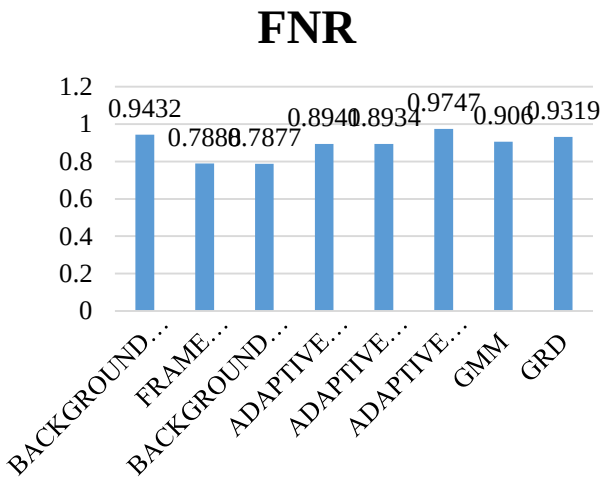



Fig. 18. FNR of Different Methods on Base Line Video Frames.

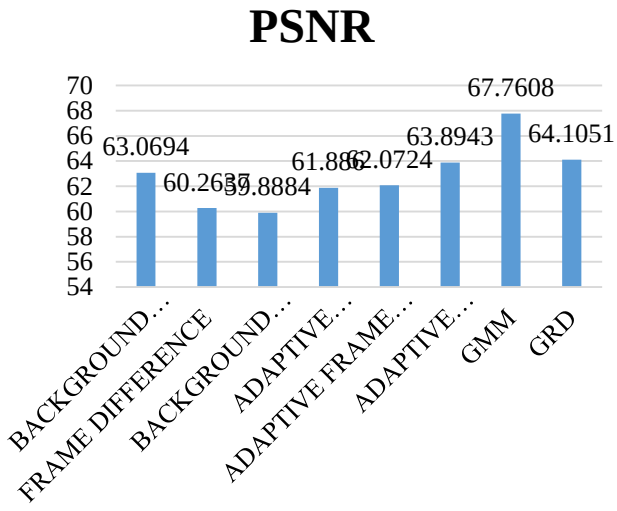



Fig. 19. PSNR of Different Methods on Base Line Video Frames.

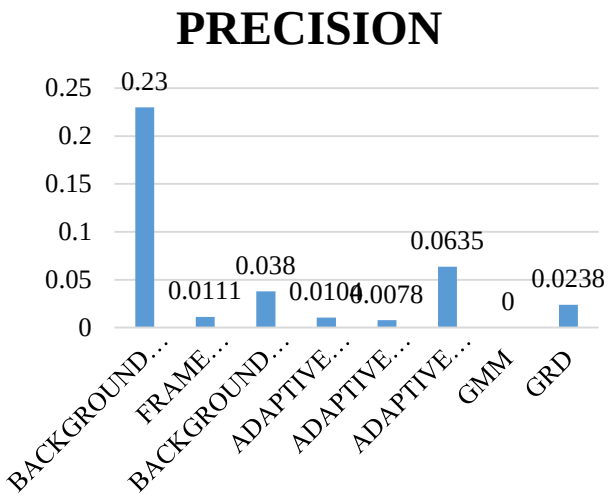



Fig. 20. Precision of Different Methods on Dynamic Background.

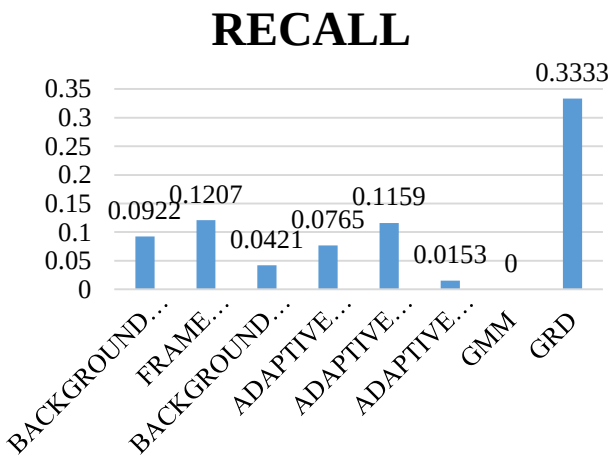



Fig. 21. Recall of Different Methods on Dynamic Background.

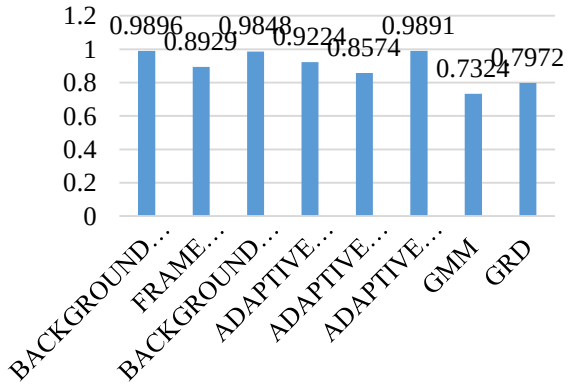



Fig. 22. Accuracy of Different Methods on Dynamic Background.

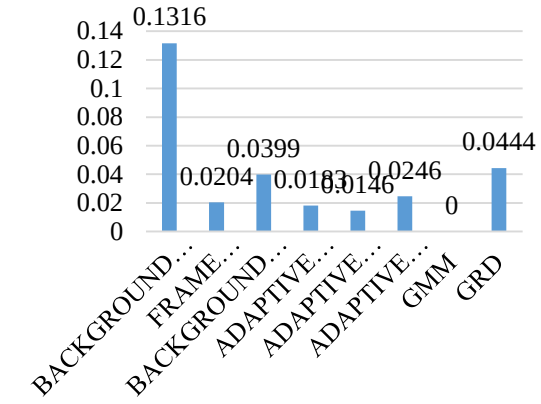



Fig. 23. F-Score of Different Methods on Dynamic Background.

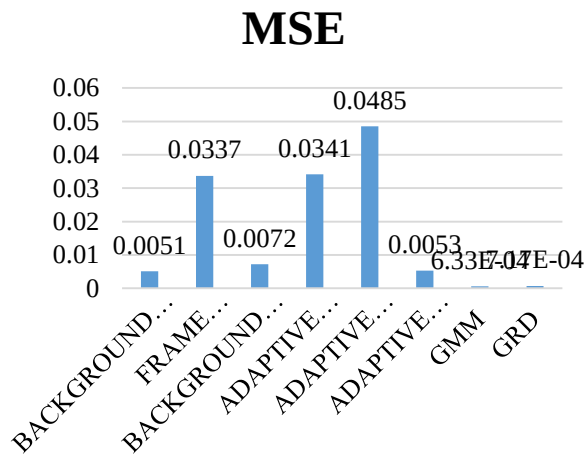



Fig. 24. MSE of Different Methods on Dynamic Background.

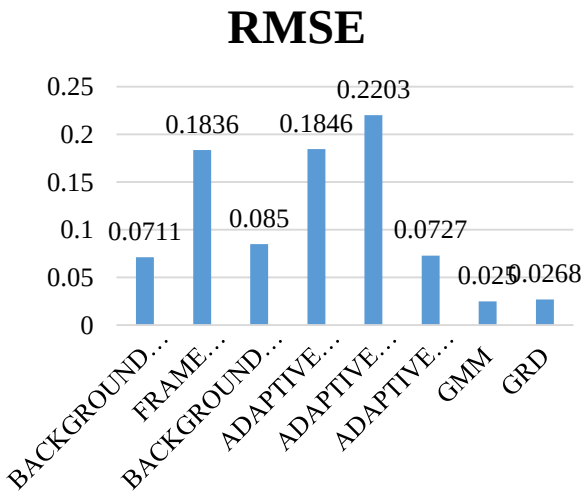



Fig. 25. RMSE of Different Methods on Dynamic Background.

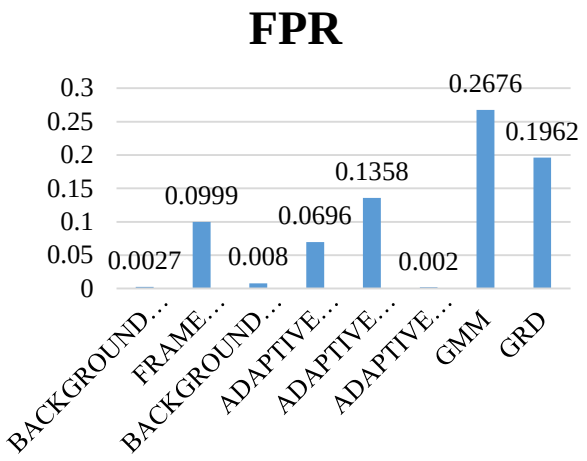



Fig. 26. FPR of Different Methods on Dynamic Background.

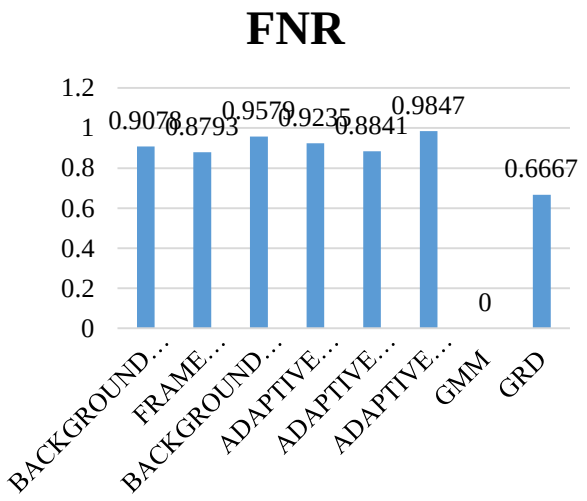



Fig. 27. FNR of Different Methods on Dynamic Background.

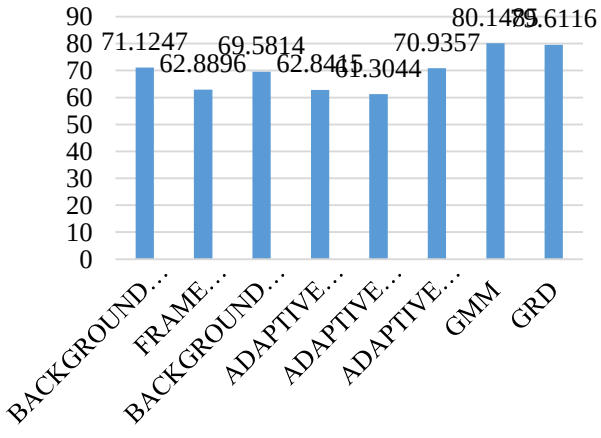



Fig. 28. PSNR of Different Methods on Dynamic Background.

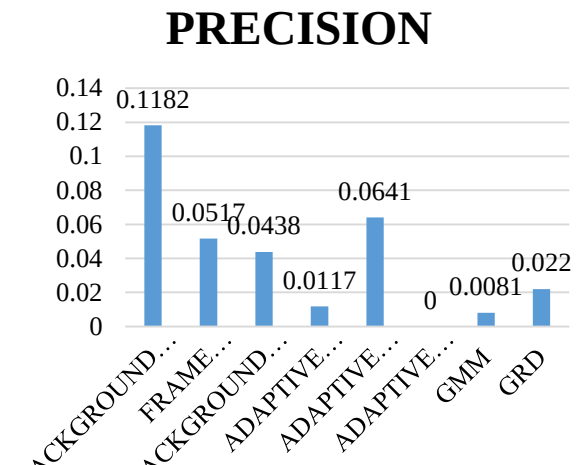



Fig. 29. Precision of Different Methods on Shadow.

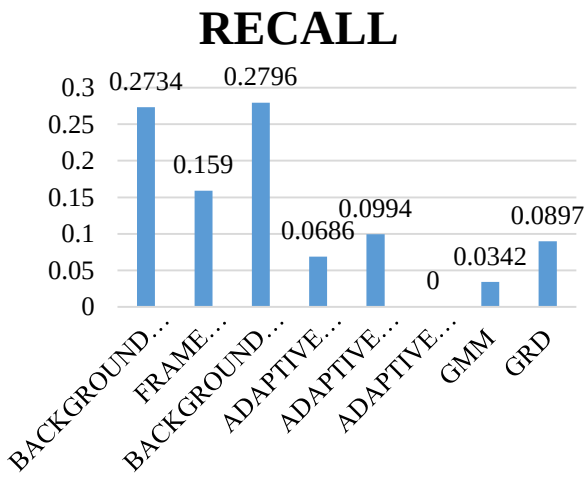



Fig. 30. Recall of Different Methods on Shadow.

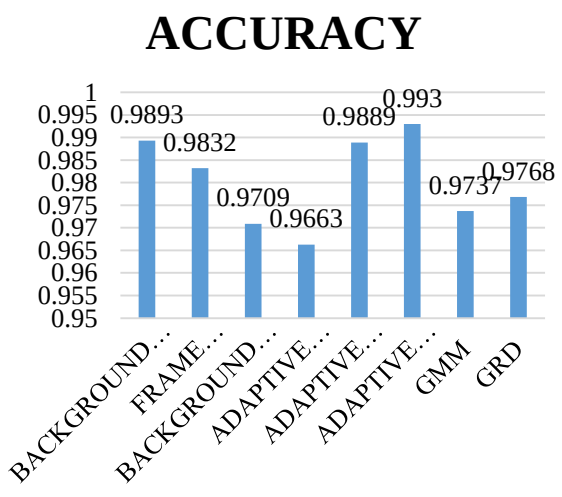



Fig. 31. Accuracy of Different Methods on Shadow.

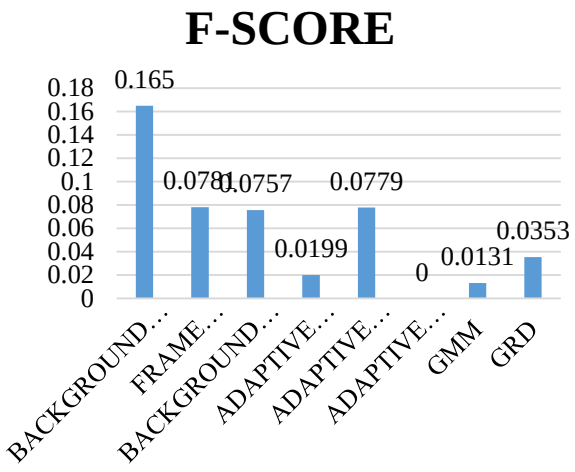



Fig. 32. F-Score of Different Methods on Shadow.

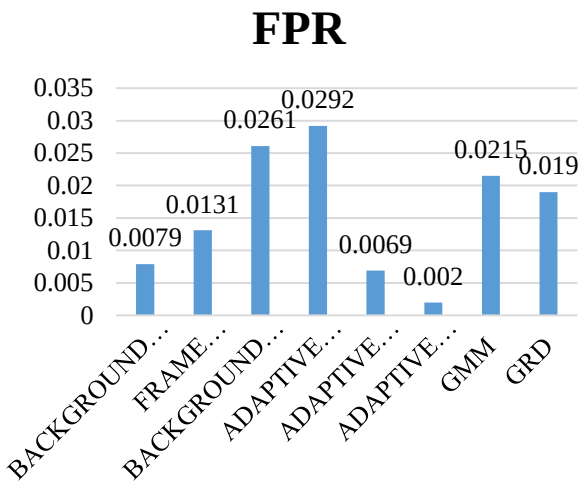



Fig. 35. FPR of Different Methods on Shadow.

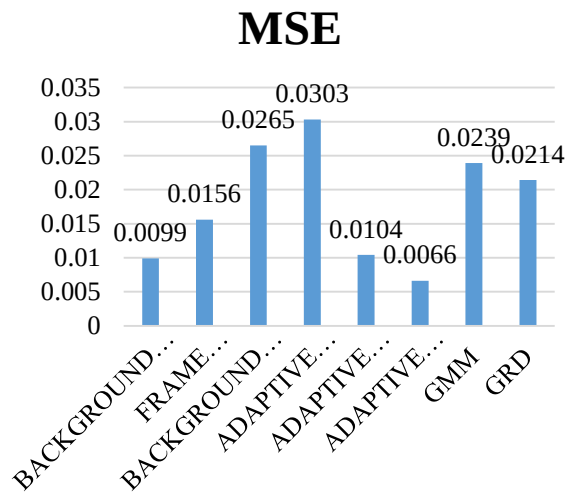



Fig. 33. MSE of Different Methods on Shadow.

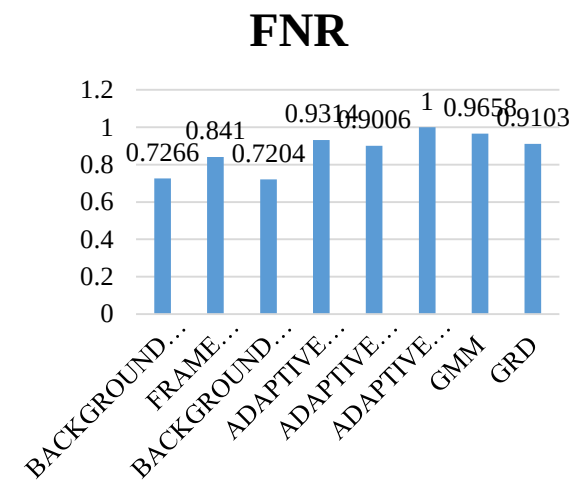



Fig. 36. FNR of Different Methods on Shadow.

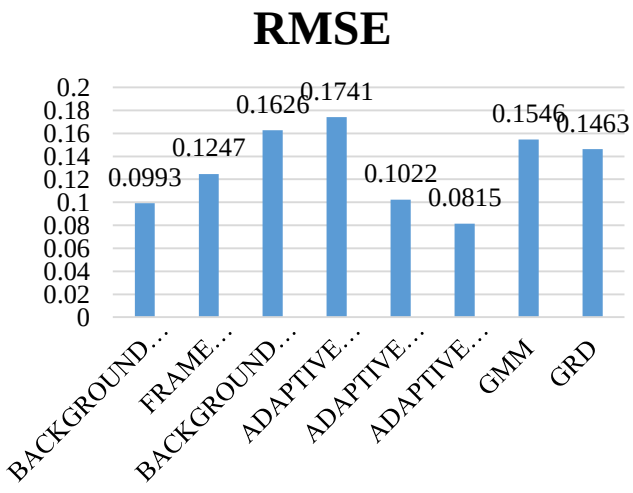



Fig. 34. RMSE of Different Methods on Shadow.

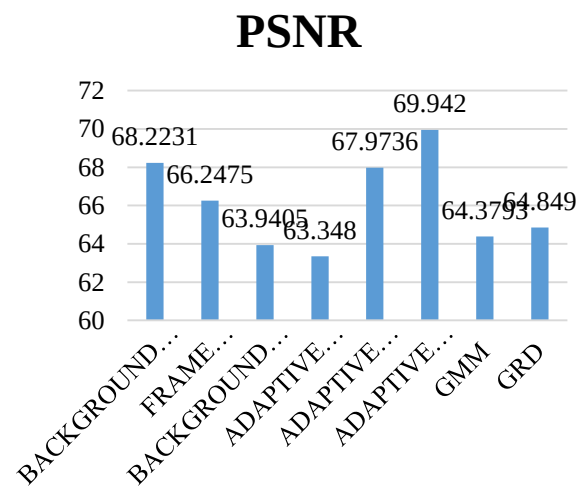



Fig. 37. PSNR of Different Methods on Shadow.

## VIII. Conclusion and Future Work

In this article, a model based on Generalized Rayleigh Distribution (GRD) is presented for the tracking of video images. The model is compared with that of the existing models based on GMM using the metrics Precision, Recall, Accuracy, F-Score, MSE ,RMSE ,FNR, FPR, PSNR. The results are derived and presented in the table -I to table -IV and Figures 1-37. From the tables,, it can be clearly observed that the proposed model perform well with respect to all the parameters and the results , when compared to the existing model based on the GMM, showcases better performance accuracy. This method can be applied to the practical situations, in particular to the areas of medical imaging, in particular situations, where the patient tracking is necessary. The model developed can be further extended in a distributed network scenario, where the images in a distributed environment are to be tracked. This methodology leverages the data intensive frameworks in distributed environment.

## References

[1] Xia, Haiying, Shuxiang Song, and Liping He. "A modified Gaussian mixture background model via spatiotemporal distribution with shadow detection. “Signal, Image and Video Processing 10.2, pp. 343-350, 2016.

[2] H.Zhou, G.Su and X.Jiang, —”Dynamic foreground detection based on improved Codebook model” The Imaging Science Journal, Vol. 64, Issue 2, January 2016.

[3] A. Viswanath, R. Kumari Behra, V. Senthamilarasu, K. Kutty, —”Background Modeling from a Moving Camera”, Elsevier Second International Symposium on Computer Vision and the Internet, Vol. 58, pp. 289-296, 2015

[4] Luo, Yuhan, and Hong Zhang. "Sparse Learning for Robust Background Subtraction of Video Sequences." International Conference on Intelligent Computing. Springer International Publishing, 2015.

[5] Yang, Lu, et al. "Pixel-to-model background modeling in crowded scenes."2014 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2014

[6] Chen Z, Ellis T (2014) “A self-adaptive Gaussian mixture model.”Comput Vis Image Underst122:35- 46.doi:10.1016/ j. cviu. 2014. 01.004

[7] Chien, Shao-Yi, et al. "Video object segmentation and tracking framework with improved threshold decision and diffusion distance." IEEE Transactions on Circuits and Systems for Video Technology 23.6, pp. 921-934, 2013.

[8] Liu, Wan-Chen, et al. "Real-time binary descriptor based background modeling." 2013 2nd IAPR Asian Conference on Pattern Recognition. IEEE, 2013.

[9] D. Hari Hara Santosh, P. Venkatesh, P. Poornesh, L. Narayana Rao, N. Arun Kumar “Tracking Multiple Moving Objects Using Gaussian Mixture Model” in International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-2, May 2013

[10] Huang, Guo-Hao, and Chun-Rong Huang. "Binary invariant cross color descriptor using galaxy sampling." ICPR, pp. 2610-2613, 2012

[11] Hedayati, M., Wan Mimi Diyana Wan Zaki, and Aini Hussain. "A qualitative and quantitative comparison of real-time background subtraction algorithms for video Surveillance applications." Journal of Computational Information Systems 8.2, pp. 493-505, 2012.

[12] Goyette, N., Jodoin, P.-M., Porikli, F., onrad, J., Ishwar, P.:changedetection.net: a new change detection benchmark dataset. In: Proceedings of IEEEWorkshop on Change Detection at CVPR-2012, pp. 16–21 (2012)

[13] Al-Najdawi N., Tedmori S., Edirisinghe E., and Bez H., “An Automated Real-Time People Tracking System Based on KLT Features Detection,” The International Arab Journal of Information Technology, vol. 9, no. 1, 2012.

[14] Bouwmans, T.: Recent advanced statistical background modeling for foreground detection: a systematic survey. Recent Pat. Comput.Sci. 4(3), 147–176 (2011)

[15] Stephan Kopf, Fleming Lampi, Thomas King, and Wolfgang Effelsberg, “Automatic scaling and cropping of videos for devices with limited screen resolution,” in MULTIMEDIA ’06: Proceedings of the 14th annual ACM international conference on Multimedia, New York, NY, USA, 2006, pp. 957–958, ACM.

[16] Tao Mei, Xian-Sheng Hua, He-Qin Zhou, Shipeng Li, and Hong-Jiang Zhang, “Efficient video mosaicing based on motion analysis,” Proc. ICIP 2005, vol. 1, 2005

[17] D. Farin, P. de with, and Wolfgang Effelsberg, “Video-object segmentation using multi-sprite background subtraction,” in Conference on Multimedia and Expo (ICME), 2004.

[18] Chris Stauffer and W.E.L. Grimson, “Adaptive background mixture models for real-time tracking,” Proceedings of theIEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 246–252, 1999

[19] Michael Unger, Mark Asbach, and Peter Hosten “Enhanced background subtraction using global motion compensation and mosaicing”, 978-1-4244-1764-3/08/ IEEE pp.2708-2711, 2008.

[20] Nagesh Vadaparthi, Srinivas Y, P Suresh Varma, P S Sitharama Raju, “Hierarchical Clustering technique for Medical Image Segmentation Based on Finite Skew Gaussian Mixture Model and EM Algorithm”, The International Arab Journal of Information Technology, 15(5), September 2016

[21] N. Otsu;” A threshold selection method from gray-level histograms”. Automatica, 1975, Vol. 11, Pg. 23- 27.

[22] W. Jun-Qin. 2012; “An adaptive frame difference method for human tracking”. Advances in information sciences & services sciences, 2012, Vol. 4, No. 1, Pg. 381-387

# A Survey on using Neural Network based Algorithms for Hand Written Digit Recognition

Muhammad Ramzan[1]
Department of Computer Science & IT
University of Sargodha,
Sargodha, Pakistan

Shahid Mehmood Awan[3]
School of Systems and Technology
University of Management and
Technology, Lahore, Pakistan

Hikmat Ullah Khan[2], Waseem Akhtar[4], Ammara Zamir[7]
Department of Computer Science
COMSATS University Islamabad,
Wah Cantt, Pakistan

Mahwish Ilyas[5]
Department of Computer Science & IT
University of Sargodha,
Sargodha, Pakistan

Ahsan Mahmood[6]
Department of Computer Science
COMSATS University Islamabad,
Attock, Pakistan

***Abstract*—The detection and recognition of handwritten content is the process of converting non-intelligent information such as images into machine edit-able text. This research domain has become an active research area due to vast applications in a number of fields such as handwritten filing of forms or documents in banks, exam form filled by students, users' authentication applications. Generally, the handwritten content recognition process consists of four steps: data preprocessing, segmentation, the feature extraction and selection, application of supervised learning algorithms. In this paper, a detailed survey of existing techniques used for Hand Written Digit Recognition(HWDR) is carried out. This review is novel as it is focused on HWDR and also it only discusses the application of Neural Network (NN) and its modified algorithms. We discuss an overview of NN and different algorithms which have been adopted from NN. In addition, this research study presents a detailed survey of the use of NN and its variants for digit recognition. Each existing work, we elaborate its steps, novelty, use of dataset and advantages and limitations as well. Moreover, we present a Scientometric analysis of HWDR which presents top journals and sources of research content in this research domain. We also present research challenges and potential future work.**

***Keyword—Neural network; digit recognition; segmentation; supervised learning; image classification; computer vision***

## I. INTRODUCTION

The huge increase in the data volume generated by the use of scientific instruments, handwritten documents, the use of the social media and channels has brought forth research challenges. These research challenges have attracted the researchers in several research domains and new research fields have emerged as well. The huge bulk of the data, due to the free content generation facility by the social web users, is in structured as well as unstructured format. The progress in data storage, its processing has given rise to various data related technologies such as sentiment analysis [1, 2], finding top influential users [3, 4]diverse and multilingual data analysis [5, 6]and recognition of handwritten content [7] Computer vision is an active research domain. It is serving the humanity in diverse fields such as video capture cards, scanners, sound cards, webcam and digital cameras. In the computer vision field, the researchers detect the objects from the given input, which are in the form of images and videos and apply the relevant function for recognition of various target objects. The supervised learning approach is applied to this research task which deals in various steps including data input, data pre-processing, the feature extraction, the feature selection using proper algorithms, the data selection based on data split or cross validation, application of various supervised learning algorithms and performance evaluation. Hand Written Digit Recognition (HWDR) is one of the main research problems of image processing, machine learning and computer vision. A digit can be different in style, shape, orientation and size etc. from the same letter written by another person. These variations make it difficult to recognize the digits automatically with the help of machine learning algorithms. HWDR, as a research domain, may be defined as a procedure of detecting and recognizing any digit from the given image and after processing on it, convert the digit into such form that is readable and editable for the machine.

A typical HWDR system consists of different stages. First is the image acquisition, which is a process of obtaining input image either by taking a picture by using a camera or by drawing on the piece of paper and then scanned that paper or by creating drawings using a digital pen. Second step is the pre-processing which is a process by which we enhance the quality of the image there are certain things in the

environment that may damage the quality of the obtained image. Preprocessing is a way by which we enhance the quality of the image by applying certain operations on it like filtering, normalizing, thinning, etc. These operations may vary according to our needs and according to the further uses of the image. Third steps are called segmentation which is used to decompose the image into meaningful subparts. It is a process by which we separate the objects from one another that are present in the picture. So, we can say that segmentation is the process by which we divide the image into more subparts and each subpart represent the presence of an object on it. This process simplifies the image for further processing. The next steps are features extraction which is applied for considering and ranking characteristics of the objects present in the subparts of the image this is helpful for differentiating between more than one objects. Characteristic obtained made further processing like classification easier. There are several techniques that are used for feature extraction. The next step is the classification which is applied for prediction which the object belongs to which class .In this phase, we define more than one classes, then, according to the characteristics of the object we assign them a class, for example, we have two classes human and cat now whenever an object with human characteristic is arrived, we will assign it to the class human and whenever an object with cat characteristics is arrived. We will assign it to the cat class on the basis of its characteristics. Several classifiers are used for classification purpose. This research paper focuses on various methods used for feature extraction and neural network methods for the HWDR and also shares the pros and cons of each technique.

The rest of the paper is divided as follows: Section II discusses the details or neural network, Section III reviews the contemporary relevant literature.domain, Section IV discusses the research methodology and Section V shares the potential research challenges and future directions before concluding the paper in the next section.

## II. An Overview of Neural Network and its Variations

Neural Network is also known Artificial Neural Network (ANN). ANN consists of a number of processing elements which process the given data by their external inputs. ANN is organized into a number of layers which are produced from a number of nodes. Patterns are shown with the help of an input layer. The input layer connects with the hidden layers where the actual processing takes place. Then the hidden layers connect with the output layer which presents the required output in the network. There are different types of ANN algorithms, few are discussed as follows:

### A. Feed Forward Neural Network (FFNN)

FFNN is the basic type of artificial neural network. In this type of NN, nodes do not make a cycle. First an input node is connected to the hidden node and the data goes only in the forward direction in the network. An example of FFNN is shown in Figure 1.

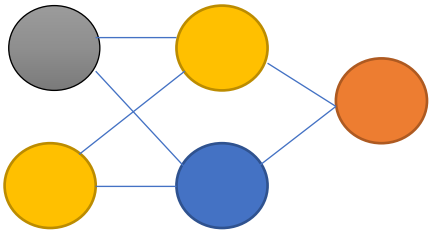


Fig. 1. [1]Feed Forward NN Architecture.

### B. Deep Feed Forward Neural Network (DFF)

The Deep Neural Network is introduced in the emerging new field of the machine learning, which is known as the deep learning. DFF is a type of FFNN, but includes a number of the hidden layers for information processing. At every step, an input layer transforms the given data into more abstract forms. The data is processed using a number of hidden layers. DFF applies on a number of fields such as computer vision, speech recognition, Natural Language Processing (NLP), etc. An example of DFF is shown in Figure 2.

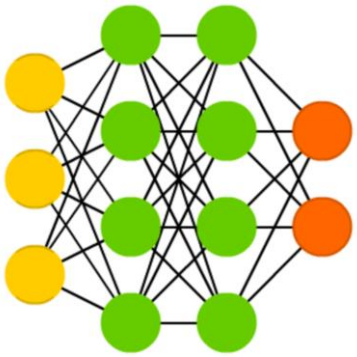


Fig. 2. [2]A Typical Architecture of Deep Feed Forward Neural Network.

### C. Recurrent Neural Network (RNN)

RNN is also a type of ANN in which nodes make use of the directed graph. RNN utilizes its internal memory to proceed with input data. Recurrent refers to two vast classes of networks having the same structure. These two classes show dynamic behavior. One is the infinite impulse whereas another is the finite impulse. The finite class uses a directed acyclic graph and infinite class uses directed cyclic graph. RNN ia applied on handwriting recognition and speech recognition. An example of RNN is shown in Figure 3.

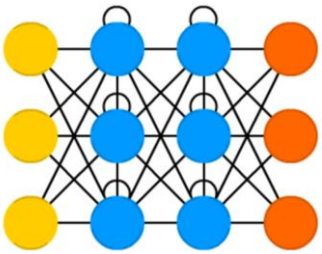


Fig. 3. [3]A Typical Architecture of Recurrent Neural Network.

[1]https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464

### D. Extreme Learning Machine (ELM)

ELM is a feed forward NN used for classification, regression, clustering, and sparse approximation. ELM consists of many hidden layers that need to be tuned for data processing. The hidden layers are not changed and can be taken from their ancestor nodes. The output weight of the hidden nodes is acquired in a single step. These weights define the ratio of learning of hidden layers. An example of ELM is shown in Figure 4.

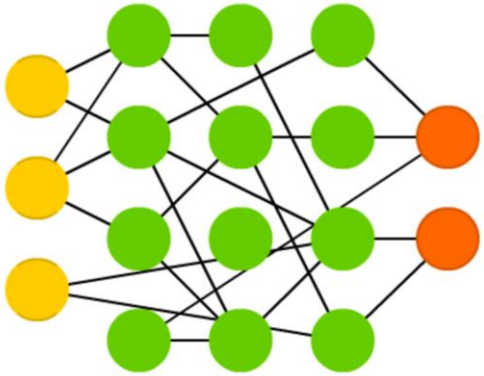


Fig. 4. [4]A Typical Architecture of Extreme Learning Machine.

### E. Convolutional Neural Network (CNN)

CNN is one of the popular classes of Neural Network. CNN also has the shift invariants. CNN is based on their shared weights. CNN is a variation of multilayer perceptron. CNN needs less processing compared to other image classification algorithms. CNN is independent of human knowledge and used for processing of visual images.

### F. Deep Convolutional Network (DCN)

DCN is a very popular class of NN. They use convolutional cells and kernels. The kernel process input data and make groups of layers to simply it. DCN is used for image recognition and visualize sub images. An example of DCN is shown in Figure 5.

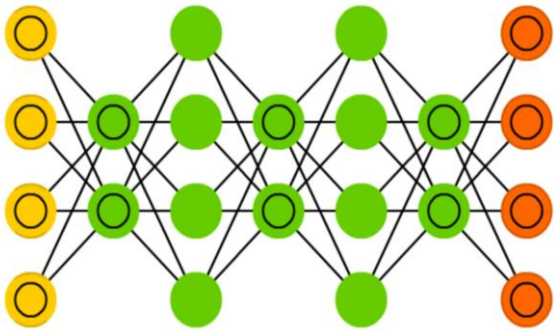


Fig. 5. [5]A Typical Architecture of Deep Neural Network.

[3] https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464

[4] https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464

[5] https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464

### G. Multilayer Perceptron (MLP)

MLP is a category of feed forward NN. MLP contains three layers. The input node is called the neuron which utilizes the nonlinear activation function. MLP uses back propagation technique for training. MLP techniques are widely used for diverse research applications such as speech recognition, character recognition and image recognition.

## III. Application of Neural Network Algorithms for HWDR

This Section reviews the existing research studies which are only related to the use of Neural Network and its variations for HWDR. We have classified the existing work into two Sections: first which focuses on the application of basic Neural Network methods while the second sub-section provides the details of the latest NN techniques for HWDR.

Pinki and Pooja [1] described a Neural Network method for the offline HWDR. In this method, the authors used conventional features with the back propagation network. For training, the classification accuracy was 100%, but recognition accuracy was 91.2%, which shows that the model was overfitted. This research work concludes that the accuracy can be improved with the addition of more features. Hayder Naser Khraibet Al-Behadili [8] has compared the performance of the different machine learning approaches such as K-nearest neighbor (KNN), Neural Network(NN) and Decision Tree(DT) which are used for the HWDR by using the MNIST dataset which is the widely used dataset in the relevant litearature. In these approaches, the Neural Network algorithm is more precise in determining the accurate results. Neural network approaches and feature extraction techniques for handwritten digits recognition. The main objective is to achieve accuracy and high recognition rate [9]. The performance of two Artificial Neural Network Models Feed Forward Neural Network (FFNN) and Recurrent Neural network(RNN) for the HWDR by using the MNIST dataset, is compared by the researchers in which the output shows that the RNN is better in the right recognition of the digits. Neural Network method for the HWDR is built in which, the researchers used binarization and skeletonized binary pixel of the different were used as a input for the multilayer perceptron network [10]. The classification and recognition result were better by using the existing features. The authors concluded that they have reached the computer to the human brain by the use of digit recognition. In a final discussion, they argued that the neural network is much better than other techniques used for digits recognition [11]. Moreover, this research study is regarded as a guide for character recognition research area. It is trusted this definite talk is be gainful understanding into different ideas included, and help additionally progresses in the region [12].

According to another research study [13], different techniques are available to recognize hand written documents. These algorithms can be used to recognize many more languages like hindi, Punjabi, etc. fuzzification with back propagation algorithm can be added to improve the efficiency and correctness of the algorithm. Furthermore, the performance of three different algorithms for HWDR is compared by using the SVM for the classification

purpose and applied on the MNIST dataset. These three algorithm given as BFSS( Binary Fish School Search) , BPSO(Binary Particle Swarm Optimization) and 3rd was ABACO (Advanced Binary Ant Colony Optimization) [14].

We find another neural network based method [15] which proposed the neural based system and surf feature extraction strategy for character recognition. The proposed approach enhanced a procedure for character recognition utilizing neural network and the the results were promising. . Backward Propogation based Neural Network(BPNN) and surf feature strategy have been utilized. Ismail *et al* [16] approach proposed arrangement method that depends on the MLP under other benchmarking datasets. Two arrangements of trials were led for testing the adequacy of the proposed procedure. The principal approach acknowledges highlightsas the data sources and the second incorporates four highlights as information sources. According to the authors, the imperative region of research is multi-lingual character acknowledgment framework. The work of Optical Character Recognition (OCR) and proposal of the frameworks for it in various applications remains a dynamic area of research [17]. According to another research study [18], a content investigation pipeline comprising of four phases: optical character acknowledgment, sentence limit identification, tokenization, and grammatical feature labeling. Utilizing a formal algorithmic model for assessing the execution of multi-arrange forms, the authors introduced trial comes about analyzing the effect of delegate OCR blunders on later stages in the pipeline.

In a recent research study, Pink *et al* [19] present a different feature extraction method along with different classifiers. To achieve high recognition of numeral text, this can be done in Classification and recognition process. The outcome of this proposed work, the neural network is proved to be the best option to recognize handwritten digits as compared to other algorithms. Muhammad *et al* [24] used a simplistic approach for recognition of handwritten digits using artificial neural networks. The knowledge base of the proposed system has been proven to be updated by feeding it new fonts of digits or feeding different patterns of earlier digits. Juan et *al* [25], applied neural networks to hand writing character recognition and they show improved results. According to the authros, Convolutional Neural Networks and Deep Neural Networks allow reaching an optimal performance measured on the MNIST Database which is the widely used data set in the HWDR. The MNIST Database contains 70,000 digitized handwritten numerals distributed in as many as ten classes. In this research study, the dataset has been divided into 60,000 images for training purposes, and the remaining 10,000 are considered asthe test set which is used to measure the accuracy of learning of the proposed model during the training phase.

The Arabic digits are usually used all over the world. We also find research work which used variations of Arabic numbers or different number sets. Alwzwazy1 *et al* [20] used a new challenging digit Arabic dataset which was prepared from different study levels of schools. The authors realized the lack of large data sets, thus they prepared a large dataset by paying vast effort for distributing and collecting digit forms over hundreds of primary, high, college students. Sara Sabour *et al* [21] utilized to conceal Markov models with Gaussian blends as yield dispersions. These models were difficult to learn; however, the results were promising. Edgar *et al* [22] compared the performance of CCN with his own model named as MNIST Model. The MNIST model works very efficiently as compared to CCN, but the authors shared their intent that in the future there could be an improvement in this model using pose matrix.

Meenu Alex *et al* [23] proposed a framework that can perceive characters from old corrupted archives like authoritative reports and change over it into machine editable shape keeping in mind the end goal to save them. The extreme exactness decrease can be seen from first stage to the second stage. In this paper, the authors proposed a strategy for the acknowledgment of Malayalam transcribed characters utilizing a blend of unique classifiers. Li Deng [24] uses MNIST dataset and computed a statis classification model for the domain area that save the problem of pre-processing. The work is novel in this regards as this uses a new dataset and work well in speech recognition research area. Zahangir *et al* [25] developed a new model for HWDR recognition. This model is based on deep neural network techniques. It used the dataset named CMATERdb. This model can give results more precisely and accurately by using other deep learning techniques.

Abdelhadi *et al* [26] proposed a neural systems for manually written digit acknowledgment from the MINST database.. McDonnell *et al* [27] proposed their framework on the basis of SLFN and achieved as good accuracy as the deep training via the help of RLM-type layer network. The authors share that the most accurate network based model is a combination of several non-iterative learning methods to define a projection from the input space to a hidden layer. While the intuitive elegance of deep networks is hard to deny, and the economy of structure of multilayer networks over single layer networks is proven, the authors argue that the speed of training and ease of use of ELM-type single layer networks makes them a pragmatic first choice for many real-world machine learning applications. The authors have demonstrated the computational and performance based advantages based on many use of different NN based algorithms. The performance of different models is compared in detail in Table 1.

TABLE I. A COMPARATIVE ANALYSIS OF RESEARCH WORK FOR HWDR USING NN

| # | Methodology | Pre-processing method | Segmentation | Feature Extraction | Classification model | Dataset | Advantages |
|---|---|---|---|---|---|---|---|
| [1] | Algorithm (gradient descent back propagation) | RGB convertor<br>-Binarization of colored/gray scale image<br>-Noise removed<br>-Median Filter<br>- Deblurring | -Standard pixels chaining technique<br>-Thresholding | Conventional features are used | Artificial neural networks | MNIST | Reduces the dimensionality<br>- Less computational complexity |
| [24] | Simplified NN Architecture<br>-Digitization<br>-Learning and Recognizing Module<br>-Comparing Stored patterns | Digitization | 10 groups segmentation | principal component analysis | Simplified ANN Architecture<br><br>-Comparing Stored patterns | 250 different fonts of Digits | The system is trained on 250 different fonts (limitation: Minor error and variation in pattern Computational Complexity) |
| [25] | Back Propagation and pattern transformations<br>-Multilayer Perceptron (MLP) | Scaling and padding | None | Block Feature, HOG Feature | Convolutional Neural Networks (CNNs) | MNIST | Reduces error rate from 0.46% to 0.34%, which is encouraging |
| [20] | -CNN consisting of 3 layers | Cropping and Wrapping | standard pixels | Robust CNN | Convolutional Neural Networks (CNNs) | 46,000 digits formed in 0-9 Arabic | The accuracy achieved in work is 95.7%. |
| [17] | NEURAL NETWORK | Pre-processing is done by applying following operations<br>RGB TO GRAY,<br>Median filter is applied,<br>skeletonization,<br>centering,<br>and thresholding is done | - Pixel Discontinuity detection,<br>-Standard pixels chaining technique<br>-Thresholding | Kernel based novel feature selection models | Classification and recognition is done using neural network techniques | MNIST (CVS files) | In Previous vol. Classification accuracy results were 100% and recognition is 91.2%. in this paper efforts are made to get higher accuracy. |
| [11] | MLP (MultiLayer Perceptron)<br>NEURAL NETWORK | Table spotting | Individual and non-overlapping text | OCR, Boundary Detection, Tokenization | Artificial Neural Network | Pen-Based Recognition of Handwritten Digits dataset | The accuracy achieved using two features is 99.8% and with four features it is up to 99.9% |
| [5] | Neural network and Back propagation neural network | Preprocessing is done by performing following operations binarization, noise reduction, thinning, skeletonization, normalization and compression | External and internal segmentation | conventional feature extraction technique | Multilayered feed forward neural network | | Recognition accuracy is 91.20% |
| [6] | Back propagation neural network | Noise removal, Normalization-scaling, Thinning and skeletonization | Only vertical segmentation is applied | This phase is not the part of project | Back propagation Neural network with three layers is used. | 1300 isolated Arabic digits used as dataset | Up to 95 % accuracy is achieved by using this technique |

### *A. Other Advanced Algorithms of NN*

There are some other methods that are used in this research. Sakshica, and Gupta[28] compared the performance of three different Neural Network approaches for the HWDR. These approaches are Single layer perceptron model, Hopfield Neural Network and back propagation algorithms. According to the result, the back propagation is more efficient, accurate and fast in digit recognition process. Another research study [29] proposed the RBF Neural Network uses decoupled Kalman filter training method for recognition of handwritten digits of different fonts. The offered method is tested on handwritten digits from 0 to 9 of 25 different fonts. It is found that this method has a very high success level in recognizing the handwritten digits. Mishra and Singh [30] proposed a slope for extraction method and grouping assignment is dealt with utilizing Radial Basis Function NN and back-engendering neural system. In this research study, the authrors used another consolidated approach of highlight extraction and neural system which increment the exactness of the proposed framework. The exactness of back-proliferation neural system is 83.66% and spiral premise work is 98.26% which is superior to other existing strategies. According to Gaurav Jain *et al* [31], a classification algorithm to recognize

handwritten digits (0to 9). Different classifiers and combination methods evaluate the caveats are included in handwritten digit recognition. The classifier based research have significance in handwritten digit recognition and use such as online handwriting recognition on computer tablets, etc. People write the same digit in many different ways.

SmallNORB is a well-known dataset for developing new shape-recognition models precisely, because it lacks many of the additional features of images. The authors implement an efficient version and test much novel models on a larger dataset. Zamani et al [32] proposed RF technique performed practically identical to the best in class strategies. They also plan to explore preprocessing stages that close the hole on RFs and the best in class techniques on this dataset. A proficient Persian manually written digit acknowledgment strategy in view of arbitrary timberland and convolutional neural system has also been presented. Broad examinations with different baselines performed on the Hoda dataset provided promising results. Viragkumar et al [33] exhibited quick, proficient simulated neural system for manually written digit acknowledgment on the GPU to diminish preparing time with PTM (Parallel Training Method). In any case, if the information dataset is bigger than GPU based parallelization is reasonable to diminish preparing the datasets. the authors inferred back proliferation calculation on GPU based parallelization ought to be favored by and large with contrasted with CPU based program. Wong et al [34] analyzed the execution of multi-layer feedforward systems with monotonic and periodical actuation capacities when connected to writing by hand digit acknowledgment. From the recreation comes about, the system with sinusoidal enactment work beats the ones with monotonic actuation work. Calderon et al [35] proposed a convolutional neural networks have been applied for handwritten digit recognition. CNNs were modified by the use of Gabor letters, which are known as good feature extractors. Furthermore, the modification of Gabor alters parameters, improvements of committee machines and development of better learning algorithms and network topologies can be studied. Ajay [36] focused on getting better accuracy rate for classification. Now, the branch of computer vision mainly depends on deep learning features like convolutional neural networks. Many other methods like angent Distance Classifier have been developed using LeNet architecture. The comparison of other models is shown in Table 2.

TABLE II. A COMPARATIVE ANALYSIS OF OTHERADVANCED NN BASED ALGORITHMS

| # | Methodology | Pre-processing | Segmentation | Feature Extraction | Classification | Dataset | Advantages |
|---|---|---|---|---|---|---|---|
| [4] | Artificial neural network with Back-propagation (BP), Single Layer Perceptron (SLP), Hopfield Neural Network (HNN). | Preprocessing is done by performing following operations thresholding, smoothing, filtering, resizing, and normalizing | Segmentation is done using some good segmentation techniques | Structural, statistical, and global transformation feature | SLP,HNN and BP is used as classifiers | Isolated handwritten digits 0-9 are collected | This paper shows that among the all three classification techniques BP is most successful technique (limitation: Problem of False Energy Minima in HNN, the single layer perceptron model does not work in case of X-OR) |
| [16] | Decoupled Kalman filter training | -RGB convertor<br>-Binarization of colored/gray scale image<br>-Noise removed<br>-Normalization<br>-Thinning and skeltonization | touching characters and overlapped characters. Line segmentation | Using Zoning method | -RBF | 0-9 of 25 different fonts | Highly Successful in recognizing digits 0-9 of 25 different fonts |
| [14] | Histogram of the oriented gradient and Haar wavelet<br>-RBFNN<br>-BPNN | -RGB convertor<br>-Binarization of colored/gray scale image<br>-Noise removed<br>-Normalization<br>-Thinning and skeltonization | A succession of digits is decayed into sub-pictures of<br><br>Individual digit. | HOG and HAAR methods are used for feature extraction | Radial basis function and back propagation classifiers | MNIST | Accuracy of BPNN is 83.66% and RBFNN is 98.26%<br>Better than other existing methods |
| [15] | Principal Component Analysis (PCA) combined with K-Nearest Neighbor | Binarization, thresholding, scaling, Thinning and skeletonization | Edge detection. thinning | Principal component analysis (PCA) | K nearest neighbor as a classifier | MINIST | With PCA AND K nearest neighbor as classifier accuracy rate 78.4% is achieved |
| [30] | Extended Kalman Filter (EKF) and modified pruning method | binarlization | normalization | Zoning, histogram, hole size, crossing count, direction, number of end points | MLFFN with periodical and monotonic activation function | MINIST | It is comparison of different techniques and results shows that network of sinusoidal activation function perform well and EKF with pruning lead faster convergence limitation: Network with Monotonic activation |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | function is not more good than network of sinusoidal function) |
| [27] | Backpropagation with MLP is used for classification and ANN is used for recognition | -gray to binary conversion<br>-skeletonization<br>-inverting the image and<br>-scaling | standard pixels | novel Features Extraction Method | Backpropagation with MLP is used for classification | MINIST and HDDIL | it is faster method and high accuracy rate is achieved using this technique. |
| [26] | Random Forest (RF) and Convolutional Neural Network (CNN) | Preprocessing is divided into two stages Scaling Stage and<br>Padding Stage | Standard pixel size | Two techniques are used for feature extraction namely HOG feature and Block features | Random forest classifier | Hoda dataset | High speed can be achieved when we use it with HOG feature extractor (limitation: Required an appropriate hardware to speed up the task, great confusion is found in some digits forms) |
| [19] | -CNN<br>-LeNet Architecture<br>- | No pre-processing step | Standard pixel size | Convolutional layer and Pooling layers | -LeNet architecture | MNIST | Obtains best accuracy rate<br>classification |
| [18] | -Backpropagation algorithm for training<br>- Gabor with Convolutional Network (GCNN) | Normalization RGB to GRAY | Standard pixel size | Gabor filters | Boosting method | MNIST | Effectively performs pattern recognition. |

## IV. Scientometrics Mapping and Investigation

Scientometric and Bibliometric analysis deals in the study in which research related data analysis is carried out. One of the novel approach is that we present investigative system which presents both computational and manual undertakings. In this paper, we likewise played out a computational investigation of the information and processed diverse pointers as characterized in standard Scientometrics papers. We present analysis using notations such as TP denotes Total Papers, TC denotes Total Citations, ACPP denotes Average Citations per Paper, RGR denotes Relative Growth Rate, DT denotes Doubling Time and ICP denotes International Collaborative Papers.

The computational examination utilizing Scientometric approach went for distinguishing year-wise research yield on HWDR using NN, rate of development, and establishment showed conveyance of distributions, global community oriented paper occasions, top production sources, etc. We have then performed content investigation of all exploration papers to distinguish real catch phrases happening in them and their event blasts. The significant subject catchphrases are chosen and a topical thickness plot is additionally produced for the exploration yield information got. We present the vital scientometric markers processed through computational examination of the information. The subsections underneath display points of interest of different markers registered and tables and figures representing the resultant qualities.

As a matter of first importance, the quantity of distributed papers on Handwritten digit is estimated using the Neural system strategies for every one of the years 2010 to 2017. Table 3 demonstrates the quantity of distributed papers in HWDR on a year-wise plot. We can watch that this bend has been pretty much level till 2011, after which there is a lofty ascent. From 2015 to 2017, the quantity of distributed papers has diminished by around five times.

We have additionally registered the relative development rate (RGR) and multiplying time (DT) for HWDR distribution information acquired. While, RGR is a measure meaning the rate of development regarding time, the parameter DT measures the time required for the quantity of distributions in a specific year to end up twofold. The parameters RGR and DT are characterized as takes after as shown in eq (1)

$$RGR = \frac{(Ln\ N2 - Ln\ N1)}{(T2 - T1)} \quad (1)$$

TABLE III. Year-wise research output and growth pattern

| S.no | Year | NOP | Cumulative | RGR | DT |
|---|---|---|---|---|---|
| 1 | 2002 | 1 | 1 | 0 | 0 |
| 2 | 2003 | 1 | 2 | 0.5 | 1.38 |
| 3 | 2007 | 1 | 3 | 0.33 | 2.10 |
| 4 | 2008 | 1 | 4 | 0.25 | 2.77 |
| 5 | 2009 | 3 | 7 | 0.42 | 1.65 |
| 6 | 2011 | 2 | 9 | 0.22 | 3.15 |
| 7 | 2012 | 1 | 10 | 0.1 | 6.93 |
| 8 | 2013 | 4 | 14 | 0.28 | 2.47 |
| 9 | 2014 | 3 | 17 | 0.17 | 4.07 |
| 10 | 2015 | 6 | 23 | 0.26 | 2.66 |
| 11 | 2016 | 7 | 30 | 0.23 | 3.01 |
| 12 | 2017 | 9 | 39 | 0.23 | 3.31 |
| 13 | 2018 | 1 | 40 | 0.02 | 3.45 |

Where, T1 and T2 are two chronological time periods and N1 and N2 T2. In our case, the values are calculated on an annually, hence accordingly re-written as: are number of publications at time periods T1 and (T2 − T1) can be taken as 1 and equations can be given as shown in Eq (2) and Eq (3)

$$RGR = (\ln N2 - \ln N1) = \ln(\frac{N2}{N1}) \quad (2)$$

$$DR = \frac{\ln 2}{RGR} \quad (3)$$

We have processed both these parameters for a HWDR look into production information. Table 3 presents the registered qualities for RGR, Mean RGR, DT and Mean DT for HWDR explore distributions amid the period 2010– 2017. We can watch that the RGR in 2012 (0.46) is twofold to that of the incentive in 2013 (0.28). We can likewise watch that as RGR builds, DT will be lessening and the other way around. The mean RGR and mean DT are likewise ascertained for the entire time frame and we can see from the table that these qualities are 0.24 and 7.00, separately.

The table shows four distinct pointers, in particular TP (Total Papers), TC (Total Citations), ACPP (Average Citations Per Paper) and h-file for the for HWDR look into yield starting from different establishments. Here, it would be worth saying that ACPP esteem is characterized as given in Eq (4)

$$ACP\ P = \frac{TC}{TP} \quad (4)$$

The h - list metric measures both the efficiency and effect of the distributed work of a researcher or a researcher. It is characterized as: a researcher has list h, if h of his/her N p papers have at any rate h references each, and the other (N p - h) papers have at most h references each (Hirsch, 2005). The h - file can be figured for people, establishments, diaries and so forth. We can watch that most creating establishments are generally referred to. For instance, the Microsoft Research and Circuit &system Intel inquire about lab contribute a most astounding number of research papers with their association, address additionally have accomplished the most astounding number of references.

We have contemplated the HWDR's examination production information to locate the significant distribution sources (essentially diaries) where HWDR's exploration work has been accounted for amid 2010– 2017 period. We figured the aggregate number of research articles distributed in every one of the unmistakable diary names found in the downloaded information. We have likewise ascertained TC, ACPP and h-file esteems for every one of the diaries for the HWDR's examination distribution information.

We have likewise considered the examination distribution on HWDR information to locate the most gainful and most referred to creators. We display a rundown of 15 most gainful creators in Table 4.

The results show that the pattern recognition is the top journal according to the proposed framework. Moreover, the top 12 journals include the “Information Sciences”, “Neural Networks” and other journals. The value of TP for Pattern recognition is 6 while the TC score is 161.64 and the value of ACPP is 26.94. This means that average citations per paper for the Pattern recognition journal is 26.94. Similarly, the “Computer Vision and Image Understanding” is also among the top journal in the field as with the total number of 4 papers in the journal, the total citation counts are 129 that means the average citation per paper are 32. Whereas the “Journals of Visual Communication and Image Representation” have a total number of 100 citations that means average citation per paper for 4 papers are 25. These relevant research topics in different journals shows that there are as low as 3.27 average citation per paper and as high as 32.47 citations per paper while among them the average citations among all the journals is 12. This means the number of citations are very high.

TABLE IV. LIST OF TOP JOURNAL SOURCES FOR RELEVANT RESEARCH PUBLICATIONS

| S.No | Journal name | TP | TC | ACPP |
|---|---|---|---|---|
| 1 | Computer Vision and Image Understanding | 4 | 129 | 32.47 |
| 2 | Pattern recognition | 6 | 161 | 26.94 |
| 3 | Journals of Visual Communication and Image Representation | 4 | 100 | 25.2 |
| 4 | Information Sciences | 3 | 46 | 15.64 |
| 5 | Applied Soft Computing | 4 | 29 | 7.43 |
| 6 | Knowledge-based System | 4 | 32 | 6.41 |
| 7 | NeuroImage | 5 | 31 | 6.31 |
| 8 | Information Fusion | 3 | 17 | 5.99 |
| 9 | Medical Image Analysis | 2 | 11 | 5.69 |
| 10 | Neural Networks | 3 | 15 | 5.15 |
| 11 | Engineering Application of Artificial Intelligence | 3 | 11 | 3.74 |
| 12 | Computer and Electronics in Agriculture | 7 | 22 | 3.27 |

## V. RESEARCH CHALLENGES AND FUTURE DIRECTIONS

The latest trend has witnessed the use of the latest techniques in hand written digit in diverse languages such as Persian language and Arabic language. The existing work of the Persian language target to detect and recognize the personal language handwritten digit using a combined feature metho [37]. A one good potential research work may be to detect the Persian language hand written detection using the diverse neural network based algorithms such as Convolutional neural network. The CNN has already been used for HWDR in the Arabic language [38]This research work already combines the Restricted Boltzmann and CNN and achieves better results. Another potential work can be used the other advanced neural network based algorithms such as Recurrent Neural Netowrk (RNN) and You Only Look Once (YOLO). The use of NN based deep learning algorithms has also been applied for HWDR of the Bangla language [25]. The comparative analysis of deep learning approaches for the Bangla language is a potential future work.

We find diverse work related to HWDR related to Asian languages, but to the best of our knowledge, there is no HWDR related to Urdu handwritten digit recognition. A good potential work can be to use the latest neural network based algorithms for Urdu HWDR. This may be good research work as Urdu Digits are different from Arabic and other digits and

also people use the Arabic and Urdu Digits in a same document as well, the difference is shown in the Table 5.

TABLE V. A COMPARISON OF ARABIC AND URDU NUMBERS

| Arabic Numbers | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Urdu Numbers | ۰ | ۱ | ۲ | ۳ | ۴ | ۵ | ۶ | ۷ | ۸ | ۹ |

It is also interesting to plan a research study about detection of handwritten digit written with right hand or left hand. This may be interesting if such studies can distinguish the writers as these and similar research studies may be helpful in criminal detection and recognition as well finding the authentic owner of a handwritten document. In addition, detection of the gender of the writer of handwritten digits may also be interesting as well which may also be beneficial in recognition of the actual writer in case of ambiguity or in the course of finding the actual writer.

## VI. CONCLUSION

This paper explores the details of past work done in the field of HWDR using NN and its variations. The basic as well as advanced NN algorithms have been applied in the relevant literature to detect and recognize the manually handwritten digits in various forms. The diversity of the human nature and write up of digits in diverse forms present a challenging research problem which has become an active research domain. This survey is unique in the sense that it is limited to HWDR only while a number of surveys are related to OCR. In addition, it is also limited to the use of NN algorithms for HWDR.

## REFERENCES

[1] H. U. Khan, "Mixed-sentiment classification of web forum posts using lexical and non-lexical features," Journal of Web Engineering, vol. 16, pp. 161-176, 2017.

[2] H. U. Khan and A. Daud, "Using Machine Learning Techniques for Subjectivity Analysis based on Lexical and Nonlexical Features," International Arab Journal of Information Technology (IAJIT), vol. 14, 2017.

[3] H. U. Khan, A. Daud, U. Ishfaq, T. Amjad, N. Aljohani, R. A. Abbasi, and J. S. Alowibdi, "Modelling to identify influential bloggers in the blogosphere: A survey," Computers in Human Behavior, vol. 68, pp. 64-82, 2017.

[4] H. U. Khan, A. Daud, and T. A. Malik, "MIIB: A Metric to identify top influential bloggers in a community," PloS one, vol. 10, p. e0138359, 2015.

[5] A. Mahmood, H. U. Khan, F. K. Alarfaj, M. Ramzan, and M. Ilyas, "A Multilingual Datasets Repository of the Hadith Content," INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, vol. 9, pp. 165-172, 2018.

[6] R. Khan, H. U. Khan, M. S. Faisal, K. Iqbal, and M. S. I. Malik, "An Analysis of Twitter users of Pakistan," International Journal of Computer Science and Information Security, vol. 14, p. 855, 2016.

[7] A. P. Castaño, "Handwritten Digit Recognition," in Practical Artificial Intelligence, ed: Springer, 2018, pp. 461-478.

[8] H. N. K. AL-Behadili, "Classification Algorithms for Determining Handwritten Digit," Iraqi Journal for Electrical And Electronic Engineering, vol. 12, pp. 96-102, 2016.

[9] A. Gattal and Y. Chibani, "SVM-based segmentation-verification of handwritten connected digits using the oriented sliding window," International Journal of Computational Intelligence and Applications, vol. 14, p. 1550005, 2015.

[10] S. B. Hallale and G. D. Salunke, "Offline handwritten digit recognition using neural network," International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, vol. 2, pp. 4373-4377, 2013.

[11] S. A. Al-Omari, P. Sumari, S. A. Al-Taweel, and A. J. Husain, "Digital recognition using neural network," Journal of Computer Science, vol. 5, p. 427, 2009.

[12] V. L. Sahu and B. Kubde, "Offline handwritten character recognition techniques using neural network: a review," International journal of science and Research (IJSR), vol. 2, pp. 87-94, 2013.

[13] P. Yadav and N. Yadav, "Handwriting recognition system-a review," Analysis, vol. 114, pp. 36-40, 2015.

[14] L. M. Seijas, R. F. Carneiro, C. J. Santana, L. S. Soares, S. G. Bezerra, and C. J. Bastos-Filho, "Metaheuristics for feature selection in handwritten digit recognition," in Computational Intelligence (LA-CCI), 2015 Latin America Congress on, pp. 1-6, 2015.

[15] R. Verma and R. Kaur, "An Efficient Technique for character recognition using Neural Network & Surf Feature Extraction," International Journal of Computer Science and Information Technologies, vol. 5, pp. 1995-7, 2014.

[16] I. M. Keshta, "Handwritten Digit Recognition based on Output-Independent Multi-Layer Perceptrons," HAND, vol. 8, 2017.

[17] N. Islam, Z. Islam, and N. Noor, "A Survey on Optical Character Recognition System," arXiv preprint arXiv:1710.05703, 2017.

[18] D. Lopresti, "Optical character recognition errors and their effects on natural language processing," International Journal on Document Analysis and Recognition (IJDAR), vol. 12, pp. 141-151, 2009.

[19] J. M. Alonso-Weber, M. P. Sesmero, G. Gutierrez, A. Ledezma, and A. Sanchis, "Handwritten Digit Recognition with Pattern Transformations and Neural Network Averaging," in International Conference on Artificial Neural Networks, pp. 335-342, 2013.

[20] H. A. Alwzwazy, H. M. Albehadili, Y. S. Alwan, and N. E. Islam, "Handwritten digit recognition using convolutional neural networks," International Journal of Innovative Research in Computer and Communication Engineering, vol. 4, pp. 1101-1106, 2016.

[21] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in Advances in Neural Information Processing Systems, pp. 3856-3866, 2017.

[22] E. Xi, S. Bing, and Y. Jin, "Capsule network performance on complex data," arXiv preprint arXiv:1712.03480, 2017.

[23] M. Alex and S. Das, "An Approach towards Malayalam Handwriting Recognition Using Dissimilar Classifiers," Procedia Technology, vol. 25, pp. 224-231, 2016.

[24] L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the web]," IEEE Signal Processing Magazine, vol. 29, pp. 141-142, 2012.

[25] M. Z. Alom, P. Sidike, T. M. Taha, and V. K. Asari, "Handwritten bangla digit recognition using deep learning," arXiv preprint arXiv:1705.02680, 2017.

[26] A. Lotfi and A. Benyettou, "Using probabilistic neural networks for handwritten digit recognition," Journal of Artificial lntelligence, vol. 4, pp. 288-294, 2011.

[27] M. D. McDonnell, M. D. Tissera, T. Vladusich, A. Van Schaik, and J. Tapson, "Fast, simple and accurate handwritten digit classification by training shallow neural network classifiers with the ‘extreme learning machine’algorithm," PloS one, vol. 10, p. e0134254, 2015.

[28] D. Sakshica and K. Gupta, "Handwritten digit recognition using various neural network approaches," Int. J. Adv. Res. Comput. Commun. Eng, vol. 4, 2015.

[29] P. Subhashini and V. Prasad, "Recognition of Handwritten Digits Using RBF Neural Network," Int. J. Res. Eng. Technol, vol. 2, pp. 393-397, 2013.

[30] A. Mishra and D. Singh, "Handwritten digit recognition using combined feature extraction technique and neural network."

[31] G. Jain and J. Ko, "Handwritten digits recognition," Multimedia Systems, Project Report, University of Toronto, pp. 1-3, 2008.

[32] Y. Zamani, Y. Souri, H. Rashidi, and S. Kasaei, "Persian handwritten digit recognition by random forest and convolutional neural networks," in Machine Vision and Image Processing (MVIP), 2015 9th Iranian Conference on, pp. 37-40, 2015.

[33] V. N. Jagtap and S. K. Mishra, "Fast efficient artificial neural network for handwritten digit recognition," International Journal of Computer Science and Information Technologies, vol. 5, pp. 2302-2306, 2014.

[34] K.-w. Wong, C.-s. Leung, and S.-j. Chang, "Handwritten digit recognition using multilayer feedforward neural networks with periodic and monotonic activation functions," in Pattern Recognition, 2002. Proceedings. 16th International Conference on, pp. 106-109, 2002.

[35] A. Calderón, S. Roa, and J. Victorino, "Handwritten digit recognition using convolutional neural networks and gabor filters," Proc. Int. Congr. Comput. Intell, 2003.

[36] A. T SIVA, "Handwritten Digit Recognition Using Convolutional Neural Networks," 2017.

[37] M. J. Parseh and M. Meftahi, "A New Combined Feature Extraction Method for Persian Handwritten Digit Recognition," International Journal of Image and Graphics, vol. 17, p. 1750012, 2017.

[38] A. A. Alani, "Arabic Handwritten Digit Recognition Based on Restricted Boltzmann Machine and Convolutional Neural Networks," Information, vol. 8, p. 142, 2017.

# Comparative Analysis of Support Vector Machine, Maximum Likelihood and Neural Network Classification on Multispectral Remote Sensing Data

Asmala Ahmad[1], Ummi Kalsom Mohd Hashim[2], Othman Mohd[3], Mohd Mawardy Abdullah[4]
Centre for Advanced Computing Technology (C-ACT)
Fakulti Teknologi Maklumat Dan Komunikasi (FTMK)
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

Abd Wahid Rasib[6]
Tropical Map Research Group, Department of Geoinformation
Faculty of Geoinformation and Real Estate
Universiti Teknologi Malaysia
Johor, Malaysia

Hamzah Sakidin[5]
Department of Fundamental and Applied Sciences
Universiti Teknologi PETRONAS
Perak, Malaysia

Suliadi Firdaus Sufahani[7]
Fakulti Sains Gunaan dan Teknologi
Universiti Tun Hussein Onn Malaysia
Johor, Malaysia

***Abstract*—Land cover classification is an essential process in many remote sensing applications. Classification based on supervised methods have been preferred by many due to its practicality, accuracy and objectivity compared to unsupervised methods. Nevertheless, the performance of different supervised methods particularly for classifying land covers in Tropical regions such as Malaysia has not been evaluated thoroughly. The study reported in this paper aims to detect land cover changes using multispectral remote sensing data. The data come from Landsat satellite covering part of Klang District, located in Selangor, Malaysia. Landsat bands 1, 2, 3, 4, 5 and 7 are used as the input for three supervised classification methods namely support vector machines (SVM), maximum likelihood (ML) and neural network (NN). The accuracy of the generated classifications is then assessed by means of classification accuracy. Land cover change analysis is also carried out to identify the most reliable method to detect land changes in which showing SVM gives a more stable and realistic outcomes compared to ML and NN.**

***Keywords*—Land cover; change detection; remote sensing; training set; supervised classification***

## I. Introduction

In the early days, land cover information was obtained by manual surveying on foot or land vehicles. This approach had been adopted in many parts of the world for decades but later was found impractical for some circumstances. For instance, for large and remote areas, such approach requires a lot of time and logistically expensive.

Aerial photography was next introduced where land cover information was captured using a camera mounted on an aircraft. This had allowed land cover information to be recorded in a much shorter time. However, such approach was found weather-dependent besides exposing aircraft operators to air accidents. With advancement of recent space technologies, remote sensing satellite was then introduced where land cover information can be captured using sensors mounted on satellites. This is a far better option than the aerial photography as huge land cover information are now able to be obtained continuously, globally and with a relatively cheaper cost.

Such technology enables monitoring of land covers to be effectively carried out at different times [1],[2],[3]. Consequently, this allows changes in land cover that occurred in a certain period of time to be detected. Changes in land cover associated with human activities and natural phenomenon is an important indication for accurate decision making to be made particularly related to agricultural, environmental and urban management [4],[5],[7].

Land cover classification is an essential process in assessing land cover changes. Classification task is performed by assigning each pixel in an image to the correct land cover type. Classification based on supervised methods, which utilise training pixels, has been preferred by researchers in detecting changes in land cover at different parts of the world [7],[8],[9]. Nevertheless, in most cases, only a single supervised method has been considered and evaluated whereas not many attempted to use more than one methods and comparatively analysed their performances [10],[11],[12]. Consequently, the performance of a method with respect to others is not known in-depth leading to a naïve understanding on the issue [13],[14],[15]. This is particularly true for Tropical land covers such as those found in Malaysia [16],[17],[18].

This study attempts to perform land cover classification using maximum likelihood (ML), neural network (NN) and support vector machines (SVM) to assess land cover changes in Klang District, located in Selangor, Malaysia. The performance of each of these methods is to be evaluated based on classification accuracy and land cover change analysis

[19],[20],[22],[23]. Comparative analysis is eventually performed among these methods by making use of these performance measures.

## II. Land Cover Classification

Classification is the process of assigning a pixel to a particular type of land cover [24], [25],[26]. Classification uses data which are mathematically known as a measurement vector or feature vector from an acquisition system. It aims to assign a pixel associated with the measurement at a specific position to a particular class. The classes are defined from supporting data, such as maps and ground data for areas of interest. Two approaches of classification can be considered i.e. unsupervised and supervised classification. Unsupervised classification is a two-step operation of grouping pixels into clusters based on the statistical properties of the measurements, and then labelling the clusters with the appropriate classes [21]. Supervised classification starts from a known set of classes, learns the statistical properties of each class and then assigns the pixels based on these properties [22]. In this study supervised classification was chosen due to the following criteria:

- Simplicity – the practicality of using a large amount of data. This should involve a smaller number of procedures but should produce reasonably accurate and standard results,
- Accuracy – the ability to select important land covers with an acceptable accuracy, i.e. each pixel will be assigned to the correct land cover on the ground. The performance of the method should not be easily affected by factors such as the complexity of land covers, topographic conditions, etc. and
- Objectivity – not involving tuning by a user to improve performance. The generated classification works straight away without needing any adjustment in terms of the number of classes, training pixels, etc.

The three supervised classification schemes to be considered in this study are support vector machine (SVM), maximum likelihood (ML) and neural network (NN).

### A. Support Vector Machine (SVM)

SVM is performed by making use of an efficient hyperplane searching technique that uses minimal training area and therefore consumes less processing time [3],[7]. It is a non-parametric method but capable of developing efficient decision boundaries and therefore can minimize misclassification. SVM works by identifying the optimal hyperplane and divides the data points into two classes. There will be an infinite number of hyperplanes and SVM will select the hyperplane with maximum margin. The margin indicates the distance between the classifier and the training points.

### B. Maximum Likelihood (ML)

In ML, the distribution for each class in each band is assumed to be normal and the probability a given pixel belongs to a specific class [4],[27] is calculated based on this assumption. Each pixel is then assigned to the class that has the highest probability. Classification is performed by calculating the discriminant functions for each pixel in the image.

### C. Neural Network (NN)

In NN, classification is carried out in the conditions where land covers are not linearly separable in the original spectral space. This is performed by making use of multiple nonlinear activation functions at different layers [5]. The training pixels help in identifying the threshold and weight vector connected in the network.

## III. Methodology

This study involves three phases i.e. data pre-processing, data processing and land cover change analysis. Landsat satellite data were obtained from the Malaysian Remote Sensing Agency (MRSA) and United States Geological Survey (USGS) involving Landsat data acquired in 1998, 2000 and 2005. In data pre-processing, we initially calibrated the data where pixel's raw digital number is converted into radiance:

$$L = gain * DN + bias \quad (1)$$

Where L is the pixel value in radiance, DN is the cell value digital number, gain is the gain value for a specific band, and bias is the bias value for a specific band. Atmospheric correction is the process of removing the effects of the atmosphere on the reflectance values of images taken by satellite. Atmospheric Effects are caused by scattering and absorption of EM radiation in the atmosphere and have significant effects mainly on visible and infrared bands that tend to affect processing and interpretation of images. Geometric correction is the process of correcting the data for geometric distortion due to non-systematic error occurred. This was done by initially applying geometric correction on a base-data selected from one of the Landsat data and then registering all other data onto the base-data. Subset was carried out for the selected area within the image, since satellite data usually covers a very large area.

Subsequently, we performed a preliminary assessment to understand the performance of ML, SVM and NN when the size of the training pixels was varied. Such situation may occur when carrying out land cover change detection later due to haze and cloud issues [6],[7]. For this purpose, 1998 Landsat-5 bands 1, 2, 3, 4, 5 and 7 that sense in visible and near infrared wavelengths were used. Band 6 that senses in thermal wavelengths was omitted due to its irrelevancy to this study. Visual interpretation of the Landsat data, aided by a land cover map, was carried out and 11 main classes were identified, viz. coastal swamp forest, dryland forest, oil palm, rubber, industry, cleared land, urban, coconut, bare land sediment plumes and water. Regions of interest (ROIs) associated with the training were determined by choosing one or more polygons for each class based on visual interpretation of the land cover map and Landsat data. This was assisted by region growing technique in which pixels within polygons were grown to neighbouring pixels based on a threshold, i.e. the number of standard deviations away from the mean of the drawn polygons. Pixels for the 11 classes of land cover were determined based on the land cover map. Sampling was carried out by means of stratified random sampling technique. This was done by dividing the population (the entire classification image) into

homogeneous subgroups (the ROI for individual classes) and then taking a simple random sample in each subgroup. 11 training sets were extracted based on percentage of pixels within the ROIs, viz. 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90%. Each of training sets was fed into each of the classifiers i.e. ML, NN and SVM where the accuracy of the classification was assessed by means of percentage classification accuracy [6], [7].

In data processing, we applied the ML, SVM and NN classification to the 2000 and 2005 data to determine land cover changes throughout these dates [28],[29]. Similarly, 11 land covers were considered i.e. coastal swamp forest, dryland forest, oil palm, rubber, industry, cleared land, urban, coconut, bare land, sediment plumes and water.

The results can be categorised into three phases, i.e. data pre-processing, data processing and land cover change detection analysis, as follows.

### *D. Data Pre-processing*

The classification results for 10% through 90% training set size were evaluated by using confusion matrices to assess the capability of SVM, ML and NN in classifying the 11 predefined land covers. In order to better see these classes, suitable colours were assigned to the land covers: coastal swamp forest (green), dryland forest (blue), oil palm (yellow), rubber (cyan), industry (thistle), cleared land (purple), urban (red), coconut (maroon), bare land (orange), sediment plumes (dark green) and water (white). Classification and reference (ground truth) data set were compared among all cases.

Fig.1 shows the classification result by applying ML, NN and SVM method for two extreme cases 10% (the smallest) and 90% (biggest) training set sizes. Visually, based on qualitative visual analysis of the land cover colour distribution, it is obvious that ML and SVM are able to classify more land covers compared to NN for both cases. For 10% training set size, NN recognizes most land covers as oil palm in which is not the case. Similarly, for 90% training set size, NN produces unrealistic scenario by assigning most land covers as rubber.

For ML, it is noticeable for both cases, coconut is found far too abundant along the sea side areas and encroaches markedly towards the inland areas in which likely to be an ambiguous case. In other words, it is likely that misclassification occurs between oil palm and coconut in ML classification. It is likely that these results are due to the similarities of spectral properties between oil palm and coconut. It is also found that there is a discrepancy between the far abundant coconut near the dryland forest for the 10% compared to the 90% training pixel. For SVM, it can be seen that the distribution of classes is rather consistent for the 10% and 90% training pixels indicating that the performance of SVM is not much influenced by the training set size.

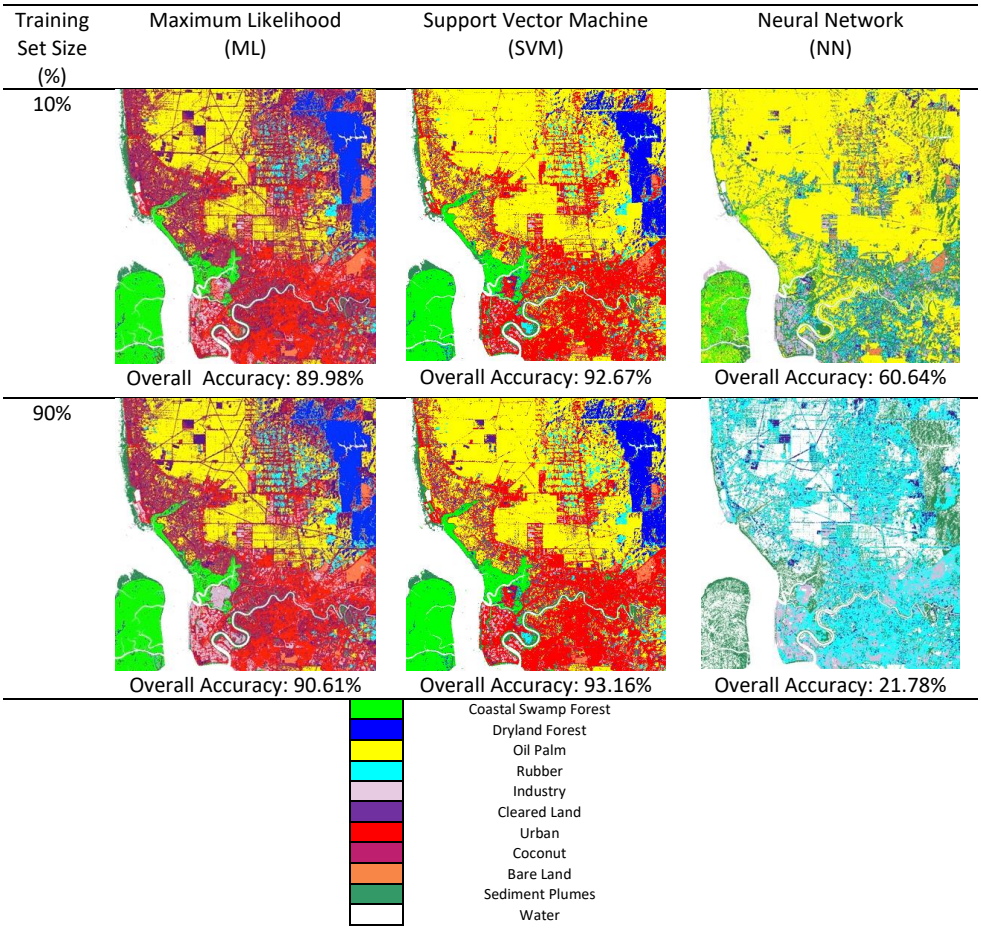



Fig. 1. Land Cover Classification using ML, NN and SVM when using 10% and 90% Training Set Sizes.

In terms of quantitative analysis, for both (10%, 90%) training set sizes, SVM (92.67%, 93.16%) has the highest overall accuracy, followed by ML (89.98%, 90.61%) and NN gives the lowest accuracy (60.64%, 21.78%). SVM and ML have a similar performance trend where the classification accuracy for 90% is higher than the 10% training set size. However, the performance is vice versa for NN. The accuracy differences of the extreme cases for SVM, ML and NN are found to be 0.49%, 0.62% and 38.87% respectively. This shows that SVM has a higher stability when making use of relatively small numbers of training data sets compared to ML and NN. Thus, ML can be regarded as the method that depends much on the accuracy and sufficiency of the training pixels. NN has been known as a method that not only depending on training pixels or learning the rules but its process is also affluence by the network topology that encompasses the hidden layer and interconnections.

To understand further the trend of the classification accuracy with respect to the training pixel size, linear regression analysis was applied to each of the classifications. Fig. 2(a) shows a plot of classification accuracy versus training set size for ML. Although fluctuating, there is somewhat an increasing trend when classification accuracy is plotted against training set size. The linear regression analysis gives $R^2$ of 0.1681 indicating weak positive correlation between the classification accuracy and training set size. Fig. 2(b) shows a plot of classification accuracy versus training set size for NN. It can be seen there is a decreasing trend between classification accuracy and training set size. The regression analysis gives $R^2$ of 0.7516 indicating a somewhat strong negative trend between the classification accuracy and training set size. Fig. 2(c) shows a plot of classification accuracy versus training set size for SVM. There is a noticeable increasing trend between classification accuracy and training set size. The regression analysis gives $R^2$ of 0.7117 indicating a rather strong positive correlation between the classification accuracy and training set size. Fig. 2(d) shows plot of classification accuracy versus training set size for ML, NN and SVM. Clearly, SVM and ML have the higher stability compared to NN in which the accuracy drops drastically as training size increases. However, SVM noticeably outperforms ML due to much higher $R^2$ besides having the least difference in classification accuracy as training set size increases.

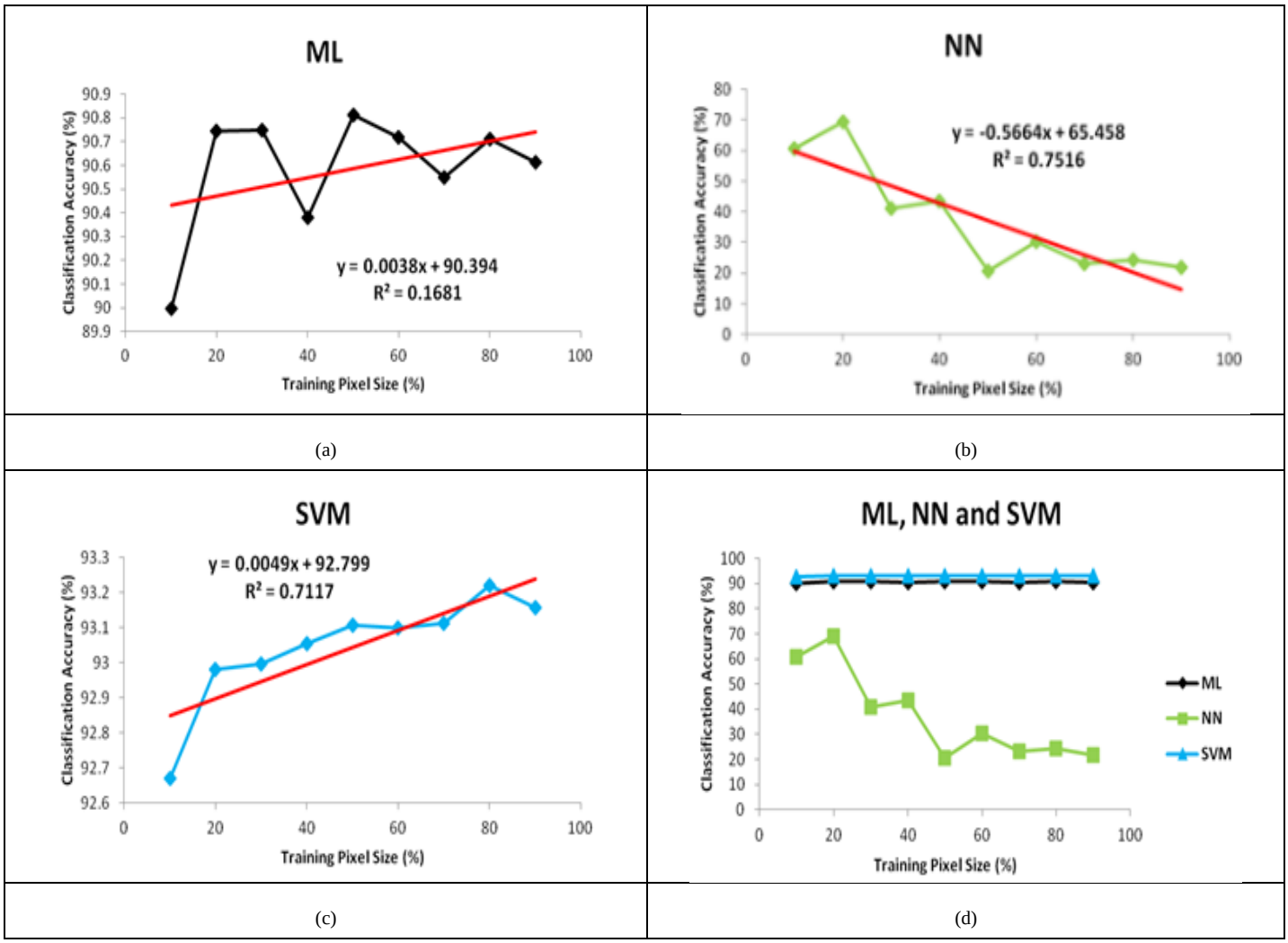



Fig. 2. Relationship between Training Set Size and Classification Accuracy for (a) ML, (b) NN, (c) SVM and (d) the Combination of all.

ML and SVM shows a more realistic classification of land covers compared to NN. However, SVM was found more stable due to not much being affected by varying training set size compared to ML when the size of the training pixels was varied; its accuracy is not much affected compared to ML. SVM also have shown a more realistic land cover area distribution changes compared to ML related to the real scenario.

### *E. Data Processing*

Table I shows land cover area in $km^2$ classified using SVM, ML and NN for the 2000 and 2005 data while Table II shows land cover changes in $km^2$ from 2000 to 2005 based on SVM, ML and NN classification. The land covers are classified into 11 classes; coastal swamp forest (CSF), coconut (C), urban (U), industry (I), dryland forest (DLF), oil palm (OP), bare land (BL), rubber (R), cleared land (CL), water (W) and sediment plumes (SP). For SVM classification, the major conversions are the bare land area and urban area. During the 5-year period, the bare land has decreased by 118 $km^2$. Urban area experienced the highest increase, i.e. 107 $km^2$. This is followed by the coastal land forest 15 $km^2$, oil palm 29.3 $km^2$increase, cleared land 13.5 km increase and industry 6.5 $km^2$increase. Other significant changes have been declines in the sediment plumes 3.9 $km^2$, water 3.4 $km^2$, rubber 0.93 $km^2$and coconut 0.8 $km^2$.

TABLE I. LAND COVERS CLASSIFIED USING SVM, ML AND NN FOR 2000 AND 2005

| Land Cover | Total of Area ($km^2$) | | | | | |
|---|---|---|---|---|---|---|
| | 2000 | | | 2005 | | |
| | SVM | ML | NN | SVM | ML | NN |
| CSF | 45.80 | 37.60 | 0.00 | 61.55 | 48.20 | 331.25 |
| SP | 17.93 | 34.89 | 0.00 | 14.00 | 18.19 | 0.00 |
| U | 52.81 | 79.25 | 5.66 | 160.32 | 161.36 | 0.00 |
| I | 2.80 | 29.57 | 0.00 | 9.27 | 45.96 | 9.83 |
| W | 54.65 | 41.72 | 52.21 | 51.26 | 46.74 | 0.11 |
| DLF | 72.43 | 44.70 | 122.57 | 27.90 | 18.30 | 143.19 |
| BL | 131.85 | 84.76 | 2.39 | 12.87 | 20.37 | 15.82 |
| CL | 3.90 | 9.40 | 0.00 | 17.39 | 14.62 | 39.97 |
| OP | 150.99 | 100.51 | 145.98 | 180.29 | 164.76 | 0.00 |
| R | 3.02 | 10.21 | 32.97 | 2.09 | 1.11 | 0.00 |
| C | 4.12 | 67.70 | 178.53 | 3.37 | 0.68 | 0.14 |

Fig. 3 shows land cover changes from 2000 to 2005 based on area for SVM, ML and NN. SVM and ML shows quite a similar trend that is more realistic but not for NN. Further analysis is carried out to analyse closely the changes based on each land cover from 2000 to 2005.

TABLE II. LAND COVER CHANGES DETECTED USING SVM, ML AND NN FOR THE YEAR 2000 TO 2005

| Land Cover | Land Cover Changes 2000-2005[a] | | |
|---|---|---|---|
| | SVM | ML | NN |
| CSF | 17493<br>15.74<br>34.37 | 11779<br>10.60<br>28.19 | 368053<br>331.25<br>NA |
| SP | -4355<br>-3.92<br>-21.87 | -18556<br>-16.70<br>-47.86 | 0.00<br>0.00<br>NA |
| U | 119460<br>107.51<br>203.59 | 91230<br>82.11<br>103.61 | -6286<br>-5.66<br>NA |
| I | 7189<br>6.47<br>231.01 | 18219<br>16.40<br>55.46 | 10918<br>9.83<br>NA |
| W | -3773<br>-3.39<br>-6.21 | 5584<br>5.03<br>12.05 | -57889<br>-52.10<br>-99.78 |
| DLF | -49486<br>-44.53<br>-61.48 | -29329<br>-26.40<br>-59.06 | 22915<br>20.62<br>16.83 |
| BL | -132205<br>-118.98<br>-90.24 | -71538<br>-64.38<br>-75.96 | 14921<br>13.43<br>562.42 |
| CL | 14993<br>13.49<br>345.75 | 5801<br>5.22<br>55.57 | 44407<br>39.97<br>NA |
| OP | 32456<br>29.30<br>19.41 | 71392<br>64.25<br>63.93 | -162205<br>-145.98<br>-100.00 |
| R | -948<br>-0.93<br>-30.85 | -10110<br>-9.10<br>-89.09 | -36628<br>-32.97<br>-100.00 |
| C | -824<br>-0.76<br>-17.96 | -74472<br>-67.02<br>-1.09 | -198206<br>-178.39<br>-0.41 |

[a] Values are in number of pixels, area ($km^2$) and percentage (%)

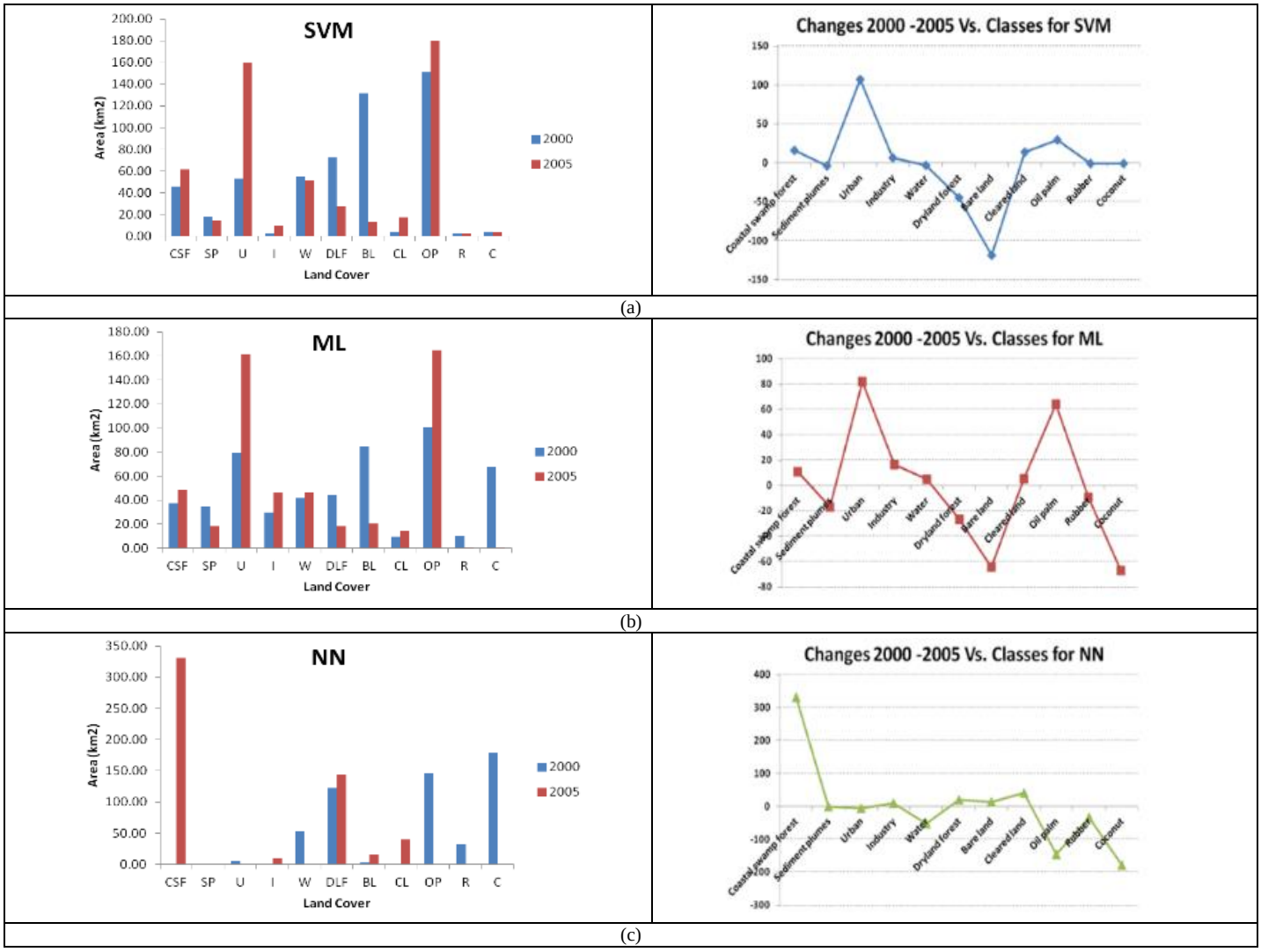



Fig. 3. Land Cover Changes from 2000 to 2005 Detected using SVM (top), ML (middle) and NN (bottom).

### *F. Land Cover Change Detection Analysis*

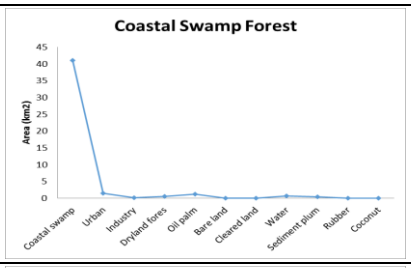
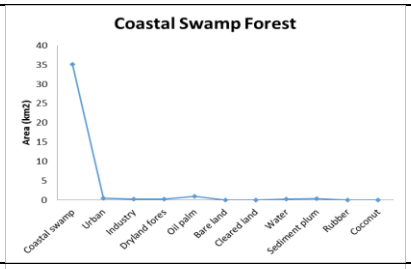
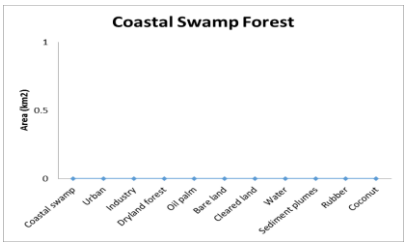
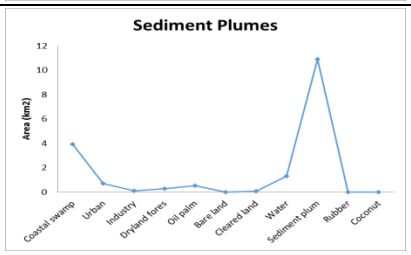
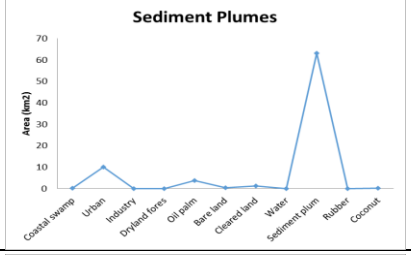
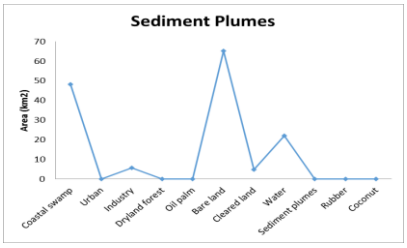
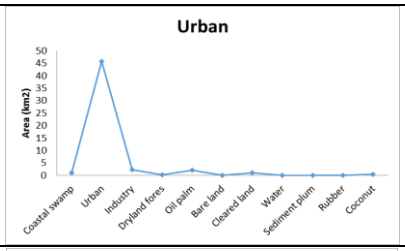
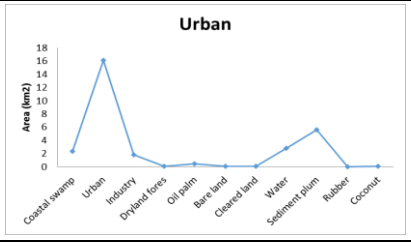
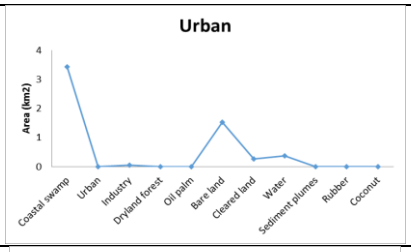
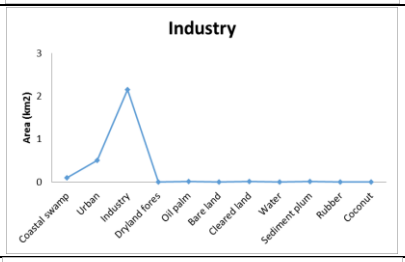
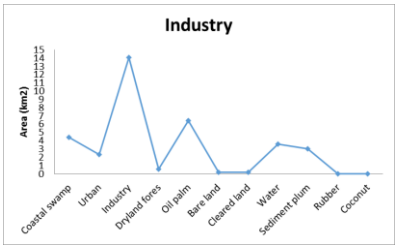
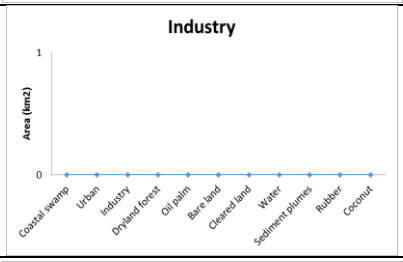
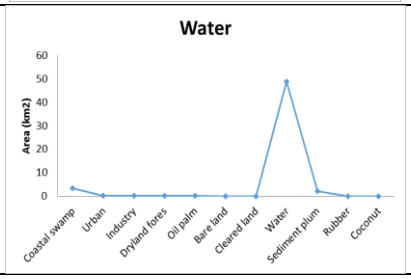
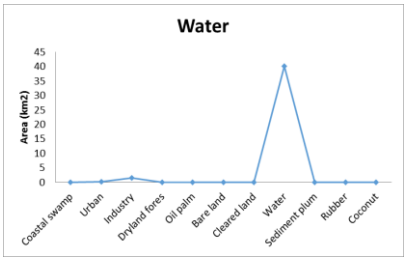
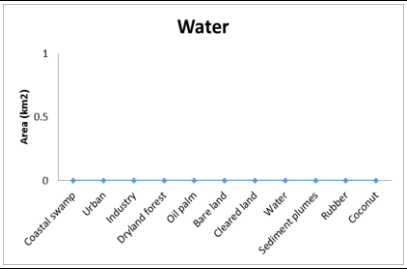
Fig. 4 shows land cover areas for the year 2000 versus 2005 classified using SVM, ML and NN for (a) coastal swamp forest, (b) sediment plumes, (c) urban, (d) industry, (e) water, (f) dryland forest, (g) bare land, (h) cleared land, (i) oil palm, (j) rubber and (k) coconut. For coastal swamp forest, SVM and ML gives a more realistic results compared to NN due to the conversion mainly from coastal swamp forest to oil palm which is the most important commercial crop for Malaysia. For sediment plumes, conversion to coastal swamp forest shown by SVM and NN are seen more realistic, due to the nature of both land covers that is located near to water body, compared to conversion to mainly urban shown by ML. For urban, SVM shows a more realistic outcome due to the very little conversion to oil palm as a highly commercial crop compared to conversion to sediment plumes and bare land by ML and NN respectively. For industry, a little conversion to urban by SVM is more realistic compared to sediment plumes and water in ML and non-existence of industry in NN. For water, a very little conversion to coastal swamp forest by SVM is more realistic compared to conversion to industry in ML and non-existence of water in NN. For dryland forest, conversion to urban in SVM is more realistic than conversion to sediment plumes in ML and non-existence of dryland forest in NN. For bare land, conversion to mainly urban and oil palm in SVM is more realistic than conversion to mainly sediment plumes in ML and conversion coastal swamp forest in NN. For cleared land, conversion to urban and oil palm in SVM is more realistic compared to conversion to sediment plumes that occurred in ML and conversion to coastal swamp forest that occurred in NN. For oil palm, conversion to urban in SVM is more realistic compared to sediment plumes in ML and conversion to coastal swamp forest in NN. For rubber, conversion to oil palm and urban in SVM is more realistic compared to conversion to sediment plumes that occurred in ML and conversion to coastal swamp forest that occurred in NN.

We carried out a different test to determine the realistic scores for each of the methods. '1'score is given for each land cover that gives sensible changes that took place during 2000 to 2005. Based on the land changes that are detected by three methods, SVM possess 100% scores due to the sensible

changes detected for all land covers i.e. coastal swamp forest, dryland forest, oil palm, rubber, industry, cleared land, urban, coconut, bare land sediment plumes and water. ML and NN score 27% each due to the sensible changes for coastal swamp forest and water for the former while sediment plumes and coconut for the latter.

The superiority of SVM over ML and NN is mainly due to less being influenced by the size of training sets. This is likely due to the behaviour of SVM that performs classification by assigning pixel to the correct land cover type by making use of the hyperplane and maximum margin. This leads to the assignment of pixels to the correct land covers and therefore misclassification is minimised. The classification process consequently affects the change detection assessment.

Further research is required to determine how these methods would perform on study areas that differ in terms of land use, vegetation cover, topography, and other variables.

| | SVM | ML | NN |
|---|---|---|---|
| (a) |   |   |   |
| (b) |   |   |   |
| (c) |   |   |   |
| (d) |   |   |   |
| (e) |   |   |   |

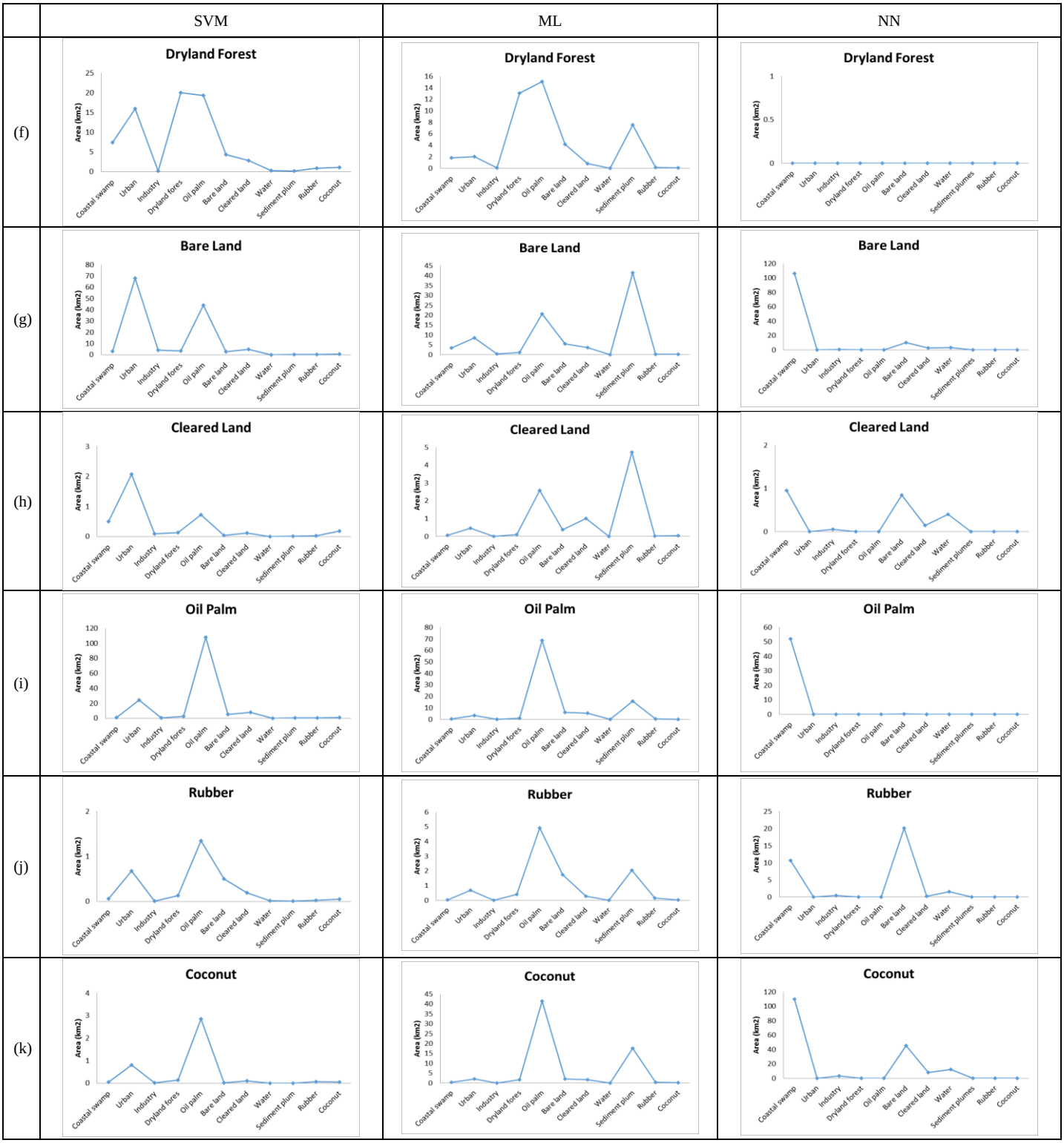



Fig. 4. Land Cover areas for the Year 2000 Versus 2005 Classified using SVM, ML and NN for (a) Coastal Swamp Forest, (b) Sediment Plumes, (c) Urban, (d) Industry, (e) Water, (f) Dryland Forest, (g) Bare Land, (h) Cleared Land, (i) Oil Palm, (j) Rubber and (k) Coconut.

## IV. Conclusion

In this study, a comparative analysis of SVM, ML and NN has been performed by means of classification accuracy and change detection analysis. Landsat satellite images of Klang have been used in the study. The study have encompassed three main process: data pre-processing, data processing and land cover change detection analysis. In data pre-processing, Landsat images have been chosen due to its multispectral capability. Spatial subset and spectral subset have been

performed to resize the images based on the study area and identify the suitable bands in the images respectively. Three methods have been performed in the classification process i.e. SVM, ML and NN. In data processing, the classifications have been performed to the year 2000 and 2005 data where SVM, ML and NN are used to identify the changes in land cover throughout this period. Finally, land cover change analysis has been implemented by analysing the distribution of land cover areas for the year 2000 and 2005 classified using SVM, ML and NN. Overall, SVM produces a more sensible and realistic results compared to ML and NN mainly due to less being influenced by the size of training sets.

## ACKNOWLEDGMENT

The Authors are grateful to Universiti Teknikal Malaysia Melaka (UTeM) for providing the student research funding through the UTeM Zamalah Scheme.

## REFERENCES

[1] A. Singh, "Review Article Digital change detection techniques using remotely-sensed data," International Journal of Remote Sensing, vol. 10, no. 6, pp. 989-1003, 1989.

[2] K.R. Manjula, J. Singaraju J. and A.K. Varma, "Data preprocessing in multi-temporal remote sensing data for deforestation analysis," Global Journal of Computer Science and Technology Software & Data Engineering, vol. 13, no. 6, pp. 1-8, 2013.

[3] N.IS. Bahari, A. Ahmad and B.M. Aboobaider, "Application of support vector machine for classification of multispectral data," 7th IGRSM International Remote Sensing & GIS Conference and Exhibition. IOP Conf. Series: Earth and Environmental Science 20, 2014.

[4] A. Ahmad and S. Quegan, "The effects of haze on the spectral and statistical properties of land cover classification," Applied Mathematical Sciences, vol. 8, no. 180, pp. 9001-9013, 2014.

[5] A. Ahmad and S. Quegan, "Haze modelling and simulation in remote sensing satellite data," Applied Mathematical Sciences, vol. 8, no. 159, pp. 7909-7921, 2014.

[6] M.F. Razali, A. Ahmad, O. Mohd and H. Sakidin, "Quantifying haze from satellite using haze optimized transformation (HOT)," Applied Mathematical Sciences, vol. 9, no. 29, pp. 1407 – 1416, 2015.

[7] U.K.M. Hashim and A. Ahmad, "The effects of training set size on the accuracy of maximum likelihood, neural network and support vector machine classification," Science International-Lahore, vol. 26, no. 4,. pp. 1477-1481, 2014.

[8] K. Islam, M. Jashimuddin, B. Nath, T. Kumar Nath, "Land use classification and change detection by using multi-temporal remotely sensed imagery: The case of Chunati wildlife sanctuary, Bangladesh," The Egyptian Journal of Remote Sensing and Space Science, vol. 21, no. 1, pp. 37-47, 2018.

[9] K. Islam, M. Jashimuddin, B. Nath and T.K. Nath, "Quantitative assessment of land cover change using landsat time series data : case of Chunati Wildlife Sanctuary (CWS), Bangladesh," Int. J. Environ. Geoinformatics, vol. 3, pp. 45-55, 2016.

[10] J. Jayanth, T. Ashok Kumar, S. Koliwad and S. Krishnashastry. "Identification of land cover changes in the coastal area of Dakshina Kannada district, South India during the year 2004–2008," Egypt. J. Remote Sens. Space Sci., vol. 19, pp. 73-93, 2016.

[11] L.N. Kantakumar and P. Neelamsetti, "Multi-temporal land use classification using hybrid approach," Egypt. J. Remote Sens. Space Sci., vol. 18, pp. 289-295, 2015.

[12] A.M. Lal and S.M. Anouncia, "Semi-supervised change detection approach combining sparse fusion and constrained k means for multi-temporal remote sensing images," Egypt. J. Remote Sens. Space Sci. vol. 18, pp. 279-288, 2015.

[13] H.M. Mosammam, J.T. Nia, H. Khani, A. Teymouri and M. Kazemi, "Monitoring land use change and measuring urban sprawl based on its spatial forms: the case of Qom city," Egypt. J. Remote Sens. Space Sci., vol. 20, no. 1, pp. 103-116, 2017.

[14] J.S. Rawat and M. Kumar, "Monitoring land use/cover change using remote sensing and GIS techniques: a case study of Hawalbagh block, district Almora, Uttarakhand, India," Egypt. J. Remote Sens. Space Sci., vol. 18, pp. 77-84, 2015.

[15] S. Sinha, L.K. Sharma and M.S. Nathawat, "Improved Land-use/Land-cover classification of semi-arid deciduous forest landscape using thermal remote sensing," Egypt. J. Remote Sens. Space Sci., vol. 18, pp. 217-233, 2015.

[16] C. Liping, S. Yujun and S. Saeed, "Monitoring and predicting land use and land cover changes using remote sensing and GIS techniques—A case study of a hilly area, Jiangle, China," PLoS ONE, vol. 3, no. 7, pp. 1-23, 2018.

[17] S.K. Singh, P.B. Laari, S. Mustak, P.K. Srivastava and S. Szabó, "Modelling of land use land cover change using earth observation data-sets of Tons River Basin, Madhya Pradesh, India," Geocarto Int., pp. 1-34, 2017.

[18] A.K. Hua, "Land use land cover changes in detection of water quality: a study based on remote sensing and multivariate statistics,"Journal of Environmental and Public Health, pp. 1-12, 2017.

[19] Y. Gao, Z. Liang, B. Wang, Y. Wu and P. Wu, "Wetland change detection using cross-fused-based and normalized difference index analysis on multitemporal Landsat 8 OLI," Journal of Sensors, vol. 2018, pp. 1-8, 2018.

[20] B. Wang, J. Choi, S. Choi, S. Lee, P. Wu and Y. Gao, "Image fusion-based land cover change detection using multi-temporal high-resolution satellite images," Remote Sensing, vol. 9, no. 804, pp. 1-19. 2017.

[21] D. Renza, E. Martinez, I. Molina, and D.M.L. Ballesteros, "Unsupervised change detection in a particular vegetation land cover type using spectral angle mapper," Advances in Space Research, vol. 59, no. 8, pp. 2019–2031, 2017.

[22] R. Singh, M. Singha, S. Singh, D. Pal, N. Tripathi and R. Singh, "Land use/land cover change detection analysis using remote sensing and GIS of Dhanbad distritct, India," Eurasian Journal of Forest Science, vol. 6, no. 2, pp. 1-12, 2018.

[23] S.M. Arafat, K. Abutaleb, E. Farg, M. Nabil1 and M. Ahmed, "Identifying land use change trends using multi-temporal remote sensing data for the New Damietta City, Egypt," Journal of Geography, Environment and Earth Science International, vol. 14, no. 3, pp. 1-12, 2018.

[24] A, Sahar, "Performance of spectral angle mapper and parallelepiped classifiers in agriculture hyperspectral image," International Journal of Advanced Computer Science and Applications, vol. 7, no. 5, pp. 55-63, 2016.

[25] A, Sahar, "Hyperspectral image classification using unsupervised algorithms," International Journal of Advanced Computer Science and Applications, vol. 7, no. 4, pp. 198-205, 2016.

[26] A. Kulkarni and A. Shrestha, "Multispectral image analysis using decision trees," International Journal of Advanced Computer Science and Applications, vol. 8, no. 6, 11-18, 2017.

[27] A.K. Thakkar, V.R. Desai, A. Patel and M.B. Potdar, "Post-classification corrections in improving the classification of land use/land cover of arid region using RS and GIS: The case of Arjuni watershed, Gujarat, India," The Egyptian Journal of Remote Sensing and Space Science, vol. 20, no. 1, pp. 79-89, 2017.

[28] M.D.C. Janiola and G.R. Puno, "Land use and land cover (LULC) change detection using multi-temporal landsat imagery: A case study in Allah Valley Landscape in Southern, Philippines," J. Bio. Env. Sci., vol. 2, no. 2, pp. 98-108, 2018.

[29] M.Z. Iqbal and M.J Iqbal, "Land use detection using remote sensing and gis (A case study of Rawalpindi Division)," American Journal of Remote Sensing, vol. 6, no. 1, pp. 39-51, 2018.

# Mobile Data Collector Routing Protocol Scheme for Scalable Dense Wireless Sensor Network to Optimize Node's Life

Farhan A. Siddiqui[1], Jibran R. Khan[2], Muhammad Saeed[3]
Department of Computer Science - UBIT, University of Karachi, Pakistan

M. Arshad[4]
Deanship of e-Learning and Information Technology, Jazan University, Jazan, Kingdom of Saudi Arabia

Dr. Nasir Touheed[5]
Institute of Business Administration (IBA), University of Karachi, Pakistan

***Abstract*—Wireless Sensor Networks (WSN) is a special kind of network communication architecture which has a very wide range of application and the cost-effectiveness of this architecture boosts its adaptability and usability. The erratic use of WSN, its rapid advancement has encouraged the research community to report several standing problems with WSN, among them is the concern of network life and node energy management in a dense network. This paper presents the experimental outcomes of using MDC multi-tier approach in dense network environments. Besides the node density, experiments also consider the inter-agricultural field hurdles that cause communication disturbance among the nodes that exist in ground level, or at some height above the farming field. The simulated experiment shows the noteworthy results, which comparatively enhance the network lifetime, efficiently utilizing individual node energy, and maximizing the content delivery.**

***Keywords*—WSN; MDC; node density; energy efficient sensor networks; agriculture; robust networks***

## I. INTRODUCTION

During the last few years, WSN stands to be the prominent cost-effective solution of many prevailing challenges. The flexibility and adaptability of WSN rapidly capture a wide range of applications including business, industry, healthcare, and security. The continuous evolution of WSN, makes it compatible with traditional network architectures, interoperation ability, and compatibility[1]. Every technology, in its early age, possesses some holes and leaks that are identified with the passage of time. These loopholes depend on the nature of the architecture it is used in, traffic volume and velocity, and complexity of deployment. In the swift of WSN utilization among many problems, the core challenge that discourages affordability and deployment is to prolong the network lifespan in dense or complex network environment[2]–[4].

A WSN can be deployed as per the business and application requirement, either loosely or tightly coupled (normal to dense or highly complex) scenarios. The connectivity and communication among the deployed sensor nodes are supported by routing protocols[5]–[8]. Unlike traditional routing protocols, WSNs are much more flexible in such that they can be deployed in various formations such as fixed or mobile nodes where each individual node acts as a router and performs computation and operation to receive and send data to the corresponding node. Thus, the volume and communication traffic velocity of a node determine the life of the node.

Many techniques and routing protocols have been designed to minimize node interactivity to overcome the problem of battery diminution and maximize the life of a node which eventually prolongs the network life. These techniques are broadly categorized into two types of WSNs: Homogenous and Heterogeneous network. In Homogeneous WSN architecture, all participating nodes equally perform all tasks and act as a node to sense environmental data; and act as a router to transmit a packet from another node to the next hop, at the same time[8]–[11]. While in Heterogeneous WSN architecture, not all nodes are responsible to receive the environmental feeds. Instead, they are responsible to collect data from other nodes in its region and then deliver aggregated data to the upper hierarchical node of the same kind, these node are known as Head node or Cluster head [1], [5] . While other nodes, despite extensive communication with each other, kill themselves early and halt the network, they communicate with their master or head node to prevent energy depletion. In WSNs, it is also possible that some or all of the nodes are in motion or moveable, and the network can be of one of the two kinds discussed above: Homogeneous or Heterogeneous[12]–[15]. We have already discussed what factors influence the node energy consumption in a static or fixed sensor nodes environment. With node mobility, the energy utilization will also increase with respect to the frequency of node displacement, updating routing information, availability of path or node, the election of head node, and determining member nodes in a specific region and cluster environment. [16].

The problem becomes more complex and challenging when we talk about the dense network scenarios. The node density in a network elevates challenges of wireless sensors networks, such as routing information management, communication management, nodes' energy management (especially mobile node) etc. [17] These challenges make WSN a questionable choice for a cost-effective, prominent, and advanced solution. Also, some question its usability

limitations i.e. where and in what environment WSN should be considered.

The rest of the paper is distributed as follows: The literature review discusses recent progress and approaches to diminish WSNs routing challenges; then we present our novel methodology along with result discussion with simulated outcomes which enhance network life; and then we finally summarize paper with a conclusion and future work discussion.

## II. Related Work

After Among all, there was one literature in previous study that specifically discusses the agriculture application need [18] of managing battery depletion which enriches the network lifecycle and impact of node density in agriculture fields. Based on the studied work, some prominent work taken as the reference point of problem and our prior knowledge and efforts in agro-economic applications of WSNs.

Salim et. al. in [19] used the LEACH protocol introducing an Intra-Balance technique which further advances the LEACH protocol by naming it IBLEACH. The idea works similarly to LEACH itself but differs in the communication scheme in which cluster head and their corresponding member nodes share the energy load based on various parameters that include remaining energy, energy required to transmit frame, and distance in each round. Besides CH, aggregated cluster members list is also determined in each round which is responsible to send only one frame by an aggregator. The simulated environment resulted in empowering the network life comparative to LEACH measured performance and minor improvement as compared with other approaches derived from LEACH.

P. Varshney et. al. [20] measure the performance of two proactive routing protocols OLSR and STAR in a dense Ad-hoc environment. Their experiment was conducted with five different variations as 10, 20, 30, 40, and 50 nodes that gradually increased the number of nodes in a 500 square unit environment. The simulation was done on the QualNet, and results found that OLSR outperforms the STAR protocol in mobility scenarios. As the number of nodes increases with each iteration STAR's performance tends to go down and OLSR keep felicitating and increasing the network performance.

D. Elmazi et. al. [21] proposed the fussy-based system approach for a dense network to measure the performance in WSNs environment. This approach uses the wireless sensor actor network (WSAN) scheme. The author used two fuzzy based actor selection systems (FBASS) which were distinguished by the number of fussy rules. System 1 (FBASS1) used a comparatively low number of rules, while System 2 (FBASS2) had complex and high number of rules defined. The environment comprised of 3 elements:

1) Sense node
2) Actor node
3) Sink node.

The WSAN F system, when it transmits data calculates the distance between two actors each time and sends it to the nearer actor, and then the actor is responsible for delivering it to sink node. The actor selection in this scheme is based on a few things. These factors include the job type, distance, remaining energy which determines the selection probability of actor node (low, medium and high). Between the said two systems, the FBASS2 performed well due to the high number of predefined rules.

## III. Challenges

With reference to study [18], it has been noticed that beyond the standing problems of network life, individual node energy and smart cluster head or actor selection are issues that need to be solved as well. The agriculture fields have another key obstacle in WSNs deployment which is communication among the ground sensor, above the field sensors, and sensors that are deployed at minimal or average height of field. The normal deployment troubled the agri-field hurdles and sensor data failed to be delivered at sink node. The problems that have been highlighted are:

- What if these dense sensors network is deployed in dense agriculture fields?
- How to keep the network performance high or consistent?
- How to prolong the network lifetime? And what technique should facilitate or overcome these issues?

## IV. Simulation Environment

The proposed solution is simulated under OPNET network modular environment to determine its worth and compare the results the previously presented result by using the same parameters of [18] listed in Table 1.

TABLE I. Simulation Parameters

| Parameters | Values |
|---|---|
| **Number of nodes** | 40 |
| **Simulation area** | 1 $km^2$ |
| **Transmission electronics ($E_{TX-elec}$)** | 50 nj/bit |
| **Receiver electronics ($E_{RX-elec}$)** | |
| **Transmit amplifier (εamp)** | 100 pj/bit/$m^2$ |
| **Node energy** | 2 joules |
| **Number of MDCs** | 2 |
| **MDC beacon message rate** | 5 s |
| **MDC velocity** | 0.054 m/s |
| **MDC energy** | 30 joules |
| **Packet size** | 160 bits/packet |

In this experimental work, number of 40 nodes used that randomly deployed in 1 km2 area. Where, each sensor node (N) has the limited energy of 2 joules and MDC that is mobile in nature has 30 joules, which assure that MDC should not die before the sensor node. MDC displacement velocity is set as 0.054 m/s. MDC speed keep relatively slow which allows enough time to bi-directional communication between CH-MDC and MDC-BS, and the size of each data packet is consider as 160 bits/packet.

## V. Results and Discussion

In our previous work [18] we have presented the new methodology and communication approach which significantly outperform the LEACH and other derived techniques by utilizing the hybrid multi-tier Mobile Data Collector (MDC) approach. The basic concept of MDC architecture shown in Fig.1.

Fig. 1 shows the sample agricultural or farm field with obstacles that could lead to inter-cluster and inter-node communication and data delivery issues. Where basic functionality of each participating element is as follows; NDCs are node data collectors which are static or fixed nodes that sense the environmental data and store it, CDCs are the cluster data collectors which are responsible for collecting the data from the NDCs periodically, MDCs are the mobile data collectors which, as per their name, are mobile in nature and can be displaced from one point to another in the region and these MDCs predesignated to get aggregated data from CDCs in its region. Whereas SDC is sink data collector point which is the destination point of sensed environmental data that is attached with the application to measure and store the information. CR represents the cluster region which specifies the nodes' associated membership with CDC in a logical region. Field Area (FA) denotes the experimental area of an open farming land. DOPs are the data obstacle points which cause communication disturbance among the NDCs, CDCs, and MDCs.

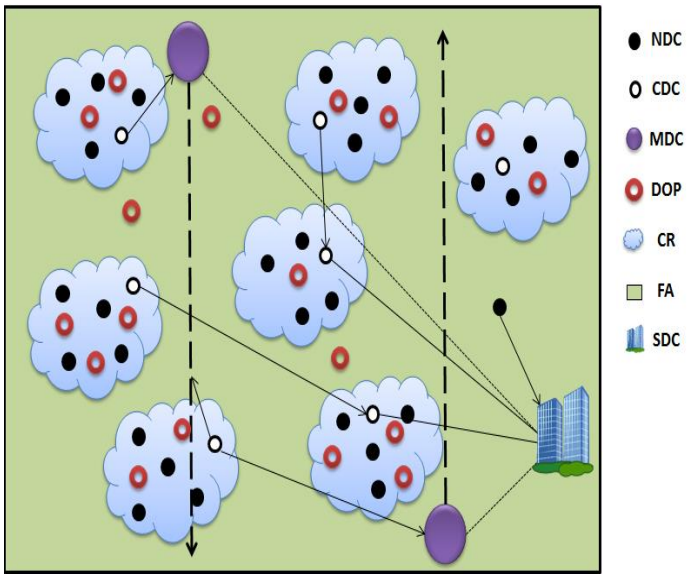



Fig. 1. Multi-Tier MDC Architecture.

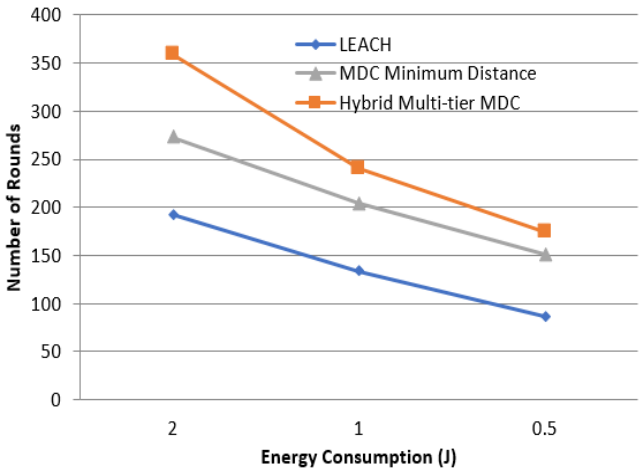



Fig. 2. First Node Die after 350 Rounds.

About 200 rounds (Not Sensor Nodes) including mobile and non-mobile are deployed in an open farming land to measure the performance of the network. Congregated results of simulation show a significant performance and high amount of data received in sample dense network. This experiment determines the energy consumed when the first node dies or has utilized 100% of its battery, maximum energy used when the last node dies, traffic received against node density, and total energy disbursed by nodes.

From the above Fig. 2 shows the state when first node consumed its 100% energy with compared to other mechanisms. It clearly observed that the presented scheme enhanced the node's lifetime and significantly consumed very less amount of energy with compare to LEACH and MDC minimum distance techniques. Where in LEACH, first node dies before the 200 rounds and utilized ¼ of energy in the earliest 90 rounds. On the other hand, in MDC minimum distance, first node relatively life longer than LECH and stays participating till 270 rounds and utilized ¼ of energy in the first 150 rounds. On contrary to LEACH and MDC minimum distance, Hybrid multi-tier MDC empower the nodes' health and prolong the its life that keeps it stay a life even after 350 rounds. Also, whereas LECH first node's consumed its 100% energy and MDC minimum distance first node used its 50% of energy, the presented scheme used only 25% of its energy that lasts for early 180 rounds.

From the above Fig. 3, it can be observed that LEACH network dies in 800 rounds which has no match to proposed scheme. While MDC minimum distance network lived comparatively much longer than LECH and stays for about 2400 rounds and slightly energy curve depicts the irregular or sudden energy falls in some rounds that cost more energy. Whereas, the proposed scheme, significantly outperform other mechanisms and stays longer. The Hybrid Multi-tier MDC uses the consistence minimal amount of energy throughout the network that prolong the network life that showed by the straight decline energy slope. It stays a life for about 3500 rounds; which lives about 3000 rounds more than of LEACH and 1000 rounds more than of MDC minimum distance.

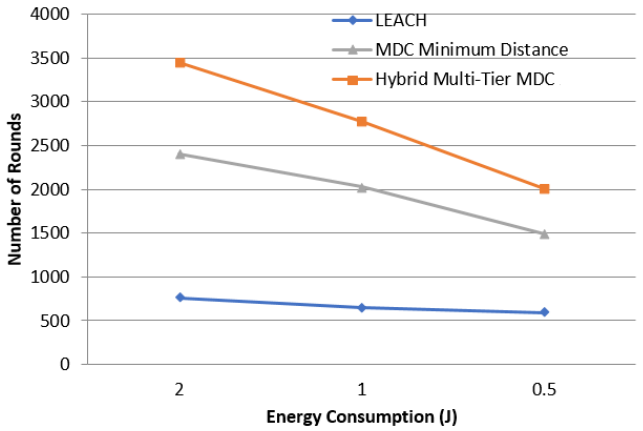



Fig. 3. Last node Compartive Status of LEACH, MDC Minimum Distance and Hybrid Multi-Tier MDC.

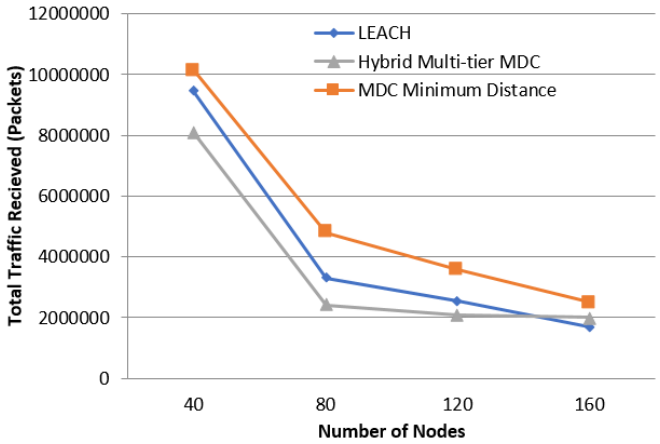



Fig. 4. Total packet Received in Dense Network Environment.

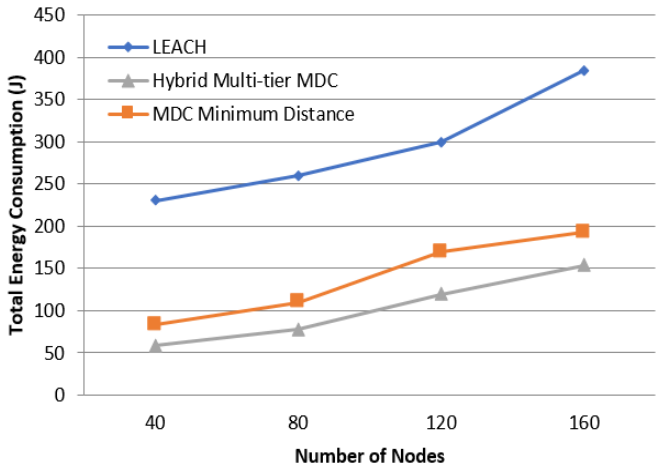



Fig. 5. Total Energy Consumption W.R.T Density.

The Fig. 4 shows the network traffic received in the presence of 40, 80, 120, and 120 nodes network environment. The result gradually shifting towards it favoritism from MDC minimum distance to our proposed scheme of Hybrid multi-tier MDC with the increasing density of network. In the above Fig., we can observed that by increasing the number of nodes or network density, the graph of other two schemes continuously decline while the proposed scheme graph getting stable and consistent with the density of network without losing any more network traffic packets in environment.

Beyond the individual node's life, the Fig. 5 represents total network energy utilization of said mechanisms with the increasing density of network. LEACH disbursed minimum 240J energy in 40 nodes network and maximum about 400J energy in the presence of 160 nodes network. Also, rising LEACH graph reveals incompatibly of dealing with dense network. Although, there is negotiable energy shift in between 40 nodes to 120 nodes network that could be use as density trade off value. But, sudden rise in energy consumption after 120 make it unfavorable for dense network. On the other hand, MDC minimum distance graph shows comparatively better than LEACH. If we notice the graph, it used minimum 40J energy with 40 nodes and maximum 195J with the 160 nodes network. Besides this min and max energy utilization, the graph also tells an interesting fact that require some better scheme to overcome the gap that is the sudden rise in energy requirement after specific amount of node i.e. we can say after each 80 nodes. The energy utilization of between 40 to 80 and 120 to 160 nodes network can be deal as trade-off energy of node density, but the abrupt energy requirement rise right after 80 node. On contrary to both mechanisms, the proposed scheme showed much better results and consistency. The Hybrid multi-tier MDC graph represents its utilized minimum 50J energy in 40 nodes network and maximum 150J in 160 nodes network. Beyond the energy consumption index, also there is a gradual and consistence rise in energy need can be observed by the graph which can be deal as density trade-off i.e. rise in 4 times of nodes (40x4=160) as compare to minimum nodes amount rise only 3 times of energy need (50x3=150). Each add-ons of 40 nodes in network require only half amount of energy required by network.

## VI. Future Work

Due to financial constraints, the studied work is done under simulated environment with some planted huddles that shows provisioning results. Although, simulated work out perform the other exiting techniques, but it does not includes the real agriculture field impact on network life and cross field signal propagation huddles. In future, work will carry out in real agriculture filed to measure the proposed performance and gather more accurate results. Actual farm field experiment may unfolds the other parameters and better options to improve the communication scheme.

## VII. Conclusion

WSNs is a rapidly growing solution which is being adopted across a wide range of industries and businesses as well. Agricultural is a key domain in which WSNs offers remarkable solutions to eradicate agriculture field problems. In previous studies, we found two issues: the first is cross-field communication disturbance, and the second is early network dissolution. This paper uses the proposed MDC multi-tier technique to experiment and determine the network performance and lifetime in a dense open farmland. The artificial obstacles ingress in the simulated environment as cross- field obstacles with high number of nodes with respect to the size. Experimental results show a significant

improvement when compared with prior methodologies. Presented MDC multi-tier approach prolongs the network lifecycle by saving individual node life due to which communication among the nodes becomes more reliable and only insignificant losses are measured in the presence of cross-field communication hurdles.

## REFERENCES

[1] G. Dhand and S. S. Tyagi, "Data Aggregation Techniques in WSN:Survey," Procedia Comput. Sci., vol. 92, pp. 378–384, 2016.

[2] S. Jain, R. C. Shah, W. Brunette, G. Borriello, and S. Roy, "Exploiting mobility for energy efficient data collection in wireless sensor networks," Mob. Networks Appl., vol. 11, no. 3, pp. 327–339, 2006.

[3] P. Siripongwutikorn and B. Thipakorn, "Mobility-aware topology control in mobile ad hoc networks," Comput. Commun., vol. 31, no. 14, pp. 3521–3532, 2008.

[4] G. V Rama and L. V Srikanth, "Location-Based Routing Protocol in Wireless Sensor Network-A Survey," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 5, no. 4, 2015.

[5] P. S. Nithya Darisini and N. S. Kumari, "A survey of routing protocols for VANET in urban scenarios," Proc. 2013 Int. Conf. Pattern Recognition, Informatics Mob. Eng. PRIME 2013, pp. 464–467, 2013.

[6] S. Halder, P. P. Meta, and S. Banerjee, "MOBILITY AWARE ROUTING PROTOCOL IN AD-HOC NETWORK," CS IT, vol. 4, pp. 17–31, 2012.

[7] M. Arshad, N. Armi, N. Kamel, and N. M. Saad, "Mobile data collector based routing protocol for wireless sensor networks," Sci. Res. Essays, vol. 6, no. 29, pp. 6162–6175, 2011.

[8] Y. Khamayseh, G. Obiedat, and M. B. Yassin, "Mobility and Load aware Routing protocol for ad hoc networks," J. King Saud Univ. - Comput. Inf. Sci., vol. 23, no. 2, pp. 105–113, Jul. 2011.

[9] A. Kumar, H. Y. Shwe, K. J. Wong, and P. H. J. Chong, "Location-Based Routing Protocols for Wireless Sensor Networks: A Survey," Wirel. Sens. Netw., vol. 9, no. 1, pp. 25–72, 2017.

[10] M. ARSHAD, M. Y. AALSALEM, and F. A. SIDDIQUI, "Energy efficient cluster head selection algorithm in mobile wireless sensor networks," J. Eng. Sci. Technol., vol. 9, no. 6, pp. 728–746, 2014.

[11] M. Di Francesco, S. K. Das, and G. Anastasi, "Data Collection in Wireless Sensor Networks with Mobile Elements: A Survey," ACM Trans. Sen. Netw., vol. 8, no. 1, p. 7:1-7:31, 2011.

[12] I. Mahgoub and M. Ilyas, Sensor Network Protocols. CRC Press, 2016.

[13] A. Norouzi, "An Integrative Comparison of Energy Efficient Routing Protocols in Wireless Sensor Network," Wirel. Sens. Netw., vol. 4, no. 3, pp. 65–75, 2012.

[14] A. Bokare, M. and Ralegaonkar, "Wireless Sensor Network," Int. J. Comput. Eng. Sci., vol. 2, no. 3, pp. 55–61, 2012.

[15] D. Goyal and M. R. Tripathy, "Routing Protocols in Wireless Sensor Networks: A Survey," 2012 Second Int. Conf. Adv. Comput. Commun. Technol., pp. 474–480, 2012.

[16] Muhammad Arshad, "Mobile data collector based routing protocol for wireless sensor networks," Sci. Res. Essays, vol. 6, no. 29, Nov. 2011.

[17] M. M. Chitlange, V. S. Deshpande, and S. Member, "Effect of node density on Congestion in WSN," Int. Conf. Pervasive Comput., vol. 0, no. c, pp. 4–6, 2015.

[18] F. A. Siddqui, M. Saeed, J. R. Khan, M. Arshad, and N. Touheed, "WSN Multi-Tier MDC GRID based Routing Protocol to mitigate Financial Impact," IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 17, no. 11, pp. 20–27, 2017.

[19] A. Salim, W. Osamy, and A. M. Khedr, "IBLEACH: Intra-balanced LEACH protocol for wireless sensor networks," Wirel. Networks, vol. 20, no. 6, pp. 1515–1525, 2014.

[20] P. K. VARSHNEY, G. S. AGRAWAL, and S. K. SHARMA, "Relative Performance Analysis of Proactive Routing Protocols in Wireless Ad hoc Networks using Varying Node Density," Invertis J. Sci. Technol., vol. 9, no. 3, pp. 1–9, 2016.

[21] D. Elmazi, M. Cuka, T. Oda, M. Ikeda, and L. Barolli, "Effect of Node Density on Actor Selection in WSANs : A Comparison Study for Two RE," in 31st International Conference on Advanced Information Networking and Applications Effect, 2017, pp. 865–871.

# IMouse: Eyes Gesture Control System

Syed Muhammad Tahir Saleem[1], Sammat Fareed[2], Farzana Bibi[3], Arsalan Khan[4], Shahzad Gohar[5]
Department of Computer Science and Software Engineering
Al-hamd Islamic University, Islamabad, Pakistan

Hafiz Hamza Ashraf[6]
Department of Computer Science and Software
Virtual University, Islamabad,
Pakistan

***Abstract*—A high number of people, affected with neuro-locomotor disabilities or those paralyzed by injury cannot use computers for basic tasks such as sending or receiving messages, browsing the internet, watch their favorite TV show or movies. Through a previous research study, it was concluded that eyes are an excellent candidate for ubiquitous computing since they move anyway during interaction with computing machinery. Using this underlying information from eye movements could allow bringing the use of computers back to such patients. For this purpose, we propose an imouse gesture control system which is completely operated by human eyes only. The purpose of this work is to design an open-source generic eye-gesture control system that can effectively track eye-movements and enable the user to perform actions mapped to specific eye movements/gestures by using computer webcam. It detects the pupil from the user's face and then tracks its movements. It needs to be accurate in real-time so that the user is able to use it like other every-day devices with comfort.**

***Keywords*—IMouse; eyes gesture control system; eye tracking systems; mouse cursor; eye mouse; webcam; eye movement**

## I. INTRODUCTION

Many people nowadays are falling victim to diseases that impair them physically, like Paraplegia, as a result of which the person is unable to use his body from neck down. In most of the OECD (Organization for Economic Co-operation and Development) countries, females are more affected by disabilities than males [1]. The only organ that can generate different actions is their eyes. In 7 billion population**,** the total of 518 million people reported having a disability in Census 2011. Currently, in Feb 7, 2018, around 10% (about 650 million) of the world's population individuals live with disabilities.

A considerable portion of people affected with Amyotrophic lateral Sclerosis (ALS) [3], or those paralyzed are not able to use computers for basic day to day tasks. Even when it comes to eating they need help from another individual to feed them. These individuals need assistance for their day to day activities. Currently, individuals with disabilities usually type on the computer keyboard by holding long sticks in their mouth. The technique that we present, will help handicaps to be independent in their lives. It will give them a chance to entertain, socialize and work in their lives.

Creative and advance techniques of HCI are being produced rapidly. Many experts are actively working in this research field. Human eyes bear a lot of information which can be obtained and can be utilized in various applications [2] (i.e. interacting with Computers). Eye gesture shows an individual's point of interest. Eye gesture tracking aims to keep track of gestures from a human eye. By capturing Eye movements and using it as control signals to enable direct interaction with interfaces without requiring input from the keyboard or mouse.

Existing computer input devices like a mouse, keyboard, and other type of input devices have been used for interaction with digital instruments. Individuals with disabilities cannot use these computer input devices by themselves. In this research work, a computer input device that is controlled only by human eyes is developed for individuals suffering from disabilities and also for wearable computing [4]. Furthermore, such information could be used to produce necessary outputs for controlling a computer like moving commercially available robotic machinery such as the robotic arm or wheelchairs to enable these patients to feed themselves. This will physically enable them and make them contributing members of the society.

The purpose of this research is to explore and improve upon existing avenues in the eye gesture tracking system. Particularly those areas which can help physically disable individuals, enabling them to use computers and programmable controlled systems. Thus, such individuals could still take on their responsibilities, improve the quality of their lives and continue with their day to day tasks often without the need for a helping hand. In present times, most eye tracking systems utilizes the use of real-time video-based tracking of the pupil. We have adopted the same technique, technologies and improved upon them developing a more robust and accurate system. We used a high-definition, small, portable Microsoft LifeCam HD-6000 this is easily available at low cost. This camera can easily attach with any computer or laptop through the USB port.

### *A. Proposed System*

The eye gesture control system directly interacts with the vision of the human eyes and then controls the system. Eye gesture, a real-time gesture assurance programming which controls a computer mouse cursor by using the user's eye gestures [5]. Only requirement to operate the imouse system is, individuals having at least one eye with good vision and ability to control the computer. Its users can be grown-ups and youngsters with cerebral paralysis, spinal rope wounds, mental wounds, ALS, different sclerosis, Brainstem strokes, and so on. Eye gesture control system can be utilized as a part of homes, workplaces, schools, healing centers, and long-haul mind offices. By looking at the control of a system that is display on a screen, a person can run a computer software,

operate a computer mouse, and access the internet and also email.

### B. Advantages of the Purposed System

- Hands-free mouse cursor control system.
- Facilitating the incapacitated to use computers.
- Mouse pointer control through eye movements.
- Real time eye tracking and eye gaze estimation is achieved through eye based human computer interaction provide.
- Simulating mouse functions, performing different mouse functions such as left click, right click, double click and so on using their eyes.

This paper is organized as follows: Section II elaborates the literature we studied, Section III presents questions regarding this research, Section IV describes the methodology of the proposed system, Section V presents various Tools and techniques that are utilized in our system, Section VI illustrate the algorithms that are used in the proposed system, Section VII describes the design of the proposed system, Section VIII presents implementation, results and tests for the proposed system, Section IX presents contributions of this research work, Section X is the conclusion of our system and finally the last Section XI shows recommendations and future work for the proposed system.

## II. Literature Review

The literature was studied to address the aims, understanding of the research area, focus on the research questions, planning of the data collection approach, clarification of the meaning of the terms and proper identification of the framework. The most important task was to understand the research domain in which eyes detection and cursor movement of a mouse is involved.

Going through the literature, the focus was on how to develop a system which can fulfill the needs of physically impaired individuals and the system should be very easy to understand.

A group over at MIT [6] has created a system titled “The sixth sense”, the system aims to enhance human-computer interaction by using gestures from the hands and eyes. The entire system is mountable on the user’s head, so that it can be projected on to smooth surfaces (like walls) and used anywhere in the world. The problem is that, it doesn’t provide enhanced assistance and accessibility to the disabled nor does it produce a system that can interact with other compatible devices.

Though Drewes, Heiko (2009) [7], presents a comprehensive overview, it was noted that most algorithms needed further refinement as they took tedious and long-winded approaches to calibration. A no-nonsense, agile approach was defined Schmidt, Jochen [8], in using structure from motion algorithm for human-computer interaction. This was exploited by Kassner, Moritz Philipp, and William Rhoades Patera [9], in using the same sfm (structure from motion) algorithm, optimizing it and extending its use as an efficient algorithm for pupil tracking. In order to achieve this goal, they developed a framework by the name of PUPIL, to critically inquire the relationship between a human subject and space to visualize this unique spatial experience and to enable its use for gaze gesture tracking.

In 2018 [10], an eye tracking algorithm based on Hough transform was developed. This system detects the face and eyes of a person. It uses a webcam to detect user’s face and eyes. The system is based on Matlab. The issue in this system is of real-time tracking and time-speed issue. The system is quite slow [11] and it needs a high-quality computer system to work properly which is costly.

In 2017 a better system was introduced by the authors. This system is developed for the paralytic patients. This system uses webcam through MATLAB and moves the mouse cursor by using the pupil of a person [12], [13]. The issue in this system is that it takes a lot of time in detecting the pupil of a person. It uses a lot of algorithms and techniques to detect the pupil.

In 2016 [14] a Vision-based wearable eye-gaze tracking system was introduced. This system works using a high infrared camera. It detects the eyes of the person through the infrared cam. The issue in this system was that it is slow and costly [15].

In 2015 [16], a Pupil center coordinate detection using the circular Hough transform technique was introduced. In this system, the webcam uses Hough Transform Techniques to detect the pupil of a person [17]. The issue in this system was that it takes a lot of time and is not a real-time system. It first captures the body after that, it move to face then eyes and finally to the pupil taking a lot of time.

In 2014, a face and eye-controlled system were developed which were based on MATLAB [18], [19]. It uses a webcam to control the mouse by eye and face movement. The issue in this system is that this system only works in a few centimeter radius.

In 2013 [20], a system was developed which used eye tracking system, this system is based on the pictogram selection. It uses different eye-tracking techniques to make the system reliable. The issue in this system is that if any liquid is found in eyes, it will not work. Like female use eyeliner or mascara in their eyes, so the system stops working in those situations.

### A. The Human Eye

The eye of a human [21], works on a two-lens system in a fluid called the vitreous humor, to project the waves of light from different objects in the world onto the retina exterior. Fig. 1 gives the structure of the human eye with the various aspects labeled. There is an area on the retinal surface of the eye that is thickly pressed with cones in the extent to poles. Once the rays pass through the lens, they cast on the retinal surface after traveling through the vitreous humor. The fovea is measured as 1/4000th of the retinal surface area [22] or a rawboned distance of between 0.3 degrees and 2 degrees. This zone is called fovea and appears as a minor yellow segment on the retinal outside.

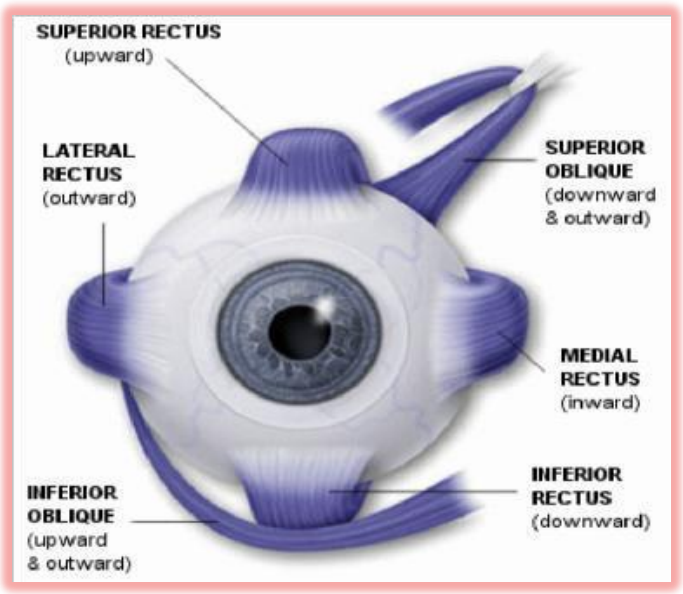



Fig. 1. Structure of the Eye

Fig. 1 shows the structure of a human eye.

The retinal surface contains two sorts of photoreceptors, bars and cones [23]. The fovea is thickly stuffed with cones, roughly 161,900 for each square millimeter, taking into account high-determination shading vision. The composition of a retinal exterior demonstrates to us that around is just a little portion of our visual field that we can resolve in high resolution [24]. The rest of the retina isn't visually impaired as the separation increments, the fovea thickness of pinecones and visual sharpness are enormously decreased.

The first [24], optical element is the Cornea, to be precise, a thin layer of tear fluid that covers the curved corneal surface. The encompassing zone is populated by bars, thickly pressed around the fovea. One function of the aperture is to control how much light can pass into the lens or camera. This gap can revolute in the estimate, becoming bigger – expansion in little light circumstances to permit all the further.

The eye focal point can change, fit as a fiddle or distort keeping in mind the end goal to center around light originating from various depths [24]. The retinal surface is secured with photoreceptor cells, so that the density of the light should be such that it is not too high that it affects the eye and cause the eye lens to contract and also it should not be too low that the camera becomes unable to track the movements of the pupil.

### *B. Analysis of Literature Review*

TABLE I. ANALYSIS OF LITERATURE REVIEW

| Years | Papers | Tools | Issues |
|---|---|---|---|
| 2018 | [1,2] | webcam Matlab Camera Circular Hough Transform | Iris tracking is slow. Web camera connectivity issue. Slow capturing during the video. |
| 2017 | [1,2] | Optical image through the webcam. HD camera using Matlab Viola-Jones, Kanade-Lucas-Tomasi (KLT) algorithms. | Less pupil detecting (KLT) algorithms fail in this paper Need high ram minimum 8 GB. |
| 2016 | [1,2] | A webcam mounted on the glasses. Monte Carlo approach is used. Wavelets Transform (WT). AdaBoost as a machine learning algorithm. | Eye-gaze tracking system with a high-cost webcam. Ada boost algorithm |
| 2015 | [1,2] | infrared video camera MATLAB Circular Hough transforms technique. Infrared camera. MATLAB. | cursor stability higher noise pupil detection algorithm error |
| 2014 | [1,2] | Optical axis. LabVIEW application MATLAB. Web camera. | navigate the mouse pointer, noise of the EOG signals The Distance between users and camera error. Real-time controlling |
| 2013 | [1,2] | Pictogram selection is performed. Eye-tracking techniques, Webcam, MATLAB. Iris tracking, MATLAB. | Any liquid on eyelashes affects the algorithm performance. Stop working when image quality is low. Stop working when the external light is low. |

TABLE I shows analysis of literature that we studied.

## III. RESEARCH QUESTION

TABLE II. RESEARCH QUESTIONS

| Sr.no | Research Questions | Motivation |
|---|---|---|
| 1 | What are the benefits of Such system? | This question will elaborate on the pros and cons of eyes gesture control System |
| 2 | What are the research Contributions? | This research motivation is to provide a complete system for disabled people to control computers with eye gestures. |

TABLE II shows questions related to this research work.

## IV. Methodology

It falls in the field of Human-Computer Interaction (HCI) and shows that by improving upon existing open source frameworks utilized for the purpose of Computer Vision and HCI, a cheap eye tracking solution can be produced for the benefit of disabled patients. Fig. 2 shows the system overview and model of the system.

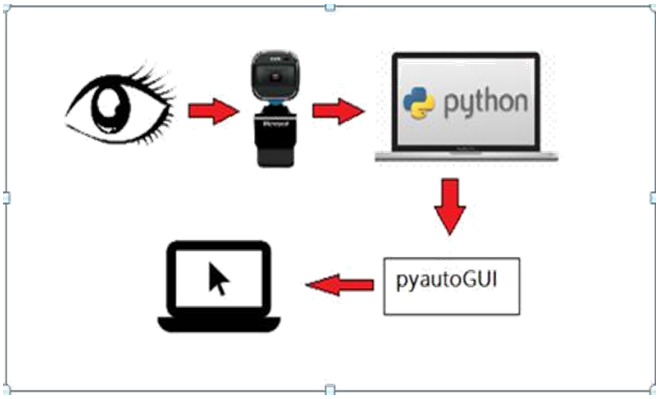



Fig. 2. Block Overview of the Designed System

The System prototype takes input from a camera and recognizes the user's pupil and tracks it in real-time [25]. This "tracking" information can then be used by computers or micro-controllers to perform various tasks, some of these tasks that the project aims to achieve is to track the pupil-movement [26] and then store that tracked eye movement to control the mouse pointer of a computer, so that someone with a disability like say Amyotrophic Lateral Sclerosis can use, to communicate with others.

It includes a carefully positioned high definition web camera, an open-platform and easy to install software module compatible with all window-based operating systems and easily installed on all modern laptops or desktop personal computers. This system can be thought of as a holistic transition from the concept to design to proof of concept. It consists of part research paper implementation and part working with the open-source community on designing and then building the prototype, all the while ensuring that only open-source, cheap, readily available and commercially off the shelf (COTS) products are used.

### A. Aims and Objective

*1) Provide a cheap eye-tracking system.*

*2) To control the cursor of a computer with eyes.*

*3) Allow physically disabled people to use computers.*

*4) To control a computer and communicate with other systems.*

*5) To provide a real-time accurate eyes gesture control system.*

*6) To provide a hand free mouse control system.*

*7) To provide a complete generic eye-gesture mouse control system.*

*8) To provide a complete wire free mouse control system.*

*9) Easy to control cursor movement of a mouse.*

## V. Tools And Techniques

### A. Python

Python, an average unique programming language, is progressively utilized as a part of numerous application areas. Dynamic features in python enable programmers to change code at runtime [27]. Some unique features, for example dynamic compose checking, have a functioning impact in performance. Along these lines dynamic component code is frequently changed to provide efficient programming development.

Some particular features of python are as per the following:

- Deciphered programming. As opposed to e.g. C or FORTRAN, one does not arrange python code before executing it. Likewise, Python can be utilized intelligently: numerous python translators are accessible, from which contents can be executed.
- A free programming release under an open-source permit: Python can be utilized and disseminated complimentary.
- An extremely discernable language with the clear non-verbose linguistic structure
- A language for which an expansive assortment of top-notch bundles are accessible for different applications, from web structures to logical processing.

In our system we use various packages of python which are discussed below.

*1) NumPy*

NumPy [28], is a low-level library written in C++ (and FORTRAN) for exceptional state scientific capacities. NumPy cunningly beats the issue of running slower algorithms on Python by utilizing multidimensional model and structures that works on clusters. Any algorithm would then be able to be communicated as a structure of models, enabling the algorithm to run rapidly.

*2) Scipy*

Scipy [28], is a library that utilizes NumPy for more scientific work. SciPy utilizes NumPy model as the essential information structure, and accompanies modules for different tasks in logical programming, including straight variable based math, joining (analytics), normal differential condition settling and flag processing (eyes signals preparing through webcam)), straight variable based math, joining (analytics), normal differential condition settling and flag processing (eyes signals preparing through webcam)).

*3) OpenCV*

At first all the new modules ought to be produced independently, and distributed in the opencv_contrib archive [29]. Afterward, when the module develops and picks up estimations, it is moved to the focal OpenCV archive and in this manner they should not be freed as a piece of control for OpenCV [30] dispersion, since the library keeps up parallel similarity, and attempts to give fair execution and stability improvements.

*4) PyautoGUI*

The reason [31], for PyAutoGUI, is to give a cross-stage Python module for GUI automation for developers. The API is intended to be as basic as conceivable with reasonable errors. PyAutoGUI can simulate moving the mouse, clicking the mouse, dragging with the mouse, pressing keys, pressing and holding keys, and pressing keyboard hotkey combinations. On Windows, PyAutoGUI has no dependencies (other than Pillow and some other modules, which are installed by pip along with PyAutoGUI). It does not need the pywiin32 module installed since it uses Python's type modules.

## VI. Algorithms

In eyes, gesture control system uses two main different algorithms.

### A. Harr-cascade Algorithm

Haar-cascade is a protest detection algorithm used to find faces, people on foot, items, outward appearances in a picture and primarily utilized for face detection. In Haar-course, the framework is given a few quantities of constructive pictures (like appearances of changed people at various foundations) and pessimistic (pictures that do not face but rather can be whatever else like the seat, table and divider and so forth).

Human face detection [32], noted to be a testing problem in the zones of image handling and patter acknowledgment. Another algorithm for human face detection by crude Haar course algorithm, joined with three extra frail classifiers is presented in this work. The three powerless classifiers depend on tint histogram coordinating, detecting a mouth and detecting eyes. To start with, pictures of individuals are handled by a crude Haar course classifier, without false human face dismissal (low rate of false negative), yet with few false acknowledgment (false positive). Besides, a powerless classifier a view of face skin tint histogram coordinating is connected and a lion's share of non-human countenances are expelled, to dispose of falsely acknowledged non-human faces. Next, another powerless classifier a view of eyes detection is attached and other lingering non-human appearances are resolved and dismissed.

Finally, a mouth detection activity is used for the rest of the non-human countenances and the possibility of false positive is additionally diminished. With the assistance of OpenCV, test starts to run on pictures of individuals under impediments and understanding and some level of preface and turns, in both preparing set and test set demonstrate that the proposed algorithm is successful and accomplishes best in class execution. Moreover, it is proficient as a result of its ease and effortlessness of execution.

### B. Hough Transform Algorithm

The Hough transform [33], algorithm is a worldwide strategy for discovering straight lines, picture examination, computer vision, and advanced imaging preparing. The reason for this method is to discover flawed occurrences of courses classified, a specific lesson of figures through a polling technique. The Hough transforms are connected for the inquiry of a typical focus of round or mostly roundabout segments exhibit in a picture. The effectiveness has been significantly improved by a persistent refresh of a rundown of a polling focuses, in conjunction with the advancement of the aggregator size and position. The technique was executed as a module for the logical open source picture handling bundle Image. Although at first intended for X-beam, diffraction examination, various different applications are cited in various other logical fields, in picture estimation strategies, machine vision and biometric, i.e. for iris restriction.

## VII. System Design

Imouse system is design in python and following modules of python are imported for working of this system.

- **NumPy:** is a Python extension module. It provides rapid and efficient operations on arrays of compatible data.
- **Scipy:** an open-source Python library which is used for technical and scientific computing.
- **OpenCV:** is a library of programming functions mainly focused on real-time computer vision.
- **PyautoGUI:** is a cross-platform GUI automation module that works on Python. In this, you can control the mouse and keyboard as well as you can perform basic image recognition to automate tasks on your computer.

### A. Use Case Diagrams

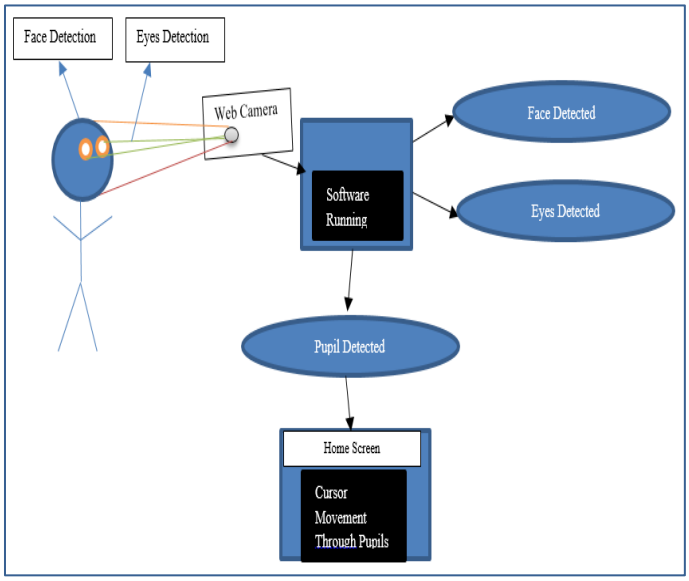



Fig. 3. System use case diagram

In 0 shows use case diagram; the system accomplishes to have the following steps.

*1) Software running.*

*1) Open the webcam on the laptop and show the image of a person.*

*2) Face detection action is performed.*

*3) The system detects the eyes of a person.*

*4) After the above action system move on to the next operation.*

*5) In the next step the system detects eyes and face through webcam of a laptop.*

*6) Pupils detected now a person can control mouse cursor through his eye movements. Curosr mevement is shown on the home screen of the computer.*

*B. Activity Diagram*

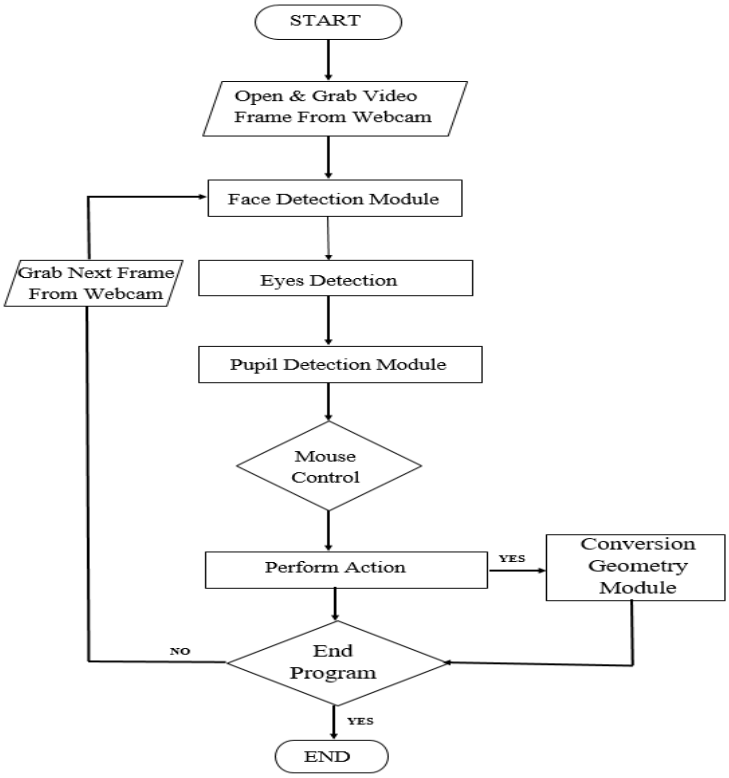



Fig. 4. Activity Diagram

Fig. 4 shows the sequence in which activities are held in the system.

Fig. 4 shows how the system operates. The following are the steps from start to end of the iMouse software.

*1) Open the webcam and capture the video.*
*2) The system performs an action and detects the face.*
*3) The system performs eyes detection.*
*4) The system detects the eye pupil.*
*5) With only the image of face from the webcam, the system will locate the eyes and perform geometry translations.*

*6) Perform action*

Mouse Control: detects a gesture, moves the mouse cursor and translate the coordinates to the user's screen. Then perform the following the action.

*1) Scroll vertical*
*2) Scroll horizontal*
*3) Scroll diagonal*
*4) Repeating the whole system cycle and the system pause or system end.*

*C. System Sequence Diagram*

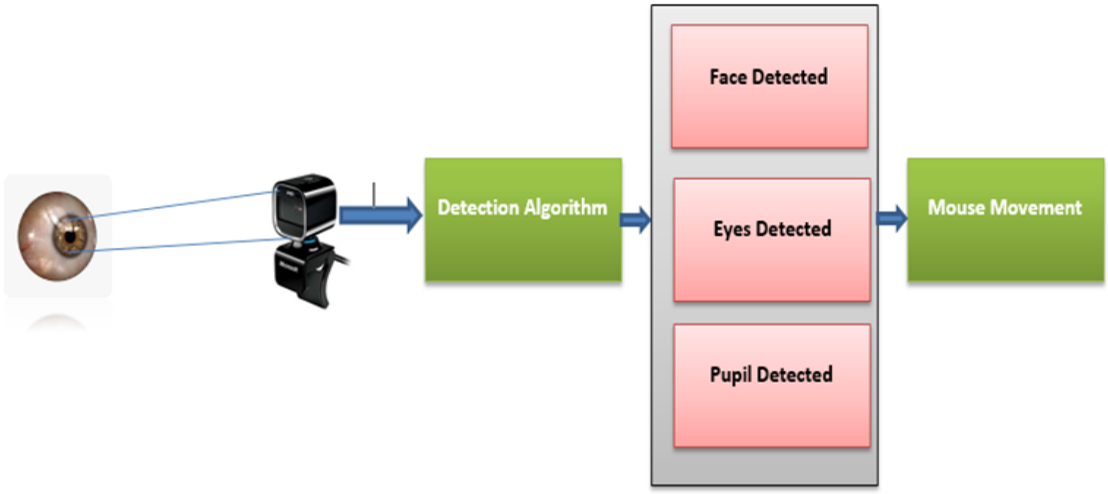



Fig. 5. Sequence Diagram

Fig. 5 show the sequence in which the user interacts with the system.

In Fig. 5 the system sequence diagram, elaborates the six basic modules of our system. In the first module, the system detects the pupil of a person through the webcam using detection algorithms. Then the system detects the face. After that the system detects and captures the eyes. Then the system detects the pupils. In the last module, the system starts moving the mouse cursor by tracking pupil movements.

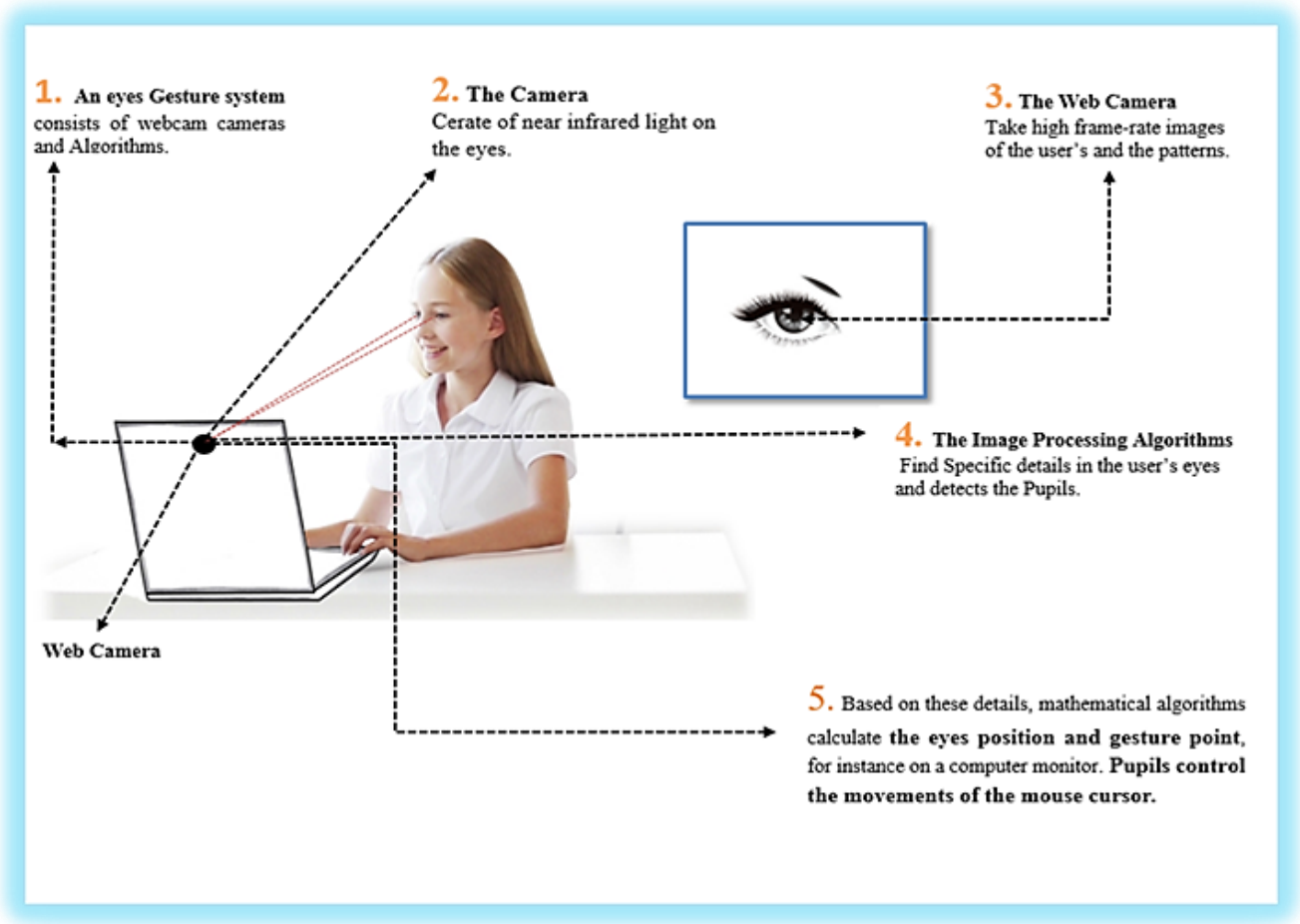



Fig. 6. System Functionality Diagram

Fig. 6 Shows, how the system functions.

## VIII. IMPLEMENTATION

### A. Work On System

In this approach, a system is developed using python ,named imouse. First it will open the camera and start capturing video, the video consists of frames, it will select a frame and then convert it to a grayscale image as it converts images to binary form and it becomes easy to detect objects from that image. It will then detect the face from it using Haar-cascade. Haar-cascade is a cascade function which is trained from a lot of positive and negative images and then it detects objects from other images. It will detect the face from the given frame and then crop that frame and will pass on for further processing. Then from the obtained frame it will detect eyes using Haar-cascade and crop that from. Haar-cascade that detects eye is called eye-cascade. A four variable array (i.e supported by numpy) is defined that will give us four values at a time x, y, w, and h. x and y are for the starting point from where the camera detects eyes, w is for the width of eye detected and h gives us the height. By using these variables we draw a rectangle around the eye and then crop that image as shown in Fig. 7. The rectangle starts from the point x and ends at ends at x+w in the horizontal direction, and start from y and ends at y+h in the vertical direction.

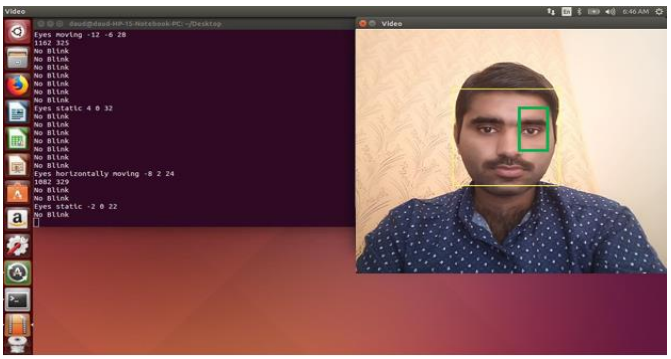



Fig. 7. Output of the system

The Output is shown in the Fig. 7. The resulting image will be of a very small size, the program will blur it to make it clear for our further processing. After that, it will detect the edges of the pupil with the help of the Hough Transform and then draw a circle around the pupil of the eye. But, sometimes the camera recognizes dark circles near the eyes as a pupil.

To overcome this problem, the system detects the dark spots that are only circular in shape and are found in the center of the rectangle drawn. This way the camera will only detect the pupil of the eye and track its movements. Define two variables x and y using pyautogui for cursor movement. Y is for the vertical movement, x is for horizontal movement of the mouse. At the start, both of them are given a random value so when the code will run the mouse will start moving from that point. To detect the movement of the eye, note the position of the pupil from two different frames and check if the distance is increased. The human eye is not still, a little bit of movement is always there. To overcome this problem a check is performed that if the difference in the position of the eye is less than 5 pixels then consider eye is still.

If difference in the horizontal position of the eye is greater than 5 pixels and in vertical direction less than 5 pixels then this means the eye is moving in the horizontal direction, If difference in the vertical position of the eye is greater than 5 pixels and in horizontal direction less than 5 pixels then this means the eye is moving in the horizontal direction and If difference in both horizontal and vertical position of the eye is greater than 5 pixels then this means the eye is moving diagonally. When we move the eye vertically, the mouse cursor will move in the upper/down direction and when we will move our eye horizontal then the cursor will move right/left and when we move our eye in diagonal then the cursor will move in diagonal.

### B. The System Analysis & Evaluation

We start by presenting the results of the working contour detecting algorithm that works by using Haar-cascade functions and Hough Transform.

### C. Harr-cascade Algorithm

In this algorithm, the system performs two functions detect the face and eyes of the persons Fig. 8 illustrate the face-cascade function.

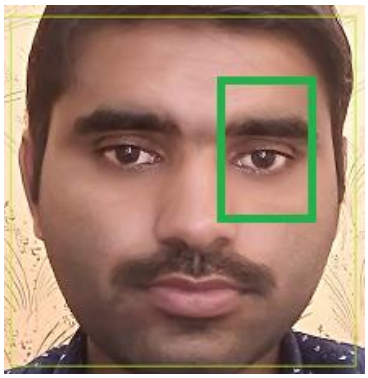


Fig. 8. Face-Cascade

It first detects the face of the user from an image using face-cascade. It draws a box around the face and crops the Image for further processing. After we extract the face from the image it detects the eyes of the user from that image. As shown in the Fig. 9 below.

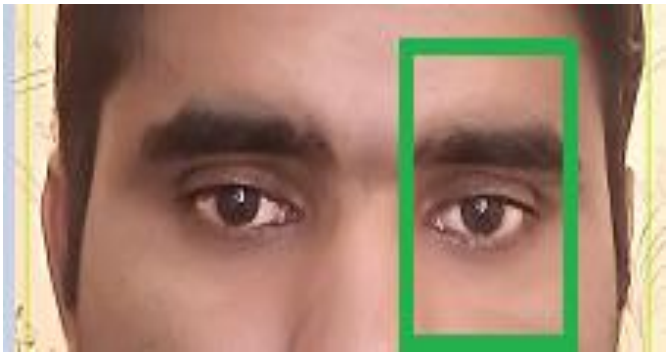


Fig. 9. Eye-Cascade

As both eyes move at the same time, so the tracking can be done by analyzing the movements of only one eye. This program picks left eye (right eye in the image as the image

flips in a camera) of the user and tracks its movement. It draws a box around it crops this image.

### D. Hough Transform Algorithm

In Hough Transform pupil is detected from this cropped image and a circle is drawn around it to track its movements. As you can see in the Fig. 10 below that this system detects the pupil and ignores other dark spots and tracks only the pupil of the eye.

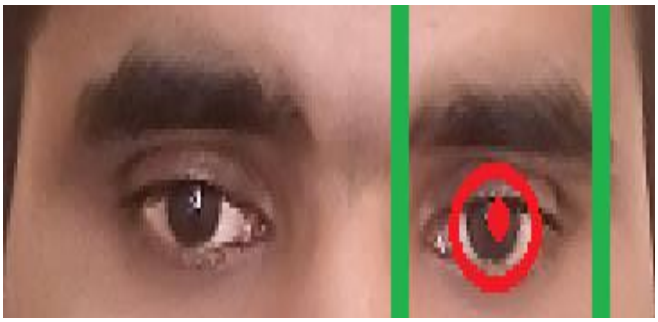


Fig. 10. Hough-Transform

Eye coordinates are then located in each frame. If there is a change in the coordinates of the eyes then the eye is considered to be moving otherwise the eye is considered still. The delay of all this processing is less than one second. Which means this project can be used in real life situations.

### E. Histograms for Mouse vs Eye

The histograms for mouse vs. eye control are shown below, please note that the readings were taken for ideal lighting conditions and may vary in other situations if pupil detection is erroneous. This has been illustrated in the 0and 0.

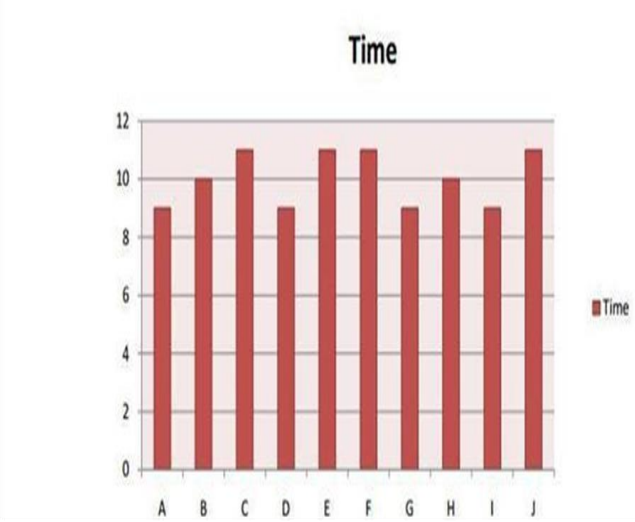



Fig. 11. Timings of actions A through J for mouse control

0 shows mouse control action time.


Fig. 12. Timings of actions A through J for eye control

Do note that in 0 and in 0 the results are rounded off to the nearest millisecond. We see that once proficiency has been achieved, the user can perform the same tasks in approximately the same amount of time as can be done with the mouse.

After tracking of the pupil, the coordinates are stored in the variables and mouse cursor is moved according to change in these values and blink is detected. The data transmitted can be seen in Fig. 13:

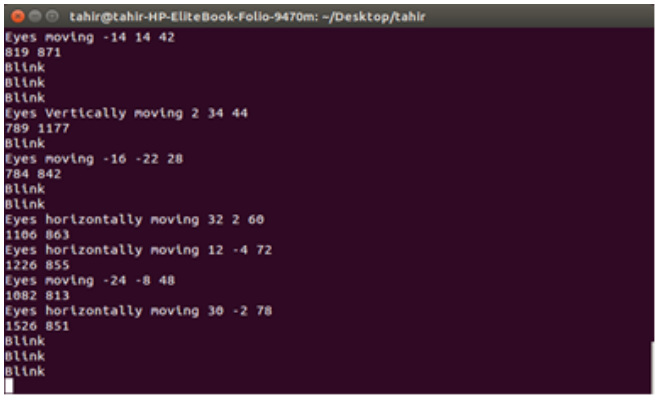



Fig. 13. Pupil coordinate values and blink detection

Note that during a blink, the transmitted coordinates are null, signifying the absence of a pupil, this fact is exploited in the click gesture, which has been defined in two ways. One, where the user can perform a click by staring at a point for a couple of seconds, or two, blink for an abnormal duration.

## IX. Result and Discussion (RQ1)

Individuals having a disability, in most cases they are not able to perform basic tasks using computers, in this situation the system will detect the pupil of a person automatically. It detects the pupil of a person and starts a connection between the mouse and eye gestures. The mouse cursor will start move as pupil move. The clicking of the mouse is based on the eye blinking.

For our system, we conduct some tests by changing the eyes position and adding liquid to the eyes. We also test this system through short and long distances. The results were different in term of distance. The result can be seen in TABLE III.

TABLE III. DISTANCE COVERAGE USING IMOUSE SYSTEM

| Distance | Max | Min | Average |
|---|---|---|---|
| 10cm | 6.458533cm | 4.50689cm | 1.0965423cm |
| 20cm | 6.890525cm | 4.90598cm | 0.5898252cm |
| 30cm | 7.588226cm | 5.25595cm | 0.4281392cm |
| 40cm | 7.895558cm | 5.59855cm | 0.3373527cm |
| 50cm | 8.063359cm | 5.95268cm | 0.2803207cm |

TABLE III shows the results from various distances. We measure the results at different distances. The maximum distance of the system is 50cm, our system works properly and in real-time at that distance. At 10cm the system is working very fast, its max value at this point is 6.458533cm while its min value is 4.50689cm. The mouse is moving at an average of 1.0965423cm.

At 20cm the average value of the mouse movement is 0.5898252cm. At that point, the mouse is working properly. At 30cm the average value is 0.4281392cm and mouse is working properly. As we getting far from the system the mouse speed is reducing. At 50cm the speed of mouse is little low but it works slowly at 50cm.

## X. RESEARCH CONTRIBUTION (RQ2)

Many systems exist that are using eye pupil to move the mouse cursor and also some embedded systems. We aim to implement such a system which is reliable and also have better performance. We developed a complete hand free mouse control system. The system uses webcam of a computer and tracks the eye pupil of a person in real-time to move mouse cursor. The system uses a webcam and detects the face of a person after detecting face it moves toward the eye than it detects the eye pupil. The mouse is moving through pupil movement using webcam infrared. The system provides help to the handicapped and other diseases patients who are not able to use their hands properly.

The recommended system will provide many additional performances than the prevailing systems.

- Suitable for all handicapped persons.

According to the survey, almost 17.7% of the world population is handicapped. Our system is specially designed for these type of individuals, so they can use computer systems by themselves.

- Directly detects the pupil of the user.

The webcam directly targets the pupil of the user. It focuses directly on the pupil of the user. In other systems, the webcam did not detect the pupil of the user

- Real-time pupil tracking.

The pupil of an individual is tracked in the real time. There is no delay while tracking a user's pupil.

- Fast and easy mouse cursor movement.

The movement of the mouse is fast and accurate and all the movement is done in real time.

- Fast clicking of the mouse

The clicking is done fast and immediate. The clicking is done by using pupil through the webcam

## XI. CONCLUSION

As a final review in this research, the system aims to deliver a low-cost eye-tracker that will allow the user to control the mouse cursor of a computer system. The designed system is low cost and efficient, utilizing only a camera (Microsoft LifeCam Hd-6000), software modules coded in C++ and Python Programming language. If required, the spatial field of view history can also be drawn on the world process, showing eye movements and where the user spent a lot of time looking, to tweak the interface or to just retrieve spatial attention data for the purposes mentioned in the "future applications" section. Finally, we note that the project is operable in variable environmental conditions, only a few tweaks in the brightness and contrast setting need to be applied for it to maintain its robustness. This is an impressive feat for such a low-cost eye-tracking system.

## XII. RECOMMENDATIONS FOR FUTURE WORK

### A. Present Work

Eye tracking has become an important field of research recently. Sony is working on developing and incorporating eye tracker in their popular PlayStation platform. A Berlin developer introduced this technology through SensoMotoric Instruments [34]. It uses RED-OEM platform among its different functions which highlights the things you are looking at, enlarging the list of options which helps to simply track and respond to your queries. Oculus Rift is a virtual reality HCI interface which is the brainchild of Kickstarter.com and relies on heavy user gaze data to reposition the field of view to give the illusion of reality.

### B. Potential Future Areas

#### 1) Search Engine Optimization

Search engine optimization (SEO) for websites and businesses can be done effectively by employing eye tracking based information and data analytics. This potential area of research is being spearheaded by Microsoft and Google [35], [36]. Results from Google search engine can be seen in Fig. 14 and in 0.


Fig. 14. Results from google of eye-tracking in Transaction


Fig. 15. Results from Google of eye-tracking in the informational query

*2) Market Research and Advertising Testing*

Eye-tracker can play a major role in marketing. To understand the demands of the customer eye trackers are placed in marker and they determine for how much time the user is starting at the product and what is its reaction to the qualities or price of the of the product. This will help the manufacturers to make products that meet with the customer's requirements and in return, their revenue will increase.

A commercial eye tracker scans and zeros in on which products the customer is scanning and where on those products is his gaze concentrating. An example of a dishwasher is shown in 0 with blue and pink markers showing where the customer's gaze concentrated and the radius of marker depicts the duration for which the customer's gaze was hooked on to that particular position. We can see the widest markers and most markers, in general, are concentrated towards the product label and this gives important information to manufacturers and advertisers alike that they should make the labels visually appealing, among other customer behavior information.

Fig. 16. Customer gazing upon a product, points of visual attention marked by the eye tracker

ACKNOWLEDGEMWNT

We are greatly thankful to our family for their encouragements, support, assistance and their prayers to ALLAH for our success. We would like to thank our Class fellows for this research, for their commitment and inspiration to help us in achieving our research goals. A Special thanks to Majid Abbas for providing us with essential advice and extraordinary guidance.

We further thank Alhamd Islamic University administration and staff for providing us with an opportunity to have quality education and research exposure, to be successful in our life.

Last but certainly not least a special thanks to our mothers, as their moral support and spiritual uplifting has inspired us. This research would not have been completed without their help.

REFERENCES

[1] Disabled persons Ratio [online] Available At: https://www.disabled-world.com/disability/statistics/ (Accessed 2 Jan 2018)

[2] The Eye Writer Project [online] Available At: http://www.eyewriter.org (2009-Present)

[3] Muhammad Usman Ghani, Sarah Chaudhry "Gaze Pointer: A Real Time Mouse Pointer Control Implementation Based On Eye Gaze Tracking"/http://myweb.sabanciuniv.edu/ghani/files/2015/02/GazePointer_INMIC2013.pdf [6 February 2014]

[4] Schmidt, Jochsen, Vogt and Nieman "Calibration–free hand–eye calibration: a structure–from–motion approach." Pattern Recognition. Springer Berlin Heidelberg, 2005. 67-74.) (2005)

[5] Prajakta Tangade, Shital Musale "A Review Paper on Mouse Pointer Movement Using Eye Tracking System and Voice Recognition" http://www.ijeert.org/pdf/v2-i8/20.pdf (08/11/2014)

[6] A Wearable Gestural Interface:www.pranavmistry.com/projects/sixthsense/ (05.04.2017)

[7] Heiko Drewes and Albrecht Schmidt. "Interacting With Computers Using Gaze Gestures" https://link.springer.comm/chapter/10.1007/978-3-540-74800-7_43 (2009)

[8] PUPIL: constructing the space of visual attention. Diss. Massachusetts Institute of Technology, 2012.) Kassner, Phillip, and Patera (2012)

[9] IEEE paper on "Differences in the infrared bright pupil response of human eyes" by Karlene Nguyen, Cindy Wagner, David Koons, Myron Flickner (2009)

[10] "An eye tracking algorithm based on Hough transform" https://ieeexplore.ieee.org/document/8408915/ (12.5.2018)

[11] Eye gaze location detection based on iris tracking with web camera: https://ieeexplore.ieee.org/document/8404314/ (5.7.2018)

[12] "An image-based eye controlled assistive system for paralytic patients" https://ieeexplore.ieee.org/document/8066549/ (2017)

[13] "Eye-ball tracking system for motor-free control of mouse pointer" 'httpps://ieeexplore.ieee.org/document/8299921 (24.3.2017)

[14] EOG signal processing module for medical assistive systems upload 2016:http://ieeexplore.ieee.org/document/7533704/?anchor=disclaim

[15] "The research paper is as follows "https://www.ijarcce.com/upload/2016/march_16/IJARCCE%20182.pdf

[16] "Pupil center coordinates detection using the circular Hough transform technique" https://ieeexplore.ieee.org/document/7248041/ (2015)

[17] Pupil detection algorithms for eye tracking applications https://ieeexplore.ieee.org/document/7342317/ (2015)

[18] EOG-based system for mouse control httpps://ieeexplore.ieee.org/document/6985240 (15.12.2014)

[19] "Face and eye tracking for controlling computer functions" https://ieeexplore.ieee.org/document/6839834/ (2014)

[20] "Eye tracking mouse for human-computer interaction" https://ieeexplore.ieee.org/document/6707244/ (2013)

[21] "The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research" by Charlie Ware and Henry Mikaelian (1987)

[22] Artificial Intelligence for Human Computing: ICMI 2006 and IJCAI 2007 International Workshops Lecture Notes edited by Thomas S. Huang, Anton Nijholt, Maja Pantic, 2006

[23] PUPIL: constructing the space of visual attention. Diss. Massachusetts Institute of Technology, 2012.) Kassner, Phillip, and Patera (2012)

[24] IEEE Paper on "Performance of input devices in FPS target acquisition" by Poika Isokoski, Benoit Martin.

[25] Arsalan Khan, Farzana Bibi, Muhammad Dilshad, Salman Ahmed, Zia Ullah and Haider Ali, "Accident Detection and Smart Rescue System using Android Smartphone with Real-Time Location Tracking" International Journal of Advanced Computer Science and Applications(IJACSA),9(6),2018. http://dx.doi.org/10.14569/IJACSA.2018.090648

[26] Drewes, Heiko. "Eye gaze tracking for human-computer interaction." Diss. lmu, 2010

[27] "Python for scientific computing" [online] Available At: http://www.scipy lectures.org/intro/language/python_language.html [Accessed 4 Feb 2018]

[28] "The Hitchhiker's Guide to Python"[online] Available At: https://docs.python-guide.org/ [ Accessed 25 Feb 2018 ]

[29] OpenCV (n.d) [online] Available At: http://opencv.org/ [Accessed 5Jan 2018]

[30] Github (n.d) opencv/opencv_contrib [online] Available At: https://github.com/opencv/opencv_contrib [Accessed 20 Jan 2018 ]

[31] "Introduction PyautoGUI" [online] Available At : http://pyautogui.readthedocs.io/en/latest/introduction.html [Accessed 10 Jan 2018]

[32] Li Cuimei, Qi Zhiliang, Jia Nan, The Human face detection algorithm via Haar cascade classifier combined with three additional classifiers" https://ieeexplore.ieee.org/document/8265863/ (23.1.2018)

[33] Julien Cauchie Valérie Fiolet Didier Villers "Optimization of an Hough transform algorithm for the search of a center" https://www.sciencedirect.com/science/article/pii/S003132030700316 0

[34] OpenFrameworks [online] Available At: http://www.openframeworks.cc/ [Accessed 25 Dec 2017]

[35] "An Adaptive Algorithm for Precise Pupil Boundary Detection" by Chihan Topel and Cuneyt Akinlar (2012)

[36] "Eye gaze Patterns while Searching vs. Browsing a Web site" by Sav Shrestha & Kelsi

# Calculation of Pressure Loss Coefficients in Combining Flows of a Solar Collector using Artificial Neural Networks

Shahzad Yousaf[1], Imran Shafi[2], Jamil Ahmad[3]
[1,2]Department of Mechanical Engineering, CEME NUST Rawalpindi
[3]Department of Computer Engineering, Abasyn University, Islamabad

***Abstract*—The paper presents a novel technique for determination of loss coefficients due to pressure by use of artificial neural network (ANN) in tee junctions. Geometry and flow parameters are feed into ANN as the inputs for purpose of training the network. Efficacy of the network is demonstrated by comparison of the ANN and experimentally obtained pressure loss coefficients for combining flows in a Tee Junction. Reynolds numbers ranging from 200 to 14000 and discharge ratios varying from minimum to maximum flow for calculation of pressure loss coefficients have been used. Pressure loss coefficients calculated using ANN is compared to the models from literature used in junction flows. The results achieved after the application of ANN agrees reasonably to the experimental values.**

***Keywords*—Artificial neural network; pressure loss coefficients for solar collector; combining flow***

## I. Introduction

Due to industrialization and automation in almost every walk of life, energy consumption is on the rise and scientists are finding new alternatives to conserve the fossils reservoirs. In this regard, the sun is the sole largest source of energy that can provide enough energy to mankind without being explored like conventional energy. If only 0.1% of solar energy reaching earth at a very low efficiency of 10% converts into useful energy, it has the potential to generate about 4 times the mankind electricity demands [1]. Solar energy based absorption refrigeration system uses a solar collector to harness the freely available solar energy to heat the working fluid. The hot working fluid circulates in the thermal loop of the cycle and runs absorption/desorption cycles. The efficiency of the solar refrigeration is highly dependent upon the efficiency of solar collector [2]. Efficiency of the system can be increased considerably through an efficient solar collector design. Absorption refrigeration and solar collector systems are shown in Fig. 1 and 2, respectively.

The solar collector has a separate manifold for inflow and outflow. A side pipe identified in this work as riser connects both these manifolds in a parallel way such that an array of pipes is made. In a solar collector, loss of pressure occurs when working fluid flows through its pipes. These losses can be classified as minor and major losses. The major losses occur due to frictional resistance in the flow of fluid while the minor losses occur due to change in momentum of the fluid due to some restrictions in its way. Minor losses are not always lower than the major losses e.g. in the event of solar collector where riser tubes are connected in an array of few hundred tubes adding much higher resistances to flow of the working fluid and hence the momentum loss in the fluid flow occurs. Such situations decrease the efficiency of the solar collectors but proper designing of tee junction may avoid the situation and the system performance increases as a whole.

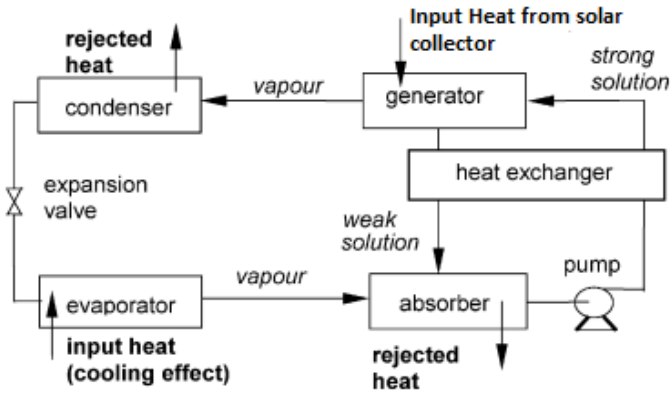



Fig. 1. Solar absorption collector with solar collector as an integral part.

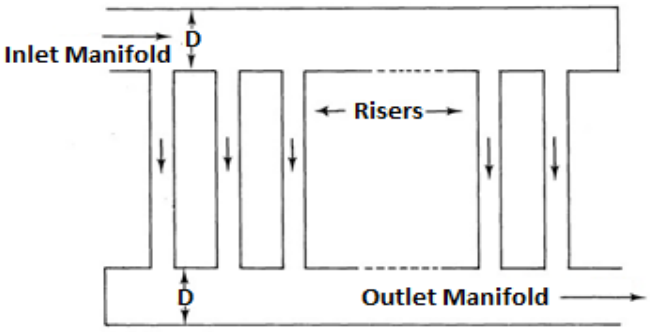



Fig. 2. Geometry of solar collector.

Back Propagation based artificial neural network, due to its guaranteed conversion in solutions of nonlinear equations, is put in use in this work. The multi-input and output nature of ANN and its less computational cost make it first choice for nonlinear equations solution [3].

## II. Literature Review

Pressure loss coefficients (PLC) are ratio of stagnation pressure difference of upstream and downstream pipe in riser/straight pipes. It represents the energy loss in flow between upstream and downstream. Therefore, there will be two loss coefficients, one for riser and second for the straight pipe. Mathematically it is given by the following formula:

$$k_{\mathrm{i}} = \Delta \mathrm{p}_{\mathrm{i}} \left[ \left( \frac{\rho \mathrm{V}^2}{2} \right) \right]^{-1} \quad (1)$$

Whereas subscript $i = st, s$

$\Delta P =$ Loss of stagnation pressure in ith pipe

$\rho =$ Heat transfer fluid density

$V =$ Velocity of flow in combined pipe

$K_{st} =$ PLC in straight pipe

$K_s =$ PLC in riser

In combining flow, energy loss in flow is dependent on flow ratios in riser ($Q_S/Q_C$) and straight pipes ($Q_{st}/Q_c$) [15]. According to the frictionless Bernoulli theorem, the velocity and momentum of working fluid decreases as the fluid approaches the junction making an increase in the. But when the effects of frictions are considered at this point, the pressure decreases and hence pressure loss occurs [4]. In a converging flow, the losses in pressure and momentum occur due to fluid mixing from both the legs of the tee junctions as well as due to 90-degree bend at the junction [5].

Pressure loss coefficients are usually calculated by experiments that require waste of resources [6]. Gardle calculates pressure losses due to the geometry of junctions by considering different discharge ratios in both the legs of the tee junctions by using an empirical formula. The empirical formula is not applicable to all types of flows [7]. Through an experimental study for both the combining and diving flow, Bingham and Blair studies the flow distribution and the pressure losses in a tee junction [8]. They developed a single equation from their experimental data. The disadvantage in work of Bingham is that data for dividing and combining flow has been mixed. Hagers investigate the effects of an area ratio of unity on the flow in a tee junction [9]. Flamang takes into account the effects of lateral angle for junction flow along with the area ratios effects [10]. But satisfactory correlation has not been made to all of the flow types in a junction pipe. For unsteady flow in engine manifolds, modeling have been done successfully by assuming that pressure loss in unsteady and steady flow is same [4]. The system of simultaneous equations has been solved by Weitbrecht *et al.* for flow distribution and pressure loss calculation [11]. For one riser pipe in a solar collector, the system contained 28 nonlinear equations. The work of Jones and Lior does not give many results and not much insight is given about effects of various flow parameters on flow. They investigated pressure losses for tee junctions by developing a discrete model using nonlinear algebraic equations [12]. Bajura and Jones studied different parameters for flow in junctions. Experimental results from their work show that flow distribution in laminar flow differs from that of a high Reynolds number flows [13]. Idelchik and Miller independently studied various parameters like area ratios, discharge ratios for different lateral pipe angles that affects loss coefficients of pressure in flow involving junctions [15], [16]. Both of the studies do not take the effects of Reynolds number into account. On a commercial CFD package, the idea of Weitbrecht *et al.* [11] has been used by Badar *et al.* for the estimation of loss in pressure due to change in momentum and flow distribution in a solar collector. He concludes that nonconformity in results and experiments is mainly due to incapability of the CFD package to capture the minute details of flow and boundary conditions that might not depicts the true nature of flow [14]. A comparative study of pressure loss in tee junction and elbows has been performed by Moujaes on Fluent- a CFD commercial software for a fixed discharge ratio of 0.5. Finding of the investigation is that tee junction offers more resistance to the flow through it as compared to elbow [15]. Bassett *et al* studied tee junction flow in combining and dividing flow. The flow is energized using shock wave for high Reynolds number flow and developed equations for all type of flows in junctions [17]. In [17] the flow reverses as soon as the energizing shock wave diminishes thus affecting the loss coefficients values and also it does not give any specific range of Reynolds number in which the equation holds true. Similarly, Wahab *et al* used SIMPLE algorithm on CFD package to predict the pressure losses in junction flow [18]. The results achieved are higher than literature results for constant diameter pipes. More recently, Bawa determined numerically the total pressure loss in a solar collector [19]. But no information has been presented regarding loss of pressure in the collector due to geometry of the solar collector under consideration.

## III. Proposed Method

We are suggesting a new novel technique, to the best of our knowledge, based upon ANN for evaluation coefficients of pressure loss in combining flow. The ANN evolved after the functioning of the human brain was studied. ANN is nonlinear statistical models. They do not compete in complexity to brain and central nervous system of human. Network is formed when neurons connects to each other through synaptic connections with the ability to convey data through various layers. This network is capable of estimation of energy loss in flow in the form of pressure loss coefficients. These networks are regarded as highly efficient system due to their adaptive nature. In this network, information is transmitted through neurons as by giving them input feed, which is further connected through a specific activation function. Data transmission is done in layers of signals with weights exciting or inhibiting the signal and finally, the output is retrieved in the form of pressure loss coefficients.

ANN performs its function in the following steps (Fig. 3):

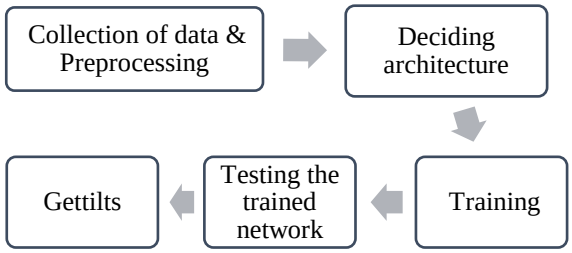



Fig. 3. Design stages of an artificial neural network.

### A. Data Collection and Preprocessing

For training purpose, an artificial neural network needs a set of data. Before the training process begins, data set refinement and preprocessing is done. In the current work, experimental data has been gathered from a tee-junction setup as shown in Fig. 4. Sensors are installed at points a, b and c that records stagnation pressure and from which loss coefficients of pressure are calculated using (1).

Geometry specifications for experimental setup are as under (Table I):

TABLE I. GEOMETERY SPECIFICATION OF TEE JUNCTION SET UP

| Parameter | Specifications |
|---|---|
| Diameter of manifold, D | 16.6 mm |
| Diameter of riser, d | 7.2 mm |
| Entry length of inlet pressure sensor, a | 560 mm |
| Entry length of outlet pressure sensor, b | 990 mm |
| Length of riser pressure sensor from outlet, c | 17 mm |
| Riser angle, α | 90° |

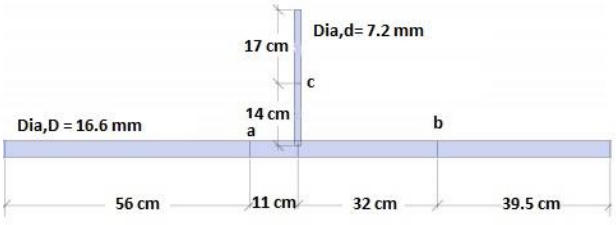



Fig. 4. Geometry specification of tee junction used.

The experimental results obtained are processed for ANN training. It is pertinent to mention that water is used as heat transfer fluid with constant density neglecting the effects of buoyancy.

The 30 features are recorded while performing experiments on the tee junction set up. Among them only relevant features that affect the pressure loss coefficients are selected using Relief algorithms for ANN training [20] as shown in Table II. Feature relevancy obtained after application of relief is shown in Table III.

TABLE II. EXPERIMENTAL CONDITIONS FOR COLLECTION OF DATA

| Factors | ISO 9086 Conditions | Experimental condition |
|---|---|---|
| Temperature range (°C) | 18 -22 | 18 -22 |
| Fluid | Water or mixture of water/glycol (60/40) | Water |

TABLE III. FEATURE RELEVANCY OBTAINED AFTER APPLICATION OF RELIEF

| Name of feature | Score of feature (%) |
|---|---|
| Penetration depth | 30 |
| Mass flow rate | 20 |
| Volume flow rate | 20 |
| Reynolds number | 12.5 |
| Velocity of flow | 12.5 |
| Entry length | 5 |

### B. Network Architecture Selection

Architecture selection for training purpose of the network has been done by hit and trial method because there is no fix rule available for network architecture specifications [21]. This work uses a network architecture of 6-30-2 neurons in input-hidden-output layers.

### C. Training

The selected network is trained using learning rule known as Marquardt-Levenberg back propagation rule. The layers and neurons are interconnected using tan-sigmoid function as an activation function. Training stoppage criterion is also chosen that includes maximum number of epochs, mean square error and time. This training criterion helps in preventing the network from overfitting and thus the ability to predict the unseen conditions improves.

### D. Testing

The network trained is tested upon the unseen data. The efficacy of the network can be seen when the ANN and experimental data are compared to each other for different situations of flow as shown in the next section.

## IV. RESULTS AND DISCUSSION

Reynolds number starts from a very low laminar region (Re =200) to a very turbulent region (Re=14000) for the current work. The discharge ratios range from 0 to 1 for combining flow in tee junction. Much focus has been on the laminar flow region due to the scant literature which has been identified by [5], [12], [13], [16]. For combining flows, the junction has two inlets pipes and one outlet pipe. A number of data points are tested and results are plotted in Fig. 5 for both manifold and riser of tee junction.

Fig. 5 shows the variations in loss coefficients due to pressure and its dependency on another non dimensional quantity-Reynolds number. Velocity of fluid in tee junction is kept as low as .01 m/s to 0.2 m/s in the laminar flow region. For the case of turbulent region, Reynolds number has been

kept high so that it may be investigated on a highly turbulent flow and it ranges from 7000 to 14000 in pipe outlet with a lower velocity in combined pipe ranging from 0.3 m/s to 0.9 m/s. In the same figure, the same property i.e. Reynolds number has been studied for side pipe (riser) which is shown by subscript 's'. For side pipe, the values of pressure loss coefficients are higher than straight pipe due to the fact the there is greater loss of pressure resulting from mixing of two the streams as well as from the bending of pipe at an angle of 90 degrees. Few values of pressure loss coefficients are negative for riser portion of the tee junction. This situation arises when the volume flow rate and Reynolds number are kept high in side pipe; the pressure difference becomes negative giving a negative value for loss coefficients. For combining flow, Fig. 5 shows a trend of decrease in values of loss coefficients considerably as the Reynolds number increases.

The second important parameter is flow ratio in both the manifold and riser pipes. The discharge ratios in both the pipes of the tee junction are varying from 0 to a maximum of 1. Fig. 6 shows plot of manifold and riser pipes for the junction flow with respect to discharge ratios for pressure loss coefficients comparison between ANN and experimental values

The observed trend for pressure loss coefficients in manifold is that for increasing discharge ratio the pressure loss coefficients decreases significantly while for the case of riser, it is exactly opposite. The pressure loss coefficients increase as the discharge ratio increases from zero to 1.

Fig. 5 and 6 shows that proposed method of ANN is following the experimental results for both parameters of Reynolds number and discharge ratios. Few of the models discussed in the literature are used for comparison in Fig. 7, 8 and 9. Fig. 7 shows the comparison of ANN with experimental values, Bassett's [18], Idelchik's [5] and Badar's [16] equations for manifold pressure loss coefficients plotted against the discharge ratios. It is evident that ANN and experimental values are close to each other while the rest are following the trend of decreasing pressure loss coefficients values as the discharge ratio increase from minimum to maximum.

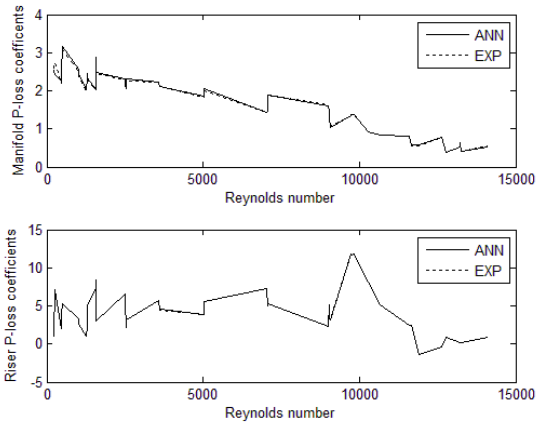



Fig. 5. Pressure Loss coefficient comparison between experiment and ANN for different Reynolds numbers.

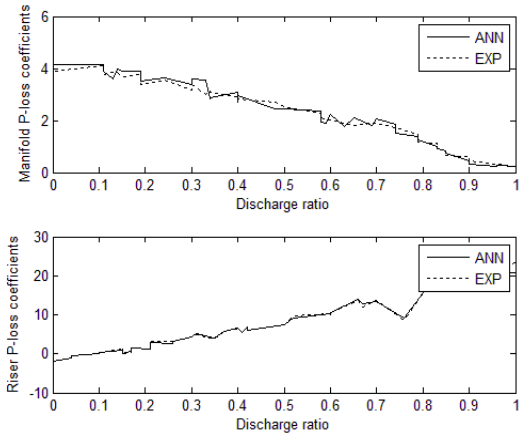



Fig. 6. Pressure loss coefficient comparison between experiment and ANN for different discharge ratios.

In Fig. 9, the comparison has been done for pressure loss coefficients in riser pipe of the tee junction. Here, the literature models and the proposed model along with the experimental values are reasonably close to each other and trend of increasing pressure loss coefficient is followed as the discharge ratio increases in the riser portion of tee junction. In Fig. 8, Badar's equations [16] for the pressure loss coefficient in riser are compared but the trend of experiment and other literature is not followed.

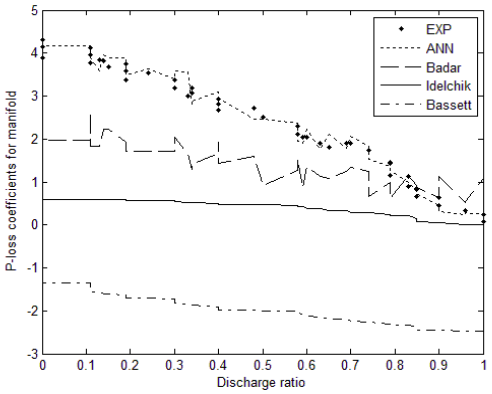



Fig. 7. Literature models comparison in manifold of solar collector's tee junction.

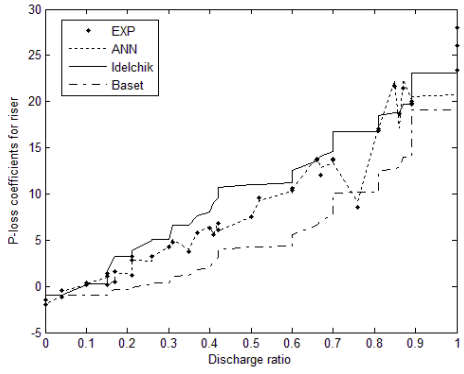



Fig. 8. Literature model comparison for riser of a solar collector's tee-junction.

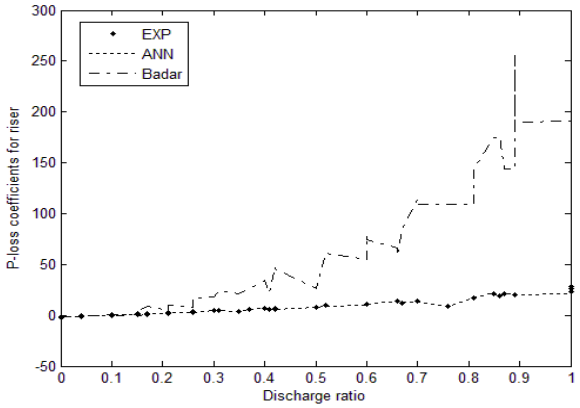



Fig. 9. Literature model comparison for riser of a solar collector's tee-junction.

## V. CONCLUSION

In this paper, effect of Reynolds number along with varying discharge ratio from zero to maximum has been studied for both the riser and manifold of tee junction for combining flow loss coefficients. Loss coefficients due to the pressure change in manifold have lower values then riser pipe due to large volume flow rate and straightness. The general trend observed is that with an increase in Reynolds number and discharge ratio in the manifold, the values of loss coefficients decreases. The relationship for loss coefficients in the riser with discharge ratios and Reynolds number is complex. The values increase with Reynolds number while decreases with discharge ratios in the riser. Plots show that ANN is determining pressure loss coefficients reasonably with an average error value of 4% in pressure loss coefficient for straight pipe and error value of 2.24% in side pipe pressure loss coefficient between experimental and proposed method. The ANN model proposed agrees satisfactorily with the experimental values and thus demonstrating the ability to replace the expensive experimental and other analytical techniques requiring high performance and large memory computers.

## REFERENCES

[1] M. Thirugnanasambandam, S. Iniyan, and R. Goic, "A review of solar thermal technologies," *Renewable and Sustainable Energy Reviews*, vol. 14, no. 1, pp. 312–322, Jan. 2010.

[2] A. Syed *et al.*, "A novel experimental investigation of a solar cooling system in Madrid," *International Journal of Refrigeration*, vol. 28, no. 6, pp. 859–871, 2005.

[3] A. Azzini, "A new generic approach for neural network design and optimization," PhD, University of Milan, Milan, 2006.

[4] D. E. Winterbone and R. J. Pearson, *Design techniques for engine Manifolds: Wave action methods for IC engines*. Professional Engineering Publishing, 1999.

[5] I. E. Idelchik and E. Fried, "Handbook of hydraulic resistance: Second edition (book)," Hemisphere Publishing,New York, NY, 1986.

[6] E. Ruus, "Head losses in Wyes and Manifolds," *Journal of the Hydraulics Division*, vol. 96, no. 3, pp. 593–608, 1970.

[7] Gardel, André. Les pertes de charge dans les écoulements au travers de branchements en té. Éd. de la Soc. du Bulletin Technique de la Suisse Romande, 1957.

[8] J. F. Bingham and G. P. Blair, "An improved branched pipe model for multi-cylinder automotive engine calculations," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 199, no. 1, pp. 65–77, Jan. 1985.

[9] W. H. Hager, "An approximate treatment of flow in branches and bends," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 198, no. 1, pp. 63–69, Jan. 1984.

[10] Flamang, P. Studie van de stroming in juncties van uitlaat--systemen van opgeladen dieselmotoren. Diss. PhD thesis, Rijksuniversiteit, Gent, 1989.

[11] V. Weitbrecht, D. Lehmann, and A. Richter, "Flow distribution in solar collectors with laminar flow conditions," *Solar Energy*, vol. 73, no. 6, pp. 433–441, Dec. 2002.

[12] G. F. Jones and N. Lior, "Flow distribution in manifolded solar collectors with negligible buoyancy effects," *Solar Energy*, vol. 52, no. 3, pp. 289–300, Mar. 1994.

[13] R. A. Bajura, " A Model for Flow Distribution in Manifolds," *Journal of Engineering for Power*, vol. 93, no. 1, pp. 7–12, Jan. 1971.

[14] A. Badar, R. Buchholz, Y. Lou, and F. Ziegler, "CFD based analysis of flow distribution in a coaxial vacuum tube solar collector with laminar flow conditions," *International Journal of Energy and Environmental Engineering*, vol. 3, no. 1, p. 24, 2012.

[15] S. F. Moujaes and S. Deshmukh, "Three-Dimensional CFD Predications and experimental comparison of pressure drop of some common pipe fittings in turbulent flow," vol. 132, no. 2, Aug. 2006

[16] D. S. Miller, "REPIDISCA-Internal flow: A guide to losses in pipe and duct systems," British Hydromechanics Research Association, 1971.

[17] M. Bassett, D. Winterbone and R. Pearson, "Calculation of steady flow pressure loss coefficients for pipe junctions", *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 215, no. 8, pp. 861-881, 2001.

[18] M. Abdulwahhab, N. Kumar, Injeti, and S. F. Dakhil, "Numerical prediction of pressure loss of fluid in a t-junction," *International Journal of Energy and Environment*, vol. 4, no. 2, pp. 253–264, 2012.

[19] F. Bava and S. Furbo, "A numerical model for pressure drop and flow distribution in a solar collector with u-connected absorber pipes," *Solar Energy*, vol. 134, pp. 264–272, Sep. 2016

[20] J. Jia, N. Yang, C. Zhang, A. Yue, J. Yang, and D. Zhu, "Object-oriented feature selection of high spatial resolution images using an improved relief algorithm," *Mathematical and Computer Modelling*, vol. 58, no. 3-4, pp. 619–626, Aug. 2013.

[21] I. Shafi, J. Ahmad, S. I. Shah, and F. M. Kashif, "Computing Deblurred time-frequency distributions using artificial neural networks," *Circuits, Systems & Signal Processing*, vol. 27, no. 3, pp. 277–294, May 2008.

# Interface of an Automatic Recognition System for Dysarthric Speech

Brahim-Fares Zaidi[1]
Laboratory of Speech Communication and Signal Processing (LSCSP)
University U.S.T.H.B
Algiers, Algeria

Sid-Ahmed Selouani[3]
Laboratory of Research in Human-System Interaction (LARHSI)
University of Moncton, Shippagan Campus
Moncton, Canada

Malika Boudraa[2]
Laboratory of Speech Communication and Signal Processing (LSCSP)
University U.S.T.H.B
Algiers, Algeria

Djamel Addou[4]
Laboratory of Speech Communication and Signal Processing (LSCSP)
University U.S.T.H.B
Algiers, Algeria

***Abstract*—This paper addresses the realization of a Human/Machine (H/M) interface including a system for automatic recognition of the Continuous Pathological Speech (ARSCPS) and several communication tools in order to help frail people with speech problems (Dysarthric speech) to access services providing by new technologies of information and communication (TIC) while making it easier for the doctors to achieve a first diagnosis on the patient's disease. In addition, an ARSCPS has been improved and developed for normal and pathology voice while establishing a link with our graphic interface which is based on the box tools Hidden Markov Model Toolkit (HTK), in addition to the Hidden Models of Markov (HMM). In our work we used different techniques of feature extraction for the speech recognition system in order to improve the dysarthric speech intelligibility while developing an ARSCPS which can perform well for pathological and normal speakers. These techniques are based on the coefficients of ETSI standard Mel Frequency Cepstral Coefficient Front End (ETSI MFCC FE V2.0); Perceptual Linear Prediction coefficients (PLP); Mel Frequency Cepstral Coefficients (MFCC) and the recently proposed Power Normalized Cepstral Coefficients (PNCC) have been used as a basis for comparison. In this context we used the Nemours database which contains 11 speakers that represents dysarthric speech and 11 speakers that represents normal speech.**

***Keywords*—Automatic Recognition System of Continuous Pathological Speech (ARSCPS); ETSI standard Mel frequency Cepstral Coefficient Front End (ETSI MFCC FE V2.0); Hidden Markov Model Toolkit (HTK); Hidden Models of Markov (HMM); Human/Machine (H/M); Technologies of Information and Communication (TIC); Mel Frequency Cepstral Coefficients (MFCC); Perceptual Linear Prediction (PLP); Power Normalized Cepstral Coefficients (PNCC)***

## I. Introduction

An interface for the control of Automatic Recognition System of Continuous Pathological Speech (ARSCPS) [1] can be very useful for speakers with dysarthria which is a neurological disorder of speech that affects millions of people. A dysarthria man has significant difficulty in communication, according to Aronson [2]; Dysarthria [3] covers various speech disorders resulting from neurological disorders. These disorders are related to the disturbance of the brain and stimulus nerves of the muscles involved in the production of speech. The characteristics of this pathology are slow, weak, imprecise or uncoordinated speech musculature movements [4]; the result is unintelligible speech.

The main objective of this work is to help the dysarthria people with the proposed interface that allows doctors to make their first diagnosis according to the type of dysarthria. Thus, patient intelligibility rate can be classified and other clinical assessments can be given to patients.

Our platform includes an ARSCPS based on the Hidden Models of Markov (HMM) [5] built by us. Furthermore, a pathological database NEMOURS [6] is integrated for improving the speech intelligibility of dysarthria patients.

This paper is outlined as follows: The following section presents the NEMOURS database. Section III explains the realization steps of an ARSCPS with Hidden Markov Model Toolkit (HTK) [5], [7]. Section IV, V and VI shows the results of the recognition with different techniques of features speech extraction. Section VII presents the different parts of our own ARSCPS which is the principal blocks of our H/M interface. We conclude this work in Section VIII.

## II. Nemours Database

The Nemours database is constituted by 74 sentences spoken with varying degrees of dysarthria for each one of the 11 male speakers.

The composition of the NEMOURS database is 11 male speakers with different degrees of dysarthria and 11 normal male speakers; 74 sentences spoken for each one of the 11 male speakers; as a result, we have 814 sentences for pathological voice and 814 sentences for normal voice; look at Table 1.

TABLE I. NEMOURS DATABASE

| NEMOURS DATABASE | Normal | Pathology |
|---|---|---|
| Number of Speakers | 11 | 11 |
| Number of Sentences For Each Speaker | 74 | 74 |
| Number of Total Sentences | 814 | 814 |

## III. STEPS OF REALIZATION OF AN ARSCPS WITH HTK

Our adapted ARSCPS for the NEMOURS database is based on the monophone models. These models require several successive treatment steps for their realization (see Fig. 1). However, during the development of our system, an additional step, which we call Step 0 has been necessary for its realization.

- **Step 0:** Transcription of texts;
- **Step 1:** Grammar of ARSCPS;
- **Step 2:** Dictionary;
- **Step 3:** Sound data;
- **Step 4:** Creating transcript files;
- **Step 5:** Acoustic parameterization of the system and data encoding (MFCC [8], ETSI FE V2.0 [9], [10], PLP [8], and PNCC [11], [12]);
- **Step 6:** Creation of HMM models;
- **Step 7:** The pause model;
- **Step 8:** Realignment of the training data;
- **Step 9:** Recognition of test corpus.

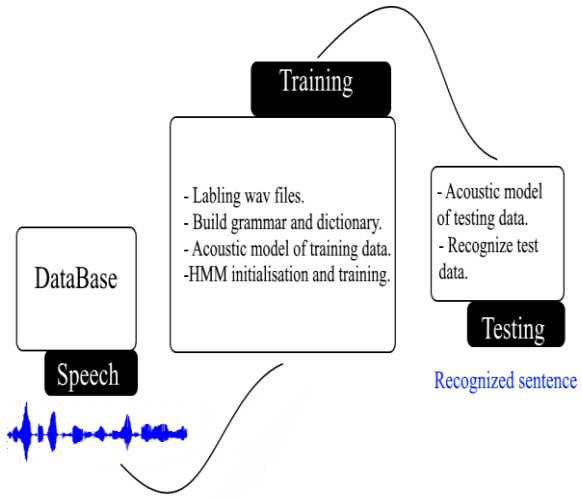



Fig. 1. Steps of an automatic speech recognition system [13].

For training steps (Fig. 2) we used 814 sentences, (for example from 1 to 74 sentences for each speaker), for testing steps (Fig. 3), we tested our ARSCPS with 352 sentences, (for example from 1 to 32 sentences for each speaker); look at Table 2.

TABLE II. TOTAL NUMBER OF SENTENCES FOR TRAINING/TESTING

| NEMOURS DATABASE | Normal | Pathology |
|---|---|---|
| Total Number of Sentences for Training (70%) | 814 | 814 |
| Total Number of Sentences for Testing (30%) | 352 | 352 |

We note that the sentences taken for the test are those that are already exist in the training, because our goal is the realization of an interface H/M that allows doctors to make their first diagnosis; and to make this diagnosis on the patient's disease, the patient have to pronounce sentences imposed by the doctor.

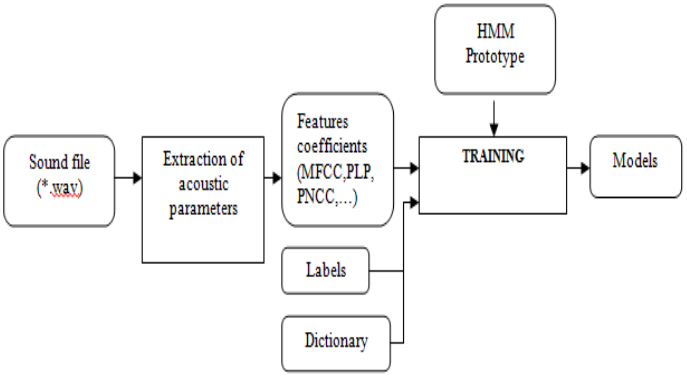



Fig. 2. Training phase.

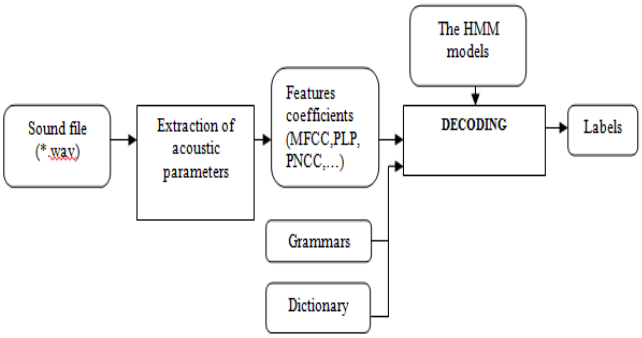



Fig. 3. Recognition phase.

## IV. EVALUATION OF RECOGNITION RESULTS

The generation of the test data acoustic model has been done, the deletion errors (D), the substitution errors number (S) as well as the number of insertion errors (I) are calculated (see Fig. 4).

Three types of recognition rates that correspond to sentence correction (1); word correction (2)' and word accuracy (3) are calculated.

The equations used are respectively:

- For SENT:

$$\text{Sentence Correction } \% = \frac{N-S}{N} \times 100 \qquad (1)$$

N: Is the overall number of sentences.

- For WORD:

$$\text{Word Correction } \% = \frac{N-D-S}{N} \times 100 \quad (2)$$

$$\text{Word Accuracy } \% = \frac{N-D-S-I}{N} \times 100 \quad (3)$$

N: Is the overall number of words.

```
==================== HTK Results Analysis ====================
  Date: Mon Mar 20 12:55:54 2017
  Ref : C:\Users\ZAIDI\Desktop\ASRN\RECOGNIZER\labels\N1.MLF
  Rec : C:\Users\ZAIDI\Desktop\ASRN\test_new/result_clean.mlf
---------------------- Overall Results -------------------------
SENT: %Correct=73.86 [H=260, S=92, N=352]
WORD: %Corr=96.54, Acc=93.09 [H=2039, D=45, S=28, I=73, N=2112]
================================================================
```

Fig. 4. HTK Results of ARSCPS.

To achieve this step of performance, a lot of scripts have been used for building our system.

## V. Acoustic Model of the Training/Testing Data

The objective of the extraction of the useful information done by the speech of NEMOURS database is to solve the problems of the speech recognition. ETSI MFCC FE, PLP, PNCC, and MFCC features are extracted from the speech.

## VI. Experimental Results

Table 3 shows the sentence correction, word correction, and the accuracy of the word from the four techniques of feature extraction for the system of speech recognition with voice normal samples. The best sentence correction, word correction, and the accuracy of the word with the proposed features can be found using ETSI FE V2.0 (MFCC_0_D_A) which is 86.65 %, 99.15 % and 97.40 %, respectively.

For the samples of voice pathological, Table 4 shows the sentence correction, word correction, and the accuracy of the word by using four techniques of feature extraction. The best sentence and word correction as well as the word accuracy are 73.86%, 96.54% and 93.09%, respectively which is for applying ETSI FE V2.0 (MFCC_0_D_A) feature with 39 coefficients. The sentence and word correction as well as the word accuracy of MFCC_0_D_A and PLP_0_D_A with 39 coefficients are almost the same, the word correction of MFCC_0_D_A and PNCC_0_D_A with 39 coefficients are the same.

In addition, for PNCC_0_D_A with 39 coefficients, the accuracy is 84.33%, and represents the worst result. For testing the system performance, we have applied the four feature extractions for speech recognition system with normal and pathological voice as shown in Tables 3 and 4. The ETSI FE V2.0 (MFCC_0_D_A) outperforms the previously techniques in terms of sentence correction, word correction and word accuracy.

TABLE III. Experiments with Voice Normal Samples

| **Normal voice (NEMOURS Database)** | | | | |
|---|---|---|---|---|
| **Features extraction** | **MFCC** | **PLP** | **ETSI FE V2.0** | **PNCC** |
| **Sentence correction (%)** | 85.23 | 81.25 | 86.65 | 83.52 |
| **Word correction (%)** | 99.10 | 98.11 | 99.15 | 98.72 |
| **Word accuracy (%)** | 96.73 | 96.16 | 97.40 | 96.54 |

TABLE IV. Experiments with Voice Pathological Samples

| **Pathological voice (dysarthria) (NEMOURS Database)** | | | | |
|---|---|---|---|---|
| **Features extraction** | **MFCC** | **PLP** | **ETSI FE V2.0** | **PNCC** |
| **Sentence correction (%)** | 67.05 | 66.19 | 73.86 | 62,50 |
| **Word correction (%)** | 95.60 | 95.98 | 96.54 | 95,60 |
| **Word accuracy (%)** | 86.84 | 86.41 | 93.09 | 84.33 |

## VII. Automatic Recognition System Interface of Continuous Pathological Speech (ARSCPS)

The interface includes an ARSCPS based on HMM models and the HTK toolbox [14].

We describe the different parts of an ARSCPS control interface using the DELPHI software and the PASCAL programming language [15].

### A. Part 1: Management of the NEMOURS Database Textual and Sound Files

For realizing this part (Fig. 5), we have written programs under DELPHI for:

- Loading sound files which allowed us to create the text files containing the links for access to the NEMOURS database.
- Listening to the sound files with extension ".wav".
- The listened Files transcription.
- Display the automatic recognition result of the listened continuous speech files to text format (Speech to Text).
- Synthesis of the recognized text files (Text to Speech).

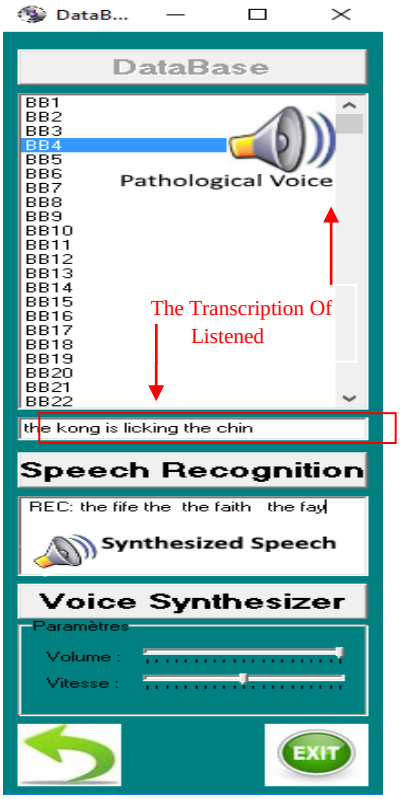



Fig. 5. DataBase (NEMOURS), speech recognition and voice synthesizer.

## B. Part 2: Result of the Voice Recognition

In this part (Fig. 6) of the interface, our system:

- Make the link with Part 1.
- Allowing the passage between ARSCPS and our interface.
- Compare the recognition of the NEMOURS database sentence.
- Show of the obtained results in percentage format.


Fig. 6. Automatic recognition system result of continuous pathological speech (ARSCPS) with ETSI FE V2.0 (MFCC_0_D_A).

## C. Part 3: Automatic Speech Recognition System in Real Time

The objective of this interface part (Fig. 7) is to realize an automatic dictation system and attempt to make the system independent of the NEMOURS database speakers and always working in real time.

### 1) Part 3.1: Interface of Sound File Recording

To do this, a recording module has been programmed with different acquisition parameters:

- The type of channels (mono or stereo).
- Bit (16 Bit).
- Sampling frequency (8000 Hz ou 16000 Hz).

### 2) Part 3.2: Control Interface of ARSCPS in Real Time

In this part, our ARSCPS in real time:

- Makes the link with interface of Sound file recording.
- Allows passage between the new ARSCPS in real time and our interface.
- Compares recorded sentence with sentences of NEMOURS database.
- Displays the recognized sentence.

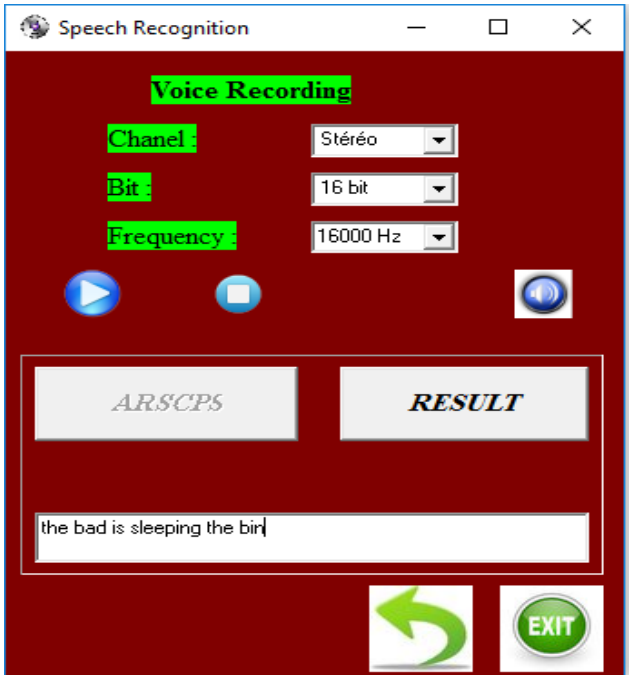



Fig. 7. Speech recognition in real time.

## VIII. Conclusion

In this paper, we proposed an ARSCPS using HTK. For the feature extraction we use ETSI FE, PLP, PNCC, and MFCC. For 22 speakers (for both normal and pathology) 114 words and 46 phonemes are applied in the experiments. The highest accuracy 97.40% is obtained ARSCPS by HTK using ETSI FE feature in the normal voice case. For the pathological voice, we achieved the highest accuracy 93.09% using ETSI FE features.

The ARSCPS that we wanted to implement must satisfy certain conditions which are:

- Continuous speech.
- Independent to speakers (Multi-speakers).
- Large vocabulary.
- Especially dedicated to the recognition of pathological speech (Dysarthria).

The proposed interface allowed us:

- The text and sound files management of the database.
- Listening to the sound files of extension ".wav".
- The transcription of listened Files.
- Display the result of ARSCPS of listened files to text format (Speech to Text).
- Synthesis of recognized text files (Text to Speech).
- Display of results of ARSCPS in percentage format.
- Make the system independent from speakers and make it working in real time.

Each of these objectives is a challenge to overcome. Therefore, realizing a complex system that treats a large number of complications requires: The writing of several scripts (XML, HTML, PASCAL, and HTK) [16]-[18], manipulation and management of a large number of scripts lines, and the management of sound files, text and binary data.

Today our challenge is to provide an interactive platform including the new TIC, making them accessible to people with communication disorders and proposing to the doctors a diagnosis system of the disease severity in order to take a quick decision.

## REFERENCES

[1] M. J. Alam, P. Kenny, P. Dumouchel, and D. O'Shaughnessy, "Robust feature extractors for continuous speech recognition," IEEE Xplore. European Signal Processing Conference (EUSIPCO), Lisbon, Portugal, November 2014.

[2] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," Journal of Speech, Language, and Hearing Research, vol 12, pp. 246-269, June 1969.

[3] K. T. Mengistu, and F. Rudzicz, "Comparing humans and automatic speech recognition systems in recognizing dysarthric Speech," Springer, C. Butz, and P. Lingras, Berlin, Heidelberg, vol 6657, pp. 291–300, 2011 [Canadian Conference on Artificial Intelligence, p. 291, 2011].

[4] K.M Yorkston, D.R.Beukelman, K.R Bell, "Clinical management of dysarthric speakers," Journal of Neurology, Neurosurgery, and Psychiatry (JNNP), Boston, vol 51, issue 11, p.1467,1988.

[5] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book," version 3.1, pp. 1-277, 2006.

[6] X. Menéndez-Pidal, J B. Polikoff, S M. Peters, J E. Leonzio,and H T .Bunnell, "The Nemours database of dysarthric speech," IEEE Xplore. Fourth International Conference on Spoken Language Proceedings (ICSLP), Philadelphia, PA, USA, August 2002.

[7] Y. J. Kim, "A simple example of generating a recognition system using HTK, for a word level recognition", Hanbat National University, pp. 1-41.

[8] A. H. Gharbi, "Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole- Selection of acoustic parameters relevant to speech recognition," PhD thesis, 09 december 2012.

[9] L. Guo, X. He, Y. Lu, and Y. Zhang, "A low cost robust front end for embedded ASR system," ISCA Archive, Kent Ridge, Singapore, 2006 [International Symposium on Chinese Spoken Language Processing, December 2006].

[10] Nitisha, and A. Bansal "Speaker recognition using MFCC front end analysis and VQ modeling technique for Hindi words using MATLAB," International Journal of Computer Applications, Sonipat, Haryana, India, Vol 45, May 2012.

[11] C. Kim, and R. M. Stern, "Power Normalized Cepstral Coefficients (PNCC) for robust speech recognition" IEEE Transactions. Audio, Speech, and Language Processing, Vol 24, pp. 1315-1329, July 2016.

[12] M. J. Alam, P. Kenny, and D. O'Shaughnessy, "Robust feature extraction for speech recognition by enhancing auditory spectrum", INRS-EMT, INTERSPEECH, University of Quebec, Montreal, Quebec, Canada, 2012.

[13] A. Mohammed, A. Mansour, M. Ghulam, Z. Mohammed, T. A. Mesallam, K. H. Malki, F. Mohamed, M. A. Mekhtiche, and B. Mohamed, "Automatic speech recognition of pathological voice," Indian Journal of Science and Technology, Vol 8, November 2015.

[14] M. Dua, R. K. Aggarwal, V. Kadyan, and S. Dua, "Punjabi automatic speech recognition using HTK," International Journal of Computer Science Issues (IJCSI), vol 9,issue 4, July 2012

[15] Club des développeurs et IT pro- Club developers and IT pro. Les meilleurs cours et tutoriels sur la programmation et l'informatique professionnelle- The best courses and tutorials on programming and business computing (Cours et tutoriels sur la programmation Delphi- Courses and tutorials on Delphi programming). http://delphi.developpez.com/cours/?page=langage[January, 20th, 2016].

[16] B. F. Zaidi, S. Selouani, M. Boudraa, and G. Hamdani, "Human/machine interface dialog integrating new information and communication technology for pathological voice," IEEE Xplore. Future Technologies Conference (FTC), San Francisco, CA, USA, January 2017.

[17] M. Nebra, "Apprenez à créer votre site web avec HTML5 et CSS3- Learn how to create your website with HTML5 and CSS3'', pp. 1-248, June 2013.

[18] Voxeo An Aspect Company, "XML DevelopmentLanguages Documentation", W3C, 685 Clyde Avenue Mountain View, CA 94043, version 2.1, pp. 1-254, January, 20th, 2015.

# Personalized E-Learning Recommender System using Multimedia Data

Hayder Murad
School of Engineering
University of Portsmouth
Portsmouth, United Kingdom

Linda Yang
School of Engineering
University of Portsmouth
Portsmouth, United Kingdom

***Abstract*—Due to the huge amounts of online learning materials, e-learning environments are becoming very popular as means of delivering lectures. One of the most common e-learning challenges is how to recommend quality learning materials to the students. Personalized e-learning recommender systems help to reduce information overload, which tailor learning material to meet individual student's learning needs. This research focuses on using various recommendation and data mining techniques for personalized learning in e-learning environment.**

***Keywords*—*E-Learning; recommender system; data mining***

## I. PURPOSE

The explosive growth in online learning materials has generated an urgent need for new techniques and access mechanisms to reduce information overload. eLearning recommender system has been paid a great attention by the researchers in the field of recommender system to solve the problem of information overload in learning materials and personalisation [1]-[3]. This research aims to develop a personalized e-learning recommender system to undergraduate students based on students' level of knowledge, and students' profiles with the use of online video learning materials.

The objectives of the project include:

- Developing a framework for supporting the development of the personalized eLearning recommender system.
- Developing a classification method to classify online video learning materials.
- Developing a profiling method to extract student profile and level of knowledge.
- A prototype will be designed as an intelligent tutor to save the time and effort of the lecturer and also increasing self-learning skills.

In this paper, we report on the design of the framework, which is able to detect students' profiles and their level of knowledge and automatically recommend online video learning material that meets the student needs.

## II. BACKGROUND/SIGNIFICANCE

UK universities accept thousands of international students every year, some of those students get accepted directly into second or third year through agreements between institutions which called 'Direct Entry programme'.

Direct Entry programme accepts student according to specific requirements like qualifications and grade certificates without taking into account the learning styles in their countries, differences between learning materials in UK and their countries and knowledge level of the students in the courses which they applied to. These may cause a problem to the student during his study in the University, also the instructor will suffer from this situation as he/she will teach students have lack of information about the courses which they teach. From this problem the proposed idea tries to solve such problems by offering a personalised e-learning recommender system to help direct entry student to reduce the lack of information by estimating their knowledge level and after that online video learning materials will be recommended to them.

The goal of a Recommender System is to generate meaningful recommendations to a collection of users for items or products that might interest them. Suggestions for books on Amazon, or movies on Netflix, are real world examples of the operation of industry strength Recommender systems [4]. In other words, Recommender systems are information agents that attempt to predict which items out of large pool a user may be interested in and Recommend the best one to the target user [5].

Due to the huge amounts of online learning materials, eLearning environments are becoming very popular as means of delivering lectures. The challenge of eLearning environments with the existence of large number of learning materials is how to recommend quality learning materials to the students when they have limited time to view and study [6]. Also, due to the large development in internet in term of speed and the wide use of mobile phones and tablets with video capabilities, all these reasons made the use of video in eLearning very popular. Learning via video became one of the popular ways of education and brings new dimensions to the learning methods. For example, if certain course provide contents that include practical skills like building a PC or designing a network using simulator, these part of the course will be most benefit to see it in video rather than explained by text and static images [7].

Accordingly, eLearning recommender system has been paid a great attention by the researchers in the field of recommender system to solve the problem of information overload in learning materials and personalisation [1]-[3].

This project is sponsored by Ministry of Higher Education and Scientific Research in Iraq.

eLearning recommender system aims to predict suitable learning materials from large amount of learning data based on student preferences [8], [9].

Research contributions organised as following:

- Define a framework to support the development of personalised eLearning recommender system.
- Develop a dynamic hierarchical classification method/algorithm to classify online video learning materials using machine learning and information retrieval techniques.
- Develop a user profiling method to model student's learning need and context determinations.
- Design personalised recommender system algorithm to filter learning materials to meet students' personal needs.

## III. METHOD

### A. Design of the Proposed Framework

As shown in Fig. 1, the proposed framework consists of two sides: the client side and the server side.

The server side:

- Student profiling module: processes student context information whether it's explicit or implicit.
- Knowledge estimation module: gathers data to produce suitable information about a student's knowledge level.
- Feedback module: deals with feedback retrieved from students to refine the recommendation process.
- Assessment module: provides dynamic assessment.
- Recommendation module: recommends learning material to individual student based on student's knowledge and profile.
- Video searcher and classifier: classify and categorize video learning materials.

The client side:

- User interface module: it's responsible for all transactions that can be happened between the web and computer.
- Student profile: it's responsible for observe student usage behaviour and store it in student profiling module.

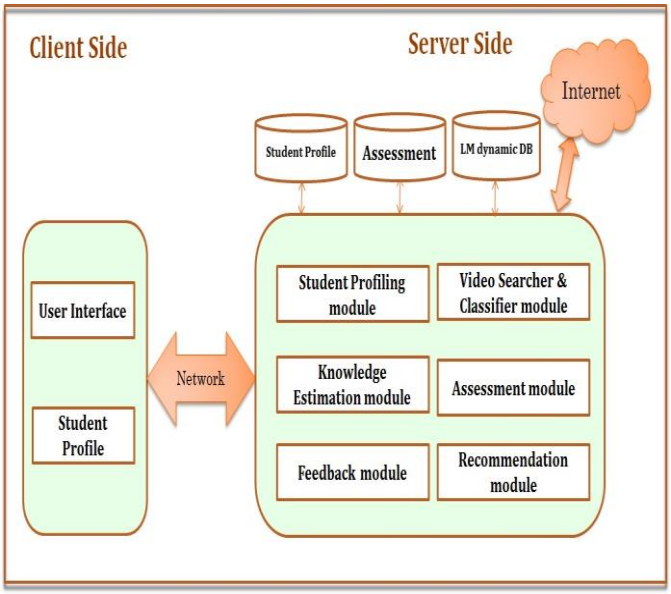



Fig. 1. Proposed framework.

### B. Data Collection

The required online learning materials will be retrieved from online video websites (initially YouTube) and classified and saved its links in dynamic learning materials repository to be recommended to the student.

The metadata of YouTube online videos are extracted using video searcher and classifier module. This metadata includes uploader dependent metadata (titles and descriptions) and uploader independent metadata (comments, No. of views, likes, dislikes … etc.) as shown in Fig. 2.

The proposed methodology to extract the metadata from YouTube videos and classify them based on the extracted metadata by applying text mining techniques into topics and subtopics are shown in Fig. 3.

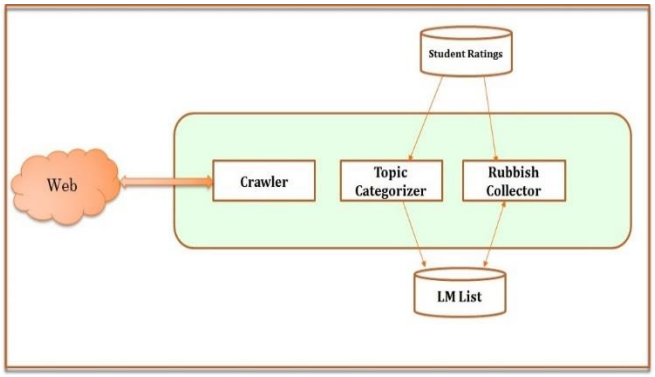



Fig. 2. Video searcher and classifier.

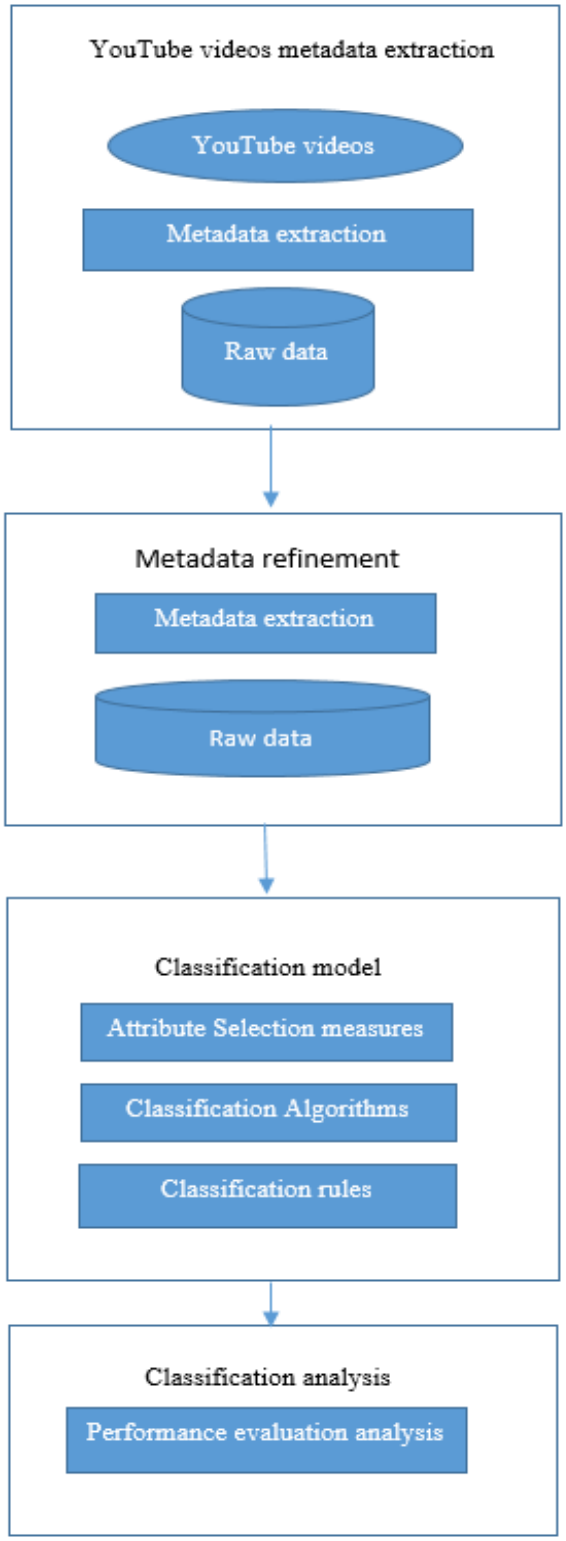



Fig. 3. Methodology.

## IV. Conclusion and Future Work

This proposed framework supports the development of personalised eLearning recommender system combining the techniques of student profiling, knowledge estimation, assessment and feedbacks in order to improve students learning process by recommending online video learning materials based on student profile and knowledge level. This paper explains the work process which is going to be implemented using algorithms that will be designed to support the system. Previous researches have been focused on text - based learning material. What we are doing is to build the foundation for processing and mining of online multimedia data, and put this work into engineering practices. Based on the proposed framework an initial prototype and a classification method have been developed. Our work is just at the beginning and far from accomplishment. In the near future, we would like to improve a dynamic Hierarchical classification method to classify online video learning material, as well as develop a user profiling method to model student's learning need and context determinations. We believe that personalised recommender is necessary for tailoring learning materials to meet individual student's learning need. We will try to build such a personalised recommender system algorithms to filter learning materials to meet students' personal needs.

## Acknowledgment

This project is funded by Ministry of Higher Education and Scientific Research in Iraq (MOHE) and the represented of the ministry is the Iraqi Cultural Attach in London.

## References

[1] C. Chen, L. Duh, and C. Liu, "A personalized courseware recommendation system based on fuzzy item response theory," IEEE Int. Conf. e-Technology, e-Commerce e-Service, 2004. EEE '04. 2004, pp. 305–308, 2004.

[2] Y. Gu, "Research on Personalized Recommendation System in E-Learning Based on Semantic Web," pp. 1046–1053, 2013.

[3] M. Holenko Dlab and N. Hoic-Bozic, "Recommender system for Web 2.0 supported eLearning," IEEE Glob. Eng. Educ. Conf. EDUCON, no. April, pp. 953–956, 2014.

[4] P. M. V.Sindhwani, "Recommender systems," Commun. ACM, pp. 1–21, 2010.

[5] S. Akshita, "Recommender System: Review", vol. 71, no. 24, pp. 38–42, 2013.

[6] K. I. Ghauth and N. A. Abdullah, "The effect of incorporating good learners' ratings in e-learning contentbased recommender system," Educ. Technol. Soc., vol. 14, no. 2, pp. 248–257, 2011.

[7] Epignosis LLC, "e-learning Concepts, Trends, Applications," Book, p. 5, 2014.

[8] N. Soonthornphisaj, E. Rojsattarat, and S. Yim-ngam, "Computational Intelligence: International Conference on Intelligent Computing, ICIC 2006 Kunming, China, August 16-19, 2006 Proceedings, Part II," D.-S. Huang, K. Li, and G. W. Irwin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 518–523.

[9] K. Verbert, N. Manouselis, X. Ochoa, M. Wolpers, H. Drachsler, I. Bosnic, and E. Duval, " Context-aware recommender systems for learning: A survey and future challenges," IEEE Trans. Learn. Vol.5, no.4, pp.318-335,2012.

# Level of Confidence in Software Effort Estimation by an Intelligent Fuzzy - Neuro - Genetic Approach

Poonam Rijwani[1], Dr. Sonal Jain[2]
Research scholar[1], Associate professor[2]
JK Lakshmipat University
Jaipur, Rajasthan

***Abstract*—Organizations are struggling to deliver the expected software functionality and quality in scheduled time and prescribed budget. Despite availability of numerous advanced effort estimation techniques overestimation and underestimation occur on a vast scale and results in project failures and significant loss to the organization. The paper proposes machine learning based approach to calculate the optimized effort and level of confidence. Genetically trained neural network evaluates the optimum effort for given COCOMO II variables. The level of confidence is evaluated by fuzzy logic and indicates the percentage that the predicted effort will not exceed the limits.**

***Keywords*—COCOMO II; artificial neural networks; genetic algorithm; fuzzy logic***

## I. Introduction

Human dependency on software is increasing continuously. Today most of the goods and services are realized with software systems. Software has become major driving force for progress even in domains that were traditionally reserved as completely mechanical or hardware systems, for instance major advances in automobile industry are being realized with software development. Companies spend 4-5 percentage of the revenue on software development [1]. The figure is as high as 10 percent in highly IT dependent sectors, for instance telecommunications and finance [1].Thus, functionality and complexity of software systems is increasing manifold. Simultaneously time to market and cost should be reduced to stay competitive. In United States 250 billion dollars are spent each year on IT development [2]. Project management and Effort estimation are key factors for success of a software project. Despite much research and technological advancement in effort estimation techniques, proportion of failed software projects is huge[2,3].According to the Chaos report submitted by Standish Group[2] only 16.2 percent of the projects are successful ,57.2 percent projects are over-budget and provides lesser functionalities than specified and 31.1 percent of the projects are cancelled during their development cycle. The percentage of the successful, challenged (over-budget with less functionality) and impaired (cancelled) projects is shown in Fig 1.

Project failure can be defined as combination of cost overruns, late deliveries, poor quality, and/or developing a product that does not get used. The two crucial reasons for failure of most software projects are Overestimation and Underestimation of the software effort [4]. Most projects either cost more than they return or fail to deliver required projects in the expected time. Both the scenarios lead to huge loss or may also result in termination of the whole project. R. Charette [3] suggested unrealistic project goals and inaccurate estimates of needed resources as principal factors that lead to project failure. A. Trendowicz and etal [2] pointed that most of the effort estimation techniques provides point estimates with hardly any support for risk management if project overruns the expected cost. Moses and etal[4] in their research concluded that in addition to estimates the effort estimation should also specify a Level Of Confidence associated with the calculated effort in order to compensate the uncertainty associated with effort estimation.

## II. Related Work

### A. What research has been Conducted so far?

For software effort estimation, numerous methods have been examined specifically data driven soft computing methods such as artificial neural networks, regression trees, evolutionary computing, rule-based induction, fuzzy logics etc. These methods exhibit many advantages like regression over other standard methodologies. Literature of software effort estimation endorse that important product feature characteristically reflects the software size which exactly impact efforts. Basically it is used to build cost models.

Initially, almost all models are based on the size of metrics which contemplates numerous coding lines coded for a software project i.e. lines of code (LOC) or thousands of source line code (KLOC), as shown in COCOMO [5], or function points (FP) which is there in models like Albrecht's FP i.e. Function Point Analysis [6].

Many researchers analyze the feasibility of evolving software effort estimation methods exhausting various methodology, parameters, datasets, etc. In the comparative analysis study given by [7], amalgamation of estimation methods may generate more reliable, accurate cost estimation for software development as it is displayed that no method is good or bad in all the situations. Review papers given by [8] [9] grasp a complete description of such studies. In review paper [10], the effort estimation was assessed by back propagation Artificial Neural Networks on datasets such as Desharnais and ASMA, generally through system size to determine the correlation of size with effort.

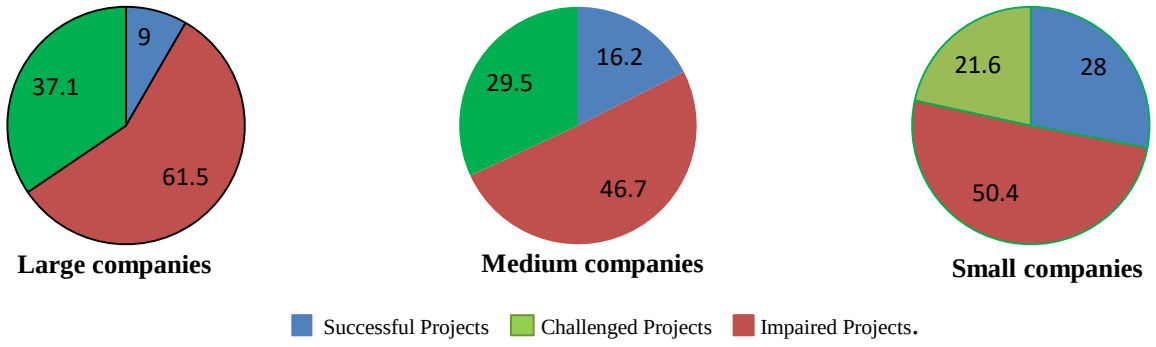



Fig. 1. Percentage of successful, challenged and impaired projects in large, medium and small organizations

The method produced reassuring estimates represent that the model need an additional methodical advance method to develop the architecture and parametric settings in order to achieve improved outcomes. In paper [11], the effort estimation evaluation of the relationship between effort and size analyzed using Genetic Programming technique in which advanced tree composition displaying number of power, linear, quadratic type common equations. The methodology lengthened to appropriate desirable stages of prediction correctness through hardly the size attribute but also approved to achieve additional enhancements.

Kumar et al. proposed a model exhausting Particle Swarm Optimization (PSO) for tuning the elements of primary COCOMO model to compute the effort accurately considering hardly KLOC factor [12].

Finnie et al. [13] conferred a comparison of statistical regression based model with other artificial Intelligence based estimation models for evaluation of software development effort. The researcher establish that statistical regression model underperformed for difficult and complex software projects as the Artificial Intelligence based models gives satisfying evaluation results. They studied dataset amidst Projects from 17 organization and Desharnais. As an estimation criterion MMRE was used. In 2002 another researcher Heiat [14] examined Feed Forward Neural Networks with function point and Radial Basis Neural Network with Source Lines of Codes for various datasets including projects of different generation languages. The results embodied that artificial neural network method is prudent with regression though a third generation language data set is used. P.Rijwani et-al [15] used hit & try method to determine best network architecture, in an experiment for the training network using back propagation. For software effort estimation a FLANN was proposed by Trimula Rao [16] which generates effort and hence processes the final layer output. It has a drawback that the relation between input and output is not equitable.

Investigation on back propagation Artificial Neural Network of 2-2-1 design based on dataset of NASA that includes eighteen projects. The inputs were development methodology and KDLOC and output was effort.

Attarzadeh [17] in which 17 cost drivers and 5 Scale factors were used as inputs. Sigmoid activation function is utilized while creating the network to achieve post architecture of COCOMOII model. The COCOMO algorithm is compared using Pred (0.25) and the results are shown in terms of MMRE. An innovative software development for effort estimation was proposed by Attarzadeh [24], exhausting neural networks, in which weights of the network were adjusted in such that it resulted in COCOMO II model. The neural network method suggested gives better result when related to COCOMO model after proper training. Even though back-propagation for neural networks is focused, complexity arises for adjusting weight and bias net parameters during training like slow convergence, sensitive to arbitrary initial weights, entering local minima, difficulty in selecting explicit optimum network configuration. The preeminent efficiency of genetically trained neural networks is mentioned in various researches. For instance After a series of experiments and simulations Shukla[22] concluded that the genetically trained neural networks outperforms back propagation trained and quick propagation trained neural networks in software effort estimation. The recommended model evaluates software development effort in the function of seventeen cost drivers and five scale factors. Thus it is evident that over the years improving the accuracy of software effort estimation has remained main concern of research with little attention being paid to quantifying the uncertainty involved in effort estimation techniques.

### *B. Loopholes in Existing Research*

Many systematic surveys have been conducted on software effort estimation. Moløkken and Jorgensen [30] provide an exhaustive review of surveys in software effort estimation. They concluded that most of the projects (60-80 percent) suffer from effort and/or schedule overruns. However the percentage overrun (30-40%) is significantly lower than suggested by Chaos report by Standish Group (80%). Jorgensen and Sheppard [29] identified and reviewed 304 research papers published in 76 journals. They found that majority of the research in software effort estimation has remained concentrated on effort estimation methods and less research is done on uncertainty assessments, data set properties and measures of estimation performance. The authors also

proposed that most of the research relies on historical dataset for evaluation and validation of effort estimation models, only a few provide real life evaluations. The different estimation methods used in industries are shown in Fig 2.Presently Expert judgment is the prominent estimation method used by organizations [26, 40].The reason is it cannot be ascertained if formal methods are better or weaker than expert judgment [27]. Identification the human factors affecting effort estimation [28] and development of practical guidelines are crucial to get benefitted from expert judgment. Applications of software effort estimation is summarized in Fig 3.

Fig 4 shows the number of papers published from the year 1996 to 2016 related to software effort estimation J.Moses[4] suggested that Effort estimation using algorithmic models, statistical prediction systems or machine learning approaches, e.g. Case Based Reasoning (CBR), Artificial Neural Networks (ANN) or Rule Induction (RL), under perform when compared to subjective estimation given by human estimators. The author also added that the prediction methods does not provides any support for decision making in case if the actual effort is greater than or less than the predicted effort. The author developed an approach by using Bayesian Inference to improve effort estimation consistency. B. Clark and etal[20] developed a multiple regression based effort calibration strategy. D.Yang and etal [21] proffered that software industry suffers from frequent cost overruns, and the software cost estimation remains a challenging issue. The authors developed a model for accounting the uncertainty based on Bayesian belief networks.

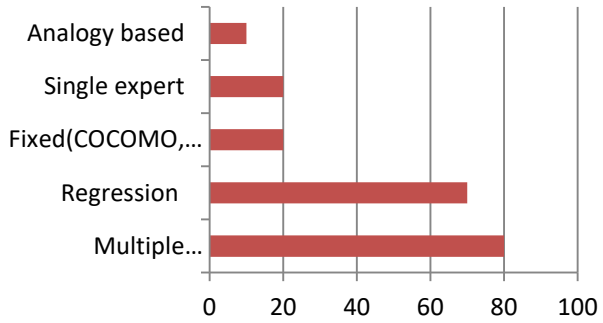



Fig. 2. Effort Estimation techniques used in industries

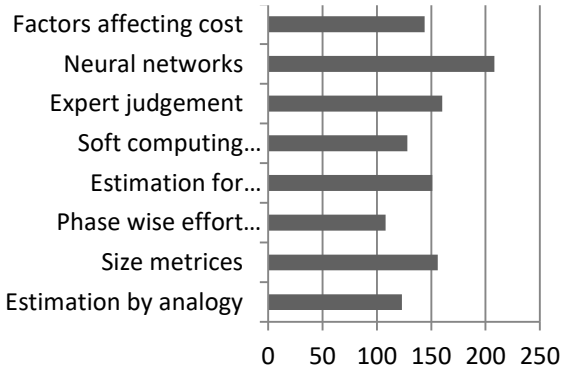



Fig. 3. [25]Papers published over past years about different estimation techniques.

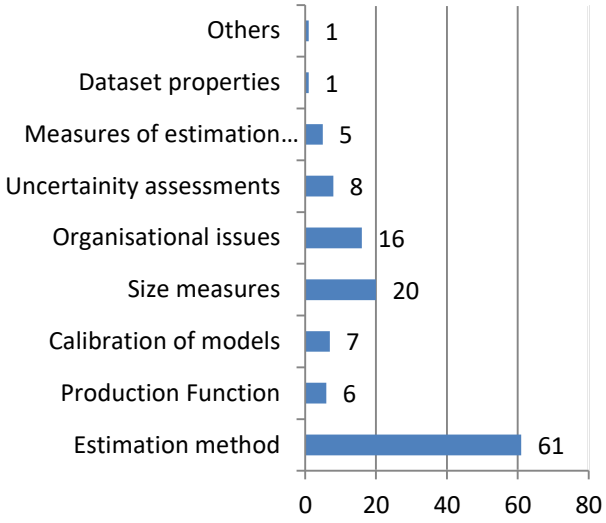



Fig. 4. Papers published over past years about different estimation techniques

## III. Implementation Details

The block diagram of the adopted methodology is shown in Fig4. Genetic Algorithm is used for improving the output of neural network. Fuzzy inference and genetically trained neural networks are employed independently to evaluate level of confidence and optimized effort. The main form or GUI constructed in MATLAB is shown in Fig 5. . The GUI prompts user to input COCOMO II variables. The user can click on the optimized effort button on the calculate panel for the GANN [3] predicted effort in person months. Similarly level of confidence button is pressed to obtain the probability that the effort will not exceed the specified limits.

### A. Dataset Generation

For the present work COCOMO 81 dataset available on PROMISE repository [18] was converted to COCOMO II dataset using the tool Rosetta stone [33].The tool is developed at IBM research in order to make COCOMO estimates functional with COCOMO II model. The output is development effort measured in man-months. The above mentioned COCOMO 81 dataset was established from exploration of sixty three developed software projects.

### B. Network Topology

The model is created with one hidden layer in MLF() neural network .One hidden layer with arbitrary units is sufficient for "Universal Approximation Property"[23].There is no rule of thumb for determining number of hidden units to be used, as it depends upon critical factors such as number of training cases and complexity of classification and learning. A convenient way is to try many different networks, calculate the generalization error for each network and select the network with minimum generalization error [23].Following the above rule, we performed a number of hit and trial experiments from 2 to 20 nodes.

The optimum topology was found to be 23-10-1 i.e. 23 input nodes, 10 nodes in hidden layer and 1 node layer (Fig 5).

The 23 neurons in input layer correspond to the 23 input variables of COCOMO II model (kloc, 5 Scale Factors, 17 Effort Multipliers). The node in output layer represents the optimized effort. The input to neural network as scale factors and effort multipliers is be represented by binary vectors$x_j (j \in [1,23])$. The COCOMO II variables are converted into binary vectors by range normalization such that such that $x_j (x_j \in [0,1])$ The vector O represents output obtained in person-months.$w_i$ And $w_{ji}$ are weight parameters or synaptic strength connecting hidden layer to output layer and input layer respectively.

$$O\left(\sum_{j=1}^{23} x_j\right) = f\left(\sum_{i=1}^{10} w_i\, y_i\right) \quad (1)$$

$$y_i = f\left(\sum_{j=1}^{23} w_{ij} x_j\right) \quad (2)$$

### C. Training The Network

Genetic algorithm is employed for training the neural network. Fig 6 represents the algorithm of training neural network by genetic algorithm. The preeminent efficiency of genetically trained neural networks is mentioned in various researches. For instance After a series of experiments and simulations Shukla [22] concluded that the genetically trained neural networks outperforms back propagation trained and quick propagation trained neural networks in software effort estimation The suitable values for control parameters of genetic algorithm have been found by running various simulations and have been listed in Table 1. We have used binary string chromosomes. Six features (very low- vl, low-l, nominal-n, high-h,very high-vh, extra high-xh) are considered for each cost driver and subsequent weights are encoded with 3 bits(0-n,1-vl,2-l,3-n,4-h,5- vh,6-xh,7-n).0 and 7 are assumed default values. Fitness function is reciprocal of MMRE as genetic algorithm maximizes the fitness function and a low Roulette Wheel selection:

Following are the steps for Roulette Wheel selection:

*1) Evaluate the sum of the fitness value of all individuals in given population...*

*2) Calculate probability of selection of a particular individual by dividing its chromosome's fitness by the total fitness values of the population.*

*3) Divide the roulette wheel into sectors based on probabilities calculated in the second step.*

*4) Spin the wheel 'n' number of times. The individual corresponding to the sector pointed by the pointer is selected.*

The probability that an individual is selected from a population of n individuals is given by equation, where is fitness value of element.

$$MMRE = \frac{1}{N}\sum_{J=1}^{N} \frac{EstEff_j - ActEff_j}{ActEff_j} \quad (3)$$

$$Fitness\ function = \frac{1}{MMRE} \quad (4)$$

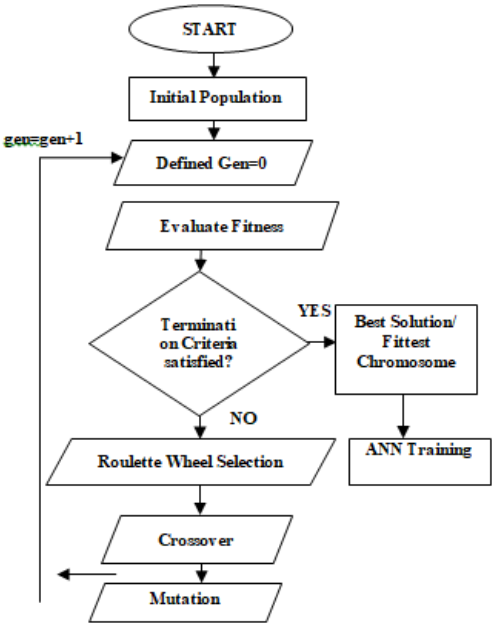



Fig. 5. Genetically trained neural network.

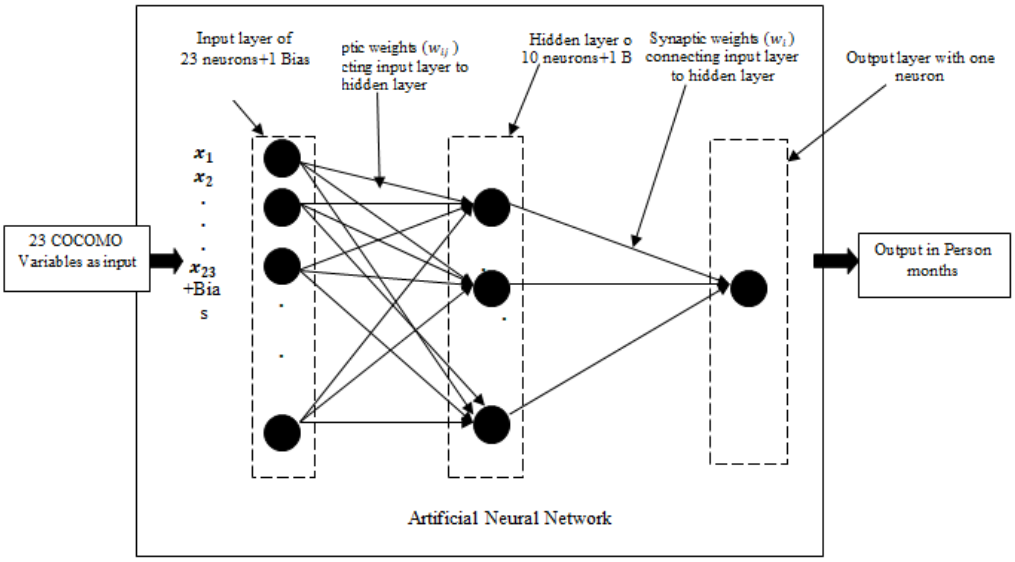



Fig. 6. Network Topology

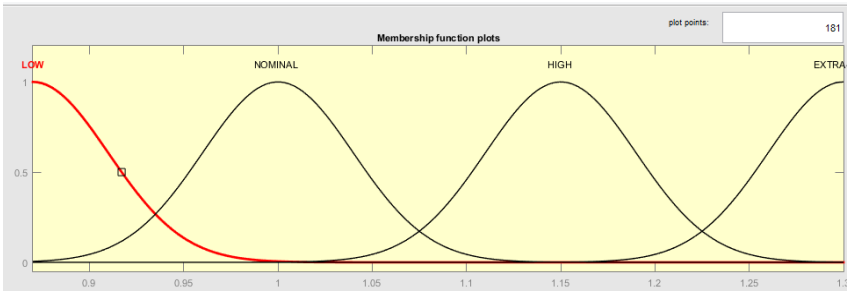



Fig. 7. Membership functions for STOR

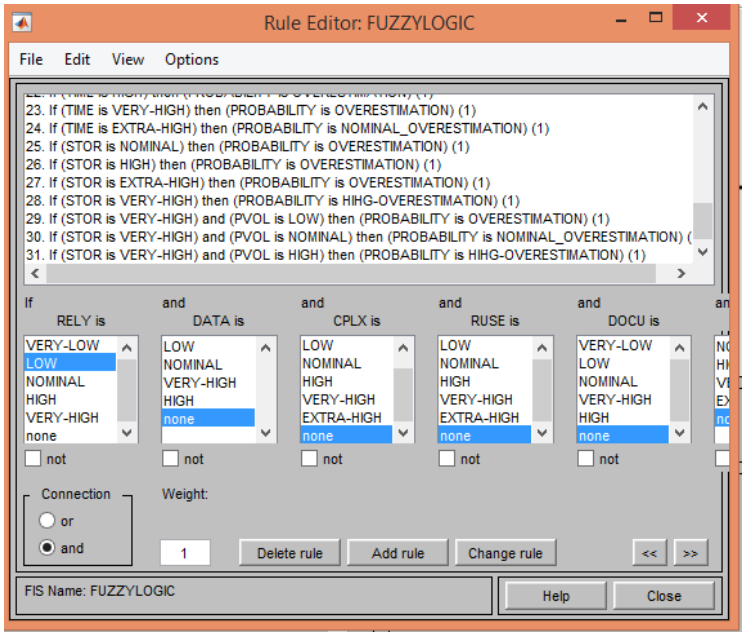



Fig. 8. Fuzzy rules Fuzzy Inference

Fuzzy logic is close to human interpretation of truth. This property can be utilized in effort estimation practices to balance the inherent imprecision with uncertainty to determine the level of confidence. The level of confidence indicates the uncertainty or the probability of overestimation or underestimation. The uncertainty can not only be used to improve the estimation consistency but can also be used in making statements to client indicating the chances of overestimation or underestimation.

Literature reveals that there are numerous approaches for incorporating fuzzy logic in effort estimation models. However available research on organization specific effort dependencies is scarce. For instance a company 'A' is facing frequent underestimation on its past projects. It is possible that the value of COCOMO variables used by them make it inclined to Underestimation. Suppose the assigned value of Programmer capability is high but due to some discrepancies the actual programmer capability is lower than expected. Such conditions can occur frequently while calculating software effort by formal methods because of following reasons:

*1) A huge amount of information is required in the starting phase.*

*2) Error due to human factor.*

The developed fuzzy inference system calculates probability of overestimation. This acts as a warning system for effort estimators and prompts them to review the process and/or set appropriate risk factors (Table2).

***Dataset preparation***: The historic dataset is converted into probability distribution functions of 23 input variables. In the proposed model conditional probability is used as basis for forming fuzzy rules. By probability theory conditional probability is defined as probability of occurrence of an event (A) by assertion that another event (B) has already occurred. The event A is hypothesis and the event B is observed evidence. It is expressed mathematically by equation

$$\mathbf{P(A/B) = \frac{P(A \cap B)}{P(B)} \quad , P(B) \neq 0} \qquad (5)$$

Here A denotes occurrences of overestimation and B denotes instances of values of input variables. For instance P (O/STOR=1) denotes the probability of overestimation provided that the selected value of STOR is nominal. The model is based on chances of overestimation as underestimation and overestimation are mutually exclusive events. Thus, if probability of one is known the probability of other can be calculated easily.

$$P(A \cup B) = 1 \qquad (6)$$

*A and B are mutually exclusive events*

The conditional probability for every value of each input variable is calculated and subsequently mapped into linguistic fuzzy rules. The developed fuzzy model uses 76 fuzzy rules to calculate output.

***Input and Output Variables:*** In this next step the range of 23 input variables as effort multipliers scale factors and lines of code is defined. The output variable is probability of overestimation and it lies between 0 and 1.

***Membership Functions:*** The membership functions are defined for each input as well as output variables. In our analysis we have considered Gaussian membership functions as it demonstrated by Kushwaha and Suryakane [31] that Gaussian membership function smoother transition in its intervals, and the achieved results were closer to the actual effort. The Gaussian membership function is governed by following equation. The Gaussian membership function for KLOC is given in Table3 [32] and STOR is shown in Fig7.

$$f(x;\sigma,\mu) = \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \quad (7)$$

***Fuzzy Rules:*** Fig 8shows the subsequent fuzzy rules.

## IV. Results And Discussions

In our project historical dataset of 63 projects is considered. The estimated effort and Mean Relative Error using COCOMO model, neural network model with back propagation and the genetically trained neural network model is shown in Table4 and Table5. Fig 10 demonstrates the error histogram obtained after training process. The histogram is centered on zero error Thus our selected topology is appropriate. Fig 12 shows the comparison between the two models. Mean square error found is 0.0124682 after 50 generations using 10000 populations. Fig 9 substantiates that the genetically trained neural network model outperforms the COCOMO model as well as BPNN by significant difference. The data of BPNN is taken from the authors' past research [19].

TABLE I. Genetic Algorithm Parameters

| SNO | Control parameter | Value |
|---|---|---|
| 1 | Population Size | 10000 |
| 2 | Elite Count | 4 |
| 3 | Crossover | 0.8 |
| 4 | Generations | 50 |
| 5 | Initial Population | 10000*251 double |
| 6 | Selection | Roulette Wheel |
| 7 | Crossover | Heuristic |
| 8 | Number of variables | 251 |
| 9 | Mutation | 0.01 |

TABLE II. Risk Factor

| Probability (Fuzzy Output) | Overestimation | Risk |
|---|---|---|
| 0-0.3 | Very low | low |
| 0.3-0.5 | low | Moderate |
| 0.5-0.7 | High | High |
| 0.7-1 | Very High | Very High |

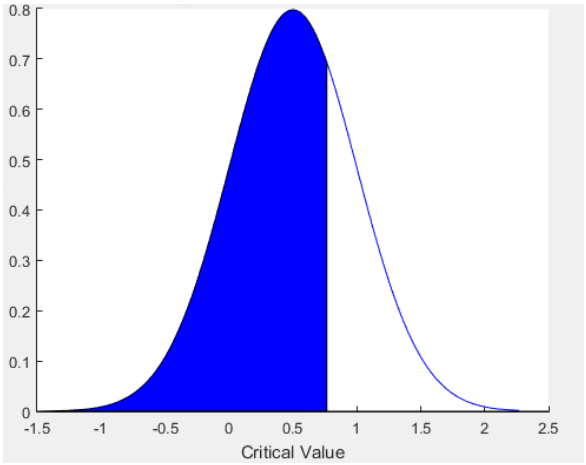



Fig. 9. Fuzzy Output.

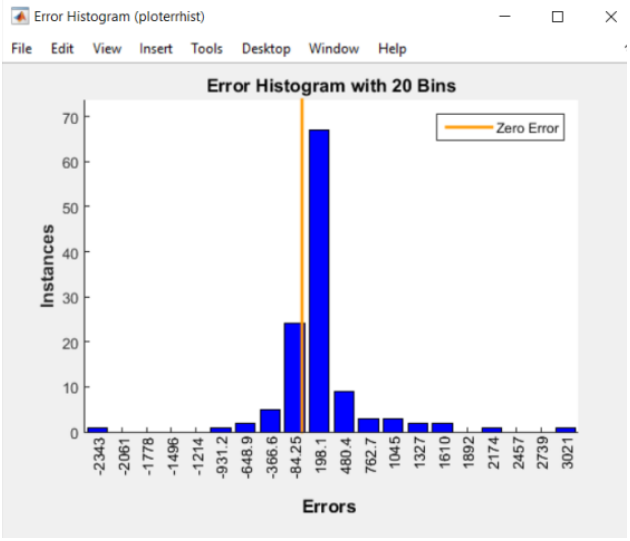



Fig. 10. Error Histogram.

TABLE III. MRE Comparison

| KLOC | Value |
|---|---|
| 0-50 | small |
| 50.1-128 | Moderate |
| 128.1-512 | High |
| 512.1-up | Very high |

TABLE IV. MRE Comparison

| SNO | Control parameter | Value |
|---|---|---|
| 1 | Population Size | 10000 |
| 2 | Elite Count | 4 |
| 3 | Crossover | 0.8 |
| 4 | Generations | 50 |
| 5 | Initial Population | 10000*251 double |
| 6 | Selection | Roulette Wheel |
| 7 | Crossover | Heuristic |
| 8 | Number of variables | 251 |
| 9 | Mutation | 0.01 |

TABLE V. EFFORT COMPARISON

| SNO | Actual Effort | Effort using COCOMO model | Effort using Hybrid model (GANN) |
|---|---|---|---|
| 1 | 117.6 | 180.8134786 | 127.4616 |
| 2 | 117.6 | 168.9821569 | 124.3451 |
| 3 | 31.2 | 43.65971251 | 34.8585 |
| 4 | 36 | 37.9393787 | 37.9985 |
| 5 | 25.2 | 49.47453739 | 32.8746 |
| 6 | 8.4 | 10.38903049 | 9.4435 |
| 7 | 10.8 | 19.08279263 | 14.8734 |
| 8 | 352.8 | 449.5306924 | 399.7262 |
| 9 | 72 | 45.9777269 | 68.6443 |
| 10 | 72 | 287.291445 | 112.9243 |
| 11 | 24 | 14.08522464 | 20.6532 |
| 12 | 360 | 287.291445 | 381.3423 |
| 13 | 36 | 50.44249685 | 41.9123 |
| 14 | 215 | 686.648327 | 300.3533 |
| 15 | 48 | 51.69199249 | 50.5451 |
| 16 | 360 | 615.4091318 | 401.2176 |
| 17 | 324 | 670.4677889 | 400.2432 |
| 18 | 60 | 162.0906366 | 80.6453 |
| 19 | 48 | 51.69199249 | 50.4434 |
| 20 | 60 | 207.7411868 | 112.5542 |

TABLE VI. MRE COMPARISON

| SNO | Actual Effort | MRE using COCOMO model | MRE using Hybrid model (GANN) |
|---|---|---|---|
| 1 | 117.6 | 0.53723 | 0.08385714 |
| 2 | 117.6 | 0.436923 | 0.04735629 |
| 3 | 31.2 | 0.39935 | 0.11726962 |
| 4 | 36 | 0.053872 | 0.05551389 |
| 5 | 25.2 | 0.963275 | 0.30454762 |
| 6 | 8.4 | 0.326789 | 0.12422619 |
| 7 | 10.8 | 0.766925 | 0.37716667 |
| 8 | 352.8 | 0.27418 | 0.13301077 |
| 9 | 72 | 0.361976 | 0.04646806 |
| 10 | 72 | 2.990159 | 0.56839306 |
| 11 | 24 | 0.413107 | 0.13945 |
| 12 | 360 | 0.201968 | 0.05928417 |
| 13 | 36 | 0.40118 | 0.16423056 |
| 14 | 215 | 2.193713 | 0.396999209 |
| 15 | 48 | 0.076917 | 0.05302292 |
| 16 | 360 | 0.70947 | 0.11449333 |
| 17 | 324 | 1.069345 | 0.23531852 |
| 18 | 60 | 1.701511 | 0.34408833 |
| 19 | 48 | 0.076917 | 0.05090417 |
| 20 | 60 | 2.462343 | 0.87590333 |

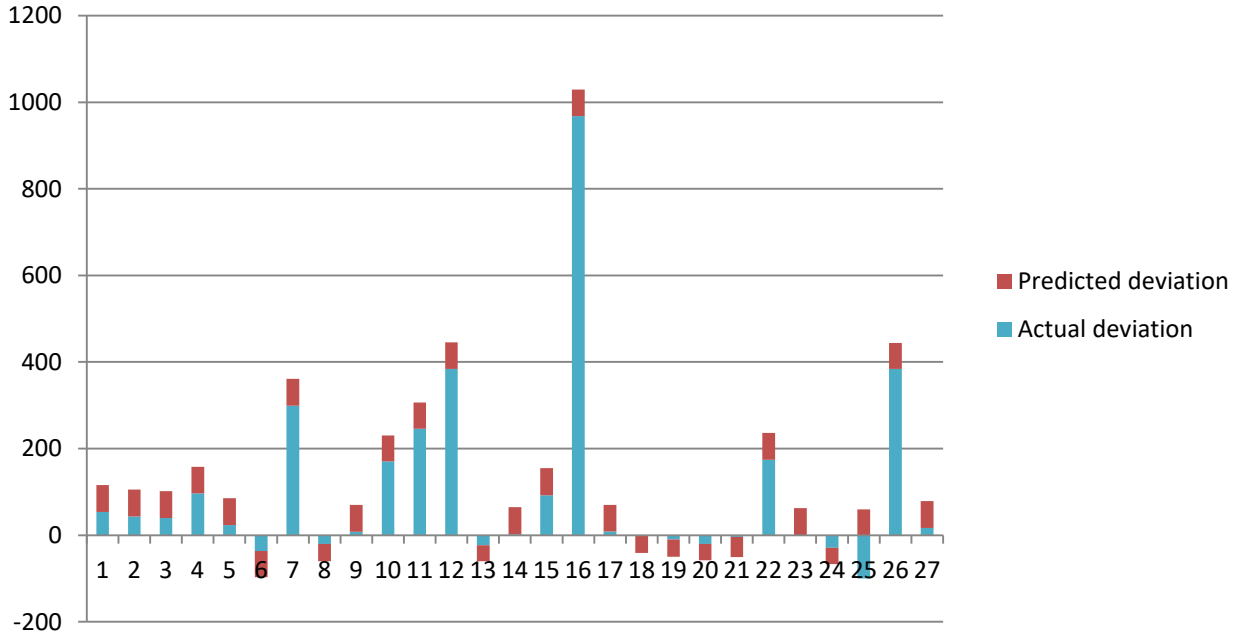



Fig. 11. Actual Deviation v/s Predicted deviation.

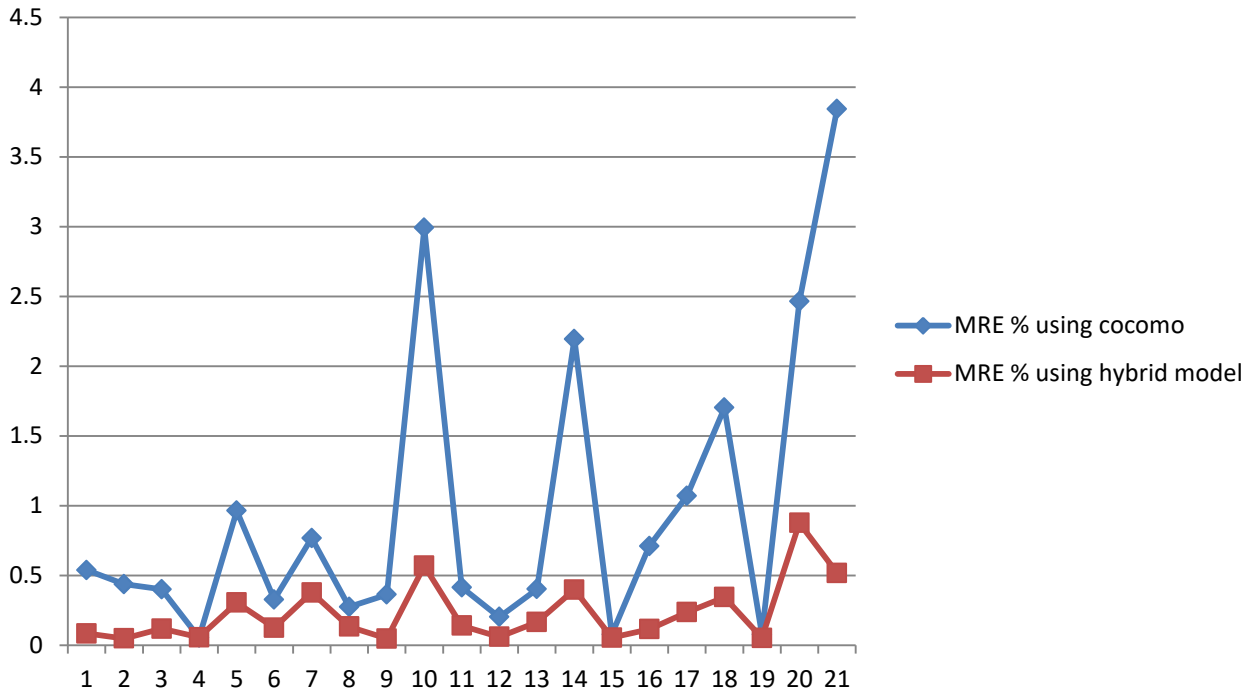



Fig. 12. MRE comparison

Fig 11 demonstrates the actual and predicted deviations from the actual effort. Where actual deviation denotes the actual percentage of overestimation/underestimation and predicted deviation denotes the predicted overestimation/underestimation. The Figure confirms that the developed model is accurate and provides optimistic view of uncertainty by effectively covering all range of deviation. Table 5 presents the computed effort values using COCOMO and proposed model. Subsequently, Table 6 presents the comparison of MRE values when effort is computed using COCOMO and with the hybrid model.

## V. Conclusion and Future Scope

The learning exposes that the suggested fuzzy logic based COCOMO II model incapacitates the uncertainty and vagueness in the inputs that is present in the conventional COCOMO and hence increases the accurateness of software effort estimation. By determining additional appropriate fuzzy rule sets and by arraying technologies like type-2 fuzzy improbability can be handled further closely and hence more precise software effort estimation is thinkable.

## References

[1] A. Trendowicz, J. Munch, and R. Jeffery, Adam Trendowicz, J¨urgen M¨unch, and Ross Jeffery, Software Engineering Techniques, Springer Heidelberg Dordrecht London NewYork,2008.

[2] The Standish Group: CHAOS Chronicles. Technical report, The Standish Group International, Inc. (2007).

[3] R. N. Charette,Why Software Fails. IEEE Spectrum ,September 2005.

[4] J.Moses., Measuring Effort Estimation Uncertainty to Improve Client Confidence, Software Quality Journal, 10, 135–148, 2002.

[5] Boehm, B.W., Abts, C., Clark, B., Devnani-Chulani, S.: COCOMO II Model Definition Manual. The University of Southern California (1997)

[6] Albrecht, A.J., Gaffney, J.R.: Software Function Source Lines of Code, and Development Effort Prediction: A Software Science Validation. IEEE Transactions on Software Engineering 9(6), 639–648 (1983)

[7] Rijwani, P., Jain, S., &Santani, D. (2014). Software Effort Estimation: A comparison based Perspective. *International Journal of Application or Innovation in Engineering and Management (IJAIEM)*, *3*(12), 18-29.

[8] Briand, L.C., Wieczorek, I.: Resource Modeling in Software Engineering. Encyclopedia of Software Engineering 2 (2001)

[9] Jorgensen, M., Shepperd, M.: A Systematic Review of Software Development Cost Estimation Studies. Software Engineering, IEEE Transactions on Software Engineering 33(1), 33–53 (2007)

[10] Wittig, G., Finnie, G.: Estimating software development effort with connectionist model. Information and Software Technology 39, 469–476 (1997)

[11] Dolado, J.J.: On the Problem of the Software Cost Function. Information and Software Technology 43(1), 61–72 (2001)

[12] Anil Kumar, C. S. Yadav et al 2012. Parameter tuning of COCOMO Model for software effort estimation using PSO. ICIAICT ISBN 978-93-81583-34-0 pp 99-105

[13] G.R. Finnie and G.E. Wittig, ―A Comparison of Software Effort Estimation Techniques: Using Function Points with Neural Networks, Case-Based Reasoning and Regression Models,‖ Journal of Systems and Software, vol. 39, pp. 281-289, 1997

[14] Heiat, Abbas. "Comparison of artificial neural network and regression models for estimating software development effort." *Information and software Technology* 44.15 (2002): 911-922.

[15] Rijwani, P., & Jain, S. (2016). Enhanced Software Effort Estimation Using Multi Layered Feed Forward Artificial Neural Network Technique. *Procedia Computer Science*, *89*, 307-312.

[16] B. T. Rao, et al., A novel neural network approach for software cost estimation using Functional Link Artificial Neural Network (FLANN), International Journal of Computer Science and Network Security, (9) (2009), pp. 126-131, 2009.

[17] I. Attarzadeh and S. H. Ow, Proposing a new software cost estimation model based on artificial neural networks, in 2nd International Conference on Computer Engineering and Technology (ICCET), (2010) pp. V3-487-V3- 491.

[18] http://promise.site.uottawa.ca/SERepository/datasets/ cocomo81.arff.

[19] P. Rijwani, and S .Jain, "Comparison and Analysis of Various Artificial Neural Networks for Software Effort Estimation". *International Journal of Advanced Research in Computer Science and Software Engineering*.

[20] Clark, Bradford, Sunita Devnani-Chulani, and Barry Boehm. "Calibrating the COCOMO II post-architecture model." Proceedings of the 20th international conference on Software engineering. IEEE Computer Society, 1998.

[21] Yang, Da, et al. "COCOMO-U: An extension of COCOMO II for cost estimation with uncertainty." *Software Process Workshop*. Springer, Berlin, Heidelberg, 2006..

[22] Hota, H. S., Shukla, R., & Singhai, S. (2015). Predicting Software Development Effort Using Tuned Artificial Neural Network. In *Computational Intelligence in Data Mining-Volume 3* (pp. 195-203). Springer, New Delhi.

[23] C. S. Reddy and K. V. S. V. N. Raju, "An optimal neural network model for software effort estimation," International Journal of Software Engineering, vol. 3, no. 1, pp. 63–78, 2010.

[24] I. Attarzadeh, et al., Proposing an Enhanced Artificial Neural Network Prediction Model to Improve the Accuracy in Software Effort Estimation, in Fourth International Conference on,Computational Intelligence, Communication Systems and Networks (CICSyN), (2012), pp. 167-172.

[25] S. K. Sehra, Y. S. Brar, N. Kaur and S. S. Sehra, Research Patterns and Trends in Software Effort Estimation, INFSOF 5836.

[26] M. Jørgensen, A review of studies on expert estimation of software development effort, Journal of Systems and Software 70 (1) (2004) 37–60. doi:10.1016/ S0164-1212(02)00156-5.

[27] M. Jørgensen, T. M. Gruschke, The impact of lessonslearned sessions on effort estimation and uncertainty assessments, IEEE Transanctions on Software Engineering 35 (3) (2009) 368–383. doi:10.1109/TSE. 2009.2.

[28] A. Magazinius, S. B¨orjesson, R. Feldt, Investigating intentional distortions in software cost estimation - an exploratory study, Journal of Systems and Software 85 (8) (2012) 1770–1781. doi:10.1016/j.jss.2012. 03.026.

[29] Magne Jørgensen and Martin Shepperd, A Systematic Review of Software Development Cost Estimation Studies, IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 33, NO. 1, JANUARY 2007.

[30] Kjetil Moløkken and Magne Jørgensen, A Review of Surveys on Software Effort Estimation, Proceedings of the 2003 International Symposium on Empirical Software Engineering (ISESE'03) 0-7695-2002-2/03.

[31] N.Kushwahal and Suryakane, Software Cost Estimation using the Improved Fuzzy Logic Framework, 978-1-4799-3064-7/14,IEEE.

[32] E. Manalif , L.F. Capretz, and D. Ho, Fuzzy Rules for Risk Assessment and Contingency Estimation within COCOMO Software Project Planning Model, DOI: 10.4018/978-1-4666-4785-5.ch006.

[33] Reifer D.J., Boehm B.W. and Chulani S., "The Rosetta Stone: Making COCOMO 81 files work with COCOMO II", University of South California.

# A Controlled Environment Model for Dealing with Smart Phone Addiction

Irfan Uddin[a], Adeel Baig[b], Abid Ali Minhas[b]
[a]Institute of Computing, Kohat University of Science and Technology, Kohat, Pakistan
[b]College of Engineering , Al Yamamah University, Saudi Arabia

***Abstract*—Smart phones are commonly used in most parts of the world and it is difficult to find a society that is not affected by the smart phone culture. But the usage of smart phone is crossing the limit of being used as a facility towards high level of abnormal dependency on the phone. This dependency can reach to the point where we have no longer control on the over-use and hence the negative impacts it can cause to our lives. The worst situation is that people do not even consider that this dependency is actually a type of addiction and we need to find some solutions to deal with it. In this research paper, we identify symptoms that show the existence of smart phone addiction and demonstrate that this addiction has an effect on the quality and even quantity of people' lives and it can ultimately affect the whole society. We propose solutions to deal with smart phone addiction and propose the design of a smart phone application to reduce the level of abnormal dependency on smart phones.**

***Keywords*—Smart phone addiction; Abnormal use of smart phone; Healthy society; Dealing with smart phone addiction**

## I. Introduction

Smart phones have become one of the most omnipresent communication devices within the last decade. These devices allow people to connect "any-time", "anywhere" and with "anybody", with the added benefit of mobility and portability. A smart phone operates not only as a mobile phone but as a computer, mp3 player, video player, internet browser, book reader, gaming device, entertaining device, etc. [1], [2]. People are rapidly adapting smart phones which can be attributed to many factors such as portability, declining costs, everything at one place and rich features. It is not possible to find a society that is exempted from smart phone culture. With the advancement of technology, smart phones are becoming smarter every generations. In a short time span we find many new applications and features added to smart phones that allow us to perform different functions suitable to our lifestyle [3]. For example, the development of *Siri* in Apple's phones have allowed people to perform different tasks by just talking to the phone. Thumb impression features have allowed to add security in different operations. Payments through NFC have reduced the dependency on credit cards, and many other features. We do not know what features we will have in smart phones after four or five years.

Smart phones are rapidly adapted, but do they have any negative social impacts? For example, the most common negative consequences of smart phone use include the dangers of driving while using smart phone. This overuse can be termed as *smart phone addiction* and it can cause detrimental damage both to individuals and to society [1]. Addicted people have a physical or psychological dependence on an entity which disable their functionality within society. Smart phone addicted people have a high dependence on theirs phone and may not able to perform simple tasks without using their smart phone. This situation can be termed as *smart phones but dumb people*. The study of dependence on smart phones to an extent that the individual's social life is effected has gained little academic attention. The dependency on the smart phones is so much integrated in our lives that people do not even recognize that they are addicted to an entity and they need to find some remedies.

Smart phones are extraordinary devices. We can communicate with each other without time and space restrictions. However, people feel that they can not live a normal life without it [2]. Therefore, the main objective of this paper is to provide some means of making an empirical distinction between normal and problematic smart phone use. This study states that if problematic smart phone use exists, and what activities lead to the abnormal dependency on smart phones. We attempt to identify how the problematic use of smart phone has an effect on the quality or even quantity of our lives. We propose solutions to deal with the problematic use of smart phones through environmental or behavioral programming. We want to defend ourselves against smart phone addiction by using smart phones. Therefore, we propose a model that can automatically determine the abnormal usage pattern of the smart phone and enable or disable different features to get the user out of the danger zone of smart phone addiction.

The rest of the paper is organized as follows. In section II we give definition of addiction and the symptoms of smart phone addiction. We demonstrate how this addiction can affect the quality and/or quantity of ours lives in section III. We present related works in section IV and present a case study that demonstrates the dependency on smart phones in section V. We identify activities on the smart phones that can possibly lead to addiction and present some solutions that advices the user to deal with the addiction in section VI. We present the proposed model which can automatically detect the abnormal behavior of the user on the smart phone and enable or disable different notifications, messages, email or other applications in section VII and conclude the paper in section VIII.

## II. The Signs of Smart Phone Addiction

An Addiction can generally be defined as *the repeated use of a substance despite the negative consequences suffered by the addictive individual* [4]. Traditionally addiction has been related to only substance addiction [5]. Examples of substance addiction include smoking, drugs, medication, alcohol

etc. More recently addiction has been expanded to include behavioral addiction [6]. This type of addiction is defined as *the repetition of a particular behavior despite the negative consequences suffered by the addictive individual*. Examples include gambling, excessive eating, exercise, internet or cell phone usage etc. More generally we can state that *any entity that can produce a pleasurable sensation can be addictive* [7]. Cell phones usage can potentially become addictive [8]. With the advancement of technology, all cell phones are becoming smart phones and it is becoming rare to find a portable phone that is not smart phone. In this paper we particularly discuss the addiction of smart phones (not traditional cell phones).

We have lost control over the habitual smart phone use. For example, when we are bored, the first thing we reach is generally our smart phone. Or we have heard many stories of driving while using smart phone that resulted into injuries or possibly the loss of lives [9], but we continue to use smart phones while driving. Or we have heard stories of employees losing their jobs because of using smart phone excessively at work [10] but we continue to use smart phone at work. Or students receive many warnings and possibly reduced GPA because of using smart phone during lecture time, but they cannot stop [11]. This loss of control is a sure sign that we are addicted to smart phones [12]. We are repeatedly using smart phones despite the negative consequences. In this section we briefly explain six signs of smart phone addiction (originally presented in [7]).

*1) Salience: The integration of smart phone in the daily routine of an individual is termed as salience*. For example, people can be involved in using smart phone while watching movies, religious activities or intimate moments. One might involve in taking selfies at different awkward situations like funerals or changing rooms. People might sleep with their smart phone next to their bed. Many people confess that the last thing they see when they go to bed and the first thing they see when waking up from sleep is their smart phone. We have seen people taking selfies when they perform Umra or Hajj (Muslim religious activities). Not only they are taking selfies but they are constantly sharing them on the social media and while performing their Umra or Hajj they keep checking the number of responses on the social media. We have made a transition from the simple capturing of photos to sharing with people.

*2) Euphoria: The use of smart phone is often followed or preceded by the excitement or anticipation. This feeling is called Euphoria or mood modification*. We do not know what we can possibly have behind the whistle or beep of the smart phone, and feel really excited to check it immediately. For example, when we are bored we reach our smart phone in order to play games or go to online social networking. We use smart phone in order to avoid awkward situations e.g. pretend to take calls. All these activities take us from one form of mood to another form of mood where we feel excited.

*3) Tolerance: The behavior is repeated with ever-increasing need*. Research has found that the longer someone has had their smart phone the more they are likely to use it [7]. It is not the case that you own the smart phone and then after some time you get bored from it as it is normally the case with other things like a house, furniture, car etc. The desire of using the smart phone again and again increases.

*4) Withdrawal: When we are separated from our smart phone we feel anxious or depressed or even get panic*. People have found that they feel isolated or lonely when they are away from theirs phones. People have an irrational fear of losing their phone. A simple question to check the withdrawal symptom can be *how long does it take to replace when your smart phone is broken, lost or stolen?* The sure answer for many is "not very long".

*5) Conflict: When we are addicted to smart phone then the common outcome is conflict with other people*. For example, parents have arguments with their children because they do not listen to them. Children complain about their parents of not getting enough attention. Teachers are in conflict with their students, or employers are in conflict with their employee.

*6) Relapse: When we realize that using the smart phone has an effect on the quality of our lives but when we try to attempt to stop the usage, we cannot stop it*. It is like a bad habit say smoking, we try to stop it but then we relapse after a short time. The same is true for smart phone usage. You try to keep your self away for some short period of time, but then you get a relapse and use the phone for even longer time.

## III. The Effect of Smart Phone Addiction on the Quality and/or Quantity of Lives

One is not addicted to a smart phone the moment he gets a smart phone. But it starts slowly and gradually. For example, at the first place, we own a smart phone mainly for safety reasons. Then we are involved in sending and receiving text messages and then start online social networking. This continues to use smart phone in an extremely dangerous activity like texting while driving. Ultimately we reach to a point where we have no control on the negative consequences from over-using the smart phone. Because of this dependency, the quality or even quantity of our lives are compromised.

The same neural circuitry experienced with substance addiction is activated with smart phone addiction [7]. Early detection of any substance or behavioral addiction is essential for easy treatment. Because when one addiction exists the likelihood of another addiction hiding in the shadows increases [7]. It means that when a smart phone addiction exists then the chances of getting addicted to another behavior or substance increases. Similarly, in addition of treating an addiction others addictions may pop up to take its place [7]. For example, we might treat smart phone addiction but in order to fill the place of addiction in our brain we might become addicted to another entity or activity.

### A. Quality of life

*1) Relationship dissatisfaction:* Phubbing refers to ignoring someone or ignored by someone by using a smart phone [7]. We all have been in situations where we are either phubbed by someone, or we have phubbed someone. This phubbing can continue in relationship, where partners can phubb each other. It has been observed that smart phones are the major distraction in relationships, because partners spend more time with their smart phones than they should with each other. This distraction is important to consider because it can affect relationship satisfaction and ultimately satisfaction with our lives [7].

*2) Parent/Child relationship dissatisfaction:* Smart phone addiction can have an affect on the relationship between parents and children. If parents are addicted to smart phones, then they may not pay enough attention to their growing children and hence the quality of life of the children are compromised [7]. On the other hand, if adults are addicted to smart phone then it is possible that they start ignoring their older parents who might need more attention. Hence the quality of lives of the parents are compromised.

*3) Performance:* In case college students are addicted to their smart phone then it might have an effect on their performance at school which can result into lower GPA or possibly dropping from school [1]. Generally, adolescents with poor academic achievements usually receive less respect from surrounding people. Similarly, in case of employees addicted to their smart phones then it is possible that they lose promotion and even get fired from work because their performance is not to the level required by his/her employers.

*4) Multitasking:* Multitasking (i.e. performing multiple tasks at the same time) reduces efficiency and performance because the brain can only focus on one thing at a time. When we try to perform multiple activities at the same time, our brain lacks the capacity to perform multiple tasks successfully. Roughly 68 percent of the population has an IQ between 85 and 115. Because of the smart phones, people are multitasking, but this multitasking (in reality task switching [7]) has an affect on the IQ by 10 to 15 points, a worse effect than smoking marijuana or losing a night's sleep. It has been found that only 2 percent people can multitask effectively, for the rest of 98 percent multitasking can do more harm than good.

*5) Limited attention span:* The amount of concentrated time one can spend on task without becoming distracted is called the attention span. In 2000 it was measured that the average attention span of human is 12 seconds. In 2013 by Microsoft it was measured that the attention span of human is reduced to 8 seconds which is one second less than the attention span of a goldfish. This reduction is very much attributed to the development of smart phones and social media. Because, the moment we start to focus on something we receive an email or a notification and our attention is diverted to deal with that notification. We are getting less focused.

### B. Quantity of life

*1) Accidents:* It is becoming common that people are using their smart phones while driving. Despite many warnings and strict driving policies by the government in many countries it is difficult to control the use of smart phones while driving [13]. This distraction during driving can be dangerous because the driver can get into accident and put his life or other people' lives in danger [14]. This way the quantity of our lives are compromised.

*2) Apathetic bystanders:* A Bystander effect or apathetic bystander is a social psychological phenomenon that refers to cases in which individuals do not offer any means of help to a victim when other people are present [15]. In these situations the probability of help is inversely related to the number of bystanders. Because people think that someone else can better help the victim or his help will not affect anything. With the smart phone use, we are becoming apathetic bystanders, because when we use our smart phone we are glued to it in such a way that we do not notice what is happening in our surroundings. The worse case is that when we have our smart phone and some body needs our help, instead of offering our help we try to take pictures or selfies with the situation. A famous sociologist Erving Goffman likens technology users as mental patients who treat others as if they do not exist [16].

One such incident happened in September 2013 in a metro station in San Francisco state. One day, people at the metro station were waiting for metro but they were glued to their smart phones. The CCTV camera was recording the whole situation. A murderer came to the station and he had a pistol in his hand. He waved his hand and he even pointed to some people standing there and glued to their phones, but nobody even noticed. He then randomly chose a guy named Justin Valdez[1], 20 years old, and shot him. This is an extreme example of apathetic bystanders, because those people standing there could have taken some actions to save the victim, but because they were glued to their phones nobody even noticed the murderer until the guy was shot dead. Using smart phones, we are turning to apathetic bystanders and it can possibly affect the quantity of our lives.

## IV. Related Work

Technology addiction has been discussed in the literature for a long time. First paper on Internet addiction was published in 1998 [17]. Mobile devices, being the fastest growing digital devices among other digital products/services across the world, have become a primary driver of internet use today. This section presents recent studies conducted in the field of smart phone addiction.

Authors in [18] investigate the adoption of low cost smart phone and its influence on the students in a selected university in Nigeria. The study shows that mobile phones led to increase in negative social behavior among the young adolescents. In some cases, smart phone helped students to improve their academic performance as well. A study in [19] presents the smart phone addiction in Omani university students. According to the study, Oman has nearly 100% mobile phone penetration. The study shows that trend of students is towards buying smart phones and they feel uncomfortable without their smart phones. Another interesting study [20] explores the frequency and indices of smart phone addiction in a group of Saudi University students. Indices selected were overuse of smart phone, the technological dimension, the psychological-social dimension, preoccupation with smart phones, and the health dimension The results show that the addiction percentage among the participants was 48%.

Authors in [21] have analyzed the relationship between university students perceived life stress and smart phone addiction. Study concludes that academic stress had negative impact on social and learning self-efficacies, which as a result increase smart phone addiction. Study also reveals that the family and emotional stresses also contribute to smart phone addiction. A study in [22] presents the smart phone usage by Mauritians. Study shows the user trends in using different smart phone services and applications. This trend helps mobile operators

[1] http://www.seattletimes.com/opinion/death-of-justin-valdez-a-disconnect-caused-by-technology/

and other businesses in developing and providing value-added services to the Mauritians users. Another paper [23] examines the educational and social uses of smart phone by students in South Africa. The study provides insight into the usage of mobile telecommunication services for different academic and social activities from students' perceptions. The study suggests that the integration of mobile telecommunications into teaching and learning in universities will facilitate student-centered learning.

A study [24] investigates smart phone usage patterns among students in Serbia. The research aims to provide an insight into the purposes and time distribution of the students' use of their mobile devices. By summarizing the answers given by the participants, authors conclude that 35% of the students come in the category of smartphone addicts. Authors in [25] investigate how extensive exposure motivates users to use smart phone addictively. Results of an online survey with 384 respondents suggest that there is a strong relationship between convenience, habits and addiction. Convenience facilitates the formation of habit, which generates users' reactions of concentration and enjoyment, which further result in smart phone addiction.

There are different applications developed to deal with Smart phone addiction. These applications include: *Breakfree*, *Moment*, *AppDetox*,*Offtime*, *Flipdapp* etc. Breakfree is developed for both iOS and Android platforms, Moment is only available is iOS, AppDetox and Flipdapp are developed only for Android platforms. All these applications blocks certain applications in different duration of time. For example, it might disable Facebook to be used at work time and might disable Outlook during the weekend. We are taking a different approach in the sense that we do not restrict a particular usage pattern on the smart phone until it is leading us to an abnormal usage. For example, a user is allowed to use Facebook during work as long as that usage is not leading towards addiction. We may also be allowed to use work related applications during the weekend if there are requirements and it does not affect the quality of our lives. We analyze the behavior of the user during different times and recommend the usage of the smart phone in such a way that it does not bring any detrimental effects to the society.

## V. Case Study that Shows the Existence of Smart Phone Addiction

We have conducted a study to measure the existence of smart phone addiction in college students. We asked the students of Al Yamamah University[2], kingdom of Saudi Arabia to participate in the survey. The number of male and female students participation is given in table II. Most of the participants are college undergraduate students. A total of 73% are in age range of 18-24 years, while 22% are in the range of 25-35 years. 70% of the participants have stated that they have a data plan in their smart phones, and 52% of the participants have indicated that they spend more than the average expenses on the smart phones (the average expense on phone is 300 SAR in Kingdom of Saudi Arabia). The number of SMSes and calls are shown in table I. As can be seen that 81% participants made less than 5 SMSes per day and 53% participants made less than

[2]http://www.yu.edu.sa

TABLE I. The Male and Female Participation in the Study

| Sex | No. of respondents | Proportion(%) |
|---|---|---|
| Male | 48 | 75% |
| Female | 16 | 25% |
| Total | 64 | 100% |

TABLE II. Frequency of SMSes and Calls Per Day

| Frequency | SMSes | Calls |
|---|---|---|
| Less than 5 | 81% | 53% |
| In the range 5-15 | 8% | 34% |
| Greater than 5 | 11% | 13% |

5 calls per day. This indicates that the fundamental reason of owning a smart phone for security reasons is vanishing. But as can be observed later that smart phones are used for other activities and most of those activities are possible because of having a data plan in smart phones.

We have asked our participants to indicate the number of hours spent on their smart phone and the response is shown in fig. 1. As can be seen that 65% of the participants have stated that they spend more than two hours per day on their smart phone. Only 7% have indicated that they spend less than 30 minutes on their phone. This indicates a high dependency on the usage of the smart phone. We also have collected data to know the type of phone used by the participants. 70% have indicated that they use Apple's iPhone, 20% have indicated that they use Samsung and only 9% participants have indicated that they use other type of phones. This shows that the leader in the smart phone market is Apple. In our survey we have observed that 65% of the participants change their phone not less than two years.

We have collected the information about the activities performed on smart phones, and it is shown in fig. 2. As can be observed that *Youtubing* is the most common activity performed on the smart phone. The only activity that is not very popular is movie making. It can also be observed that Send/Receive text activity is shown to be performed very frequently, which is the opposite of the observation that indicated that less than 5 SMSes are sent per day. We think that participants have indicated the send and receive messages through social messages services such as WhatsApp, Line etc.

We have created a questionnaire based on the addiction

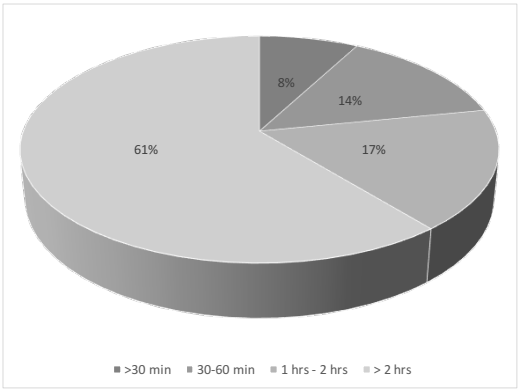



Fig. 1. The amount of time spent on the smart phone per day

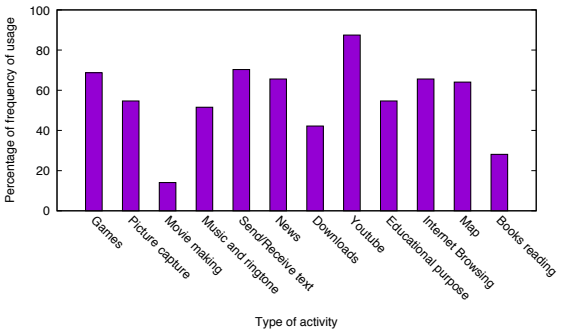



Fig. 2. The frequency of activities performed on smart phone

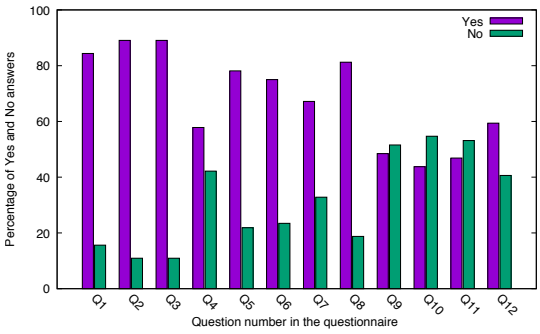



Fig. 3. The percentage of Yes and No answers for all individual questions asked in the survey

TABLE III. QUESTIONNAIRE TO TEST THE ADDICTION OF SMART PHONE

| |
|---|
| *Salience:* |
| Q1. Is the first thing to reach for after waking in the morning is your smart phone? |
| Q2. Do you sleep with your smart phone next to your bed? |
| *Euphoria:* |
| Q3. Do you use your smart phone when you are bored? |
| Q4. Have you pretended to take calls to avoid awkward social situations? |
| *Tolerance:* |
| Q5. Do you find yourself spending more and more time on your smart phone? |
| Q6. Do you think that you spend more time than you should on your smart phone? |
| *Withdrawal:* |
| Q7. Do you become irritable when you are away from your smart phone? |
| Q8. Do you have fear of losing your smart phone? |
| *Conflict:* |
| Q9. Have you been in arguments with friends/family about your smart phone usage? |
| Q10. Do you use your smart phone while driving? |
| *Relapse:* |
| Q11. Have you tried to reduce time spent on your smart phone but you could not? |
| Q12. Do you think that you should reduce the time spent on phone but could not? |

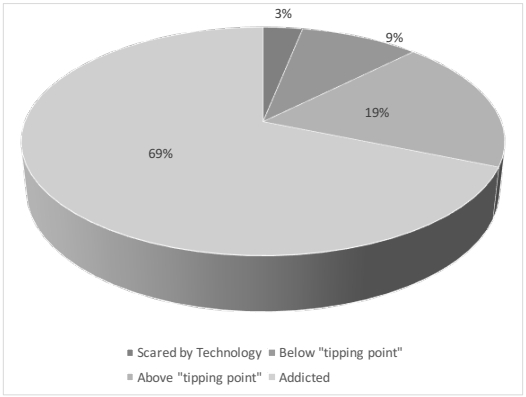



Fig. 4. Smart phone addiction in college students

level measured in [7] and asked the questions given in table III. The scale of measuring smart phone addiction is given in table IV. Those who have answered "Yes" to 2 or less questions are either not interested in technology or have a very high self control. We term them as "scared by technology". Those who have answered "Yes" between 3 to 4 questions are below the dangerous zone called the "tipping point". The "tipping point" is the point where the liking of a smart phone is converted to become a need. Answering "Yes" to 5 to 7 questions means that they have crossed the limit and have already entered the dangerous zone and are moving with full steam towards smart phone addiction. Answering "Yes" to 8 or more questions mean that the individual taking the test has become smart phone addictive.

We have recorded the number of Yes and No answers to each individual question asked in the questioner and is shown in fig. 3. For questions 1 to 8 the percentage of Yes answer is higher than No answers. The higher percentage of No answers for Q9 indicates that they do not get into arguments with friends/family. A possible assumption can be that their friends/family are also addicted to smart phone. The higher percentage of No answers to Q10 is because 50% of the population are not allowed to drive in Saudi Arabia. The worst situation is that most of the people have not even tried to reduce the over time spent on the smart phone. This can be shown by the higher percentage of No answers for Q11. But the high percentage of Yes answer for Q12 indicates that people think that they should reduce the time spent on their smart phone. This is considered a good sign for improvement.

TABLE IV. THE SCALE OF MEASURING SMART PHONE ADDICTION

| Answers | Description |
|---|---|
| 8+ "Yes" answers | You are addicted to smart phone. |
| 5-7 "Yes" answers | You have crossed the "tipping point" and are moving full-steam ahead to smart phone addiction. |
| 3-4 "Yes" answers | You have not yet reached your "tipping point" but need to carefully assess how your smart phone is impacting your life. |
| 0-2 "Yes" answers | You are either living in a monastery or at least have a very high patience and self-control. Or, technology simply scares you. |

The result of smart phone addiction based on the scale given in table IV is given in fig. 4. It can be seen that only 9% students are below the "tipping point". 19% participants are above the "tipping point" and 69% participants are actually addicted to smart phones and need a way to deal with this addiction.

## VI. ACTIVITIES ON SMART PHONES THAT CAN LEAD TO ADDICTION AND SOLUTIONS TO DEAL WITH IT

The six signs of addiction given in section II can be observed in different activities that we can perform on smart phones. The reward gained from these activities encourage higher involvement with more and more time spent in the particular behavior. These activities can generally be classified into three categories based on different theories.

*1) Escape theory:* Those activities that we get involved in order to avoid or get away from an unhappy situation. For example, students use their smart phone in the class because

they want to escape from listening to a lecture that is not interesting for them. Or in order to avoid a social awkward situation we pretend to talk on our phones. Or in order to avoid social isolation we check some things on our phones. Any activity on the phone that takes us from an unhappy situation to a happy situation can fall in the category of Escape theory.

*2) Learning theory:* Those activities on the smart phone that help us in increasing our knowledge. For example, accessing social media in order to get knowledge about people/world. Using new apps on the phone to learn something new. Or when we read a text or email we feel that we have completed a task and learned something new. The reward of learning that we gain from different activities lead to addiction.

*3) Law of effect:* When a behavior is followed by an effective reinforcer the behavior is most likely to happen again. For example, when we pass a waiting time by playing a game on our phone, it is more likely that the next time we are in waiting we are more likely to play the same or similar games again.

### A. Dealing with smart phone addiction

The use of smart phone can be both freeing and slaving. It is liberating in the sense that we are constantly in touch with our friends, family, colleagues or even strangers. We have access to an endless amount of data and continue to live our lives without restrictions of being tied to a particular location. We are master of our domain when we have access to our smart phone, because everything is just a touch away from us. On the other hand, the use of smart phone can be enslaving, because it leads to dependence and more restrictions. For example, having a smart phone, we are reachable 24/7. Our employer can reach us at any time and we are obliged to reply to his/her requests. It might be that at 2:30 AM we feel to reply to an email because it is important. Or the less number of likes to a post on social media can let us feel that no body cares about us.

To quote Richard Branson founder of Virgin Group, "Do not become a slave to technology - manage your phone do not let it manage you". We need to manage our level of smart phone usage. The good thing is that smart phone are poor masters but very good servants. In order to control our smart phones, we have to exercise some self-control. We must be aware of the usage of our smart phone habits and know what are the activities that we are spending most of the time on smart phone rather than spending in doing a healthy activity. Once we have a list of activities that leads to smart phone addiction we can make plan to cut the amount of time we are spending. For example, instead of spending one hour on social media we can spend some time with friends/family on a cup of coffee. Once we make a plan, we must try to firmly execute it. We can involve friends/spouse to help us executing our plan. For example, they can act as a judge, jury or prosecutor.

We can also program our behavior or environment. A behavioral programming is like self-praising when we achieve to control some time spending on social media. It can be a mental image of getting promotion at work or for students it can be a mental image of getting A+ in a course. We can also offer ourselves a chocolate or a meal out to reward our control. When we fail to achieve a mile stone in dealing with smart phone addiction we can criticize ourselves. We can create a mental image of losing a promotion or failing a course. We can cut the budget or time that we are going to spend on entertainment to criticize ourselves.

Environmental programming is to avoid temptation rather than resist it. For example if we are addicted to shopping the good solution is to avoid going to shopping centers. In case of smart phone addiction, it can be when the phone is in our pocket we are going to check it when it buzzes. A good solution could be to turn off the phone during the activity we do not want to reach to phone. For example, in case of driving instead of putting our phone in pocket, it is better to turn it off and put it in the trunk of the car. This way we will not be tempted to reach to our phone while driving. We can also define smart phone free zones like bed room, dining table etc.

In order to deal with smart phone addiction, we must have to go off the digital grid some times. We are not against technology, we love all the facilities that technology has brought in our lives, but we are saying that the quality or quantity of our lives should not be compromised because of smart phone addiction.

## VII. Proposed Model of Application Dealing with Smart Phone Sddiction

The traditional clinical approach for screening and assessment in behavior addictions has been surveys and interviews, and this also applied to smart phone addiction. However, that approach has serious drawbacks and needs to be complemented by objective assessment analysis. First, in interview-based methods, it is difficult to follow the change of subjects status, especially when there is a massively large number of subjects as in smart phone addiction. Second, the survey questions are often subjective and the replies from those surveys depend upon their seriousness, mood, and attitude to the survey. Hence, objective and quantitative usage needs to be recorded to perform diagnosis and treatment based on concrete and accurate data.

[3]

We propose the design of an application that observes the behavior of the person and controls the level of addiction. We call the application as *Smart Phone Addiction Controlled Environment* (Space). The application keeps analyzing all the activities on the phone. There are different software sensors which measures the level of addiction on the smart phone. As soon as an activity is measured to be addictive, the program puts that activity in observation. The program might impose restriction on using the activity or might disable completely until the person is out of the danger zone. The design of *Space* is shown in fig. 5.

## VIII. Conclusion

Smart phones have been deeply penetrated in our lives and we feel that this dependency can reach to a point where a simple benign use of smart phone can have dangerous results. We demonstrated different symptoms of the dependency to demonstrate that smart phone addiction exists and it need to be treated. We cannot afford individuals who put their lives

[3] http://talkmoretechless.com/

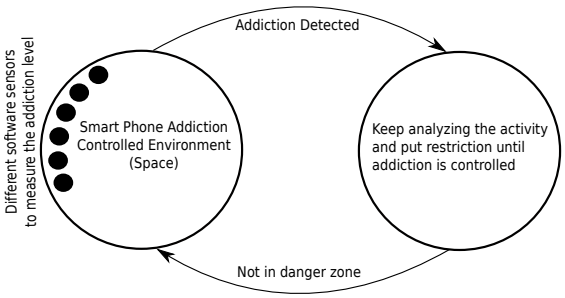



Fig. 5. Smart Phone Addiction Controlled Environment (Space)

or other people' lives at risk. We proposed different ways which help individual to get out of the danger zone of the smart phone addiction. Yet these solutions are very subjective. We think that an objective solution is required to keep people out of the danger zone of smart phone addiction and that is possible only to develop smart phone application to deal with the situation. This way we defend ourselves against the overuse of technology by using the technology. It is worth to mention a community *TalkMore TECH LESS* which is started for the cause to let people cut the over use of smart phone and connect people with each other.

## REFERENCES

[1] H. Kim, "Exercise rehabilitation for smartphone addiction," *Journal of Exercise Rehabilitation*, vol. 9, no. 6, pp. 500–505, 12 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3884868/

[2] H. Lee, H. Ahn, S. Choi, and W. Choi, "The sams: Smartphone addiction management system and verification," *Journal of Medical Systems*, vol. 38, no. 1, 2014. [Online]. Available: http://dx.doi.org/10.1007/s10916-013-0001-1

[3] M. Kwon, D.-J. Kim, H. Cho, and S. Yang, "The smartphone addiction scale: Development and validation of a short version for adolescents," *PLoS ONE*, vol. 8, no. 12, p. e83558, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3877074/

[4] T. E. Robinson and K. C. Berridge, "The psychology and neurobiology of addiction: an incentive–sensitization view," *Addiction*, vol. 95, no. 8s2, pp. 91–117, 2000. [Online]. Available: http://dx.doi.org/10.1046/j.1360-0443.95.8s2.19.x

[5] W. J. Lynch, K. L. Nicholson, M. E. Dance, R. W. Morgan, and P. L. Foley, "Animal models of substance abuse and addiction: Implications for science, animal welfare, and society," *Comparative Medicine*, vol. 60, no. 3, pp. 177–188, 06 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2890392/

[6] L. Clark and E. H. Limbrick-Oldfield, "Disordered gambling: a behavioral addiction," *Current Opinion in Neurobiology*, vol. 23, no. 4, pp. 655 – 659, 2013, 23/4 Addiction. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959438813000202

[7] J. Roberts, *Too Much of a Good Thing: Are You Addiction to Your Smartphone*. Santia publishing, 2015.

[8] P. Robert, S. Livia, S. Cisar, R. Andrea, and B. Daniel, "Your constant companion engineering students and their mobile phones," in *Intelligent Systems and Informatics (SISY), 2015 IEEE 13th International Symposium on*, Sept 2015, pp. 161–165.

[9] K. E. Beede and S. J. Kass, "Engrossed in conversation: The impact of cell phones on simulated driving performance," *Accident Analysis & Prevention*, vol. 38, no. 2, pp. 415 – 421, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457505001867

[10] W. M. Baker, E. J. Lusk, and K. L. Neuhauser, "On the use of cell phones and other electronic devices in the classroom: Evidence from a survey of faculty and students," *Journal of Education for Business*, vol. 87, no. 5, pp. 275–289, 2012. [Online]. Available: http://dx.doi.org/10.1080/08832323.2011.622814

[11] L. Leung, "Linking psychological attributes to addiction and improper use of the mobile phone among adolescents in hong kong," *Journal of Children and Media*, vol. 2, no. 2, pp. 93–113, 2008. [Online]. Available: http://dx.doi.org/10.1080/17482790802078565

[12] W. Park, "Mobile phone addiction," in *Mobile Communications*, ser. Computer Supported Cooperative Work. Springer London, 2005, vol. 31, pp. 253–272. [Online]. Available: http://dx.doi.org/10.1007/1-84628-248-9_17

[13] A. Ferreira, G. Bianchi Piccinini, S. Rôla, and A. Simões, "Gender and age-related differences in the perception of in-vehicle mobile phone usage among portuguese drivers," *Intelligent Transport Systems, IET*, vol. 7, no. 2, pp. 223–229, June 2013.

[14] G. Porter, "Alleviating the dark side of smart phone use," in *2010 IEEE International Symposium on Technology and Society*, June 2010, pp. 435–440.

[15] J. M. D. BIBB LATANÉ, "Bystander "apathy"," *American Scientist*, vol. 57, no. 2, pp. 244–268, 1969. [Online]. Available: http://www.jstor.org/stable/27828530

[16] E. Goffman, *Asylums : Essays on the Social Situation of Mental Patients and Other Inmates*. Anchor, Nov. 1961. [Online]. Available: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0385000162

[17] K. S. Young, *Internet addiction: The emergence of a new clinical disorder.* CyberPsychology and Behavior, 1998, vol. 1(3), no. 237-244.

[18] S. O. Alaba, "An empirical survey of the rate of mobile phone adoption, its influence on social behavior, schooling and academic works of adolescent students in osun state, nigeria," *Education and Management Technology (ICEMT), 2010 International Conference on*, pp. 630–635, 2010.

[19] R. Belwal and S. Belwal, "Mobile phone usage behavior of university students in oman," *New Trends in Information and Service Science, 2009. NISS '09. International Conference on*, pp. 954–962, 2009.

[20] S. S. Aljomaa, M. F. Al.Qudah, I. S. Albursan, S. F. Bakhiet, and A. S. Abduljabbar, "Smartphone addiction among university students in the light of some variables," *Computers in Human Behavior*, vol. 61, pp. 155 – 164, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0747563216302126

[21] S.-I. Chiu, "The relationship between life stress and smartphone addiction on taiwanese university student: A mediation model of learning self-efficacy and social self-efficacy," *Computers in Human Behavior*, vol. 34, pp. 49 – 57, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0747563214000296

[22] R. K. Moloo, "Mobile phone usage behaviour in mauritius," *2011 International Conference on Business Computing and Global Informatization*, pp. 607–610, 2011.

[23] O. K. Abatan and M. Maharaj, "Analyzing educational and social usage of mobile telecommunications in a south african university," *Information Society (i-Society), 2014 International Conference on*, pp. 64–69, 2014.

[24] P. Robert, S. Livia, S. M. Čisar, R. Andrea, and B. Daniel, "Your constant companion — engineering students and their mobile phones," *2015 IEEE 13th International Symposium on Intelligent Systems and Informatics (SISY)*, pp. 161–165, 2015.

[25] C. Chen, K. Z. K. Zhang, S. J. Zhao, M. K. O. Lee, and T. Cong, "The impact of mere exposure effect on smartphone addiction," *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pp. 1507–1514, 2016.

# On the Distinction of Subjectivity and Objectivity of Emotions in Texts

Manh Hung Nguyen[1,2]
[1]Posts and Telecommunications Institute of Technology (PTIT)
[2]UMI UMMISCO 209 (IRD/UPMC), Hanoi, Vietnam

***Abstract*—Emotion classification in texts is an instance of the text classification problem. It therefore could apply some existing text classifiers by considering each emotion as a label of the text. However, most of recent works does not differentiate the subjectivity and objectivity of the same emotion in the text. This paper firstly builds some datasets whose labels are emotion, in which the subject and object of the same emotion are considered as two separated labels. Secondly, this paper evaluates some existing classifiers via some scenarios on the built datasets. The results are then discussed on some difficulties of these kinds of problem.**

***Keywords*—*Text classification; Emotion classification; subjective emotion; objective emotion***

## I. Introduction

Emotion classification in text (or emotion recognition from text) is one of popular instances of the text classification in general. It could be applied into several applications. For instances, in the application of consumer sentiment analysis where the detection of emotion in the consumer feedback may help other consumer to choose the best seller or help sellers to improve their services. In the application of home-machine interaction where the computer and the user can take some conversations, recognition of user emotion in the conversation text may help the computer to improve the effect of the conversation. In the application of measurement the similarity among social network users, recognition of emotion in the posts (status), or the comments of users on some other's posts could help us to measure the similarity among users' opinion on the topic. These data may then help us to further analyse and/or predict the similarity of these users on their interests or their behavior of online shopping, etc.

As this is an instance of the text classification problem, it thus could use some existing (and popular) text classifiers. However, most of recent works does not differentiate the subjectivity and objectivity of the same emotion in the text. For instances, let's consider these two following texts:

- Text A: I am happy
- Text B: They are happy, but me not

Most of current text classifiers will assign both these texts to the label of *joy* because they do not differentiate the happiness of the teller (subjective joy), and that of other (objective joy). Meanwhile, in the case of emotion classification with the distinction of their subjectivity/objectivity, these two texts may be assigned to different labels: the text A may be assigned to the label of *subjective joy*, and the text B may assigned to the labels *objective joy* and *subjective sadness*. Intuitively, the case of emotion classification with the distinction of their subjectivity/objectivity could classify the texts more precisely than the case without distinction of emotion subjectivity/objectivity. But the second case may be more difficult than the first case.

This paper investigates in the problem of emotion classification in texts with the distinction of their subjectivity/objectivity. This paper firstly builds some datasets whose labels are emotion, in which the subject and object of the same emotion are considered as two separated labels. Secondly, this paper evaluates some existing classifiers via some scenarios on the built datasets. The results are then discussed on some difficulties of these kinds of problem.

This paper is organised as follows: Section II presents some related works. Section III presents the build of datasets of texts labelled with subjective/objective emotions. Section IV presents some preliminary experiments on the built datasets. Finally, section V is a conclusion.

## II. Related Works

Researches on emotion in computer science domain are mainly based on the appraisal and cognitive theories of emotions such as the cognitive structure of emotion of Ortony et al. [29], the cognitive pattern of emotion of Lazarus [13], [19] and the belief-desire theory of emotion (BDTE) of Reisenzein [34]. These attempts could be regrouped into three mains directions. First, the approaches to represent the concept of emotions (Van Dyke Parunak et al. [30], and Stephane [37]), to formalize some emotions in a formal logic (Meyer [21], Ochs et al. [28], and Bonnefon et al. [5], [6], [25]), and to calculate the degree of emotions (Steunebrink et al. [38], Nguyen [26]). This direction is far from our work, therefore this paper does not investigate in this direction.

Second, the approaches to recognize the emotion from facial expression (Ekman [11], Russe [36], Adolphs [1], and Busso et al. [9]). This direction is also far from our work, therefore this paper does not investigate in this direction.

Third, the approaches to recognize the emotion from text. This is an instance of the problem of text classification. Therefore, we can use any method of text classification in to the application of emotion detection. For instance, we can use any of (or extended of any ) classical classifiers such as Naive Bayes (NB) [16], Support Vector Machine (SVM) [8], k-Nearest Neighbors (kNN or IBk) [2], C4.5 [33]. Moreover, some authors could improve some classical classifier for their model. For instances, Danesh et al. proposed three improvements using Decision Template, or Voting, or

Ordered Weighted Averaging (OWA) [10]; Erkan et al. [12] proposed a model with a Harmonic function; Nigam et al. [27] proposed the expectationmaximization (EM); Kibriya et al. [17] proposed the Multinomial Naive Bayes.

Recently, some authors proposed their model for emotion detection in text. For instances, Alm et al. [3] used supervised machine learning with the SNoW learning architecture. Szpakowicz and colleagues [4], [14] used annotation scheme. Li and Xu [20] tried to infer and extract the reasons of emotions by importing knowledge and theories from other fields such as Sociology. Kralj et al. [18] investigated in EMOJIS, emotion expression in Twitter. Perikos and Hatzilygeroudis [31] used an ensemble of classifier: two are statistical (a Nave Bayes and a Maximum Entropy learner) and the third one is a knowledge-based tool performing deep analysis of the natural language sentences.

However, most of these models are not tested with the subjectivity/objectivity of the same emotion. Therefore, this paper aims to evaluate some existing classifiers on the problem of emotion classification with the distinction of their subjectivity/objectivity.

## III. Dataset

This section presents some related datasets for the problem of emotion classification; and then, builds some datasets which support the classification of subjective/objective emotions.

### A. Related datasets

There are many datasets built for the problem of emotion classification. These datasets could be divided into two groups. Firstly, group of *single label*, in which a text has only one label. For instances, Plutchik [32], CrowdFlower [24], EmoLex [23], and Semeval2017 [22]. Secondly, group of *multi-label*, in which a text may have more than one label. For instance, Brat data [7]. These are presented in Table.I.

TABLE I. Some Related Emotion Datasets

| Dataset | N. sample | N. emotion | single/multi | Sub/obj |
|---|---|---|---|---|
| Brat data [7] | 629 | 8 | multi | no |
| Semeval2017 [22] | 6755 | 4 | single | no |
| EmoLex [23] | 20000 | 6 | single | no |
| CrowdFlower [24] | 40000 | 13 | single | no |
| Plutchik [32] | 2524 | 19 | single | no |

However, most of related datasets do not support the distinction of subjectivity and objectivity of emotion. In these datasets, the default label is that subjective emotion. Therefore, these datasets could not be used for the problem of emotion classification with the distinction of their subjectivity/objectivity. This is the main reason this paper has to build some new datasets to support this problem.

### B. Built dataset

In order to build some dataset for the problem of emotion classification with the distinction of their subjectivity/objectivity, we collected several texts from several sources: status on social networks, title of news papers, idioms and quotations, lyric of songs, etc. These texts are then labelled with (subjective and/or objective) emotions. The emotions are mainly based on the cognitive definition of Frijda [13] and Lazarus [19]. The texts are divided into two dataset based on language: Vietnamese and English.

TABLE II. Distribution of Emotion Number in Two Datasets

| Emotion | Number of text | |
|---|---|---|
| | dataset VN | dataset EN |
| 1 emotion | 849 | 481 |
| 2 emotions | 547 | 307 |
| 3 emotions and more | 104 | 12 |
| **All** | **1500** | **800** |

TABLE III. Distribution of Emotion Type in Two Datasets

| Emotion | dataset VN | | | dataset EN | | |
|---|---|---|---|---|---|---|
| | Sub. | Obj. | All | Sub. | Obj. | All |
| Joy | 81 | 86 | 167 | 57 | 58 | 115 |
| Sadness | 162 | 116 | 278 | 61 | 66 | 127 |
| Hope | 54 | 50 | 104 | 25 | 25 | 50 |
| Fear | 110 | 88 | 198 | 37 | 34 | 71 |
| Satisfaction | 48 | 53 | 101 | 24 | 26 | 50 |
| Disappointment | 93 | 62 | 155 | 30 | 31 | 61 |
| Love | 58 | 65 | 123 | 27 | 24 | 51 |
| Disgust | 161 | 100 | 261 | 26 | 24 | 50 |
| Pride | 48 | 71 | 119 | 23 | 24 | 47 |
| Shame | 59 | 56 | 115 | 25 | 24 | 49 |
| Admiration | 56 | 55 | 113 | 22 | 29 | 51 |
| Gratitude | 45 | 48 | 93 | 24 | 24 | 48 |
| Anger | 71 | 67 | 138 | 24 | 24 | 48 |
| Other | 153 | 174 | 327 | 148 | 164 | 312 |
| **All** | **1199** | **1093** | **2292** | **553** | **577** | **1130** |

In the Vietnamese dataset, there are about 1500 samples. Meanwhile the English dataset has about 800 samples. The distribution of samples on each label is presented in the Table.III. And the distribution of samples on the number of label for each sample is presented in the Table.II. These are multi-label datasets: each text may have more than one label.

One of the most important feature of these two datasets is that in their labels, the subjectivity and the objectivity of the emotion are distinguished. A text may have only a subjective emotion, or objective emotion, or both subjective and objective of an emotion. For instances, the text *"My dream becomes true!"* may have two labels of *subjective satisfaction* and *subjective joy*. Meanwhile, the text *"His dream becomes true, but not mine!"* may have four labels: *objective satisfaction*, *objective joy* (for him), *subjective disappointment*, and *subjective sadness* (for the teller). In this case, the *subjective satisfaction* and *objective satisfaction* are considered as two different labels. Therefore, in these two datasets, there are only 14 different emotions, but there are 28 different labels because of the distinction of their subjectivity/objectivity.

## IV. Evaluation

### A. Experiment 1: Evaluation of the classifiers

The objective of this experiment is to find out the most suitable classifier for these datasets. The found classifier will be used in the next experiments.

*1) Scenario:* This experiment is taken with the following scenario for each dataset:

1 For each text in the dataset, remove all stop-words.

2 Split the remain character sequence into 1-gram.

3 Transform it into a vector of TF-IDF value.

4 Using the k-folds crossed-validation: Split the dataset into ten sets (10-folds). Each time, a set is used for testing (called *testing set*), and the nine remain sets are used for training (called *training set*).

5 Train and test with following classifiers[1]:
- 5.1 Support Vector Machine (SVM) [8].
- 5.2 K-Nearest Neighbors (KNN or IBk) [2].
- 5.3 C4.5 (J48) [33].
- 5.4 Rotation Forest (RF) [35].
- 5.5 Naive Bayes (NB) [16].
- 5.6 Multinomial Naive Bayes (MNB) [17].

6 Note the observed output parameters for each time of running.

7 Repeat the steps from 5 to 6 in ten times (10-folds) and take the mean values of each output parameters for all times of running.

*2) Output parameters:* Let's $O_i$, and $E_i$ are respectively the original set of label and the extracted set of label of the text $i$. And $C_i = O_i \cap E_i$ is the intersection set of $O_i$ and $E_i$. We make use of these parameters:

- The *precision* on the sample $i$ is:

$$pre_i = \frac{\mid C_i \mid}{\mid E_i \mid} * 100\% \quad (1)$$

- The *precision* on all $n$ samples in the test set is:

$$Precision = \frac{\sum_{i=1}^{n} pre_i}{n} = \frac{1}{n} * \sum_{i=1}^{n} \frac{\mid C_i \mid}{\mid E_i \mid} * 100\% \quad (2)$$

- The *recall* on the sample $i$ is:

$$rec_i = \frac{\mid C_i \mid}{\mid O_i \mid} * 100\% \quad (3)$$

- The *recall* on all $n$ samples in the test set is:

$$Recall = \frac{\sum_{i=1}^{n} rec_i}{n} = \frac{1}{n} * \sum_{i=1}^{n} \frac{\mid C_i \mid}{\mid O_i \mid} * 100\% \quad (4)$$

- The *F1-score* on all samples of the test set is:

$$F1 - score = \frac{2 * Precision * Recall}{(Precision + Recall)} \quad (5)$$

For each experiment, we consider the results on three output parameters: *Precision*, *Recall*, and *F1-score*.

TABLE IV. COMPARISON RESULTS AMONG SOME CLASSIFIERS

| Classifier | dataset VN | | | dataset EN | | |
|---|---|---|---|---|---|---|
| | Pre. | Rec. | F1. | Pre. | Rec. | F1. |
| SVM | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| C4.5 (J48) | 2.50 | 1.63 | 1.94 | 0.45 | 0.40 | 0.40 |
| RF | 2.50 | 1.63 | 1.94 | 0.55 | 0.50 | 0.50 |
| KNN | 17.41 | 18.44 | 17.89 | 10.28 | 12.28 | 11.11 |
| NB | 17.70 | 25.77 | 20.91 | 13.76 | 22.24 | 16.96 |
| MNB | 24.89 | 44.54 | 31.86 | 16.64 | 35.50 | 22.62 |

[1]These classifiers are called from API of Weka open source library [39], [15] for Java.

*3) Results:* The results are presented in the Table.IV: The classifier SVM gets the lowest value on all three output parameters, on both datasets. KNN gets higher value than SVM; C4.5 and RF get higher value than SVM; KNN gets higher value than C4.5; NB gets higher value than KNN. And the classifier MNB gets the highest value on all three output parameters, on both datasets. Therefore, the MNB is the chosen classifier for the next experiments.

### B. Experiment 2: The effects of the stop-words

The objective of this experiment is to test the effect of stop-words on the distinguish of subjective and objective emotion in texts. Therefore, this experiment will compare two strategies in pre-processing of data: remove (without) or not remove (with) stop-words from the texts.

*1) Scenario:* This experiment is taken with the following scenario for each dataset:

1 For each text in the dataset, consider two cases:
- 1.1 All stop-words are removed.
- 1.2 Do not remove stop-words.

2 Split the remain character sequence into 1-grams.

3 Transform each text into a vector of TF-IDF value.

4 Using the k-folds crossed-validation: Split the dataset into ten sets (10-folds). Each time, a set is used for testing, and the nine remain sets are used for training.

5 Train and test with the classifier of Multinomial Naive Bayes (MNB).

6 Note the observed output parameters for each time of running.

7 Repeat the step from 5 to 6 in ten times (10-folds) and take the mean values of each output parameters for all times of running.

In this experiment, three output parameters are also used: *Precision*, *Recall*, and *F1-score*.

TABLE V. COMPARISON RESULTS IN THE CASE WITHOUT AND WITH STOP-WORD

| Method | dataset VN | | | dataset EN | | |
|---|---|---|---|---|---|---|
| | Pre. | Rec. | F1. | Pre. | Rec. | F1. |
| without stop-word | 24.89 | 44.54 | 31.86 | 16.64 | 35.50 | 22.62 |
| with stop-word | 25.68 | 44.44 | 32.50 | 18.76 | 32.88 | 23.80 |

*2) Results:* The results are presented in the Table.V on three output parameters, on both datasets. At the level of *precision*, the value in the case with stop-words is higher than that in the case without stop-words, on both datasets. Meanwhile, at the level of *recall*, the value in the case with stop-words is lower than that in the case without stop-words, on both datasets. However, at the level of *F1-score*, the value in the case with stop-words is higher than that in the case without stop-words, on both datasets.

Based on these results, in the next experiment, all stop-words are not removed from the texts.

### C. Experiment 3: The effects of N-gram

The objective of this experiment is to find out the best $n$ in the n-gram extraction of texts. This experiment will consider five gram-extraction strategies: using only 1-gram, from 1 to 2-grams, from 1 to 3-grams, from 1 to 4-grams, and from 1 to 5-grams.

*1) Scenario:* This experiment is taken with the following scenario for each dataset:

1 For each text in the dataset, remove all stop-words.

2 Each time, using one of these five following gram-extraction strategies:
- 2.1 1-gram: using only 1-gram.
- 2.2 From 1 to 2-grams: using both 1-gram and 2-grams.
- 2.3 From 1 to 3-grams: using 1-gram, 2-grams, and 3-grams.
- 2.4 From 1 to 3-grams: using 1-gram, 2-grams, 3-grams, and 4-grams.
- 2.5 From 1 to 3-grams: using 1-gram, 2-grams, 3-grams, 4-grams, and 5-grams.

3 Transform each text into a vector of TF-IDF value.

4 Using the k-folds crossed-validation: Split the dataset into ten sets (10-folds). Each time, a set is used for testing, and the nine remain sets are used for training.

5 Train and test with the classifier of Multinomial Naive Bayes (MNB).

6 Note the observed output parameters for each time of running.

7 Repeat the step from 5 to 6 in ten times (10-folds) and take the mean values of each output parameters for all times of running.

In this experiment, three output parameters are also used: *Precision*, *Recall*, and *F1-score*.

TABLE VI. COMPARISON RESULTS AMONG DIFFERENT N-GRAMS

| N-gram | dataset VN | | | dataset EN | | |
|---|---|---|---|---|---|---|
| | Pre. | Rec. | F1. | Pre. | Rec. | F1. |
| 1-gram | 18.76 | 32.88 | 23.80 | 25.68 | 44.44 | 32.50 |
| 1 to 2-grams | 21.55 | 30.40 | 25.17 | 31.82 | 51.15 | 39.20 |
| 1 to 3-grams | 20.74 | 32.48 | 25.27 | 33.70 | 50.69 | 40.45 |
| 1 to 4-grams | 21.67 | 30.76 | 25.41 | 34.40 | 49.69 | 40.61 |
| 1 to 5-grams | 21.85 | 31.21 | 25.65 | 35.59 | 48.56 | 41.02 |

*2) Results:* The results are presented in the Table.VI: Generally, the higher the n-gram is up to, the higher the value of output parameters, in both datasets. However, from the value of 3-grams, the increment of output parameters is slowdown and there is no significant difference among three output parameters in the case of 3-grams, 4-grams, and 5-grams. Therefore, it is sufficient to use the case up to 3-grams.

### D. Experiment 4: The difficulty of the problem

This experiment will compare the case of emotion classification with or without distinction of their subjectivity/objectivity to see how hard the problem of emotion classification with distinction of their subjectivity/objectivity in comparing to the classical problem of emotion classification without distinction of their subjectivity/objectivity.

*1) Scenario:* This experiment is taken with the following scenario for each dataset:

1 For each text in the dataset, do not remove stop-words.

2 Split the remain character sequence into grams from 1 to 3-grams.

3 Transform each text into a vector of TF-IDF value.

4 Using the k-folds crossed-validation: Split the dataset into ten sets (10-folds). Each time, a set is used for testing, and the nine remain sets are used for training.

5 Train and test with the classifier of Multinomial Naive Bayes (MNB) in three cases:
- 5.1 Emotion only: Only emotions are differentiated. It means that the subjective and objective of an emotion are considered as the same label, that is the given emotion. For example, *subjective joy* and *objective joy* are considered as the same label of *joy*. So in this case, there are only 14 labels to classify.
- 5.2 Subjectivity/objectivity only: Only the subjectivity and objectivity of emotion are differentiated. It means that the subjective of all emotions are considered as only one label, the same for objectivity. For example, *subjective joy*, *subjective disappointment*, and *subjective anger* are considered as the same label of *subjective*. So in this case, there are only two labels to classify (*subjective* and *objective*).
- 5.3 Emotion + subjectivity/objectivity: This is the original problem of this paper. Each of 14 emotions are differentiated in their subjectivity/objectivity. So in this case, there is totally 28 labels to classify.

6 Note the observed output parameters for each time of running.

7 Repeat the step from 5 to 6 in ten times (10-folds) and take the mean values of each output parameters for all times of running.

In this experiment, three output parameters are also used: *Precision*, *Recall*, and *F1-score*.

TABLE VII. COMPARISON RESULTS AMONG DIFFERENT PROBLEMS

| Problem | dataset VN | | | dataset EN | | |
|---|---|---|---|---|---|---|
| | Pre. | Rec. | F1. | Pre. | Rec. | F1. |
| Emotion only | 38.05 | 49.16 | 42.85 | 52.28 | 65.29 | 58.03 |
| Sub/obj only | 65.40 | 67.30 | 66.31 | 72.44 | 73.31 | 72.86 |
| Emotion+sub/obj | 20.74 | 32.48 | 25.27 | 33.70 | 50.69 | 40.45 |

*2) Results:* The results are presented in the Table.VII on three output parameters. Unsurprisingly, the output values of the first problem, classification of emotion without distinction of their subjectivity/objectivity (classification of 14 labels), are much higher than those in the third problem, classification emotion with distinction of subjectivity/objectivity (classification of 28 labels), on both datasets. And the output values of the second problem, classification of the subjectivity/objectivity only (classification of 2 labels), are much higher than those in the first problem (classification of 14 labels), on both datasets.

There are two reasons to be considered. Firstly, in the case of emotion classification with the distinction of subjectivity/objectivity, the number of label is double than in the case of classical emotion classification. Generally, in the context of classification problem, the higher the number of label to classify, the more difficult the problem.

Secondly, that is the difficulty of the differentiation between the two label of the same emotion, but different subjective/objective. Let's return to the example from the introduction section:

- Text A: I am happy
- Text B: They are happy, but me not

In the case of emotion classification without subjectivity/objectivity, it could be easy to detect that both these two texts are in the label of *joy*. However, in the case with the distinction of subjectivity/objectivity, the results are totally different: the texts A belongs to the label of *subjective joy*, meanwhile the texts B belongs to two labels *objective joy* and *subjective sad*. We can see the difficulty of classify these texts among two labels *subjective joy* and *objective joy*.

These results indicate that the problem of classification emotion with distinction of subjectivity/objectivity is much more difficult than the classical classification of emotion without distinction of their subjectivity/objectivity. Consequently, the current classifiers could not reach an average value for output parameters. Meanwhile they could get an above-average value when applying them into problem of classification of emotion without distinction of their subjectivity/objectivity. This could be considered as a challenge for researches in the near futures.

## V. Conclusion

This paper considered the problem of emotion classification with the distinction of their subjectivity/objectivity. There are two datasets of text labelled with subjective/objective emotions are built and introduced, one in English, another in Vietnamese. This paper also taken some very preliminary experiments to evaluate some current statistical-based classifiers on these kind of problem. The results indicate that there are two different aspects regarding the classical emotion classification problem (without distinction of subjectivity/objectivity): first, using stop-words is better for differentiating the subjectivity and the objectivity of emotion in texts. Secondly, using current statistical-based classifiers such as SVM, KNN, C4.5, NB, RF, MNB could no more helpful in the given problem.

These difficulties are our challenges and objectives to work in the near future.

## References

[1] Ralph Adolphs. Recognizing emotion from facial expressions: Psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Reviews*, 1(1):21–62, 2002.

[2] D. Aha and D. Kibler. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

[3] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 579–586, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[4] Saima Aman and Stan Szpakowicz. Identifying expressions of emotion in text. In *Proceedings of the 10th International Conference on Text, Speech and Dialogue*, TSD'07, pages 196–205, Berlin, Heidelberg, 2007. Springer-Verlag.

[5] Jean-François Bonnefon, Dominique Longin, and Manh Hung Nguyen. Relation of trust and social emotions: A logical approach. In *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*, volume 2, pages 289–292. IET, 2009.

[6] Jean-François Bonnefon, Dominique Longin, and Manh Hung Nguyen. A logical framework for trust-related emotions. *Electronic Communications of the EASST*, 22:1–15, 2010.

[7] Lars Buitinck, Jesse van Amerongen, Ed Tan, and Maarten de Rijke. Multi-emotion detection in user-generated reviews. In *Proc. European Conference on Information Retrieval (ECIR)*, 2015.

[8] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, jun 1998.

[9] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, pages 205–211, New York, NY, USA, 2004. ACM.

[10] A. Danesh, B. Moshiri, and O. Fatemi. Improve text classification accuracy based on classifier fusion methods. In *Information Fusion, 2007 10th International Conference on*, pages 1–6, July 2007.

[11] Paul Ekman. Facial expressions of emotion: New findings, new questions. *Psychological Science*, 3(1):34–38, 1992.

[12] Gunes Erkan, Ahmed Hassan, Qian Diao, and Dragomir R Radev. Improved nearest neighbor methods for text classification. Technical report, Technical Report CSE-TR-576-11, University of Michigan. Department of Electrical Engineering and Computer Science, 2011.

[13] Nico H. Frijda. *The Emotions: Studies in Emotion & Social Interaction*. Edition de la Maison des Sciences de l'Homme. Cambridge University Press, Paris, 1986.

[14] Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Hierarchical versus flat classification of emotions in text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 140–146, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[15] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explorations Newsletter*, 11(1):10–18, November 2009.

[16] George H. John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo, 1995. Morgan Kaufmann.

[17] Ashraf M. Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. Multinomial naive bayes for text categorization revisited. In *Proceedings of the 17th Australian Joint Conference on Advances in Artificial Intelligence*, AI'04, pages 488–499, Berlin, Heidelberg, 2004. Springer-Verlag.

[18] Petra Kralj Novak, Jasmina Smailovic, Borut Sluban, and Igor Mozetic?. Sentiment of emojis. *PLOS ONE*, 10(12):1–22, 12 2015.

[19] Richard S. Lazarus. *Emotion & Adaptation*. Oxford University Press, 1991.

[20] Weiyuan Li and Hua Xu. Text-based emotion classification using emotion cause extraction. *Expert Syst. Appl.*, 41(4):1742–1749, March 2014.

[21] John-Jules Ch. Meyer. Reasoning about emotional agents: Research articles. *International Journal of Intelligent Systems*, 21(6):601–619, 2006.

[22] Saif M. Mohammad and Felipe Bravo-Marquez. Emotion intensities in tweets. In *Proceedings of the sixth joint conference on lexical and computational semantics (*Sem)*, Vancouver, Canada, 2017.

[23] Saif M. Mohammad and Svetlana Kiritchenko. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326, 2015.

[24] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[25] Manh Hung Nguyen. *A logical framework for Trust-Related emotions: formal and behavioral results*. PhD thesis, University of Toulouse III, 2010.

[26] Manh Hung Nguyen. Combination of cognitive and quantitative aspects to estimate the degree of emotions. *Journal of Computer Science and Cybernetics*, 32(3):209–224, 2016.

[27] Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2):103–134, May 2000.

[28] Magalie Ochs, Karl Devooght, David Sadek, and Catherine Pelachaud. A computational model of capability-based emotion elicitation for rational agent. In Dirk Reichardt, Paul Levi, and John-Jules C. Meyer, editors, *Proceedings of the 1st workshop on Emotion and Computing - Current Research and Future Impact*, Bremen, Germany, June 2006.

[29] Andrew Ortony, Geraldl L. Clore, and Allan Collins. *The Congnitive Structure of Emotions*. The Cambridge University Press, 1988.

[30] H. Van Dyke Parunak, Robert Bisson, Robert Matthews Sven Brueckner, and John Sauter. A model of emotions for situated agents. *AAMAS'06*, pages 993–995, May 2006.

[31] Isidoros Perikos and Ioannis Hatzilygeroudis. Recognizing emotions in text using ensemble of classifiers. *Eng. Appl. Artif. Intell.*, 51(C):191–201, May 2016.

[32] Robert Plutchik. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553, 1982.

[33] Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[34] Rainer Reisenzein. Emotions as metarepresentational states of mind: Naturalizing the belief-desire theory of emotion. *Cognitive Systems Research*, 10(1):6–20, 2009.

[35] Juan J. Rodriguez, Ludmila I. Kuncheva, and Carlos J. Alonso. Rotation forest: A new classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(10):1619–1630, October 2006.

[36] James A. Russell. Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological Bulletin*, 115:102–141, 1994.

[37] Lucas Stephane. Cognitive and emotional human models within a multi-agent framework. In *Proceedings of the International Conference of Human - Computer Interface*, Beijing, China, 2007.

[38] Bas R. Steunebrink, Mehdi Dastani, and John-Jules Ch. Meyer. Towards a quantitative model of emotions for intelligent agents. In Dirk Reichardt and Paul Levi, editors, *Proceedings of the 2nd Workshop on Emotion and Computing - Current Research and Future Impact*, Osnabrck, Germany, 2007.

[39] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

# First Out First Served Algorithm for Mobile Wireless Sensor Networks

Anouar Bouirden[1]
LTM, Faculty of Science, Ibn Zohr University
Agadir, Morocco

Maryam El Azhari[2], Ahmed Toumanari[3]
LISTI-ENSA,Ibn Zohr University
Agadir, Morocco

Ahmed Aharoune[4]
LTM, Faculty of Science, Ibn Zohr University
Agadir, Morocco

***Abstract*—Wireless Sensor Networks (WSNs) have recently gained tremendous attention as they cover a vast range of applications requiring an important number of sensor nodes deployed in the area of interest to measure physiological types of data and send it back to the base station for further analysis and treatment. Many routing protocols have been proposed to perform data routing towards the destination in accordance with energy consumption,end-to-end delay and throughput. In this paper,the First Out First Served algorithm for cluster based routing in Mobile Wireless Sensor Networks was presented. The algorithm aims to increase packet reception within the cluster in a highly constrained environment.The results prove the efficiency of the proposed algorithm in increasing the reception of data packets by the cluster head and enhancing the $Radio\_Coefficient_{Diff}$ parameter of the network.**

***Keywords*—MWSNs; WSN; Packet reception; energy consumption; convex area; Mobility Manager**

## I. Introduction

The recent evolution in electronic and robotics manufacturing has enabled the improvement of sensors architecture in order to work alongside additional devices such as actuators,which are capable of acting upon the physical environment. Such entities can be implemented to perform networking-related functionalities such as receiving ,transmitting, processing and relaying data towards the destination,they can also roam the network to some particular positions and proceed with data measurement in respect to the application requirements. A collection of these type of sensor networks constitutes a Mobile Wireless Sensor Networks (MWSNs). MWSNs [12-13] [17-21] present many advantages compared with static WSNs in terms of energy consumption;improved coverage and enhanced target tracking ,they can also include mobile sink nodes to reduce the network density to the maximum in addition to the overall network cost . Since the topology of MWSNs is constantly changing, performing data routing to the destination becomes a challenging task . WSN(Wireless Sensor Network)'s routing protocols are specifically designed for a static topology where paths establishment are performed in configuration phase or upon the occurrence of unpredictable events such as batteries depletion or sensor node location deviation. In general,the configuration phase consumes a considerable amount of energy,therefore, it is crucial to reduce the re-execution of this phase for a longer network lifetime,however when it comes to MWSNs ,the topology is changing rapidly leading to a constant network reconfiguration and a drastic energy completion in a short amount of time,another problem can be highlighted which refers to the important amount of packets being lost due to the frequent link dis-connectivity, hence it is plausible to propose a new routing paradigms to support the high mobility rate while mitigating the packet loss ration. In this article,the First Out First Served algorithm for cluster based routing in Mobile Wireless Sensor Networks was presented. FOFS algorithm aims to increase the network packet reception rate in a highly mobile environment. A detailed analysis of the forwarding connectivity condition is also presented in order to choose the forwarding sensor node that maintains the link connectivity between the sender and the receiver throughout the packet transmission time interval.

## II. State of the Art

In recent research on wireless sensor networks, there has been an important focus on improving the communication between sensors by presenting new solutions related to QoS, MAC and routing protocols. Several solutions have been exploited from which the clustering approach were distinguished. M. Akbar et al. proposed in [1] a Mobile Balanced Energy-Efficient Network Integrated Super Heterogeneous (MBEENISH), which is a hybrid protocol used in four-level heterogeneous WSN models. The protocol utilizes the benefits of both clustering and sinks mobility to maximize the network efficiency. It presents a new mathematical sink mobility model to maximize the network lifetime and guarantee a stabilized period. An improved version of LEACH named M-LEACH and LEACH-C have been proposed in order to take advantage of multi-hop routing and regulating cluster formation which have proven a better performance compared with LEACH. K. Thanigaivelu, and K. Murugan presented in [2] a novel cluster-based method. A Gridbased clustering (GBC) along with dual CHs are configured by a mobile sink node to achieve an enhanced energy balance in the network and prevent the energy hole problem for an extended network lifetime. D Seong et al. [3] proposed a novel mobile sink operation method in which the probe priority of the mobile sink is determined from data priority to increase the QoS. The authors used the mobile sink to reduce the routing hot spot. In[4],authors handled the mobility constraint by proposing an improved version of LEACH protocol named LEACH-Mobile, which consisted in

adding the membership declaration to LEACH protocol so that to guarantee the addition of sensor nodes to a particular cluster during the steady state. The mean concept of LEACH Mobile protocol is to specify whether a sensor node is able to communicate with a cluster-head during its affected time slot, if no response has been received from the cluster-heard in a threshold interval, the sensor node is declared out of the cluster and will afterwards join a new cluster after receiving acknowledgment from its cluster head. The LEACH-Mobile protocol has being proven to outperform LEACH protocol in terms of packet reception rate but still has a high energy consumption due to the increased amount of control packets being sent. Jin Wang et al. [5] proposed an energy efficient routing algorithm to handle the sink mobility and increase the WSN lifetime for real time application. The mobile CHs collect data from the network before sending it to the sink node located in the center of the interest field. All moving CHs provide connection with the base station during the report. Three schemes are adapted to CHs movement to minimize communications and improve network lifetime. The requirement of resource-rich mobile CHs is the main limitation of this system. Another cluster based routing protocol named CBR Mobile-WSN was proposed in [6]. It consisted in receiving data from cluster member and non-member sensor nodes. Cluster member send their data using TDMA mechanism whilst non-member sensor nodes can use a free time slot to communicate its data to the cluster head efficiently using the received signal strength parameter. CBR Mobile-WSN justified better flexibility when adapting to traffic rate and mobility changing characteristics. Moreover, it has been highlighted that packet loss reduction reaches a value equals to 25% compared to LEACH-Mobile routing protocol [7]. Muhammad Ali Khan et al. [8] presented a fixed mobility based reactive protocol named Mobile Sink based Data Gathering Protocol (MSDGP), using a clustering based on the amount of sensed data and residual energy. MSDGP achieved less energy consumption and provided an extended network lifetime through implementing single message CH selection and introducing a mobile sink instead of static sink. By using greedy policy and dynamic programming, H. Zhao et al. [9] proposed a tree-based heuristic topology control algorithm, called MLS to maximize the network lifetime in large scale wireless sensor networks with mobile sinks. The algorithm introduces a predefined delay tolerance level in order for a sensor node to store data temporarily and transmit it to the mobile sink at the most suitable distance to achieve an extended network lifetime. F. Tashtarian et al. [10] studied how to determine a trajectory for a mobile sink without considering any predefined rendezvous points or virtual structures. The trajectory, which is named the Continuous and Optimal Trajectory (COT) makes a significant gain in terms of the network lifetime. M. Nabi et al. [11] proposed a MAC protocol (MCMAC) to support the cluster mobility for TDMA-based Medium Access Control (MAC) protocols in Wireless Sensor Networks (WSNs). MCMAC protocol exploits a hybrid contention free and contention-based communication approach to support cluster mobility. In [12],authors defined the Exploratory Coverage problem in Limited Mobility Sensor Networks that resolve the issue of the full exploration of the target area in respect to sensor range,coverage area, the number of sensors and the mobility. The problem has been designed and the simulation results demonstrated the impact of limited mobility on the overall area coverage,however,no insight has been given to insure successful packet delivery. Similary in [13],authors proposed a zone-based routing protocol (ZBRP) for WSNs based on the considered novel design space. ZBRP considers the features of unequal clustering and edge-based routing capabilities for optimal network resources usage. It aims to achieve a balanced energy consumption for both intra-cluster i.e inter-cluster communication and extends network lifetime by reducing the overall communication cost in the network. The proposed routing protocol distributes the load uniformly with proper cluster formation across the network and enhances the sensor network lifetime. However it remains unsuitable for MWSNs for no strategy has been taking place to tackle the problem of mobility.

## III. Theoretical Approach of Connectivity

A cluster has a radius $\beta$ defined in terms of hops and it includes sensor nodes where the distance(in hops) is less than or equals to $\beta$,an example of such subdivision is presented in Fig. 1.

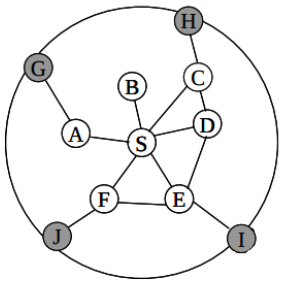



Fig. 1. The routing zone of node S

As can be seen in Fig. 1,The cluster member size can be determined by changing the transmission power,in other words,when the transmission power of a sensor node is important,more sensor nodes tend to be included within the cluster and vice versa. The cluster members could maintain information of each other pro-actively using a limited-depth,proactive link-state routing protocol.Correspondingly,the information about new routes discovery also routes maintenance can be made available via a family of reactive routing protocols. Therefore,the question to be asked is: how to find the itinerary towards the destination in a way that successful packet delivery is guaranteed to the cluster head node under the mobility constraint? In other words,what is the condition under which an intermediate node is selected in a way that packets can still be delivered when both sensor and cluster head nodes are mobile? In order to answer this question,three cases of the $< source - destination >$ couple connectivity state are to be distinguished:

the first case (Fig. 2) refers to the approximately null distance between the sensor and the receiver,in this case of study, the connectivity will be lost when the distance overcomes the communication range value dependently on the mobility model of sensors but most importantly the direction taken by both the sender and the receiver.

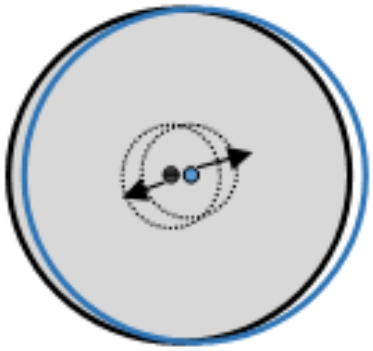


Fig. 2. The distance between the sender and the receiver is approximately null

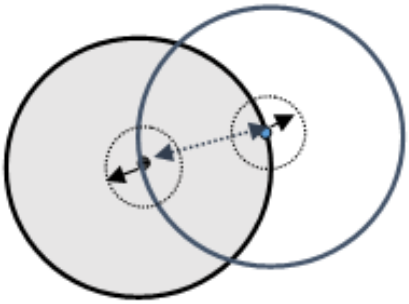


Fig. 3. Maximum connectivity distance between the sender and the receiver

The second case (Fig. 3) depicts the maximum distance at which both sensors remains connected,but once a value $\epsilon \rightarrow 0$ (where $\epsilon \neq 0$) is added to the previous distance,the connectivity is completely lost which refers to the third case scenario where the distance gets bigger in respect to a particular sensor movement directions(Fig. 4).In order to resolve the frequently interrupted connectivity,it is plausible to choose the most suitable forwarding sensor that will ensure packet delivery with respect to sensor mobility.

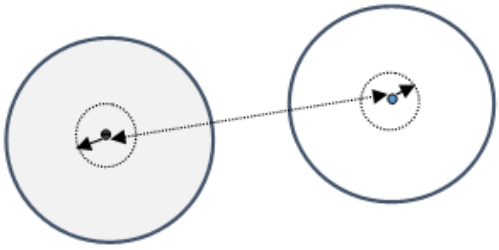


Fig. 4. Both sensor nodes are out of range

For this,the area of coverage was redrawn in a way that all sensor nodes belonging to a convex full area where vertices refer to the sensor nodes in the extremist(see Fig. 5).

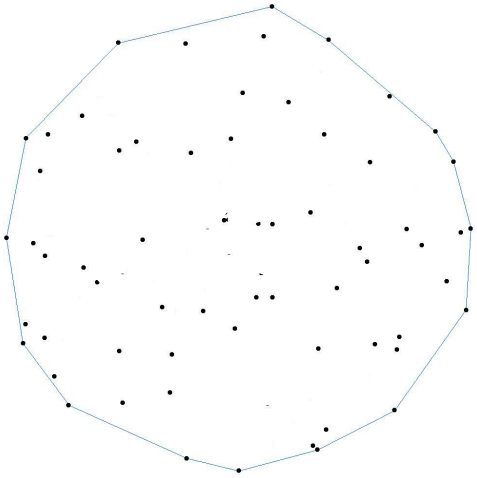


Fig. 5. Convex Hull area

For clarification purposes,Fig. 6 presents a simplified version of a convex full area $A$ where $P_i$ designates a vertex,S is the source node and D is the destination,it was proceeded to define a sub-area $A_j$ through which control packets are broadcasted in look for the destination D.

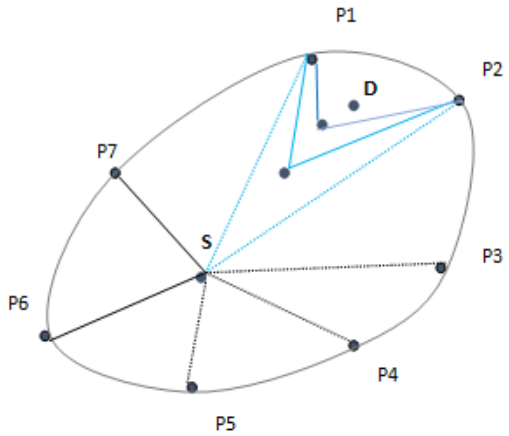



Fig. 6. A simplified version of a convex full area

The sub-area $A_j$ is defined as follows :

$$A_i = \begin{cases} P_{i\%(n+1)}, S, P_{(i+1)\%(n+1)} & \text{if } i \in \{1, ..., n-1\} \\ P_{i\%(n+1)}, S, P_{(i+1)\%i} & \text{if } i = n. \end{cases}$$

where :

- $n$ refers to the number of sub-areas in the convex full,
- $P_i$ is a convex full vertex,
- $S$ is a source node (the cluster head).

As previously mentioned,when sensor nodes are changing direction,either they get closer towards each other,or they head different directions but they stay within each other

communication range, or they maintain different positions that lead to loss of connections after an x amount of time period being exceeded. A subdivision of the cluster into sub-areas was performed,so that to speed the convergence of data packet towards the destination by choosing a forwarding node that verifies the closest distance towards the furthest vertex of an area $A_j$. The proposed method has the advantage that the search for the destination node is fast especially when it is located in the vicinity of areas vertices,moreover it reduces the number of packets roaming the network since only a small portion of sensor nodes are forwarding packets towards the destination,which will result in minimizing the number of packet been corrupted due to collision. The equation 4 refers to the condition that assures a successful delivery towards the destination,as previously mentioned,the destination can be a forwarding node or the destination itself. In order to choose the next node to forward packet to, the metric TPT was defined which refers to the round trip period located at the maximum distance. This main purpose is to guarantee a fast convergence to the destination by selecting the furthest sensor node from the source node but most likely the closest to destination.The definition of the round time packet transmission is relative to the path traveled to get to the end point.In fact,the destination location can be anywhere within the triangle $SP_iP_j$ (see Fig. 8) or identical to the vertex $P_i$ i.e. $P_j$. In this case of study,two definition of TPT were considered :the first definition consists in assuming that the length of the packet path is always below the maximum distance between the source node $S$ and the vertex $P_i$ i.e. $P_j$,which can be represented as follows:

$$max\_dist = \begin{cases} Max(d(S, P_{i\%(n+1)}), d(S, P_{(i+1)\%(n+1)})); \\ \text{if } i \in \{1, ..., n-1\}. \\ Max(d(S, P_{i\%(n+1)}), d(S, P_{(i+1)\%i})); \\ \text{if } i = n. \end{cases} \quad (1)$$

$$d_{thr_j} = \frac{C_R - d(N_i, N_j)}{2} \quad (2)$$

where : $C_R$ is the communication range,$N_i$ and $N_j$ are two sensor nodes and $d_{thr_j}$ is the threshold above which the connection is lost between $N_i$ and $N_j$.

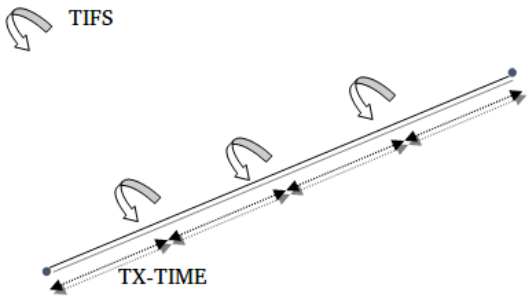



Fig. 7. Abstract assumption of $TPT_{m_1}$ from $S$ to the destination $P_i$ i.e. $P_j$

The transmission time of a packet can be estimated in accordance with the maximum distance towards the vertices of an area $A_j$.The distance is then subdivided into small chunks of distances that equal to the maximum communication range. The time needed to transmit data packet from one sensor node to another($TX\_TIME$) depends mainly on the physical data rate and the physical layer overhead while each sensor node will spend a duration equals to $TIFS$ to switch from the transmission state to reception state upon transmitting a packet(see Fig. 7).

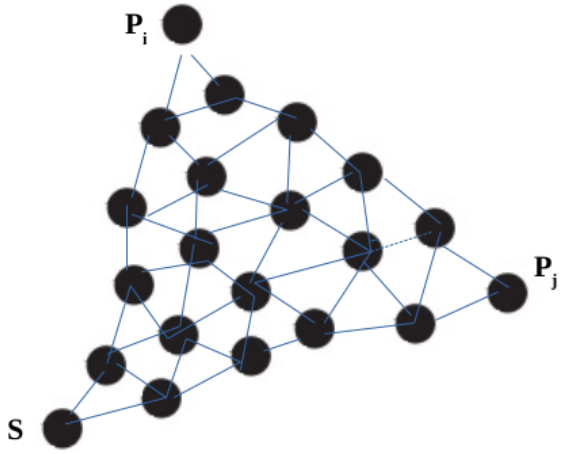



Fig. 8. Abstract assumption of $TPT_{m_2}$ from $S$ to the destination $P_i$ i.e. $P_j$

The time of transmission is represented as follows:

$$TPT_{m_1} = 2.Inf(\frac{max\_dist}{C_R}).(TX\_TIME + 2.TIFS) \quad (3)$$

where :

- $TPT$ refers to round time packet transmission;
- $TX\_TIME$ is the transmission time;
- $TIFS$ is the time needed to switch from RX to TR state

The second definition of $TPT_{m_2}$ consists in :

- Assuming that all sensor nodes are located within the triangle $SP_iP_j$.
- Finding all non-cyclical paths between the two vertices $P_i$ and $P_j$ which include all sensor nodes.

In this case,the $TPT_{m_2}$ will then correspond to the average round time traveling of each path,which is supposed to be more than sufficient to represent the maximum $TPT$ of a random sensor node in $SP_iP_j$.Therefore,the $TPT$ definition is represented as follows :

$$TPT = Max(TPT_{m_1}, TPT_{m_2})$$

Hence,in order for a communication to take place,this condition has to be verified:

$$d_{thr_i}.V_{elocity} \geq TPT_i \quad (4)$$

In this case of study,the border node can be identical to a previously defined border nodes as it can be referring to a recently defined border node which corresponds to the equation 4. The connectivity condition is repeated consecutively for

each subarea $A_j$,and it is updated for each couple $< N_i, N_j >$ of sensor nodes as follows :

$$\begin{cases} d_{thr_i}.V_{elocity} \geq TPT_i; & \text{if } i = 1. \\ d_{thr_i}.V_{elocity} \leq d_{thr_{i-1}}.V_{elocity}; & \text{if } i \geq 2. \end{cases} \quad (5)$$

where i refers to the number of hops needed to reach the destination.

The equation 5 highlights the dependency of the next destination node to be selected on the connectivity of the previously chosen border node,and this is maintained in ascendant way until the destination is found,in other words,the time required to look for the destination should be included within the time connectivity between the source node and the first border node to which packet is forwarded. The equation 5 can be reformulated with respect to $TPT_i$ (equation 6) so that at each stage of destination research,the determination of the border sensor node depends mainly on the connectivity between the sender and the receiver but also the previous round time packet transmission if it does exist.

$$\begin{cases} d_{thr_i}.V_{elocity} \geq TPT_i; & \text{if } i = 1. \\ \\ \begin{cases} d_{thr_i}.V_{elocity} \geq TPT_i; & \text{if } i \geq 2. \\ TPT_{i-1} - (TX\_TIME_{i-1} + TIFS) \leq TPT_i \end{cases} \end{cases} \quad (6)$$

The equation 6 proves the advantage presented by choosing a full convex area to proceed with the search for the destination (i.e. a cluster member). In fact,when the round transmission packet time was defined as being the maximum distance between the sender and the received,it is believed that this distance represents the maximum value that could separate two sensor nodes within the cluster It is also considered the case where the destination meets with a full convex area vertex. Once the maximum distance is extended to a certain extent, $TPT_i$ would never respect the condition of time connectivity.

### A. Prerequisites :

- It is considered that all nodes are mobile and constantly changing directions with the same velocity .
- It is also believe that the $< S, D >$ path length is less than or equals to the maximum distance between the source node and the vertices of the area $A_i$.
- It is supposed that the time for packet processing is negligible compared to transmission time.

## IV. First_Out_First_Served Algorithm for Cluster based Routing

In this section,much emphasis was put on regulating communication within a 1-hop cluster to handle the mobility constraint. As previously stated,when the sensors nodes are constantly changing location,the wireless link between the cluster head and the corresponding cluster members are frequently corrupted.Therefore,for a routing protocol that is based on TDMA mechanism,the overall performance is expected to degrade considerably,leading to a high packet loss. In order to resolve this issue,the First_Out_First_Served algorithm is proposed,which purpose is to give priority of channel access to sensor nodes that are more likely to leave the communication range of the cluster head during data transmission. Hence,the first sensor to be out of the communication range is the first to be affected a time slot to transmit data to the CH. The algorithm is based on subdividing the timeline into equal superframes and the corresponding structure for the cluster head and the cluster member are presented in Fig. 9 and Fig. 10 respectively. The superframe is composed of three phases : the setup phase, the TDMA phase and the unslotted CSMA/CA phase. During the setup phase,the cluster head and its members are defined based on different parameters such as;the remained energy,the last round during which the sensor has been selected as a cluster head and the distance between the sensor and the base station. Several cluster head selection formulas have been presented in several research works described respectively in [14][15][16],where the common goal is to choose the most suitable distribution of clusters that verifies an extended network lifetime and a high packet reception.The network lifetime refers to the time when the first or the %x of sensor nodes run out of energy. As can be seen in Fig. 9,the cluster head start by sending an advertisement packet to sensor nodes within its communication range. The setup phase duration can be adjusted by the user.Furthermore,the chosen duration aims to guarantee a vast reception of the advertisement packet by the sensors nodes in order to construct an x-hop cluster.During this phase(see Fig. 10),a cluster member receives one or multiple advertisement packets from which the closest cluster head is selected. The TDMA phase is composed of multiple slots where each slot is affected to a particular sensor node to transmit data or management packet. The last phase i.e. Unslotted CSMA/CA is used to transmit data packet from new sensor nodes joining the cluster during TDMA phase and to proceed with data transmission by the cluster members.The time needed by each cluster member during this phase depends on the length of data to be transmitted. The TDMA phase is preceded with a mini-slot where the collected data is aggregated to be forwarded to the base station. The unslotted CSMA/CA phase can be set to inactive,therefore all sensor nodes composing the cluster can put their transceiver off to lower the energy consumption for an extended period of time. A new round starts after completion of the first one and a new set of clusters is constructed.

### A. First_Out_First_Served functioning mechanism

As previously stated,in order for the communication between the cluster head and the cluster member to take place,the condition presented in equation 4 has to be verified. A number of real case scenarios consider that the sender and the receiver have different mobility models,therefore $d_{thr_i}.V_{elocity}$ value will constantly change dependently on the pr-established path designated for each deployed sensor. In this case study,a random mobility model is considered where all possible mobility models have the same probability to occur. In order to tackle the issue of connectivity,a regular verification of sensor nodes locations is considered. As described in Algorithm 1,when a cluster head send an advertisement packet,a joint packet is received from sensor nodes in the vicinity. The packet contains the current location coordinates.Based on the received informations,the cluster head defines the set of distances $Distance\_Set_i$ away from the cluster members. The $Distance\_Set_i$ is sorted afterwards in descending order so that to give priority of time slot assignment to sensor nodes that are more likely to leave the communication range of the cluster head while being affected a time slot. During the

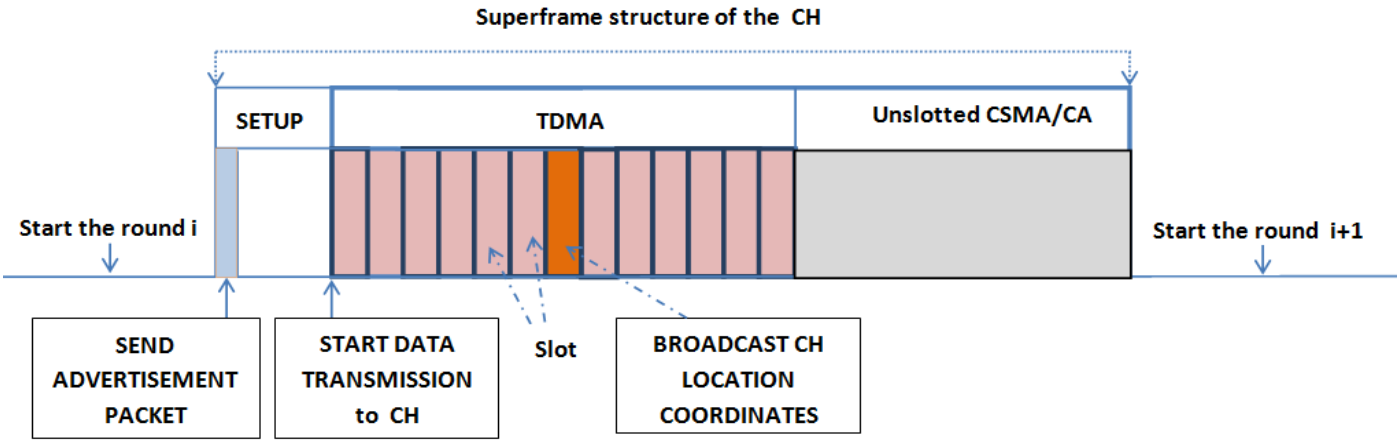



Fig. 9. Superframe structure of the cluster head.

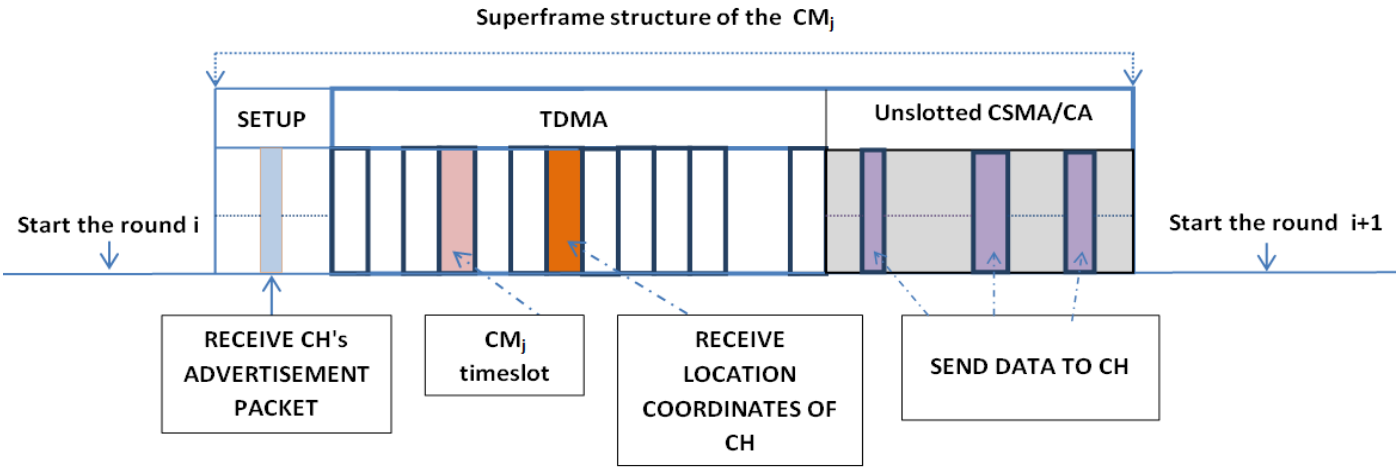



Fig. 10. Superframe structure of the cluster member.

TDMA phase,the cluster members might loose their connectivity to the cluster head, or a new sensor node might join the cluster.Therefore,it is crucial to handle packet loss and allow incoming sensors to transmit their corresponding data packets. To this end,a $Slot_{CH}$ is dedicated to the CH during the TDMA phase so that to broadcast the current location coordinates.The ranking of the affected slot is determined with respect to the communication range $C_R$ of the cluster head and the average distance between the cluster member and the cluster head. The formulas is defined as follows :

$$Slots\_Away = \left(C_R - \frac{\sum_i Distance\_Set_i}{Cluster\_Size}\right) \times \frac{1}{2 \times slotlength}$$

During $Slot_{CH}$,the cluster members set their transceivers to the reception mode to receive information related to the cluster head location coordinates and cluster size. If a cluster member doesn't receive the packet,then the connectivity is considered lost or a collision with sensors belonging to other clusters has been taken place.The third phase is automatically activated and the sensor will contend for the channel using the unslotted CSMA/CA mechanism[15] in order to transmit data packets. If a sensor node receives the packet from another cluster head then data packets are broad-casted to multiple cluster heads.This may cause packet redundancy but will guarantee a successful data packet reception in a highly critical environment.The unslotted CSMA/CA phase is also activated when sensor nodes have more data packets to be sent. After completion of data transmission,the transceiver is set to sleep mode until the next round so that to reduce the energy consumption.

## V. Simulation Setup

The simulations were performed in OMNET++ based simulation framework named Castalia 3.3 designed for WBSNs and WSNs.Three metrics were used to evaluate the efficiency of the proposed approach: the average energy consumption,$Radio\_Coefficient_{Diff}$ and the average packet received by the cluster head ($APR\_CH$).
The $Radio\_Coefficient_{Diff}$ is defined as follows :

$$\begin{cases} Radio\_Coefficient_{Diff} = \sum(A,B) - \sum(C,D) \\ State_{FOFS} \neq \varnothing \quad and \quad State_{cluster_based} \neq \varnothing, \end{cases} \quad (6)$$

**Algorithm 1**

$CH_i$: designates a cluster head i.
$Cluster_i$: designates the $cluster_i$.
$SN$ : corresponds to the number of sensor nodes in the network.
$CM_i$ :indicates the set of cluster members with $CH_i$ as cluster head.
$Num\_CM\_i$ : designates the number of cluster members with $CH_i$ as cluster head.
$Distance\_Set_i$: indicates the set of distances between the cluster head $CH_i$ and the cluster member $CM_i$
**for** s = 1 to $CM\_i$ **do** ▷ Defining the set of distances $Distance\_Set_i$
Compute the distance between $CM\_i$ and the sensor node $s$.
Insert dist(s,$CH\_i$) into $Distance\_Set_i$
**end for**
Sort($Distance\_Set_i$) ▷ Sort the computed set of distances in descending order.
Define $Avg\_Dist_i$;the average distance with respect to cluster size.
Define $Slots\_away$; ▷ Calculate the ranking of the cluster head slot during which the location coordinates and the cluster size are broadcasted.$Slots\_away$ is defined according to $Avg\_Dist_i$ and the communication range of the cluster head.

where : $State \in \{A, B, C, D\}$ and;

(A) **RxReachedNoInterference:** indicates the average number of successfully packet received without interference
(B) **RxReachedInterference:** refers to the average number of packets received with possible interference.
(C) **RxFailedInterference:** indicates the average number of packets failed because of interference.
(D) **RxFailedNoRxState:** refers to the average packets received due to the non reception state of the transceiver.

The packet generation rate is constant and equals to 5 pps (packets/sec).The initial energy level of each sensor node is assumed to be 18720J. The average temporal variation of the channel and the path loss map are defined in Castalia distribution (pathLossMap.txt and TemporalModel.txt).The radio reception mode is set to high which is equivalent to 1024 for data rate (kbps), DIFFQPSK modulation type, bandwidth (MHz), sensitivity (dBm) and power consumed (mW) equal respectively to: 20,-87 and 3.1. The size of the deployed sensors was varied to highlight their impact on the aforementioned metrics.

As can be seen in Fig. 11, FOFS algorithm improves the network performance in terms of the average number of packets received by the destination sensor nodes, which is determined by $Radio\_Coefficient_{Diff}$. A positive value of $Radio\_Coefficient_{Diff}$ indicates a successful reception of data or management packets as the average number of packets received without or despite interference surpasses the average number of packets encountering failure due to the interference with other clusters, or because the transceiver is in the transmission or the sleeping mode. A large value of $Radio\_Coefficient_{Diff}$ proves the efficient performance of

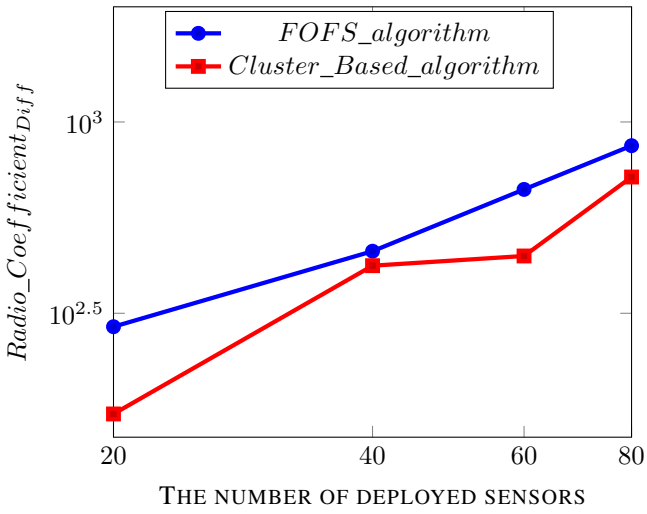



Fig. 11. $Radio\_Coefficient_{Diff}$ of $FOFS$ and $cluster\_based$ algorithms

the considered algorithm as it maximizes packet reception rate and minimizes packet loss rate.

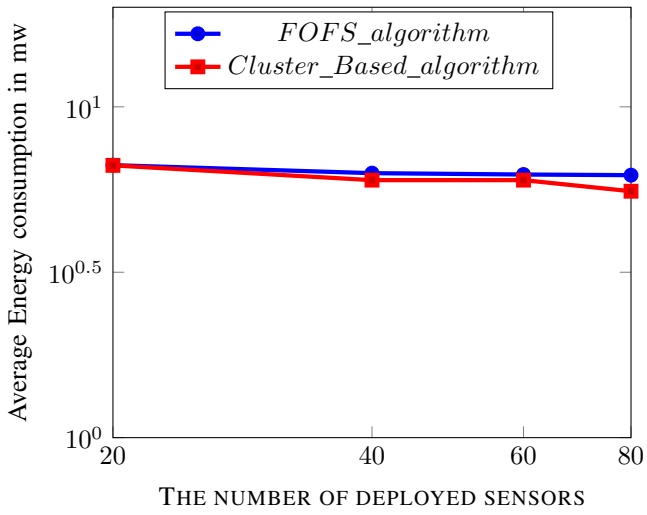



Fig. 12. The average energy consumption of $FOFS$ and $cluster\_based$ algorithms

As can be seen in Fig. 11, $FOFS$ algorithm shows a better performance compared with the cluster based algorithm, as the $Radio\_Coefficient_{Diff}$ (which corresponds to FOFS algorithm)is greater than the Cluster_based one with respect to the number of deployed sensors.

In Fig. 13, the average packets received by the cluster heads is highlighted with respect to the average number of deployed sensors. As can be seen in Fig. 13, there is no correlation between the number of deployed sensors and $APR\_CH$,which is relevant to the random mobility model of the deployed sensors. $FOFS$ algorithm increases the number of packets received by the cluster head compared with the $Cluster\_Based$ algorithm. This behavior was expected as $FOFS$ algorithm allows more data and management packets to be transmitted during the $Unslotted\_CSMA/CA$ phase. Furthermore, the

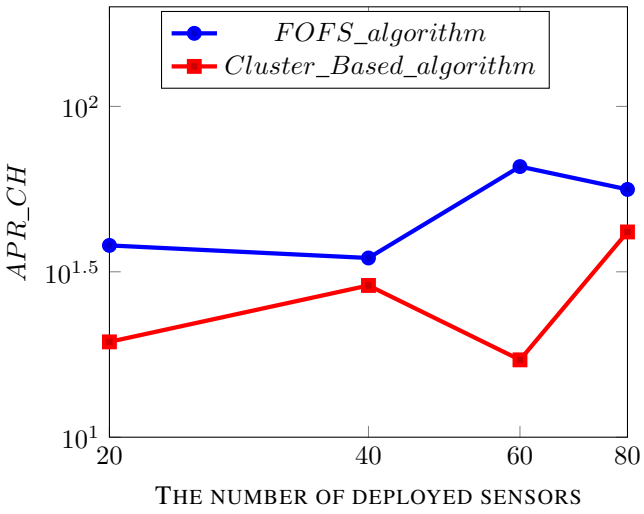



Fig. 13. The average number of received packet by the Cluster heads

connectivity between the cluster heads and their cluster members is constantly verified to determine whether the cluster members need to activate their $Unslotted\_CSMA/CA$ phase due to wireless link interruption. This enhances the successful packet reception and allows non-cluster members (i.e.sensors joining the cluster during the TDMA phase) to send their corresponding data to the new CH after reception of $Slot_{CH}$. Fig. 12 presents the average energy consumption of $FOFS$ and $Cluster\_Based$ algorithms. $FOFS$ algorithm consumes a slightly more energy compared to $Cluster_Based$ algorithm in respect of the average number of deployed sensor nodes. The maximal energy consumed by $FOFS$ algorithm is equal to 6.66 mw compared with 6.53 mw for $Cluster_Based$ algorithm. Moreover,when the number of deployed sensors increases,the average energy consumption decreases,this behavior is related to the transceiver state. In fact, the radio model used in the simulation is CC2420 which consumes 62mw when the transceiver is in the reception state versus 57.42 mw in case of transmission state. The transceivers transmission power is correlated to energy of transmission and the frequent transmission from one state adds a considerable amount to the overall energy consumption. Based on the radio model characteristics, its deducted that a network where sensor nodes are in the reception state for a prolonged period of time will eventually consume more energy compared with a very active network. The high mobility of the sensor nodes obstructs maintaining the established clusters intact during the same round. Hence, at a time $t_i$, a sensor node may be either an isolated node or a node that belongs to one or many clusters with a specific role. Therefore,when the number of deployed sensor nodes increases,more sensors are in the transmission state,which reduces the overall energy consumption in the network.The $Cluster\_Based$ algorithm allows the cluster members to put their transceivers into the sleep state for the rest of the round period in order to save the energy of the storage battery, as opposed to the $FOFS$ algorithm which activates the $Unslotted\_CMSA/CA$,hence more energy is consumed.

## VI. Conclusion

Many routing protocols have been conceived to resolve the mobility issue in mobile wireless sensor networks. The cluster-based routing protocols is considered as a subtype category of routing protocols which represents a basic design to develop routing protocols destined to overcome WSN challenges with the existence of the mobility constraint. The ongoing study focuses on presenting a new cluster based algorithm to enhance the performance of MWSN in terms of packet reception in a very constrained environment. A detailed analysis of the forwarding connectivity condition is also presented in order to choose the forwarding sensor node that is most suitable to maintain connectivity between the sender and the receiver. The connectivity condition can be integrated in the routing process as a separate cross layer module.It can be generalized to include other types of routing protocols where the mobility is considered as a real issue to tackle. As a future work,the structure of the FOFS algorithm will be enhanced to include x-hop clusters. A trade-off balance has to be found between the energy consumption and end-To-end latency in a very mobile environment.It is also believed that the proposed approach can also be extended to include Adhoc networks where nodes are highly mobile.

## References

[1] M. Akbar,N. Javaid,M. Imran,N. Amjad,MI Khan, M.Guizani, "Sink mobility aware energy-efficient network integrated super heterogeneous protocol for WSNs", EURASIP Journal on Wireless Communications and Networking, 2016;2016(1):1.

[2] Thanigaivelu, and K. Murugan, Grid-based clustering with predefined path mobility for mobile sink data collection to extend network lifetime in wireless sensor networks, IETE Tech. Rev., Vol. 29, no 2, pp. 13347, MarApr. 2012.

[3] Dhanoop K Dhanpal and Ajit Joseph."Energy Efficient Routing Protocol for Cluster Based WSN with Non Uniform Node Distribution",International Journal of Scientific Engineering Research, V(6),I(5),pp. 622–626.

[4] G. Renugadevi and M.G. Sumithra, "An Analysis on LEACH-Mobile Protocol for Mobile Wireless Sensor Networks",International Journal of Computer Applications. V(65),I(21),pp. 38–42.

[5] Jin Wang, Xiaoqin Yan, Zhongqi Zhang, Liwu Zuo, Jeong-Uk Kim, Energy Efficient Routing Algorithm for Wireless Sensor Networks Supporting Mobile Sinks , Advanced Science and Technology Letters Vol.49 (SoftTech 2014), pp.262-268.

[6] Samer A. B. Awwad,Chee Kyun Ng, Nor K. Noordin, Mohd. Fadlee A. Rasid and A. R. H Alhawari. Mobility and Traffic Adapted Cluster Based Routing for Mobile Nodes (CBR-Mobile) Protocol in Wireless Sensor Networks. Ad Hoc Networks: Second International Conference.V(49),pp. 281–296

[7] Liu, Xuxun, "A Survey on Clustering Routing Protocols in Wireless Sensor Networks",Sensors,2012,http://www.mdpi.com/1424-8220/12/8/11113.

[8] Muhammad Ali Khan, Arif Iqbal Umar, Babar Nazir, Noor ul Amin, Shaukat Mehmood and Kaleem Habib, Energy Efficient Clustering Using Fixed Sink Mobility for Wireless Sensor Networks International Journal of Advanced Computer Science and Applications(ijacsa), 7(2), 2016.

[9] Huan Zhao, Songtao Guo, Xiaojian Wang, Fei Wang, Energy-efficient topology control algorithm for maximizing network lifetime in wireless sensor networks with mobile sink, Applied Soft Computing, Volume 34, 2015, Pages 539-550.

[10] F Tashtarian, M Hossein Yaghmaee Moghaddam, K Sohraby, S Effati, On maximizing the lifetime of wireless sensor networks in event-driven applications with mobile sinks. Vehicular Technology, IEEE Transactions.64.7(2015), 31773189.

[11] Majid Nabi, Marc Geilen, Twan Basten, and Milos Blagojevic. 2014. Efficient Cluster Mobility Support for TDMA-Based MAC Protocols in Wireless Sensor Networks. ACM Trans. Sen. Netw. 10, 4, Article 65 (June 2014), 32 pages.

[12] Saad Ahmed Munir,Biao Ren, Weiwei Jiao,Bin Wang,Dongliang Xie,and Jian Ma,"Mobile Wireless Sensor Network: Architecture and Enabling Technologies for Ubiquitous Computing", AINAW '07 Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops - Volume 02 Pages 113-120 .

[13] Natalizio, Enrico and Loscr Valeria,"Controlled mobility in mobile sensor networks: advantages, issues and challenges, Telecommunication Systems,2011,V(52),pg(24112418).

[14] A. Shrestha,F. Tekiner,"On MANET Routing Protocols for Mobility and Scalability",2009 International Conference on Parallel and Distributed Computing, Applications and Technologies,(2009).

[15] Kemal Akkaya,Mohamed Younis,"A survey on routing protocols for wireless sensor networks",Ad Hoc Networks,Volume 3, Issue 3, May 2005, Pages 325349.

[16] M. Snyder,S. Chellappan and M. Thakur,"Exploratory Coverage in Limited Mobility Sensor Networks," 2013 16th International Conference on Network-Based Information Systems, Gwangju, 2013, pp. 186-193. doi: 10.1109/NBiS.2013.30

[17] S. A. Sahaaya Arul Mary, and Jasmine Beulah Gnanadurai, Enhanced Zone Stable Election Protocol based on Fuzzy Logic for Cluster Head Election in Wireless Sensor Networks, International Journal of Fuzzy Systems, 2016. View at Publisher View at Google Scholar

[18] Zygmunt J. Haas and Marc R. Pearlman. 2001. ZRP: a hybrid framework for routing in Ad Hoc networks. In Ad hoc networking. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA 221-253.

[19] M. Er-Rouidi, H. Moudni, H. Mouncif and A. Merbouha, "An Energy Consumption Evaluation of Reactive and Proactive Routing Protocols in Mobile Ad-Hoc Network," 2016 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV), Beni Mellal, 2016, pp. 437-441.doi: 10.1109/CGiV.2016.90.

[20] O. Chughtai, N. Badruddin, A. Awang and M. Rehan, "A cooperative selection reactive routing protocol for wireless sensor networks," Research and Development (SCOReD), 2013 IEEE Student Conference on, Putrajaya, 2013, pp. 495-499. doi: 10.1109/SCOReD.2013.7002640.

[21] D. F. Macedo, L. H. A. Correia, A. L. dos Santos, A. F. Loureiro and J. M. S. Nogueira, "A pro-active routing protocol for continuous data dissemination in wireless sensor networks," 10th IEEE Symposium on Computers and Communications (ISCC'05), 2005, pp. 361-366.doi: 10.1109/ISCC.2005.15.

# Developing Disease Classification System based on Keyword Extraction and Supervised Learning

Muhammad Suffian, Muhammad Yaseen Khan, and Shuakat Wasi
Department of Computer Science,
Mohammad Ali Jinnah University,
Karachi, Pakistan

***Abstract*—The Evidence-Based Medicine (EBM) is emerged as the helpful practice for medical practitioners to make decisions with available shreds of evidence along with their professional expertise. In EBM, the medical practitioners suggest the medication on the basis of underlying information of patients descriptions and medical records (mostly available in textual form). This paper presents a novel and efficient method for predicting the correct disease. Since these type of tasks are generally accounted as the multi-class classifying problem, therefore, a large number of records are needed, so a large number of records will be entertained in higher n-dimensional space. Our system, as proposed in this paper, will utilise the key-phrases extraction techniques to scoop out the meaningful information to reduce the size of textual dimension, and, the suite of machine learning algorithms for classifying the diseases efficiently. We have tested the proposed approach on 6 different diseases i.e. Asthma, Hypertension, Diabetes, Fever, Abdominal issues, and Heart problems over the dataset of 690 patients. With key-phrases tested in the range [3,7] features, SVM has shown the highest (93.34%, 95%) F1-score and accuracy.**

***Keywords*—Natural language processing; Machine Learning; Multi-Class Classification; Patient descriptions; Keyword Extraction**

## I. Introduction

The idea of Evidence-Based Medicine (EBM) caused incredible enthusiasm among well-being experts. As indicated by definition [1] Evidence-Based Medicine is the medication suggested by the doctors underlying the available health status of the patient by formulating the question or query accordingly and then applying on the corpus of medical journals to retrieve the summaries or results related to the disease. The reason for consulting the medical journals is because the medical practitioners have to get aligned with the day by day new achievements published in medical journals. The current technological advancements have revolutionised the EBM concept. This mechanism is helpful for the doctors to pick the latest curing medications for the severe type of diseases. In spite of many hurdles, Evidence-Based Medicine practice has gained the reputation over recent years due to the reasons, like the improvements in patients' health-care. Research advancements are removing the barriers in EBM and it is inferred that the boom will come with NLP techniques. Our problem is an inspiration from Sarker et al. work [2]. They discussed the problems and obstacles in evidence-based medicine faced by the practitioners. They categorised the problems in five major parts. One of those problems is related to formulate the question or query that should include all important information without ambiguity and about the information retrieval. Névéol et al. [3] had identified the opportunities and challenges to work with clinical natural language processing. They had also described the problems with different methods/algorithms with respect to language context.

Natural language processing can do helpful things for the evidence-based medicine. The current research in medical information retrieval has concentrated on query design and other facets of information retrieval to support practitioners. The sentences in form of patient descriptions spoken or written by the patient are very important for the doctor and the machine/robot to instruct/suggest/search the medication strategy from the large medical corpus or using the own skill set based on experience. The very first thing to help doctors/machines to formulate the query/strategy needs the semantic extraction or information extraction from the sentences uttered/written by the patient. Here involves the natural language processing. The second thing is to classify the patient description into a specific disease. The correct or true information searched or retrieved by the doctor/machine depends on the correctness of the formulation of query or the understanding developed by the doctor/machine from the sentence.

The first reason is that most of the doctors and machines/robots cannot formulate the correct query because of the ambiguity in sentences due to the multiple meanings of the sentence [1]. Second possible reason can be the less awareness of technology to doctors i.e. how to search or retrieve the information results from the corpus? Now this problem of query formulation and classification of patient description can be fixed using the natural language processing and machine learning techniques and in this way, the precision and recall of searched query can be increased.

The rest of the paper is organised as follows: In section II the related work is highlighted, in section III the methodology and experimental setup is discussed. The section IV provides information regarding to the data sets used in the experiment. In section V the results are shown and discussed. in the end conclusion and future work are presented in section VI and VII respectively.

## II. Related Work

An approach similar in spirit to our work is discussed and modelled in QRAQ [4]. The authors discuss user story as text and the challenging question is given to the agent that deduce the information from the text with existing ambiguities, and it should be able to answer the question. If the agent cannot answer then firstly it learns and deduces the variables from the

fact in the problem. Secondly, if the agent cannot answer the question by reasoning alone then it infers from the simulator to extract the other variables from the problem and should be relevant to question. The problem domain of this work is similar to our domain work. They used the Reinforcement Learning (RL) approach in their work and based on (RL) they presented and evaluated two memory network architectures. Our work is more towards Natural language processing machine learning.

In [5] Molla et al. built a corpus for the text processing. They have taken the data set from the clinical inquiries segment of the journal dealing with family practice [6]. They annotate the data using the annotation techniques like automatic extraction, manual annotation and the rephrasing text. The inquiry sentence is used as the query and the retrieval text then summarised to answer. The summary of the text is basically is divided into few sentence classes and the human annotation was used to classify them into according summary. They associated three evidence-based answers to each question and each answer deal with separate evidence. The criteria of suggestion are based on the score of matching to the evidence.

In the work of Molla et al. [5] one thing can increase the accuracy of the retrieved summaries that is the removal of ambiguity from the input sentence/query. In [7] Dönmez et al. formed a phrase-content finder system for the Turkish sentences. They have done this study by underlying the importance of subject, verb and object relation with actionable things. The phrase content relationship is also valuable because of its structural importance for sentence. They divided the sentence mainly into two parts, one the phrase and the other as content. In each sentence, they separated it into 8 different phrases, then if the phrase exists the concepts are determined from the database like Word-Net [8]. These phrase-concepts pairs like syntactic and semantic information of sentences have shown with matrix representation.

Avani et al. showed an Question/Answering system [9] which is built focusing on the structured and annotated knowledge base. The system is divided into three parts question processing, information retrieval and the answer extraction. The question processing part is related to my study that is divided into two parts: First, the question is given to python factoid question classifier [10] this determines the type of the question and also the category of the answer to this question. Second, the question is parsed using the Stanford dependency parser which checks the dependencies of words and POS tagging is done in parallel. In this way, they determined the focus of the question. But they also highlighted the limitation of this approach that python factoid classifier does not categorise the questions in which there is a call for action. They evaluated their Question/Answer system on TREC 2004 question data set. In [10], Kim et al. build a sentence classifier that identifies the key sentences and then classifies them with medical tags. Their classifier uses conditional random fields CRFs for the learning algorithm purposes. The classifier is trained with basically four features lexical information, semantic information, structural information and sequential information.

In lexical information feature they used the bag of words with bigrams and then applied POS tagging for the semantic similarity in two texts. In semantic information the meta-thesaurus from UMLS (Unified Medical Language Systems) was used, then directly query the thesaurus with each input token. MetaMap analyser used for sentence parsing, in this way they get the concept unique identifiers and identified the same text. The corpus was 1000 abstracts and each sentence was annotated. I highlighted only the relevant work of kim et al, their work is more towards the sentence classification retrieved from the abstracts. The features like lexical and semantic information are more related, but utilised on results after querying, the ambiguity of query and question meaning before applying on data set is not handled in their work.

Sarker et al. presented a query focused approach for text summarisation to support evidence-based medicine [11]. The query specific summaries were extracted by introducing a scoring scheme in which the score was assigned to sentence on UMLS type and the category type it contains. Semantic type information improved the extractive summarisation performance. They classified the questions in their corpus into medical topics using the approach [12]. For the better question associations with summaries, they set two semantic types for each question (a) important question semantic types that were identified during training and (b) important answer semantic types that are identified from human-authored summaries in training. They evaluated their approach using ROUGE evaluation tool, their QSpec system outperforms previous systems working on the same perspective with 96.5% percentile rank. But the (Sarket et al) also highlighted the room for improvement that can be achieved by improving intermediate steps for the feature generation in summarisation task.

In [13] Pratt et al. gave a new approach for categorising the search results was implemented with the name DynaCat system. In this, they divided the semantics of dynamic categorisation into two models (a) small query model that keeps the knowledge of the types of queries users make (b) a large domain-specific terminology model, Dynacat uses UMLS for handling large terms and their synonyms. In the query model, the algorithm takes the types of queries and check the category of relevant query types. The limitation of query model is, it independent of disease-specific terms means it generalises the query into the specific category like categorising in treatment type or adverse effect etc. This system was made for the patients and their family members with a questionnaire form to input the query data. This system was claimed better than the previous ranking based and clustering based models. In this work, the query or question from the patient was taken but the processing on it is not more to clear the sentence level ambiguities and it did not assign the category on the basis of disease.

In [14] Cao et al. developed an online system that is related to question answering in a complex clinical query environment, AskHERMES is a system that is in comparison with Google and upToDate system for complex questions to answer with beating accuracy. Their complex question handling part is the NLP and Information retrieval (IR) problem and they have handled it with UMLS and CRFs. The system worked on vast datasets like Medline, PubMed, eMedicine etc. This system limit is highlighted by the Cao et al. that is it does not integrate the complex clinical evidence identification part that is entered by upToDate manually.

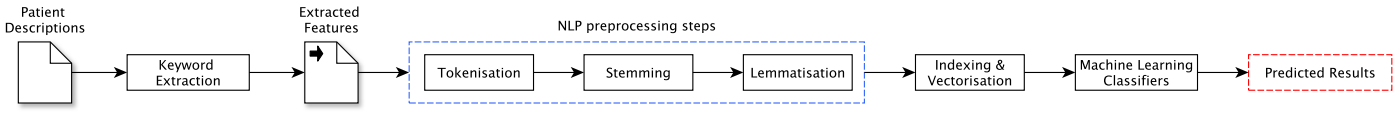



Fig. 1. Architectural scheme of query formulation and disease classification from patient descriptions.

## III. Methodology

The proposed methodology has two straight forward phases. The first phase is related to the extraction of specific key-phrases out of detailed patient descriptions/records, and application of standard NLP techniques on the extracted information followed by indexing and vectorisation of features. In the second phase, these key-phrases are then employed for the supervised learning and disease classification. These two phases are described in sections III-A and III-B respectively. The figure 1 illustrates the pipeline of the methodology and the details of both phases are given is two separate sections accordingly.

### A. Extraction, Preprocessing & Pre-classification tasks

**Keywords and keyphrases extraction.** The keywords (or extended keywords i.e. key-phrases) are the central representatives of the content in any document. It helps to identify the basic theme of any document. Hence, in spite of analysing and computing the whole bunch of documents for formulating queries, it is more easy-to-use to extract the important keywords and proceed with the rest of computational procedures in comparatively space and time efficient way. In this regard, Rose et al. [15] presented the idea for keyword extraction in which they described and compared their algorithm, Rapid Automatic Keyword Extraction (RAKE), with different NLP based methodologies and algorithms with their specific use. RAKE is an unsupervised, domain-independent algorithm that works on co-occurrence graph, it extracts the candidates for the key phrases and words from the text, then checks whether these can be declared keywords or not and then the score is assigned to each keyword. The scoring metric is quite simple [15]. Letting $w$ be the word in a corpus, the score will be a ratio i.e. $deg(w)/freq(w)$, where $deg(w)$ is the degree of word $w$ and $freq(w)$ is the frequency of the word $w$ in the given corpus. We can say the keywords extracted through RAKE are the features that are mostly in the form of $n$-grams where $n > 1$. In our experiments, we took the top $3-7$ extracted features for training and classification. Table I provides the example of extracted keywords and their scores accordingly.

TABLE I. Sample of Patient Description and Extracted Features by RAKE

| **Sample Patient Description** |
|---|
| This is a 36-year-old woman with a history of type-I, chronic renal insufficiency on hemodialysis as well as chronic skin ulcers who was at hemodialysis on the day of admission when she developed high temperature to 101, chills, and rigors. |
| **Extracted Features/Key-phrases** |
| ('chronic renal insufficiency', 9.0), ('chronic skin ulcers', 9.0), ('developed high temperature', 9.0), ('36-year-', 1.0), ('woman', 1.0), ('history', 1.0), ('type-', 1.0), ('hemodialysis', 1.0), ('day', 1.0), ('admission', 1.0), ('chills', 1.0), ('rigors', 1.0), ('101', 0) |

**Preprocessing.** The core and essential part of any task in the domain of data-sciences is preprocessing. With the extracted features in the previous phase, regular methods of case-folding, lemmatising and stemming are applied. These methods are supposed to sort words so as to group together inflected or variant forms of the same words. These methods are employed by using the Natural Language Toolkit (NLTK) module in Python [16]. The extracted features are tokenised prior to pass through these methods.

**Indexing and vectorisation of feature vectors.** The classifying ML algorithms require input in a vector format. Thus, at this stage, the main goal is the transformation and vectorisation of extracted features. These vectors are typically a boolean representation of the documents in an $n$-dimensional space, where each term resides at a separate dimension. Thus, if a term $t_i$ is present in the document $d_j$, the vector $v_j$ representing the document $d_j$ will mark 1 at the index corresponding to the term $t_i$, otherwise, there will be 0 representing the absence of the term in the document. The collection of these vectors is a matrix and often named as term-document incidence (TDI). Table II renders an example of documents in a vector space model.

TABLE II. Example of Term-Document Incidence

| | $t_1$ | $t_2$ | $t_3$ | $\cdots$ | $t_n$ |
|---|---|---|---|---|---|
| $d_1$ | 1 | 0 | 0 | $\cdots$ | 1 |
| $d_2$ | 1 | 1 | 0 | $\cdots$ | 0 |
| $d_3$ | 0 | 1 | 1 | $\cdots$ | 1 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $d_n$ | 1 | 0 | 1 | $\cdots$ | 1 |

In order to construct TDI for extracted features, we created a universe of features ($\mathbb{U}$) where each tokenised term $t$ is tagged with a unique index number. For this kind of tasks dictionary (as the data-structure) is the most suitable solution. Algorithm 1 will give the simple and robust solution for universe construction, where the tokenised features are check iteratively in the dictionary for their existence, if the result of

**Result:** A dictionary ($\mathbb{U}$) with the terms as keys and respective index number as values.
$\mathbb{D} \leftarrow$ be the set of extracted features;
$\mathbb{U} \leftarrow$ be the empty dictionary;
$\mathbb{C} \leftarrow 0$;
**for** *each document* $d$ *in* $\mathbb{D}$ **do**
  $\mathbb{T} \leftarrow$ split $d$ into tokens;
  **for** *each token* $t$ *in* $\mathbb{T}$ **do**
    **if** $t \notin \mathbb{U}$ **then**
      $\mathbb{U}_{[t]} \leftarrow \mathbb{C}$;
      $\mathbb{C} \leftarrow \mathbb{C} + 1$;
    **end**
  **end**
**end**

**Algorithm 1:** Algorithm for building universal set of distinct terms.

lookup is false (i.e. $t \notin \mathbb{U}$) then the term $t$ is added to the $\mathbb{U}$; where the term $t$ is set to the key with value $\mathbb{C}$ as a key-value pair, along with this process the counter $\mathbb{C}$ gets increment by 1 for serving as the index of forthcoming term.

Once $\mathbb{U}$ is created we can utilise the dictionary to proceed towards the construction of term-document incidence. Algorithm 2 shows the simple procedure, where the collection of preprocessed documents i.e. patient descriptions as $\mathbb{D} = \{\langle d_1, l_1, m_1\rangle, \langle d_2, l_2, m_2\rangle, \cdots, \langle d_n, l_n, m_n\rangle\}$ (where $d_i$ is the extracted features, $l_i$ is the label/class, and $m_i$ is the medication accordingly) is going to be vectorised with respect to $\mathbb{U}$, and $\mathbb{I} = \{v_1, v_2, v_3, \cdots, v_n\}$ is an empty list in which all vectors have to be appended such that $v_i$ is the corresponding vector representation of $d_i$. Letting X be the local list of zeros equal to size of $\mathbb{U}$. Iteratively each document is split into the set of tokens and each token $t$ is looked up in $\mathbb{U}$ that gives the index-value stored agains $t$, hence, the 1 will be replaced at the index $\mathbb{U}_t$ in the X. Technically in the end of procedure the correctness can be checked through $|\mathbb{I}| = |\mathbb{D}|$, and since, $\mathbb{I}$ is a non-sparse matrix/list of lists, therefore, the length of vector is equal to the size of dictionary $|v_i| = |\mathbb{U}|$.

**Result:** A non-sparse matrix showing the boolean representation of documents in n-dimensional vector space.
$\mathbb{U} \leftarrow$ be the dictionary having terms as keys and index numbers as values (generated through algorithm 1);
$\mathbb{D} \leftarrow$ be the set of extracted features;
$\mathbb{S} \leftarrow$ be the size/length of $\mathbb{D}$;
$\mathbb{I} \leftarrow$ be the empty list;
$\mathbb{C} \leftarrow 0$;
**for** *each document* $d$ *in* $\mathbb{D}$ **do**
  X $\leftarrow$ be the local list of $\mathbb{S}$ zeros;
  $\mathbb{T} \leftarrow$ split $d$ into tokens;
  **for** *each token* $t$ *in* $\mathbb{T}$ **do**
    X$[\mathbb{U}_{[t]}] \leftarrow 1$;
  **end**
  $\mathbb{I}.append([\text{X}, label_d])$
**end**

**Algorithm 2:** Algorithm for constructing Term-Document Incidence using dictionary defined in algorithm 1.

### B. Supervised Learning and Disease Classification

In a machine learning classification system, documents or text with already tagged class labels are set as an input to the ML algorithm that learns the underlying information and patterns from the data to build a predictive model. The document ($D$) typically consist of features ($f_i$) and the expected target outcome is a member of the discrete classes ($Y$) $\therefore D = \{f_1, f_2, f_3, \cdots, f_n\} \rightarrow y_j$, where $y_j \in Y$. Thus, if there are two possible answers (i.e. $|Y| = 2$) then we can say the problem is binary or binomial classification whereas, if the possible answer is more than two (i.e. $|Y| > 2$) then, it would be a multi-class or multinomial classification problem. Hence, the problem addressed in this paper is the multi-class or multinomial classification as there are 6 possible prediction outcomes (Asthma, Hypertension, Diabetes, Fever, Abdominal issues, and Heart problems). The phase in which learning for the predictive model is made is also called training phase. This predictive model is used to classify the unseen data.

In our experiment, we have used 4 different ML algorithms: (a) Random forest (RF) [17] which are counted as the ensemble learning approach in classification. (b) Iterative Dichotomiser 3 (ID3) [18] which is the classifying algorithm that works as a decision tree, (c) Support vector machine (SVM) [19] which is the linear model of classification where data is split into distinct parts in such a way that it holds maximum margin among the splits, and (d) Naïve Bayes, which is a likelihood-based probabilistic classifying function. These algorithms are employed by using the scikit-learn module for Python [20].

We have learned in the previous section about the indexing of extracted features. Utilising the indexing method, we rigorously repeated the experiment with top 3-7 extracted features contributed by RAKE.

## IV. Data Set

Availability of relevant data for EBM is a real obstruction. There is as such no evident mechanism for the digitalisation of the patient descriptions/ records in the form of text. Although, there are hospitals and medical centres where they have gathered information about patients but, ordinarily these centres do not share information with the groups who are intended to conduct research in the current domain. Thus, to handle and solve this issue of dataset we have prepared our own dataset that is the patient descriptions in the form of text. This dataset is prepared with the help of few online medical forum like patients.info [21][1] and i2b2 dataset[2] [22], [23].

The dataset comprises of total 240 records from patient.info, and 450 records from the i2b2 dataset. Thus, there are in total 690 patient records. The datasets are comprised of 6 diseases as classes in the case of classification. They are Abdominal issues, Heart, Fever, Diabetes, Asthma, Hypertension. Table III will give you the details of class distribution with respect to both of the datasets. For the sake of training-testing split, we randomised the records and set cross-validation for 10 folds. Hence, the algorithm will use 9 parts of 10 splits in training and remaining will be utilised for the test.

TABLE III. Distribution of Diseases/Classes with Respect to the Datasets

| Dataset | Asthma | Hyper-tension | Diabetes | Fever | Abdominal Issues | Heart |
|---|---|---|---|---|---|---|
| patient.info (dataset A) | 40 | 40 | 40 | 40 | 40 | 40 |
| i2b2 (dataset B) | 53 | 61 | 60 | 50 | 76 | 150 |

## V. Results

### A. Model Evaluation

In this experiment, results are evaluated on the metrics of precision (P), recall (R), accuracy (A) and F1-scores (F). In a conventional binary or binomial classification system, these metrics are calculated with the number of 'true positives' ($tp$) and 'true negatives' ($tn$) which means the classifier respectively predicts the instances *positive* that are actually positive, and *negative* that are actually negative. With these two statistics, there are two more i.e. 'false positive' ($fp$) and

[1] https://patient.info/
[2] https://www.i2b2.org/NLP/DataSets/

'false negative' ($fn$), which means the classifier mistakenly predicts a negative instance as *positive*, and a positive instance as *negative* respectively. Thus the equations for calculating these metrics in a binary classification system are given below:

$$\mathrm{P} = \frac{tp}{tp+fp} \qquad \mathrm{R} = \frac{tp}{tp+fn} \qquad \mathrm{F} = 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P}+\mathrm{R}}$$
$$\mathrm{A} = \frac{|\text{correctly predicted instances}|}{|\text{all instances}|} \tag{1}$$

Thus, w.r.t the equations for precision and recall we can say precision as positive predictive rate and recall as true positive rate. An ideal classifying system should have both high precision and recall. While, F1-scores is a harmonic mean between precision and recall, and accuracy shows the overall success of the system. Since, this paper deals with the multi-class problem therefore, the equation for metrics are altered as per the following equations:

$$\mathrm{P}_i = \frac{M_{ii}}{\sum_j M_{ji}} \qquad \mathrm{R}_i = \frac{M_{ii}}{\sum_j M_{ij}} \tag{2}$$

Where, $M$ is a $k \times k$ dimensional matrix such that $k = |\text{classes}|$, and $i$ represents a certain class to be classified. Hence, $M_{ii}$ is the number of $tp$ instances for the class $i$. Similarly, $\sum_j M_{ji}$ is the aggregate of all values for class $i$ in $j^{th}$ column in $M$. Whereas, $\sum_j M_{ij}$ is vice-versa of $\sum_j M_{ji}$. Thus, the equations in 2 shows the precision and recall for the class $i$. The precision and recall for the whole multi-class system will be an aggregate of individual precisions and recalls with respect to all classes and it can be calculated as per equation 3.

$$\mathrm{P} = \frac{1}{k} \sum_{\text{i=0, j=0}}^{k} \frac{M_{ii}}{\sum_j M_{ji}} \qquad \mathrm{R} = \frac{1}{k} \sum_{\text{i=0, j=0}}^{k} \frac{M_{ii}}{\sum_j M_{ij}} \tag{3}$$

### B. Experimental Results

Results are quite exciting and interesting. Factually, we witness the improvement in all results with the increment of keywords features form 3 to 7. Tables IV shows the results that on averages, patient.info shows $\approx$15% improvement in F1-score when keyword feature size moved to top-5 from top-3, and further $\approx$12% improvement when feature size updated from top-5 to top-7. Similarly, on the i2b2 dataset, there is $\approx$12% improvement when feature size is moved to top-5 features from top-3, and further $\approx$5% improvement on increasing feature size up to top-7 features. In comparison to the F1-scores, table V shows the improvement in average accuracies on keyword increment. Collectively, patient.info outperforms the results by yielding $+5.75\%$ and $+9.55\%$ difference in accuracy and F1-scores respectively.

TABLE IV. RESULTS OF AVERAGE F1-SCORES W.R.T THE KEYWORDS SIZE AND IMPROVEMENT

| Dataset | Average F1-scores | | | Improvement | |
|---|---|---|---|---|---|
| | Top-3 keywords | Top-5 keywords | Top-7 keywords | Keywords 3→5 | Keywords 5→7 |
| patient.info | 60.50 | 75.79 | 88.20 | 15.29 | 12.41 |
| i2b2 | 61.25 | 73.75 | 79.40 | 12.50 | 5.65 |

TABLE V. RESULTS OF AVERAGE ACCURACIES W.R.T THE KEYWORDS SIZE AND IMPROVEMENT

| Dataset | Average Accuracies | | | Improvement | |
|---|---|---|---|---|---|
| | Top-3 keywords | Top-5 keywords | Top-7 keywords | Keywords 3→5 | Keywords 5→7 |
| patient.info | 61.25% | 80.25% | 90.00% | 19.00 | 9.75 |
| i2b2 | 53.50% | 71.25% | 76.50% | 17.75 | 5.25 |

TABLE VI. RESULTS OF DIFFERENT MATRICES ON DATASETS WITH TOP-3 FEATURES

| | patient.info (dataset A) | | | i2b2 (dataset B) | | |
|---|---|---|---|---|---|---|
| Algorithm | Prec. | Rec. | F1-score | Prec. | Rec. | F1-score |
| SVM | 79 | 64 | 62 | 94 | 53 | 62 |
| Random Forest | 61 | 62 | 60 | 90 | 54 | 61 |
| Decision Tree | 74 | 61 | 60 | 80 | 57 | 60 |
| Naïve Bayes | 90 | 57 | 60 | 94 | 52 | 62 |

Tables VI, VII, and VIII show the details of precisions, recalls, and F1-scores of the algorithms on top $3-7$ keywords extracted on patient.info and i2b2 respectively. Overall in the entire experimental suit, the average lowest F1-score is $\approx$60 which is shown with top-3 features on the dataset of patient.info, the i2b2 dataset shows the second lowest F1-score ($\approx$61) i.e. $+1\%$ improvement w.r.t patient.info at the same feature setting. SVM shows the highest individual F1-score i.e. $\approx 93\%$ on patient.info with top-7 keywords followed by random forest $\approx 91\%$ which sets behind $-2\%$ in improvement.

Recall and precision as exhibits by naïve Bayes is uncanny. Almost in all experiments (i.e. except on patienet.info with top-5 and 7 features), it shows the highest value for precision $(90-97\%)$ along with the lowest value for recall $(52-67\%)$. SVM on dataset patient.info with top 7 keywords outperforms the results of all classifiers in achieving the desired high values in precision and recall.

Table IX provides the classification report of SVM on the dataset of patient.info. The values in the table correspond to the results of top-7 keywords, where SVM shows the highest performance. Similarly, for the dataset of i2b2, table X gives the classification report with decision tree as the classifying function. In both datasets, w.r.t these two tables (IX and X) we can see a the disease/class 'abdominal issues' secures the near-human predicting results. While, results for 'diabetics' $(97\%)$, 'heart' $(98\%)$, and 'fever' $(96\%)$ in patient.info (table IX), and 'diabetics' $(92\%)$ in i2b2 (table X) are also encouraging.

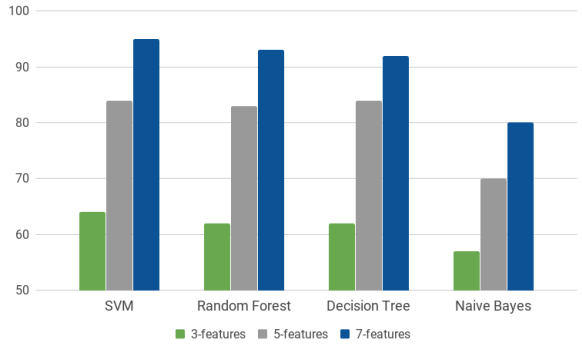
Collectively, on the basis of resulting accuracies, we can put forward that the naïve Bayes performs poor amongst the

TABLE VII. RESULTS OF DIFFERENT MATRICES ON DATASETS WITH TOP-5 FEATURES

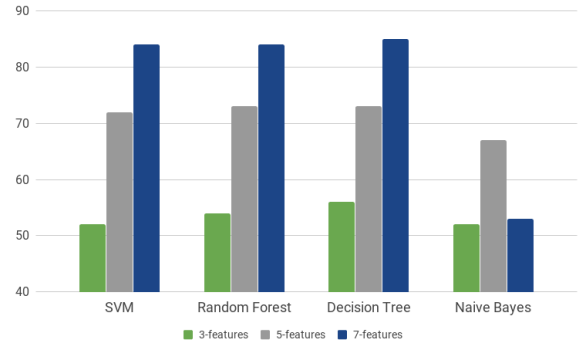
| | patient.info (dataset A) | | | i2b2 (dataset B) | | |
|---|---|---|---|---|---|---|
| Algorithm | Prec. | Rec. | F1-score | Prec. | Rec. | F1-score |
| SVM | 85 | 84 | 83 | 88 | 73 | 74 |
| Random Forest | 85 | 81 | 81 | 86 | 73 | 74 |
| Decision Tree | 86 | 83 | 82 | 82 | 74 | 75 |
| Naïve Bayes | 75 | 74 | 73 | 90 | 67 | 72 |

TABLE VIII. RESULTS OF DIFFERENT MATRICES ON DATASETS WITH TOP-7 FEATURES

| | patient.info (dataset A) | | | i2b2 (dataset B) | | |
|---|---|---|---|---|---|---|
| Algorithm | Prec. | Rec. | F1-score | Prec. | Rec. | F1-score |
| SVM | 95 | 92 | 93 | 88 | 84 | 84 |
| Random Forest | 93 | 91 | 91 | 88 | 84 | 84 |
| Decision Tree | 92 | 90 | 90 | 89 | 86 | 86 |
| Naïve Bayes | 80 | 79 | 79 | 97 | 53 | 64 |




(a) Confusion matrix of SVM on patient.info (dataset A) with top-7 features.




(b) Confusion matrix of Decision Tree on i2b2 (dataset B) with top-7 features.

Fig. 2. Confusion matrix of algorithms where they performed outstanding w.r.t patient.info and i2b2.


(a) Confusion matrix of SVM on patient.info (dataset A) with top-7 features.


(b) Confusion matrix of Decision Tree on i2b2 (dataset B) with top-7 features.

Fig. 3. Confusion matrix of algorithms where they performed outstanding w.r.t patient.info and i2b2.

TABLE IX. CLASSIFICATION REPORT OF SVM ON PATIENT.INFO (DATASET A) WITH TOP-7 FEATURES

| Disease | Precision | Recall | F1-score |
|---|---|---|---|
| Abdominal Issues | 1.00 | 1.00 | 1.00 |
| Asthma | 0.93 | 0.83 | 0.88 |
| Diabetes | 0.94 | 1.00 | 0.97 |
| Heart | 0.97 | 0.99 | 0.98 |
| Hypertension | 0.88 | 0.84 | 0.86 |
| Fever | 0.92 | 1.00 | 0.96 |

TABLE X. CLASSIFICATION REPORT OF DECISION TREE ON I2B2 (DATASET B) WITH TOP-7 FEATURES

| Disease | Precision | Recall | F1-score |
|---|---|---|---|
| Abdominal Issues | 0.99 | 1.00 | 0.99 |
| Asthma | 0.70 | 0.93 | 0.80 |
| Diabetes | 0.87 | 0.98 | 0.92 |
| Heart | 0.99 | 0.70 | 0.82 |
| Hypertension | 0.56 | 0.97 | 0.71 |
| Fever | 0.72 | 1.00 | 0.84 |

classifiers, while the SVM shows the highest figures followed by Random Forest and Decision Trees. The accuracy of the proposed system is also invigorating. Figure 2(b) shows that except the result of naïve Bayes on i2b2, every experiment shows the gradual increment in the accuracies as we move forward with the increment in keyword features. The highest accuracy yields on patient.info dataset are $\approx 95\%$ and on i2b2 is $\approx 85\%$ by SVM and decision tree respectively. These accuracies are yielded at the feature size of 7 keywords.

Figure 3 shows the errors and misclassification in the form of confusion matrices with the same experimental setting reported for tables IX and X. In figure 3, at x-axis there is the predicted and y-axis refers the actual classes. Since patient.info accounts hypertension as a sub-class in heart-related diseases therefore, we can give an empirical argument that the misclassifications are due to the nearly co-related diseases, like in figure 3(a) misclassifying 'asthma' as 'heart', 'hypertension', and 'fever', and in figure 3(b) misclassifying 'heart' as 'hypertension', 'asthma' and 'fever'. In the entire experimental suit, the highest misclassification is seen in i2b2, specifically 'heart' as 'hypertension'.

## VI. CONCLUSION

In the medical field mostly the problems need their solutions for the betterment of the society at a broader level. The natural language processing can help many things related to the text. The patient descriptions in our local context are written in the form of textual format. In this study, we have developed a

solution for the medical practitioners and the doctors. Our solution is more focused towards the processing of text and feature extraction from the plain text and then to form a query that can work both for the classification of the textual descriptions and suggest the preventions based on the information given in the description. We have employed the patient descriptions for this purpose and applied the natural language processing and machine learning techniques to provide the first aid type decision to proceed for further diagnosis. We have got good results with small datasets. Also, we have calculated the results of multiple keywords and key phrases. The results shows SVM as the classifying champion amongst naïve Bayes, Random Forest, and Decision Tree algorithms.

## VII. Future Work

The real-time and problem related local context based dataset was the bigger challenge. In future, this work can be improved with more optimal results by utilising the Named Entity Recognition (NER) with word embedding techniques and deep learning algorithms. Also, we see this work as an extension towards the chatbot form with the large dataset.

## References

[1] B. Djulbegovic and G. H. Guyatt, "Progress in evidence-based medicine: a quarter century on," *The Lancet*, vol. 390, no. 10092, pp. 415–423, 2017.

[2] A. Sarker, D. Molla, and C. Paris, "Automated text summarisation and evidence-based medicine: A survey of two domains," *arXiv preprint arXiv:1706.08162*, 2017.

[3] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, "Clinical natural language processing in languages other than english: opportunities and challenges," *Journal of biomedical semantics*, vol. 9, no. 1, p. 12, 2018.

[4] X. Guo, T. Klinger, C. Rosenbaum, J. P. Bigus, M. Campbell, B. Kawas, K. Talamadupula, G. Tesauro, and S. Singh, "Learning to query, reason, and answer questions on ambiguous texts," 2016.

[5] D. Mollá, M. E. Santiago-Martínez, A. Sarker, and C. Paris, "A corpus for research in text processing for evidence based medicine," *Language Resources and Evaluation*, vol. 50, no. 4, pp. 705–727, 2016.

[6] M. Conway, S. Doan, A. Kawazoe, and N. Collier, "Classifying disease outbreak reports using n-grams and semantic features," *International journal of medical informatics*, vol. 78, no. 12, pp. e47–e58, 2009.

[7] İ. Dönmez and E. Adali, "Extracting phrase-content pairs for turkish sentences," in *Application of Information and Communication Technologies (AICT), 2015 9th International Conference on*. IEEE, 2015, pp. 128–132.

[8] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[9] A. Chandurkar and A. Bansal, "Information retrieval from a structured knowledgebase," in *Semantic Computing (ICSC), 2017 IEEE 11th International Conference on*. IEEE, 2017, pp. 407–412.

[10] S. N. Kim, D. Martinez, L. Cavedon, and L. Yencken, "Automatic classification of sentences to support evidence based medicine," in *BMC bioinformatics*, vol. 12, no. 2. BioMed Central, 2011, p. S5.

[11] A. Sarker, D. Mollá, and C. Paris, "Query-oriented evidence extraction to support evidence-based medicine practice," *Journal of biomedical informatics*, vol. 59, pp. 169–184, 2016.

[12] H. Yu and Y.-g. Cao, "Automatically extracting information needs from ad hoc clinical questions," in *AMIA annual symposium proceedings*, vol. 2008. American Medical Informatics Association, 2008, p. 96.

[13] W. Pratt and L. Fagan, "The usefulness of dynamically categorizing search results," *Journal of the American Medical Informatics Association*, vol. 7, no. 6, pp. 605–617, 2000.

[14] Y. Cao, F. Liu, P. Simpson, L. Antieau, A. Bennett, J. J. Cimino, J. Ely, and H. Yu, "Askhermes: An online question answering system for complex clinical questions," *Journal of biomedical informatics*, vol. 44, no. 2, pp. 277–288, 2011.

[15] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," *Text Mining: Applications and Theory*, pp. 1–20, 2010.

[16] S. Bird and E. Loper, "Nltk: the natural language toolkit," in *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, 2004, p. 31.

[17] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[19] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] P. P. Limited. (2018) Symptom checker, health information and medicines guide — patient. Patient Platform Limited. Accessed: 2018-06-27. [Online]. Available: https://patient.info/

[22] i2b2. (2018) i2b2: Informatics for integrating biology & the bedside. Partners Healthcare. Accessed: 2018-06-27. [Online]. Available: https://www.i2b2.org/NLP/DataSets/

[23] J. Patrick and M. Li, "High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 524–527, 2010.

# Deep Learning based Object Distance Measurement Method for Binocular Stereo Vision Blind Area

Jiaxu Zhang
Automation School
Xi'an University of Technology
Xi'an, China

Shaolin Hu
Automation School
Guangdong University of Petrochemical Technology
Guangdong, China

Haoqiang Shi
Automation School
Xi'an University of Technology
Xi'an, China

***Abstract*—Visual field occlusion is one of the causes of urban traffic accidents in the process of reversing. In order to meet the requirements of vehicle safety and intelligence, a method of target distance measurement based on deep learning and binocular vision is proposed. The method first establishes binocular stereo vision model and calibrates intrinsic extrinsic and extrinsic parameters, uses Faster R-CNN algorithm to identify and locate obstacle objects in the image, then substitutes the obtained matching points into a calibrated binocular stereo model for spatial coordinates of the target object. Finally, the obstacle distance is calculated by the formula. In different positions, take pictures of obstacles from different angles to conduct physical tests. Experimental results show that this method can effectively achieve obstacle object identification and positioning, and improve the adverse effect of visual field blindness on driving safety.**

***Keywords*—deep learning; computer vision; binocular stereo vision; intelligent transportation**

## I. Introduction

With the improvement of peoples living standards, the number of vehicles in China has continuously increased. As of the end of 2016, the number of motor vehicles in China has reached 19.4 million. More and more cars not only make life more convenient, but also accompanied with security risks, and 2016 accidents happen 165,200 times a year in road accidents. Some of these car accidents were caused by poor back sight. Usually, during the reversing process, the driver can only judge the rear environment of the vehicle through technologies such as rear-view mirror and parking sensor. These methods inevitably have problems such as single function or blind area, which leads the driver to make wrong judgments. With the growth of automotive safety and intelligence requirements, how to effectively achieve rapid detection and positioning of the reversing environment, and distinguish obstacle types become one of the focuses of research on intelligent traffic safety.

Compared with active measurement sensors such as ultrasonic sensors, laser sensors, and infrared sensors, vision sensors have the advantages of simple installation, rich information, and high efficiency, which makes the application of visual information more and more in the field of traffic intelligence. Binocular vision technology imitates human eye structure and uses two or more cameras to collect images of different orientations of the same target. The 3D information of the target can be accurately calculated by the matching image points between the left and right images under the binocular camera model [1]. In the technology of binocular positioning measurement, how to extract the features in the image and match the left and right images is a difficult and important point in the binocular vision field. The traditional stereo feature matching method is to find corresponding points in the image pair from features such as edges, colors, and textures in the image [2]. However, these methods are computationally intensive, lacking in pertinence, and are prone to mis-matching unnecessary targets, which reduces the performance and accuracy of the algorithm. In fact, the distance detection of the blind area of the vehicle only needs to obtain the coordinates of the target to be detected in a complex traffic scene, and the target detection algorithm can achieve target recognition and segmentation. At present, there are many researches on target detection. Among them, the background difference method [3] and optical flow method [4] can be used to extract moving objects from the background in continuous video frames and apply them to the background of intelligent traffic [5], [6]. The target detection can also be accomplished by using features such as Histogram of Gradient [7] and Scale-invariant feature transform [8] on the image, and a good detection effect is achieved in visual navigation [9] and intelligent video surveillance [10]. In recent years, with the rapid development of deep learning in the field of image and video processing, the target detection algorithm has developed by leaps and bounds. Deep learning overcomes the limitations of traditional hand-designed features and can accomplish goals in complex lighting, angular rotation, and occlusion conditions. A growing number of methods based on convolutional neural networks (CNNs) [11] have been proposed in the target detection direction, such as the YOLO [12] based on the regression idea proposed in 2016, and the R-CNN based on classification thought proposed in 2014 and the later improved fast-R-CNN [13] and faster-R-CNN [14]. Compared with traditional methods, target detection based on deep learning has a great improvement in performance and speed.The fastR-CNN detection method combines RPN(Region Proposal Networks) and convolutional neural network. Compared with the traditional method, the network greatly improves the robustness of the target detection algorithm and has strong adaptability to illumination changes. The accuracy rate has been greatly improved, in addition to the advantages of fewer parameters and easier updating.

In view of the accuracy, intelligence and real-time requirements of distance measurement for traffic obstacles, a new method combining deep learning and binocular vision is proposed in this paper. Using deep learning to detect obstacles in the scene, it avoids the ineffective calculation in traditional

methods. Binocular vision is combined with the deep learning detection algorithm to achieve the distance measurement of obstacle targets. It is verified by experiments that the algorithm can realize the distance measurement task.

## II. Principles and Methods

### A. Distance Measurement System Overall Structure

In order to achieve intelligent distance measurement of blind obstacles in vehicles, an overall architecture is designed as shown in Fig 1.

The entire system consists of three parts: camera calibration module, deep learning detection module and distance detection module. The deep learning detection module in the system contains two parts: model building and training. The target detection network built in this paper can learn to accurately extract the features of the target to be detected in the continuous gradient descent iteration, and realize the target recognition and segmentation location in the image. The binocular camera calibration module is used to construct the linkage between the binocular image plane coordinates and 3D spatial position in the image pair data. In the calibration module, the data collection platform of the binocular camera is set up first, and the intrinsic and extrinsic parameters of the left and right cameras are obtained by the camera calibration, and the relationship between the image coordinate system and the world coordinates can be obtained under the binocular vision system. Finally, in the distance detection experiment, the binocular camera collects the image pairs of the different angle of the obstacle in the scene. The data input target detection module is used to find the coordinates of the target in the image. The coordinates of the matching points between the left and right cameras and the binocular vision model can be combined to restore the space position of the target. Finally, the intelligent recognition and distance measurement of the obstacles are realized. The system uses the binocular vision principle and deep learning method to quickly and accurately realize the functions of obstacle object identification, detection, and distance measurement.

### B. Object Detection Based on Deep Learning

Based on the above system design, this paper first uses the Faster R-CNN [14] model for non-contact obstacle recognition, which can complete the stereo matching task of the image plane during distance measurement. Faster R-CNN is an end to end target detection algorithm, which unify the steps of feature extraction, region proposal, bounding box regression and classification to the framework as shown in Fig 2 This structure avoids the problem that the algorithm needs extra space to cache a large number of region proposal, and reduces the difficulty of training, and has a great improvement in the accuracy and speed of detection and positioning. The test process can be divided into four steps. First, features are extracted from the original image using a convolutional neural network and combined into a more abstract feature map; then, the feature map is input into the RPN (Region Proposal Networks) to extract the candidate region of the target. Then use the ROI pooling layer to pool the target candidate area to a fixed scale connected full-connected layer, and finally, the softmax regression algorithm is used to classify the target, and the target boundary frame is obtained by using the multi task loss function. The output of the network is a 5-dimensional vector containing the target category and location information.

*1) Feature Extraction Based on Deep Learning:* The difference between deep learning and traditional pattern recognition is that it can automatically learn visual features rather than traditional manual design features. Traditional feature extraction relies on prior knowledge and parameter adjustment experience, and the performance of the system is mainly determined by the quality of the feature design. In the face of rich and diverse mass data, traditional methods cannot take advantage of the number of data. Each feature can only extract a specific feature of a small area, such as texture, color or combination of various features. In order to realize the global feature learning of the image and get pixel-level information of the image, LeCun proposed a convolutional neural network (CNN) [11] based on the traditional neural network, which is composed of the Convolutional layer and the pooling layer alternately. Among them, the convolutional layer and the pooling layer match each other, automatically discovering the hidden features in the image data, and have satisfactory results on computer vision problems. This method not only reduces the complexity of the implementation, but also has a strong scalability for different types of image data with different illumination and occlusion conditions.

The target detection in this paper requires the use of VGG convolutional neural network [15] for feature extraction, and then the extracted feature map is sent to the fully connected layer to complete the task of classification or bounding box regression. The VGG model is the same as a conventional convolutional network and consists of a convolutional layer, a pooled layer, and a fully connected layer. The network structure, as shown in Fig 3, consists of 16 layers of convolution, 5 layers of maximum pooling, 3 layers of full connectivity, and a softmax classifier. The difference compared with the traditional convolutional network is that a 77 convolution kernel is decomposed into multiple convolution kernels with a size of 33 and a stride size of 1, and the number of convolution layers per layer is doubled. Each pooling layer halved the side length of the feature map. This structure allows the network to use less parameters to get a deeper network structure, add more nonlinearity to the network, and the details of the image can be better expressed. Convolution through layers makes it possible to minimize the amount of subsequent calculations while retaining the features of the input image, providing input for region proposal networks.

*2) Region Proposal Networks:* The Region Proposal Networks (RPN) [16] takes the feature map extracted by the convolutional neural network as input, and outputs bounding boxes of the detection target , and each candidate frame corresponds to the probability score and location information of a target. The RPN network structure used in this paper, as shown in Fig 4, uses a 33 convolution sliding window to traverse the entire feature graph, and maps each sliding window to a 256-dimensional feature vector. An anchor is set for each sliding window. Each anchor corresponds to a total of k region proposals with multiple scales and different aspect ratios(This article uses a scale of 128,256,512, length, width ratio of 1:1, 1:2, 2:1 total of nine anchor). Each low-dimensional feature

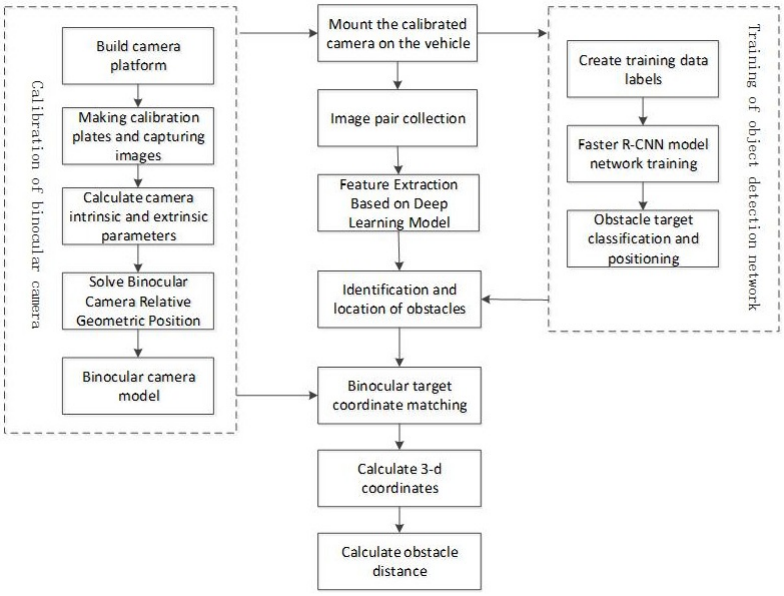



Fig. 1. System architecture diagram

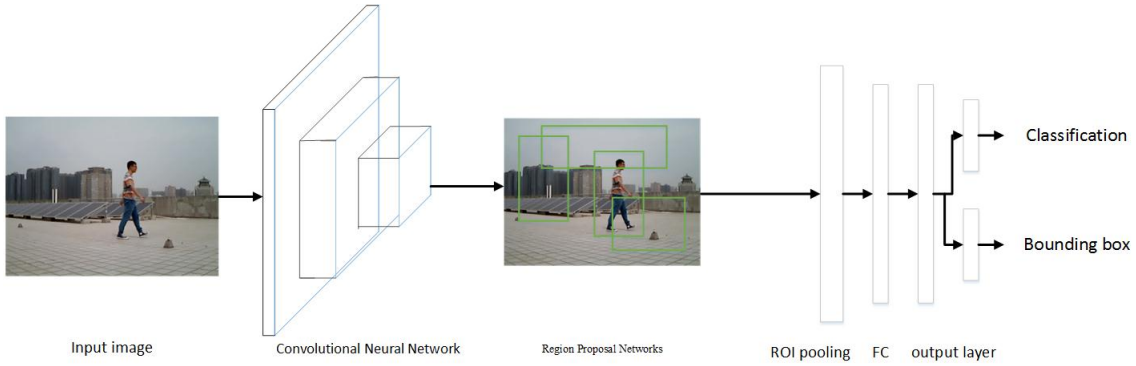



Fig. 2. Object detection model based on deep learning

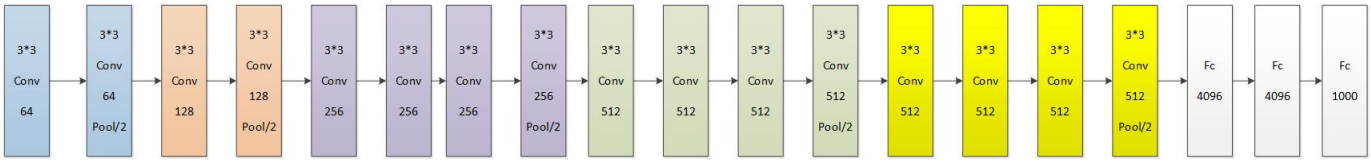



Fig. 3. VGG network structure

vector connects two parallel full-connection layers, which are used to implement bounding box regression and classification, respectively. Then all the candidate regions of the feature map are ranked according to the probability. Finally, the first 300 regions are selected as candidates for the Faster R-CNN network.

*3) Obstacle Detection Model Based on Faster R-CNN:* After the feature extraction and candidate region proposal have been performed on the input image, the candidate regions with different sizes are connected to the ROI pooling layer, and the feature maps of these candidate regions are converted into a fixed size. Finally, the fixed-size feature map is connected to the full-connection layer , and the full connection layer is divided into two branches, which are the classification layer using Softmax and the regression layer that rectifies the target bounding box. The multitask loss function used by the model is shown in Formula (1).

$$L(p_i, t_i) = \frac{1}{N_{cls}} \sum L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum p_I^* L_{reg}(t_i, t_i^*) \quad (1)$$

$$L_{cls}(p_i, p_i^*) = -\log(p_i^* p_i + (1 - p_i^*)(1 - p_i)) \quad (2)$$

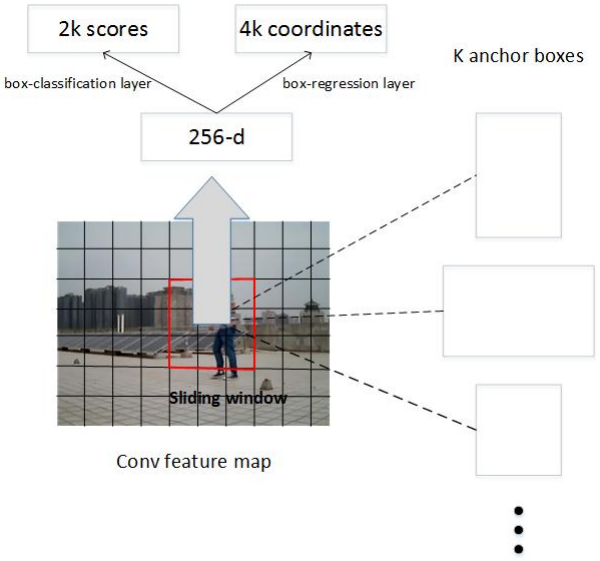



Fig. 4. Region Proposal Networks

$$L_{reg}(t_i, t_i^*) = R(t_i - t_i^*) = smooth_{L1}(t_i, t_i^*) \tag{3}$$

$$smooth_{L1}(t_i, t_i^*) = \begin{cases} 0.5(t - t^*)^2, if|t - t^i| < 1 \\ |t - t^*| - 0.5, otherwise \end{cases} \tag{4}$$

Formula (1): $p_i$ is the target prediction probability, if the region contains the target $p_i^*$ is 1, otherwise it will be 0. $t_i$ is the predicted value of bounding frame coordinates, and $t_i^*$ is the true coordinates of the bounding box. $N_{cls}$ and $N_{reg}$ are the normalized parameters of the classification and regression terms, respectively, and $Lambda$ is the balance weight. $L_{cls}$ is the cross-entropy loss of the classification and $Lreg$ is the regression loss, where $R$ is the robust loss function.

Faster R-CNN implements an end-to-end network structure to achieve target positioning and classification by sharing convolutional features. Save a lot of computing and caching needs, and achieve fast and accurate target detection.

### C. Binocular Stereo Distance Measuring Principle

Binocular stereo ranging uses two cameras with different positions to shoot the same target, and then uses the corresponding image points to get the depth information of the target. The distance measuring system is mainly realized by three steps: camera imaging, three-dimensional reconstruction and distance measurement.

*1) Camera Model:* The image acquisition in the system is completed by two cameras of the same type. The camera projects the three-dimensional scene onto the two-dimensional imaging plane through the optical lens. In order to describe the perspective mapping relationship between the plane image and the space target, this paper uses a pinhole camera model, as shown in Fig 5, to describe the camera imaging process. The image point x is formed by intersecting a straight line connecting the three-dimensional point X and the camera optical center C with the image plane. The dashed line represents

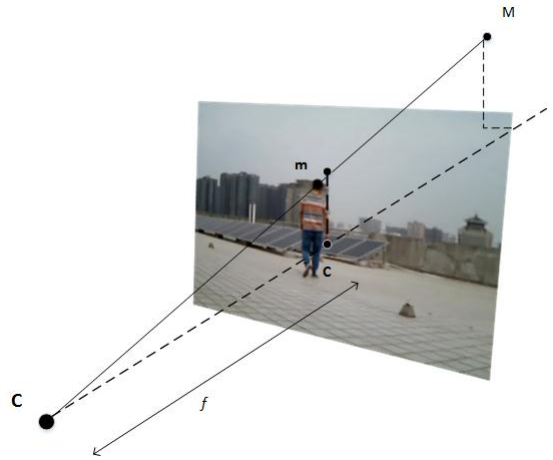



Fig. 5. Pinhole camera model

the camera's optical axis and f represents the camera's focal length.

In the pinhole camera model, the mapping relationship between the three-dimensional space points and the two-dimensional projection point coordinates is as shown in formula (5), and the homogeneous coordinate form of formula (5) is as shown in formula (6).

$$m = PM \tag{5}$$

$$\begin{aligned} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} &= K[R|t] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \end{aligned} \tag{6}$$

$m$ is the coordinate of the spatial point $M$ projected on the imaging plane. Among them, $f$ represents the focal distance, that is, the distance between the image plane and the camera center, and the $f_x = \frac{f}{d_x}$ and $f_y = \frac{f}{d_y}$ and ($d_x, d_y$ as the physical dimensions of a single pixel) represent the normalized focal length on the X and Y axis, respectively. The intersection point of optical axis and image plane is called the origin of image coordinates,$c = [c_x, c_y]$ . These parameters are called intrinsic parameters, and they are composed of $3 \times 3$ matrix $K$ called the camera's intrinsic parameter matrix to describe the projection relationship of the camera. $R$ is a rotation matrix describing the direction of the camera, t is a three-dimensional translation vector describing the center position of the camera, and a $4 \times 4$ matrix consisting of $R$ and $t$ is called an extrinsic parameter matrix.

*2) Binocular Vision Principle:* The camera can only get two-dimensional images and lose depth information in the process of photography. Therefore, in order to realize the visual distance measurement, we can imitate the human eye with two cameras in different locations to collect the images, and then use the projection relationship between the position of the

image point and the three-dimensional space point to calculate the target space position. In this experiment, the positions of two cameras are arranged as shown in Fig 6. The optical axes of the left and right cameras do not need to be parallel but the visual field overlaps. $O_L$ and $O_R$ are the optical centers of the left and right cameras respectively. The projections of the space point $M$ on the imaging planes of the two cameras are $x_1$ and $x_2$, respectively, and $Z_L$ and $Z_R$ are the optical axes of the two cameras.

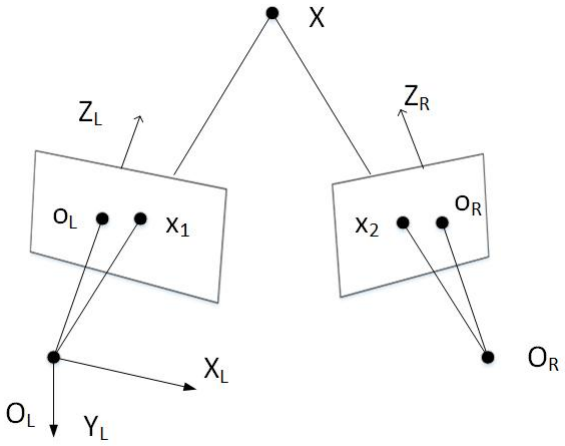



Fig. 6. Binocular 3-dimensional reconstruction model

When analyzing binocular geometric relationships, the relative positional relationship between the two cameras is represented by a homography matrix. Since only rigid body transformation is performed between the left and right cameras, that is, the coordinate system is transformed only by the homography matrix, the geometric structure between the two cameras can be represented by the rotation matrix $R$ and the translation matrix $T$. For calculation convenience, coordinate axis and origin of coordinates are aligned with the first camera coordinate system. That is, the origin of the space coordinate system is the left camera optical center $O_L$, the coordinate $Z$ axis coincides with the optical axis $Z_L$ and is perpendicular to the image plane, and the $X_L$ and $Y_L$ axes are parallel to the x axis of the image y axis. Thus, the projection matrix formula of left and right cameras can be obtained (7)(8):

$$P_1 = K_1(I|0) \tag{7}$$

$$P_2 = K_2(R|t) \tag{8}$$

According to the projection model formula (5) of the camera, we can obtain the relationship between the three-dimensional point $M$ and the two-dimensional projection point $P_L$ and $P_R$ as formula (9). The projection models of the two cameras are $p_1$ and $p_2$ respectively, and the projection points of 3D point $M$ are $x_1$ and $x_2$.

$$\begin{bmatrix} P_1 & -x_1 & 0 \\ P_2 & 0 & -x_2 \end{bmatrix} \begin{bmatrix} M \\ 1 \\ 1 \end{bmatrix} = 0 \tag{9}$$

During distance measurement, the corresponding coordinates of the left and right images are $x_1(u_L, v_L)$ and $x_2(u_R, v_R)$, respectively. The intrinsic and extrinsic parameters in the projection matrices $P_1$ and $P_2$ can be obtained by stereo camera stereo calibration. The spatial point $M(x_L, y_L, z_L)$ can be calculated by Equation (9).

*3) Distance Measurement Principle:* The binocular camera model and the deep learning method are combined to obtain the three-dimensional coordinates $(x_L, y_L, z_L)$ of the target in the left camera coordinate system, and then the distance between the obstacle object and the binocular camera device is calculated. Therefore, the formula (10) of the position coordinate of the right camera optical center $(x_r, y_r, z_r, 1)^T$ in the coordinate system can be obtained according to formula (7)(8).

$$\begin{bmatrix} X_R \\ Y_R \\ Z_R \\ 1 \end{bmatrix} = \begin{bmatrix} R_{3\times3} & t_{3\times1} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_L \\ Y_L \\ Z_L \\ 1 \end{bmatrix} \tag{10}$$

From the known three coordinates, from Cosine's Theorem (11), the angle between the left camera and the obstacle line $v_1$, the left camera and the right camera line of the line $v_2$ can be calculated, and the distance between the target and the binocular camera is obtained. The distance $D$ between the target obstacle and the vehicle is as shown in formula (12).

$$\cos\theta = \frac{v_1 v_2}{\|v_1\|\|v_2\|} \tag{11}$$

$$D = \sin\theta \cdot v_1 \tag{12}$$

## III. Test and Result Analysis

During the test process, the data acquisition system consists of two CCD cameras, model GY0814-3MP, a USB 2.0 interface CCD industrial digital camera with the same parameters and an industrial lens with 3 million pixels. Select a 210mm*297mm calibration plate, computer and corresponding software to achieve target distance measurement. The data acquisition platform is shown in Fig 7. The two cameras are fixed on the platform with adjustable spacing, angle and height to meet the needs of different distances and effective ranges. The experimental part includes two parts: The first part is to detect the corresponding image points of the target on the imaging planes of the left and right cameras through deep learning; the second part is to obtain the three-dimensional coordinates and distances by substituting the image points into the calibrated camera model.

Fig. 7. Binocular experimental platform

### A. Faster R-CNN Network Training

In the distance measurement system, we use the detection result (object centroid) of the deep learning model in Section 2.2 as the matching point between the left and right image planes. In order to verify the positioning capability of the Faster R-CNN model, video data was collected by a vehicle digital camera to verify the model performance. The data set is divided into six video sets ($Set_{L1} - Set_{L3}, Set_{R1} - Set_{R3}$).Extract 5 frames of images per second and save them as jpg images of size 800*600 for a total of 2983 images. The data set was randomly divided into training set (80%) and test set (20%) to achieve object detection network training. Each picture also has a detailed tag in XML format for recording image information. The tag contains the target category, image width, height depth and bounding box coordinate information. Finally, 580 images and corresponding tag files are obtained for training and testing models.

In the training process, the initial learning rate is 0.004. As the number of iterations increases, the initial learning rate decreases at a decay rate of 0.95. The model first unified the training set image size to $300 \times 300$, and then iterated 8,000 times using RMSProp algorithm until the loss function reached a minimum, so that obstacles with different distances and angles can be detected and identified, as shown in Fig 8 . When using the test set to verify model performance, the Precision indicator evaluation model was used. The results are shown in Fig 9. With the increase of the number of trainings, the detection accuracy of deep learning can reach 99%, which indicates that Faster R-CNN can accurately locate the obstacles.

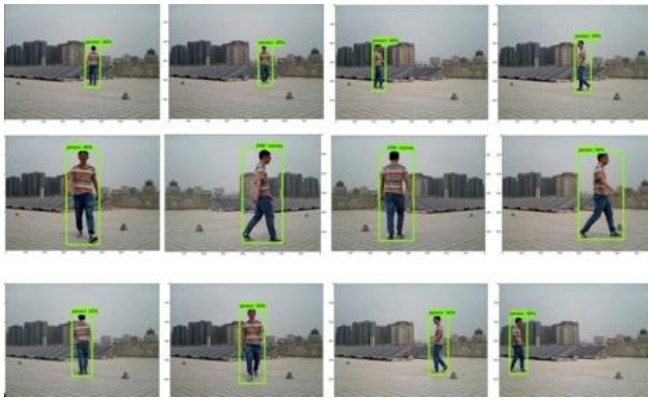


Fig. 8. Test set detection results with different distances and different angles

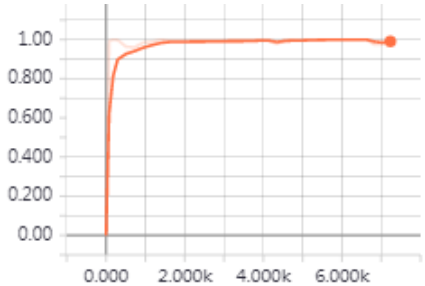



Fig. 9. Performance of object detection network

Fig. 10. Checkerboard calibration board

### B. Binocular Camera Calibration and Distance Measurement

Next, in order to achieve the calculation of three-dimensional position through two-dimensional image coordinates. The centroid matching points $x_1$ and $x_2$ are substituted into the binocular model (9) to calculate the three-dimensional coordinates. The parameters $P_1$ and $P_2$ of the camera in the formula need to be calibrated, and he calibration result has a very important influence on the measurement accuracy. The distance measuring system uses the calibration plate as a reference for binocular vision calibration to solve the camera's intrinsic and extrinsic parameter matrix. In the calibration process, the positions and angles between the binocular cameras are adjusted, and take 28 pictures from different angles on the checkerboard calibration board shown in Fig 10. Then use software Matlab calibration tool stereoCameraCalibrator [17] to extract the corner points of the calibration plate. After obtaining the corner points, the intrinsic parameters of the two cameras are solved using the calibration method [18] as shown in Table 1. Then the mapping relationship between the left and right images is calculated by stereo matching, that is, the extrinsic parameter rotation matrix $R$ and the shift vector $T$. The result is shown in Table 2.

The binocular camera calibration results are shown in Table 1 and Table 2.

TABLE I. BINOCULAR STEREO VISION SYSTEM INTRINSIC PARAMETER CALIBRATION RESULTS

| Intrinsic parameters | $f_x$ | $f_y$ | $c_x$ | $c_y$ |
|---|---|---|---|---|
| Left camera | 1083.08 | 1086.58 | 416.18 | 324.09 |
| Right camera | 1092.60 | 1096.39 | 421.79 | 257.48 |

TABLE II. BINOCULAR STEREO VISION SYSTEM EXTRINSIC PARAMETER CALIBRATION RESULTS

| | | | |
|---|---|---|---|
| $R =$ | 0.9976 | −0.0122 | 0.0677 |
| | 0.0109 | 0.9997 | 0.0198 |
| | −0.0679 | −0.0191 | 0.9975 |
| $t =$ | $2.1858 \times 10^2$ | −2.2566 | 5.7440 |

In order to verify the method proposed in this paper, physical measurement experiments were conducted. Binocular cameras take pictures of obstacles at different distances and angles. The obtained image data is input into the target detection network that has been trained in Section 3.1. The model can automatically distinguish obstacle types and centroid coordinates $x_1$ and $x_2$.Substituting the matching points $x1$, $x_2$ and the calibrated camera projection matrix $P_1$, $P_2$

into equation (9), the three-dimensional coordinate $X$ of the obstacle in space can be obtained. Using distance formula (11) (13), the distance between obstacles and vehicles can be calculated.

The distance measurement experiment was performed on a moving obstacle with a distance of 5-7.5 meters from the binocular camera. As shown in Table 3, the accuracy is analyzed by calculating the error with the actual distance. The experimental results are shown in Table 3 and Fig 11.

TABLE III. BINOCULAR STEREO VISION SYSTEM EXTRINSIC PARAMETER CALIBRATION RESULTS

| Real distance/m | Measuring distance/m | Error/m | relative error/% |
|---|---|---|---|
| 4.5 | 4.311 | 0.189 | 4.20 |
| 4.5 | 4.452 | 0.047 | 1.04 |
| 4.5 | 4.101 | 0.398 | 8.84 |
| 4.5 | 4.456 | 0.044 | 0.90 |
| 4.5 | 4.374 | 0.126 | 2.80 |
| 5.5 | 5.110 | 0.390 | 7.09 |
| 5.5 | 5.756 | -0.256 | 4.65 |
| 5.5 | 5.271 | 0.229 | 4.16 |
| 5.5 | 5.382 | 0.118 | 2.15 |
| 5.5 | 5.264 | 0.236 | 4.29 |
| 6.5 | 6.814 | -0.314 | 4.83 |
| 6.5 | 6.577 | -0.077 | 1.71 |
| 6.5 | 6.283 | 0.217 | 3.34 |
| 6.5 | 6.329 | 0.171 | 2.63 |
| 6.5 | 6.073 | 0.426 | 6.55 |
| 7.5 | 7.171 | 0.329 | 4.39 |
| 7.5 | 7.804 | -0.304 | 4.05 |
| 7.5 | 7.098 | 0.402 | 5.36 |
| 7.5 | 7.747 | -0.247 | 3.29 |
| 7.5 | 7.560 | 0.06 | 0.80 |

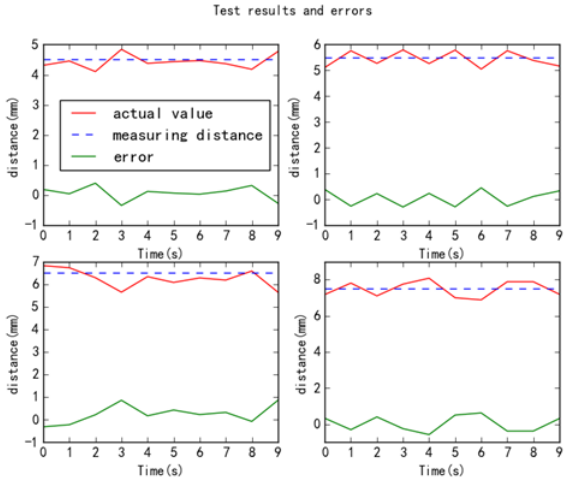



Fig. 11. Test set detection results with different distances and different angles

## IV. CONCLUSION

In order to avoid the traffic accident caused by the driver's limited vision during reversing, a system based on binocular vision and deep learning is designed to realize obstacle recognition and distance detection. In the system, the calibrated binocular camera equipment is installed in the rear of the vehicle, the image data is detected by the deep learning corresponding image points. The triangulation method and the binocular camera model are used to calculate the distance information of the target. Tests were performed on targets with different viewing angles and different distances to verify the performance of the system. Compared with traditional methods, deep learning has the advantages of robustness, accuracy, and speed, and it can achieve real-time, high-precision recognition and distance measurement through the combination of binocular stereo vision. The method proposed in this paper can effectively identify the target object and determine the distance, which ensures the safety of the driver's personal and property during driving.

Although this article has achieved some results in ranging, there are still many problems to be solved. Here are the next research plans for this article:

1: In the study of obstacle detection in this paper, only one category of pedestrians is involved, but in actual situations, there are many types of obstacles and the shape is varied. In the future research, it is necessary to select different types of obstacles for research and improve the applicability of the research content.

2: The method used in this paper can detect the distance of the obstacle, and in the future can detect more details of the target, including speed, direction, angle and so on. Achieve more accurate target detection and ensure the safety of vehicle driving.

## ACKNOWLEDGMENT

This paper is supported by the National Nature Science Foundation of China (Grant No.61473222 and No. 91646108

## REFERENCES

[1] Xiaochao Tian,Zhongke Li. Research on Computer Vision Feature Representation and Learning for Image Classification and Recognition. Electronics Optics & Control, 2015,22(03):54-57+96.

[2] Zhirui Wang,Cailiang Yan. Method of Image Feature Extraction. Journal of Jishou University(Natural Science Edition), 2011,32(05):43-47.

[3] Yu bo,Guili Wang,Yanxu Zhou,Peng. Investigation on Video Moving Target Detection Algorithm in Traffic Monitoring System. Computer Technology and Development, 2017,27(08):156-158+163.

[4] Yijing Hu, Zhengfang Li, Yueming Hu. Theories and Applications of Motion Analysis Based on Optical Flow. Computer Measurement & Control, 2007,15(2):219-221.

[5] Huiyin Zhu, Li Zhang, Yunyan Li. Intelligent Traffic Event Detection System Based on Background Difference. Journal of wut (information & management engineering ) , 2011, 33(2):224-227.

[6] Xilai Li, Aihua Li, Xiangfeng Bai. Moving Vehicles Detection in Intelligent Transportation Systems Based on Optical Flow. Electro-Optic Technology Application, 2010, 25(2):75-78.

[7] Pass G, Zabih R. Histogram refinement for content-based image retrieval// IEEE Workshop on Applications of Computer Vision. IEEE Computer Society, 1996:96.

[8] Lowe D G. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 2004, 60(2):91-110.

[9] Xiandong Xu, Bingrong Hong, Yiguan, , et al. A somparison of several feature points methods used in mobile robot visual navigation. Journal of Huazhong University of Science and Technology(Natural Science Edition), 2011, 39(s2):200-203.

[10] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Engineering & Science, ,2015,37(10):1947-1951.

[11] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.

[12] Uijlings J R R, Sande K E A V D, Gevers T, et al. Selective Search for Object Recognition. International Journal of Computer Vision, 2013, 104(2):154-171.

[13] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. 2015:770-778.

[14] He K, Zhang X, Ren S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(9):1904-1916.

[15] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Computer Science, 2014.

[16] Felzenszwalb P F, Girshick R B, Mcallester D A, et al. Object Detection with Discriminatively Trained Part-Based Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

[17] Wei Zhang, Hong Cheng, SuiWei. Design and Implementation of Camera Calibration System . Computer Engineering, 2007, 33(2):255-256.

[18] Z. Zhang, "Flexible Camera Calibration by Viewing a Plane from Unknown Orientations," Proceedings of the Seventh IEEE International Conference on Computer Vision(ICCV), Corfu, Greece, 1999, pp. 666. doi:10.1109/ICCV.1999.791289

# Light but Effective Encryption Technique based on Dynamic Substitution and Effective Masking

Muhammed Jassem Al-Muhammed
Faculty of Information Technology
American University of Madaba
Madaba, Jordan

***Abstract*—Cryptography and cryptanalysis are in ever-lasting struggle. As the encryption techniques advance, the cryptanalysis techniques advance as well. To properly face the great danger of the cryptanalysis techniques, we should diligently look for more effective encryption techniques. These techniques must properly handle any weaknesses that may be exploited by hacking tools. We address this problem by proposing an innovative encryption technique. Our technique has unique features that make it different from the other standard encryption methods. Our method advocates the use of dynamic substitution and tricky manipulation operations that introduce tremendous confusion and diffusion to ciphertext. All this is augmented with an effective key expansion that not only allows for implicit embedment of the key in all of the encryption steps but also produces very different versions of this key. Experiments with our proof-of-concept prototype showed that our method is effective and passes very important security tests.**

***Keywords*—Encryption techniques; dynamic substitution; key expansion; directive based manipulation; block masking**

## I. Introduction

In the digital era, almost all of our sensitive information is either transmitted over the network or digitally stored on machines. This information will inevitably be in a great risk if we do not properly secure them. Encryption is the *de-facto* means for keeping the security of the transmitted or stored information. Many encryption techniques have been proposed (e.g. [1][2][3][4][5][6][7][8][9][10] [11][12][13][14][15][16][17][18]). Although these methods are effective and purport to provide high levels of security, there is always a truly pressing need for new techniques that can effectively face the formidably ever-advancing hacking tools.

This paper proposes an effective encryption technique. This technique consists of three effective encryption operations along with a novel key expansion operation. First, the proposed substitution operation adopts dynamic behavior. Unlike all other encryption techniques (noticeably [3]), which use a static substitution operation, our technique uses a dynamic substitution operation whose state depends on the key and greatly sensitive to its changes. Second, the diffuse operation is highly sensitive to changes of the key or the plaintext block and is capable of greatly magnifying and then propagating these changes to all of the block's symbols. Third, the masking operation uses a novel technique that makes sharp changes to both the individual symbols of the plaintext block and to its structure. Finally, the key expansion technique proposes an innovative model that combines the lookback-based substitution with the random manipulations.

The paper makes the following contributions. First, it proposes a full-fledged encryption technique. Second, it proposes an effective dynamic substitution operation whose functionally highly depends on the key. Third, it proposes a novel key expansion technique. Forth, it proposes an effective masking operation that impacts both the block's individual symbols and the structure of the block.

We present our contribution as follows. Section II describes the technical details of the dynamic substitution operation. Sections III through VI present the details of the operations that comprise our proposed cipher. Section VII presents the technical details of the proposed cipher. Section VIII presents the related work. Section IX presents the performance analysis of our technique. We conclude and give directions for future work in Section X.

## II. Dynamic Box Substitution

The purpose of the substitution is to move from the actual block symbols to new symbols. It uses a data structure superficially similar to S–Box of AES encryption method [3], but is fundamentally different in its dependency on the key. This section describes the D–Box and discusses how it is used in substituting symbols.

### A. The D-Box and its Inverse D-Box$^{-1}$

The D-Box is conceptually $K \times K$ array. The D-Box is populated with the unicode symbols from 0 to $K^2$-1. Each symbol in the D-Box can be accessed by its column and row indexes. The positions of the unicode symbols in the D-Box are never static. Their locations depend on the encryption key. To do this, we utilize a sequence of integers $I_1 I_2 ... I_m$, which are generated in a process that involves the key. (The process for generating these integers is discussed in Appendix A.) These integers are used to reposition the symbols in the D-Box. The repositioning of the symbols is performed by swapping the symbol at the index *i* in the D-Box with the symbol at the index $I_i$.

In order to use the D-Box for encryption, it is necessary to define the inverse substitution so that the original block can be recovered. Fig. 1 illustrates the process of creating D-Box$^{-1}$ from the D-Box. Given a 16×16 D-Box, the D-Box$^{-1}$ is created as follows. For each symbol *s* in D-Box[*r*, *c*], we create an entry in the D-Box$^{-1}$ by dividing the 8 bits of *s*

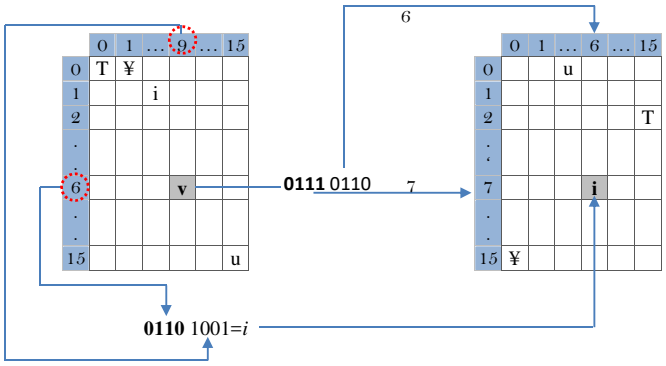



Fig. 1. the creation of D-Box$^{-1}$ (right) from D-Box (left).

into two halves.[1] The left half bits create an index $x$ for the row of D-Box$^{-1}$ and the right half bits create an index $y$ for the column of D-Box$^{-1}$. On the other hand, the 4 bits that represent $r$ and the 4 bits that represent $c$ are concatenated $rc$ to create a unicode symbols $t$. The unicode symbol $t$ is placed in the index ($x$, $y$) of D-Box$^{-1}$.

Referring to Fig. 1, the symbol **v** is at the index $r$ = 6, $c$ = 9. The 8 bits that represent **v** are "01110110" (118 in decimal). The left half of the bits "**0111**" (7 in decimal) index the row of the D-Box$^{-1}$ and the right 4 bits "**0110**" (6 in decimal) index its column. On the other hand, the row index of **v** in D-Box is $r$ = 6 ("**0110**") and the column index of **v** in D-Box is $c$ =9 ("**1001**"). The binaries of the row index and the column index are concatenated "**01101001**" to create the entry **i** (105 in decimal). The symbol **i** is placed in the D-Box$^{-1}$ [7, 6] as shown in Fig. 1.

### B. Symbol-Substitution using D-Box / D-Box$^{-1}$

We define two substitution operations. The first operation is called **Substitute**, which uses the D-Box for replacing a symbol $p$ with a new one. The second operation is called **Inverse Substitute**, which replaces a symbol $q$ using D-Box$^{-1}$. Substituting a symbol using the D-Box and D-Box$^{-1}$ is straightforward. It is a table lookup operation. Given a symbol $p$, the left half bits of $p$'s binary representation index the row of the D-Box and the right half bits index its column. Indexing D-Box$^{-1}$ is done in an identical way.

Fig. 2 shows an example of D-Box and its Inverse D-Box$^{-1}$ (right). For the sake of the simplicity, we use only the hexadecimal numbers. Since we have only 16 symbols, our D-Box is 4×4. Therefore, we need two bits to index its rows and two bits to index its columns. The two bits at the top index the column and the two bits on the leftmost index the rows.

To illustrate the substitution operation, consider a block of four symbols "9C06". The **Substitute** operation reads the first symbol "9" whose binary representation is "1001". The left two bits "10" index the row of the D-Box and the right two bits "01" index the column. The symbol at row 2 ("10") and column 1 ("01") in the D-Box is retrieved as a substitute for the symbol "9". Thus, the symbol "9" is substituted with "7". Continuing likewise, the entire block is substituted yielding the new block "7F4B".

To recover the original block from "7F4B", we use the **Inverse Substitute** operation. The binary of the symbol "7" is "0111". The left half bits "01" index the row of the D-Box$^{-1}$ and the right half bits "11" index its column. As a result, the symbol at row 1 and column 3, which is "9", is retrieved from the D-Box$^{-1}$. The next symbol is "F" whose binary representation is "1111". Thus, the symbol at row "11" and column "11", which is "C", is retrieved from the D-Box$^{-1}$. Using the same process, we recover the original block "9C06".

D-Box

| | 00 | 01 | 10 | 11 |
|---|---|---|---|---|
| 00 | **4** | **A** | **6** | **3** |
| 01 | **2** | **9** | **B** | **C** |
| 10 | **5** | **7** | **E** | **8** |
| 11 | **F** | **0** | **1** | **D** |

D-Box$^{-1}$

| | 00 | 01 | 10 | 11 |
|---|---|---|---|---|
| 00 | **D** | **E** | **4** | **3** |
| 01 | **0** | **8** | **2** | **9** |
| 10 | **B** | **5** | **1** | **6** |
| 11 | **7** | **F** | **A** | **C** |

Fig. 2. D-Box and its inverse D-Box$^{-1}$.

Input Key: $k_1k_2...k_n$
Output: manipulated key $\mathcal{W}$
$\mathcal{T}_1$= Substitute ($k_1$)
$\mathcal{W}$ = $\mathcal{W}$ + $\mathcal{T}_1$
**FOR** i = 2 to n **DO**
  $K_i^\star$ = $\mathcal{T}_{i-1}$ $\oplus$ $K_i$
  $\mathcal{T}_i$ = Substitute($K_i^\star$)
  $\mathcal{W}$ = $\mathcal{W}$ + $\mathcal{T}_i$
RETURN $\mathcal{W}$

Fig. 3. the substitution operation for a key's symbols.

## III. Key Expansion

The key expansion process expands encryption keys to an arbitrary length. The process defines two operations. The first one is the *substitution* operation whose logic is fully described by Fig. 3. The input to this operation is a key, which consists of $n$ unicode symbols $k_1k_2...k_n$. The output is a new (substituted) key $\mathcal{W}$ whose length is also $n$ symbols. As the logic clearly shows, the first symbol $k_1$ in the key is substituted with a new one using the D-Box. For all the symbols $k_i$ ($i$=2, 3, ..., $n$), the operation uses the outcome of substituting $k_{i-1}$ (or $\mathcal{T}_{i-1}$) to substitute the symbol $k_i$. Therefore, to create a new symbol in the output key $\mathcal{W}$, the operation XORes the current symbol of the input key $k_i$ and the result of the previous substitution $\mathcal{T}_{i-1}$ to create a new symbol $K_i^\star$. The symbol $K_i^\star$ is substituted to produce the new symbol $\mathcal{T}_i$ in the output key.

The second operation, called *Manipulate*, creates sharp modifications to its input string. The modifications include changes to both individual symbols and the structure of the input string. Suppose $t_1t_2...t_n$ be an input string. The manip-

[1]We assume, without losing the generality, $K$=16. The D-Box is therefore 16×16. Since we use 16×16 D-Box, the rows are indexed by 4 bits and the columns by 4 bits. Additionally each symbol in the D-Box is represented by 8 bits.

$$Manipulate(t_i) = \begin{cases} \overset{LH}{Flip}(t_i) & if\ t_i \le t_{i+1} \\ \overset{RH}{Flip\ Swap}(t_i) & if\ t_i > t_{i+1} \end{cases}$$

Fig. 4. the manipulation operation.

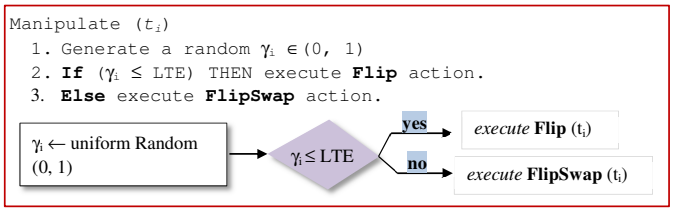



Fig. 5. the dynamic random model for action selection.

ulate operation handles each symbol $t_i$ using the two actions defined in Fig. 4.

The manipulate operation (Fig. 4) handles each symbol $t_i$ based on the lookahead symbol $t_{i+1}$. If symbol $t_i$ lexically comes before or equal to the lookahead symbol $t_{i+1}$, the left half bits of $t_i$ are flipped using the action $\overset{LH}{Flip}(t_i)$. If otherwise, the right half bits of $t_i$ are flipped and the resulting symbol is moved to a position determined by the lookahead symbol $t_{i+1}$. Consider for instance the input string "pqetreeloopc". If the current symbol is "p", then the left half bits of "p" (**0111**0000) are flipped because "p" lexically comes before the lookahead symbol "q" yielding "€" (10000000). Assume that the current symbol is "q". The lookahead symbol is thus "e". Since "q" lexically comes after "e", the right half bits of "q" (0111**0001**) are flipped (0111**1110**) yielding "∼". Additionally and based on Fig. 4, the resulting symbol "∼" is moved to the position 101 % 12 = 5 in the string. (101 is the unicode index of the lookahead symbol "e" and 12 is the length of the input string.) This operation results in "€...∼.....".

Although selecting and applying one of the two actions of the *Manipulate* operation based on the lexical order of input symbols may be effective by itself, we prefer to add more randomness to the action selection. As such, instead of using the plain lexical order of the input symbols, we introduce a random noise to the action selection. In particular, we define a dynamic random model that guides the selection process (Fig. 5). Let $T$ be the number of the so-far processed symbols and $P$ be the number of times in which $t_i \le t_{i+1}$. Based on this, we define two dynamically updated variables *LTE* and *GT* as follows.

$$LTE = \frac{P}{T} \quad and \quad GT = 1 - \frac{P}{T}$$

Where *LTE* means "less than or equal" and represents the ratio in which the condition $t_i \le t_{i+1}$ holds. *GT* means "greater than" and represents the ratio in which the condition $t_i \le t_{i+1}$ is false. Using these two variables, we redefine our *Manipulate* operation (Fig. 4) in Fig. 5.

Although neither *LTE* nor *GT* is random, their values

```
Input: Key=k1k2 ...kn
Output: Expanded-Key = Key
1.  Let L = Key, T = 1, P =0
    /*T=total number of processed symbols and P = number of times in
    which the current input symbol lexically comes before the lookahead
    symbol*/
2.  x1x2 ... xn←Substitute (L)
3.  For i=1 to n Do
4.    If (xi ≤ xi+1) P ++ //increment P
5.    T ++
6.    LTE = P/T
7.    S = S+ Manipulate (xi)
8.  Expanded-Key = Expanded-Key + S
9.  If desired length not reached yet, L = S, GOTO 2
10. Return Expanded-Key
```

Fig. 6. the key expansion process.

change (increase or decrease) based on the lexical order of the current symbols $b_i$ and the lookahead symbols $b_{i+1}$. The amount of the bias toward either action therefore does change. When *LTE* increases (*GT* decreases), the likelihood of executing the *Flip* action becomes larger than the likelihood of executing *FlipSwap* action. Likewise, when *GT* increases (*LTE* decreases), the likelihood of executing the *FlipSwap* action becomes larger than that of executing *Flip* action. As such, *LTE* increases (or decreases) the likelihood of the executing an action over the other, but the choice of the action to be executed happens randomly since it depends on the random value $\gamma_i$. For instance, if *LTE* is 0.8, the likelihood of executing *Flip* is much higher than that of executing *FlipSwap*, but the actual choice whether to execute *Flip* or *FlipSwap* depends on the current random value $\gamma_i$.

After introducing the two operations that constitute the key expansion process, we delineate this process in Fig. 6. The logic is straightforward. We start with the input key $k_1k_2...k_n$. The process creates a new version of the input key using the steps 1 through 7. The symbols of the input key are first substituted with new symbols $x_i$'s. Next, the new symbols $x_i$'s are manipulated using the manipulation operation (Fig. 5). To define the amount of the bias *LTE* (line 6), we always increment the number of processed symbols $T$ while incrementing $P$ only when the currently processed symbols $x_i$ lexically equals to or comes before the lookahead symbols $x_{i+1}$. Therefore, the ratio *LTE* is always in the interval [0, 1] and represents a likelihood of executing *Flip* action.

The new version of the key $S$ is concatenated with the original key. If the desired length has not been reached yet (step 9), the process uses the latest version $S$ of the key to create a new version.

## IV. Diffuse and Inverse Diffuse Operations

The **diffuse** operation detects changes in the input block and propagates this change to affect every symbol in the corresponding output block. To be effective, the diffuse operation must be highly sensitive to the input's change regardless of its magnitude and amplify it so that this change causes tremendous changes to the output. The **inverse diffuse** operation reverses the effect of the diffuse operation and recovers the original input.

Fig. 7 shows the algorithmic steps for the diffuse operation. As the figure shows, it performs double substitutions for the

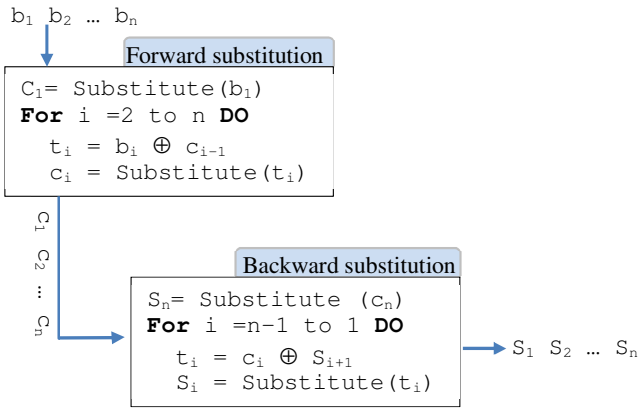



Fig. 7. the algorithmic steps of the diffuse operation.

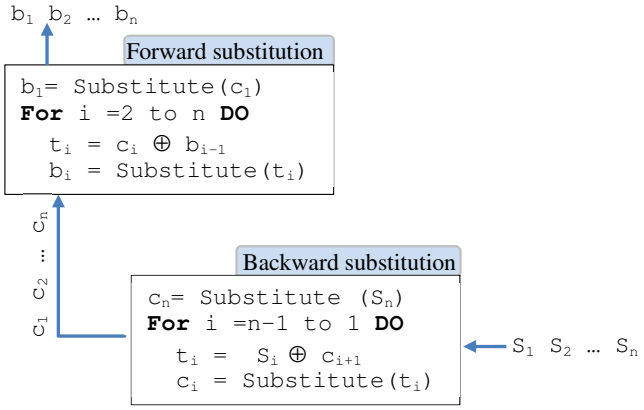



Fig. 8. the algorithmic steps of the inverse diffuse operation.

TABLE I. A SAMPLE OF THE DIFFUSE OPERATION'S OUTPUT

| Input Block | Output Block |
|---|---|
| aaaaaaaaaaaaaaaa | A8 **9B** 32 **81** D1 **91** B9 **EE** E5 **60** 1F **A6** 50 **0E** 2B **36** |
| aaaaaaaaaaaaaaa**b** | 59 **2D** B9 **BE** AB **44** 0A **88** 35 **7A** 01 **B4** D7 **8A** 08 **CD** |
| aa**b**aaaaaaaaaaaaa | F3 **3D** 00 **42** 80 **EF** D9 **0F** D8 **93** E3 **55** DC **FA** 0A **2A** |
| To be or not to be | 2A **0F** 5B **51** EA **E9** 0B **41 DC D1** 1C **30** 58 **1C** 31 **9B** 10 **76** |
| To **B**e or not to be | B5 **71** C4 **1F** A3 **4F** D9 **27** D9 **45** E0 **36** 86 **FE** 6C **08** E8 **19** |

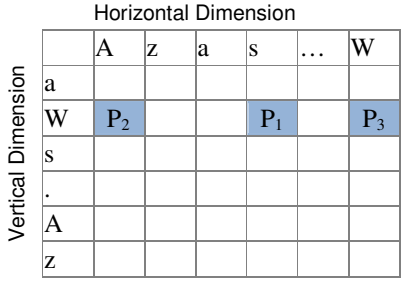



Fig. 9. the mesh.

block's symbols: forward and backward substitutions. When operates on an input block $b_1b_2...b_n$, the forward substitution effectively propagates the change in a symbol $b_i$ to all the following symbols $b_j$ ($j > i$). First, the operation substitutes $b_1$ to yield a new symbol $c_1$. For every subsequent input symbol $b_i$ ($i > 1$), the operation first XORes $b_i$ with the result of the previous substitution $c_{i-1}$ and substitutes the outcome of the XOR. That is, the symbol $t_i$ is calculated as $t_i = b_i \bigoplus c_{i-1}$ and then substituted to yield a new symbol $c_i$.

As Fig. 7 shows, the output of the forward substitution $c_1c_2...c_n$ is passed to the backward substitution. The backward substitution uses similar logic as the forward except that it starts from the end of the input block. The backward substitution substitutes $c_n$ to yield the output symbol $S_n$. For the input symbols $c_i$ ($i$=$n$-1, $n$-2, ..., 1), $c_i$ is first XORed with $S_{i+1}$ and the result of the XOR operation is substituted to yield $S_i$.

With this feedback-based forward substitution, the result of substituting a symbol $b_i$ is impacted not only by the symbol $b_i$ per se, but also by the symbol $b_{i-1}$. That is because if a symbol $b_i$ changes so does its substitution outcome. This change also impacts the substitution of the following symbol $b_{i+1}$ due to the XORing operation, which in turn impacts the substitution of $b_{i+2}$, and so on. In other words, the change in symbol $b_i$ collectively creeps to affect all the successive symbols $b_j$ ($j > i$). Similarly, the backward substitution propagates the change in the symbol $b_i$ back to the symbols $b_k$ ($k < i$). In this case, no matter where the change occurs, the forward and backward substitutions always guarantee that this change impacts every symbol of the input.

The diffuse operation can be reversed. Fig. 8 shows the algorithmic steps of the inverse diffuse operation. As the figure shows, the operation starts from the backward substitution to recover the original block $b_1b_2...b_n$. The input to backward substitution is $S_1S_2...S_n$. The backward operation yields the block $c_1c_2...c_n$. This block is passed as an input to the forward substitution operation, which recovers the original block.

Table I shows an example of the diffuse operations output. The input blocks in the first three rows differ in only a single bit. This minor change causes a very remarkable difference in the corresponding output blocks. Additionally, the position of the change does play a major role. As an example, compare between the output in row 1 and the second and third rows.

## V. DIRECTIVE GENERATOR

This section describes the directives generator. Subsection V-A discusses the mesh and Subsection V-B shows how to use the mesh to produce directives.

### A. The Mesh

Conceptually the mesh is a two dimensional array. The horizontal and vertical dimensions are populated with the unicode symbols from 0 to some integer $N$. Fig. 9 provides an example of a mesh.

Each cell in the mesh is accessed by its row and column indexes. Each move within the mesh whether along the vertical or horizontal dimension has a distance and a direction with respect to the current position. The distance of the move $d$ is the number of cells passed. The direction of the move is either toward a lower or a higher index. We capture the move toward the lower indexes by the flag "–" and toward the higher indexes by the flag "+" regardless whether this move is along the horizontal or vertical dimension. For example, if we move from the current point, say $P_1$, to $P_2$, the distance of the move

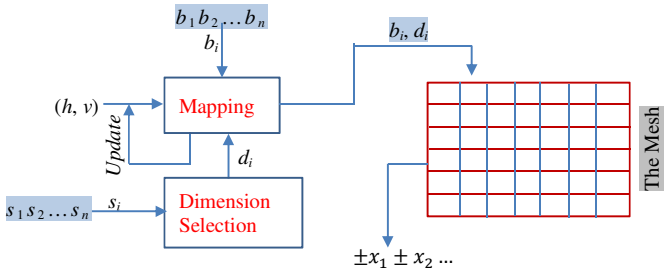



Fig. 10. mapping using the mesh.

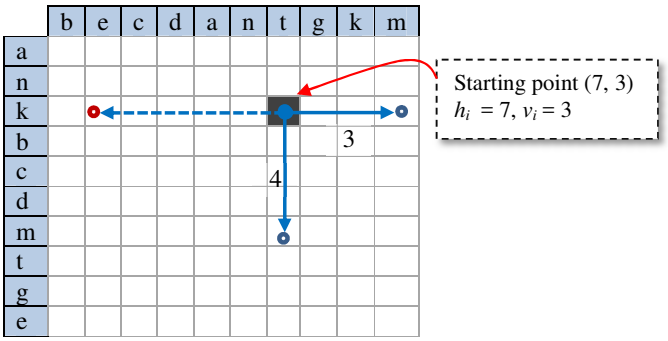



Fig. 11. example of mapping the symbol *m*.

is 3 and the direction is captured by "–" because we moved three cells to the lower indexes. As another example, the move from the current point $P_1$ to $P_3$ has a distance of 2 and the direction is "+" because we moved two cells to the higher indexes.

We can now formally define a directive as a signed integer $\pm\ x$, where $x \geq 0$ represents the distance of the move and $\pm$ represents its direction. Accordingly, moving from the current point $P_1$ to $P_2$ is captured by the directive "–3" and from $P_1$ to $P_3$ is captured by the directive "+2".

### B. Mesh-Based Mapping

Based on the mesh and directive definition, the proposed mesh-based mapping consists of two operations: Mapping and Dimension Selection (see Fig. 10). The mapping operation utilizes three inputs: a block of symbols to be mapped $b_1 b_2 \ldots b_n$, a mapping dimension $d_i$ (could be the horizontal or the vertical dimension of the mesh), and a starting point ($h$, $v$). The starting point ($h$, $v$) is a point within the boundary of the mesh and from which the mapping is started; where $h$ represents the index on the horizontal dimension and $v$ represents the index on the vertical dimension.

Initially, the starting point is created from the first symbol of the input $b_i$'s and the first symbol of the input $s_i$'s. That is, the starting point is defined as ($s_1$, $b_1$). The mapping operation updates the starting point as the mapping proceeds (after producing each directive). Additionally, for mapping any subsequent block, the latest starting point is used for the mapping.

The dimension selection operation utilizes a sequence of symbols $s_1 s_2 \ldots s_n$ and produces a dimension $d_i$, which can be either one of the two mesh's dimensions. The logic of the dimension selection operation must be based on the sequence $s_i$'s. For the purpose of this paper, the dimension selection operation uses the following simple functionality: the mapping dimension $d_i$ is the horizontal if the unicode index of $s_i$ is odd; otherwise the mapping dimension is the vertical.

Based on this, we can formally define our mapping for a symbol $b_i$ as follows. First, the dimension selection operation uses its input symbol $s_i$ to determine the mapping dimension $d_i$. The mapping operation uses the current value of the starting point and the mapping dimension $d_i$ to map the symbol $b_i$ and produce a directive $\pm\ x_i$. Specifically, the mapping begins from the starting point and moves along the mapping dimension toward the index of $b_i$ in this dimension. Both the number of cells passed and the direction with respect to the reference point (starting point) are compiled into a directive $-x$ or $+x$, where $x$ is the distance. The mapping operation updates the starting point to the new position before mapping any new symbol $b_j$.

Fig. 11 provides an example. Suppose that we want to map the symbol "m" to the mesh. Suppose that the current value of the starting point is ($h$ = 7, $v$ = 3) and the input symbol for the dimension selection operation is "a". Since the value of the symbol "a" in the unicode coding is 97, which is odd, the mapping dimension is the horizontal dimension. Therefore, the mapping operation starts from the starting point (7, 3) and moves along the horizontal dimension to the index of "m". Since the number of passed cells is 3 and the direction of the move is to the higher indexes, the mapping yields the directive "+3" as a mapping value for "m". The mapping operation updates the starting point to the new value (10, 3). To further illustrate the mapping operation, assume now that the symbol to be mapped is "m", the starting point is (7, 3), and the input symbol to the dimension selection is "b" (instead of "a"). Since the value of "b" in the unicode is 98, which is even, the mapping dimension is the vertical. Therefore, the mapping starts from the starting point (7, 3) and moves along the vertical dimension toward the index of "m". The number of passed cells is 4 and the direction of the move is toward the higher indexes. The mapping operation produces accordingly the directive "+4" as a result. The starting point is updated to (7, 7). As a final illustration, suppose that we want to map the symbol "e", where the input of the dimension selection is "a" and the starting point is (7, 3). Since the value of "a" is 97, which is odd, the mapping dimension is the horizontal. The mapping operation hence starts from the starting point (7, 3) and moves along the horizontal dimension to the position of "e". Since the distance of the move is 5 and is toward the lower indexes, the mapping operation produces the directive "–5" as a result of the mapping. The starting point is updated to (2, 3).

## VI. Mask Process

This mask process deeply alters its input block. The alteration includes (1) masking the symbols and (2) altering both the symbols and the block's structure. Fig. 12 shows the main operations of the mask process.

First, the mask process executes the *Manipulate* operation. This operation uses a sequence of the directives "$+d_1$ +

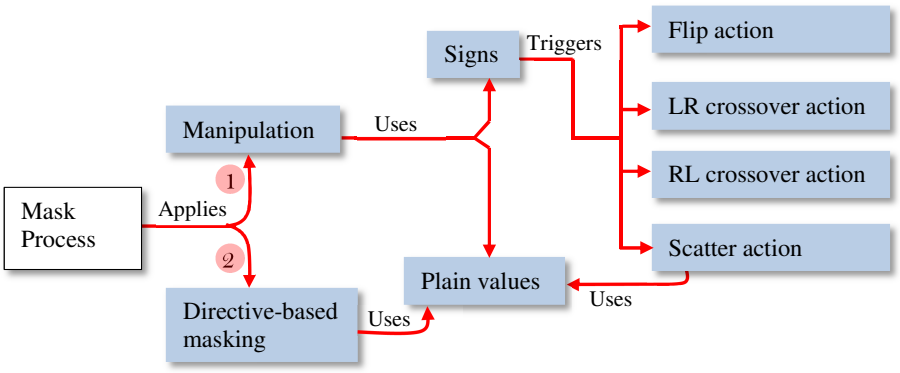
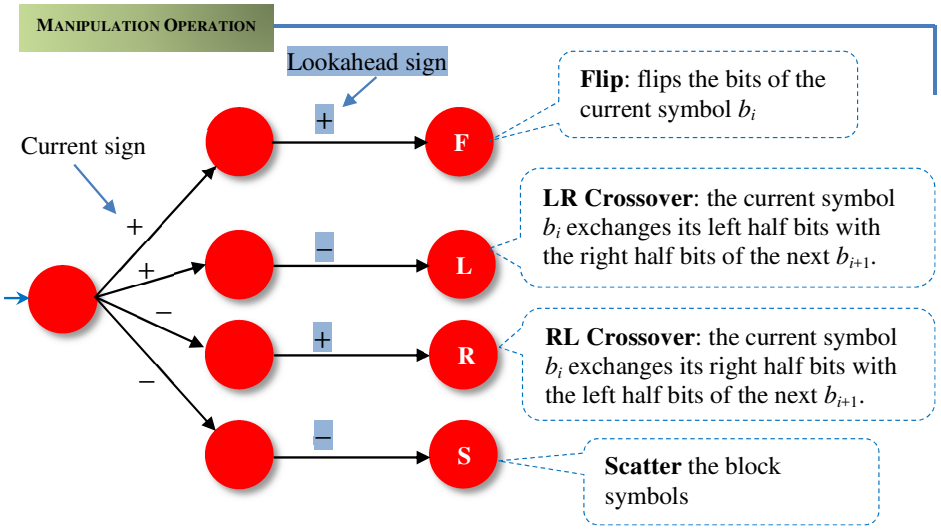



Fig. 12. the main actions of the mask operation.



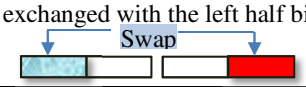
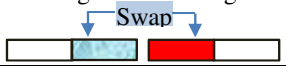

| Current sign $S_i$ | Lookahead sign $S_{i+1}$ | Action Performed on the block |
|---|---|---|
| + | – | **LR Crossover**: left half bits of the symbol $b_i$ is exchanged with the left half bits of $b_{i+1}$   |
| + | + | **Flip**: flip the bits of the current input symbol $b_i$ |
| – | + | **RL Crossover**: right half bits of the symbol $b_i$ is exchanged with the right half bits of $b_{i+1}$   |
| – | – | **Scatter**: move the symbol $b_i$ ahead $d_i$ positions, where $d_i$ is the plain directive. |

Fig. 13. the logic of the manipulate operation. The figure demonstrates how the sequences of signs trigger the manipulation actions.

$d_2 - d_3 ...$" to make two large effects on the block's symbols: breaking the structure of the block by reordering its symbols and altering the individual symbols by modifying or mixing some of their bits. To perform these effects, the *Manipulate* operation uses the signs of the directives to execute four different actions on the block. Fig. 12 shows these four actions: *Flip*, *LR* Crossover, *RL* Crossover, and *Scatter*. *Flip* action negates the bits of the symbol. *LR* Crossover action causes the current symbol $b_i$ to exchange the left half of its bits with the right half bits of the next symbol $b_{i+1}$. *RL* Crossover action causes the current symbol $b_i$ to exchange the right half of its bits with the left half bits of the next symbol $b_{i+1}$. The *Scatter* action repositions the symbol $b_i$ in the place specified by the directive $d_i$.

Fig. 13 shows the conditions under which these actions

are triggered. In particular, an action $X$ is triggered based on a specific pattern of the current sign and the lookahead sign. Referring to Fig. 13, when the current sign is "+" and lookahead sign is also "+", the *Flip* action is triggered and flips all the bits of the current input symbol $b_i$. When the current sign is "+" and the lookahead symbol is "–", the *LR* crossover action is triggered causing the left half bits of the current symbol $b_i$ to exchange with the right half bits of the next symbol $b_{i+1}$. When the current sign is "–" and the lookahead sign is "+", the *RL* crossover action is triggered causing the right half bits of the current symbol $b_i$ to exchange with the left half bits of the next symbol $b_{i+1}$. Finally, when the current sign is "–" and the lookahead sign is "–", the *Scatter* action is triggered causing the current symbol $b_i$ to move to the index specified by $d_i$.

It is clear that the former three actions (*Flip* and *LR*/*RL* Crossover) alter the individual symbols. While the latter action (*Scatter*) changes the structural relation between the symbols of the input block.

Second, the mask process introduces further masking to the block's symbols by executing the directive-based masking operation. This operation does the masking by embedding the effects of the plain directives. In particular, the plain directives $d_i$'s (without the sign) are XORed with the corresponding block symbols to yield a masked block $c_1c_2...c_n$. That is $c_i = b_i \oplus d_i$ for $i$ = 1, 2, ..., $n$.

We illustrate the mask process using an example. Suppose for the sake of simplicity the following 4–byte block "77, 61, 6C, 6C". Suppose further that the sequence of directives is "+30–7–18+3". Fig. 14 shows the steps of masking this block. The leftmost column represents the index of the current input symbol–where "0" represents is the index of the leftmost symbol in the input block. The second column shows the sequence of signs, where the currently considered sign pattern is shaded, and shows also the action performed (*LR*, *S*, etc.) based on the shaded sign pattern. The third column shows the input string (in Hex) and the input symbols which the corresponding action operates on. The rightmost column shows the sequence of directives. The manipulation process starts with the input symbol at index 0 ("77") and the first sign pattern "+ –". According to the logic in Fig. 13, this pattern triggers the *LR* Crossover action. The *LR* action therefore exchanges the left half bits of the current input symbol "77" with the right half bits of the next input symbol "61", yielding the new block "17, 67, 6C, 6C". Now, the second input symbol (at index 1) is "67" and the second sign pattern is "– –". This sign pattern triggers the *Scatter* action, which moves the second input symbol "67" to the position specified by the plain directive of the second directive "–7". That is, Scatter action moves "67" to the position 7 % 4 = 3. (We take the module % because the length of the input block is only four symbols.) Continuing likewise, the *Manipulate* operation produces the manipulated block "17, 6C, 66, 38". The manipulated block is further masked by XORing the symbols with the plain part of the directives sequence. (Note the plain directives 30, 7, 18, 3 are transformed to Hex: 1E, 07, 12, 03.) The final output of the masking operation is "09, 6B, 74, 3B".

The effect of the mask process is reversible provided that we have the directives sequence. To recover the original block, we define the inverse mask process. This process performs

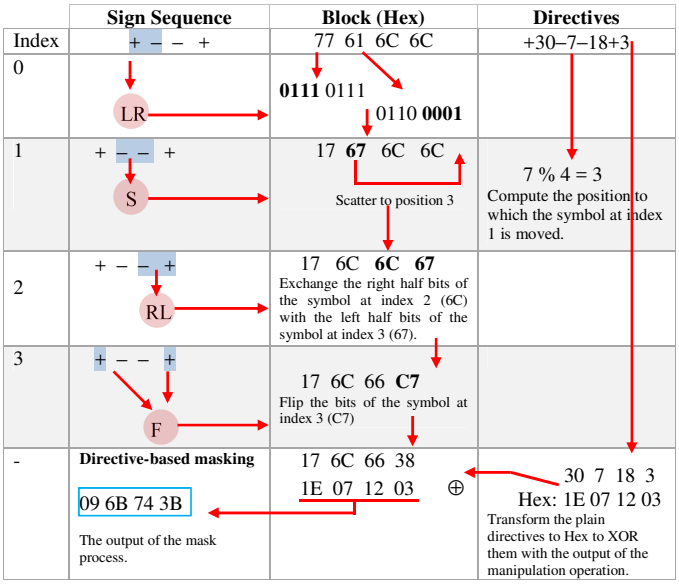



Fig. 14. an example of the mask process.

the same operations as the mask process, but in a reverse order. That is, the inverse mask process applies first the directive-based masking operation and then the *Manipulate* operation. The directive-based masking operation applies the XOR operation to the block and the plain directives $d_i$'s. The manipulation operation starts from the rightmost of the sign sequence and moves backwards (instead of starting from the leftmost). For instance, if the signs sequence is "+ – – + – –", the reverse operation starts from the rightmost pattern "– +", – –", "+ –", "– +", "– –", and "+ –". The *Flip*, *LR* Crossover, and *RL* Crossover actions do not change. The *Scatter* action functionality is changed. Instead of moving the symbol $b_i$ to the position $d_i$ % $n$, the action moves the symbol at the psition $d_i$ % $n$, to the current index $i$.

Fig. 15 shows an example of recovering the original block "77, 61, 6C, 6C" from the masked block "09, 6B, 74, 3B". The sequence of directives is "+30–7–18+3" ("+1E–07–12+03" in Hex.) As the figure shows, the masking process applies first the directive-based masking operation to the input block "09, 6B, 74, 3B", yielding "17, 6C, 66, 38". The manipulation operation is then applied starting from the last sign pattern "++" and from the end of the input block. The pattern "++" triggers the *Flip* action, which flips all of the bits of the last symbol "38", yielding the new block "17, 6C, 66, C7". The current symbol now is "66" (at index 2) and the sign pattern is "– +". This pattern triggers the *RL* crossover action, which causes the right half bits of the current input symbol "66" to be swapped with the left half bits of the following symbol "C7", yielding the new block "17, 6C, 6C, 67". The current symbol now is "6C" (at index 1). The sign pattern is "– –", which triggers the *Scatter* action. The symbol at the index 7 % 4 = 3, which is "67", is moved to the current index (position 1), yielding "17, 67, 6C, 6C". Finally, the sign pattern "+ –" triggers the *LR* crossover action, which causes the symbol at the current index 0 ("17") to exchange the left half of its bits with the right half bits of the next symbol at index 1 ("67"), yielding "77, 61, 6C, 6C"—the original block.

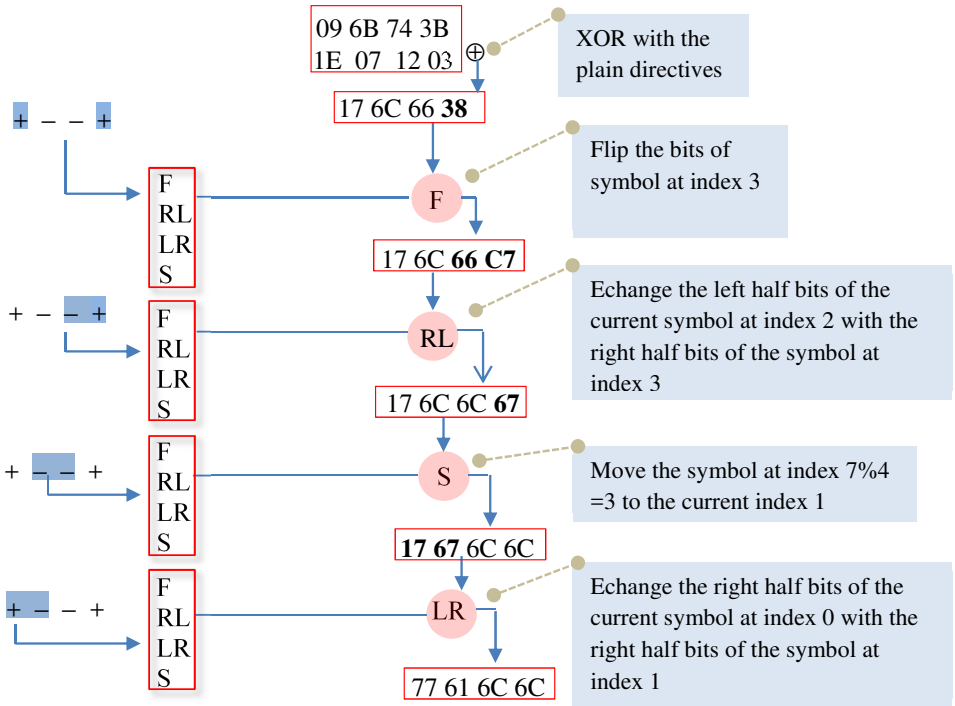



Fig. 15. example of recovering the plaintext block from the masked block.

## VII. The Cipher Technique

This section presents the technical details of the proposed block cipher. Subsection VII-A presents how to use the operations defined in the previous sections to encrypt blocks of plaintext. Subsection VII-B presents how to decrypt the resulting ciphertext.

### A. The Encryption Process

Let $B_i$ be plaintext block of size $n$ and *Key* be a key of size $m$, where $m$ is not necessarily equal to $n$. We impose no constraints on the size of the block. We also impose no specific constraints on the size of the key except those required for the key security. Therefore, a key size of 16 symbols or larger is highly recommended.

Initially, the encryption process prepares the D-Box and its inverse D-Box$^{-1}$ as described in section II. This step is necessary for the functionality of the encryption operations and the key expansion.

After the D-Box is created, the encryption process executes as Fig. 16 shows. The process encrypts the block $B_i$ using the *Key* in $t$ rounds. It first expands the *Key* to size of $t \times n$ symbols, where $n$ is the size of the plaintext block $B_i$ and $t$ is the number of rounds. Each round applies the diffuse and substitute operations to the input block $B_i$ as described in sections II and IV. The outcome is passed to the mask operation. The mask operation receives also as an input a sequence of $n$ directives obtained from the directive generator. In order for the directive generator to produce these directives, it receives two inputs: the previous plaintext block $B_{i-1}$ and sub-key of size $n$ symbols. The sub-key symbols are obtained from the expanded key (EK[x: y]). The directive generator maps the symbols of the previous block $B_{i-1}$ to the mesh using the sub-key as described in Section V.

The encryption of the first block is handled slightly differently. That is because the first block $B_1$ has no predecessor block $B_{i-1}$. In this case, we use the $n$ symbols of the *Key* instead of the block $B_{i-1}$. Therefore, the same $n$ symbols of the *Key* are used as an input for both the *Mapping* and the *Dimension selection* operations.

### B. The Decryption Process

The decryption process takes ciphertext and a key as an input and uses the key to produce the corresponding plaintext as an output. In order to accomplish this, the decryption process uses the inverse of the operations that are used during the encryption. The order in which the operations are applied is also reversed. Therefore, the decryption operations are executed in the following order: **Inverse Mask**, **Inverse Substitute**, and **Inverse Diffuse**. Fig. 17 shows the detailed steps of the decryption process. As the figure shows, the first operation to be applied to the input is the inverse mask operation instead of diffusion operation. The inverse substitute uses the D-Box$^{-1}$ to reverse the effect of the substitution operation. Finally, the inverse diffuse operation is applied to reverse the effect of the diffuse operation.

In addition as Fig. 17 shows, the expanded key is used backwards. That is, the decryption process starts from the end of the expanded key (with $t \times n$ symbols) instead of starting from the beginning. Therefore, the decryption process uses the first $n$ key symbols from the right then the second $n$ symbols, and so on until the last left $n$ symbols. The sequences of directives are also used backwards. As Fig. 17 shows, the last

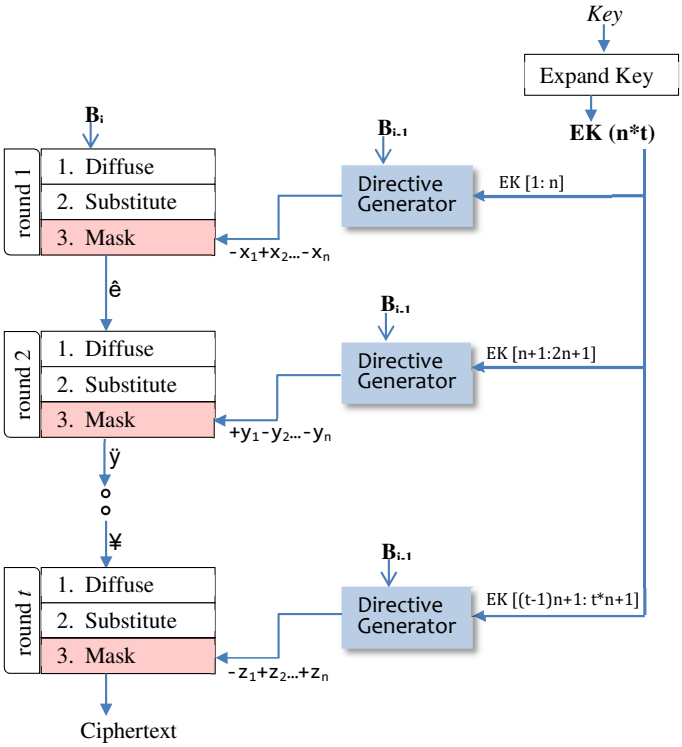



Fig. 16. the encryption process.

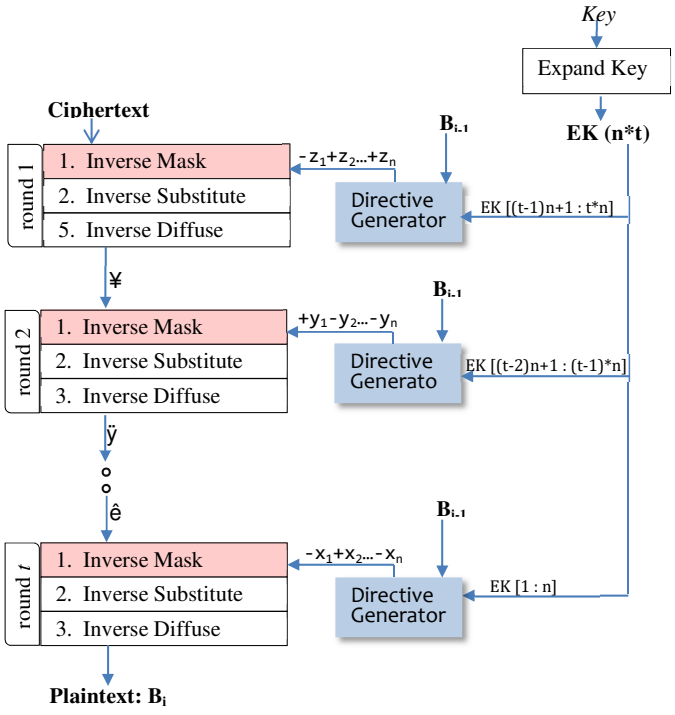



Fig. 17. the decryption process.

directive sequence in the encryption process "$-z_1+z_2...+z_n$" is used first.

## VIII. Related Work

The most related encryption technique to ours is the advanced encryption standard (*AES*) [3]. The *AES* encrypts a block of plaintext in many rounds. The number of rounds depends on the size of the key. In each round, *AES* applies different operations: key round, substitution using S-Box, row shifting, column mixing, and so on. The *AES* technique is NIST's standard [24]. Many researchers, however, reported major weaknesses. Specifically, the *AES* suffers from the weakness in the S-Box, making it more susceptible to computational attacks [5][19]. Several proposals have been suggested to improve the S-Box [6][20][21].These proposals claimed to improve the effectiveness of the S-Box. The effectiveness of the S-Box is further enhanced in [22]. As claimed in [22], the new enhancement increases the algebraic complexity of the S-Box and therefore it becomes more immune against differential and algebraic attacks. These extensions, however, have their inherent drawbacks as pointed out in [5].

Despite the importance of these improvements and the work around methods, we believe that strengthening the substitution part of the encryption algorithms (Mainly *AES*) cannot be properly done by merely adding more mathematical transformations. We think the strength of the substitution must be based on the key as proposed by the technique herein. In other words, instead of having static substitution table (S-Box), it is more effective to have a more dynamic substitution operation whose state must depend on the key and change according to the changes of the this key (D-Box).

Data encryption standard (DES) and its successive improvements such as Triple DES [9][10] are symmetric block ciphers that follow Feistel cipher structure [30]. Their encryption model depends on applying a set of substitutions and manipulation operations augmented with key evolution. Although this algorithm is used to be a standard, its security challenged [23][25]. Even with fundamental improvement such as using two keys, the algorithm still suffers and its security in question [23].

MARS, Blowfish, and Serpent [14][11][13] are symmetric block ciphers. Their encryption model depends like [9] on a set of substitutions and manipulation operations. Although their authors purport that these methods provide high levels of security, they failed to pass important randomness tests [27].

*LEX* encryption technique [16] is a stream cipher that is inspired by the one-pad encryption method (called also Vernam cipher). Although *LEX* claimed to have strong security properties, it uses the same round key repeatedly. This repeated use of the key makes it vulnerable to key-recovery attacks [28]. Our technique never uses the same key; each round uses a different version of the key and thus each block is encrypted using different keys.

Camellia encryption method [17] is a symmetric encryption technique, supporting128-bit block size and 128, 192, 256-bit key. Efficiency on both software and hardware platforms is a remarkable characteristic of Camellia in addition to its high level of security. The main important difference between our technique and Camellia is that ours has more effective diffusion operation along with a dynamic substitution box.

To conclude this section, it is worth mentioning that

our proposed encryption method possesses several important characteristics that make it unique. First, in contrast to all encryption methods, our technique implicitly uses the key in every encryption operation. Second, our technique uses the plaintext to introduce further confusion to the encryption. In particular, it embeds the effect of the previous plaintext block in the encryption of the next block. As such, changes in the previous block impact not only the encryption of the previous block itself but the encryption of the next block as well; a feature that–to the best of our knowledge–is unique to our cipher. Third, our technique triggers the manipulation actions (these actions belong to the masking process) based on the signs of the directives. Since the signs of the directives depend on both the key and the plaintext blocks, the sequence of the triggering highly relies on the plaintext blocks and keys and greatly sensitive to their changes. This makes predicting the triggering order is infeasible. Finally, the functionality of our substitution operation changes with the changes of the key. This adds additional confusion layer, making the substitution step much more effective than the static substitution used by other encryption techniques.

## IX. Performance Analysis

We present our performance analysis in this section. We first present an example that demonstrates some of our technique's features and then analytically discuss its security properties in Subsection IX-B. We present the empirical evaluation in Subsection IX-C.

### A. Encryption Example

We start our analysis by presenting examples of the technique's output (ciphertext). The examples are meant to be simple but indicative. For the sake of simplicity, we assume that the plaintext blocks and the keys are of length 16 symbols. Fig. 18 shows the ciphertexts for the corresponding plaintexts, keys, and previous blocks. Referring to the figure, one can see that the output changes drastically when the input slightly changes. Consider for instance the first five rows, which are encrypted using the same key. Although these plaintexts differ in a single bit, their respective ciphertexts are greatly different. Changing a single bit in the key causes the ciphertexts for the same plaintexts to be really different (e.g. compare between the ciphertexts in rows 6 and 7). Finally, changing the previous block highly impacts the resulting ciphertexts. For instance, a quick look at rows 1 and 9 shows that changing a single bit in the previous block causes large changes to the ciphertexts of the same plaintext block.

This property of the proposed technique is very important. From one hand, a tiny change to any of the input (plaintext, key) results in extremely large changes to the respective ciphertexts. From the other hand, a tiny change to the previous plaintext block also largely impacts the ciphertext of the following plaintext block. From the security prospective, the proposed technique causes the relationship between plaintext and ciphertext to be so complicated and untraceable.

### B. Analytical Performance Evaluation

The proposed technique has high confusion. The key is never explicitly used in the encryption; it is implicitly used via the directives. The key "trace" in the ciphertext is therefore so small to help predicting the used key. As such, the relationship between the key and the ciphertext is untraceable.

The proposed technique is highly sensitive to the changes of the input (see Fig. 18). This sensitivity provides an additional strong security guard. It makes the relationship between a key, ciphertext, and plaintext so complicated in a sense that if any of them changes, the respective ciphertext greatly changes. We attribute this high sensitivity to the changes of the input (plaintext or a key) to all the encryption operations (especially the diffuse and key expansion operations).

The state of the substitution box (D-Box) highly depends on the encryption key. Each different key results in a largely different new state and consequently results in different substitution outcomes. This means that changes of the key cause the same plaintext block to be substituted differently. This is in contract to the other encryption algorithms (especially *AES*), where the substitution of a block is independent of the key. That is, the outcome of substituting a block remains the same regardless of the used key.

The key expansion operation adopts a highly complicated computational model. From one hand, the substitution sub-operation relates the process of substituting the current symbol to the outcome of substituting the proceeding symbols. This means that the outcome of the substitution for the current symbol $k_i$ is impacted by all the previous symbols $k_j$ ($j$=1, 2, ..., $i$-1). On the other hand, the key expansion manipulation sub-operation is highly complicated process. It partially depends on the order of the key's symbols. For each new key, we have a new state (symbol order). In addition, the selection of the key's manipulation operations depends on a random process. In other words, although the order of the symbols may bias the selection toward the operation with the higher ratio, the selection of the operation to be applied is nevertheless random. Even more, the random process is always seeded with latest version of the key. These properties assure that the key manipulation operations have a greatly complicated functional behavior in a sense that the number of states is tremendously large.

The mask process has perhaps the highest impact on its input block. It makes deep changes to both individual symbols and to the structure of the block, largely diverging the resulting block from the input block. The functionality of this operation fully depends on directive sequences. The manipulation operation, which modifies the individual symbols and the structure of the block, depends on the pattern of the signs (of the directives). Since, this pattern depends on both the key and the previous block, changes to either the key or the previous block certainly results in a different pattern (regardless of the magnitude of the change) and consequently in a different functional behavior. Further the directive-based masking operation, which embeds the effect of the plain directive (value without the sign), depends also on the key and the previous block and changes according to their changes. As such, a change to the key or the previous block creates different modifications to the output of the masking operation.

| **Previous block**: 0000000000000000 | | |
|---|---|---|
| **Key** | **Plaintext** | **Ciphertext** |
| 0000000000000000 | 0000000000000000 | 59 e0 5a 42 96 f4 13 6d 46 c9 c3 c7 5d 27 70 57 |
| 0000000000000000 | **1**000000000000000 | a6 9d c4 12 55 c2 57 da 04 ba b5 62 f4 49 b0 cb |
| 0000000000000000 | 000000000000000**1** | 8e a2 13 87 55 d7 8f dc 13 ab 66 c1 97 76 65 4a |
| 0000000000000000 | 00000000**1**000000 | 67 44 cc ec e5 e4 3f 2b f8 84 9f ac 11 30 25 8a |
| 0000000000000000 | **1**00000000**1**000000 | f8 29 21 d8 33 49 2f f0 3b 57 74 e1 20 08 a3 32 |
| **1**000000000000000 | 0000000000000000 | 44 28 e3 80 4e 0b fb 20 98 70 1b d2 59 4b 57 fe |
| 000000000000000**1** | 0000000000000000 | 6d 42 3a aa ae c0 d8 91 4c 48 d6 65 d3 31 cb 22 |
| **Previous block**: **1**000000000000000 | | |
| 0000000000000000 | 0000000000000000 | 9f 66 a2 3e 34 f5 d4 96 ee 40 5d 92 89 ca c3 5c |
| **Previous block**: 000000000000000**1** | | |
| 0000000000000000 | 0000000000000000 | 71 32 c7 77 84 96 19 9f 12 3e 82 5b 55 77 fe 76 |
| **Previous block**: 0000000000000001 | | |
| ACF98IFTRmk90AGT | To beornotto be! | 5b 8c 80 61 77 cc 3a 81 b7 41 3c 39 6a 60 35 d8 |
| ACF98IF**S**Rmk90AGT | To beornotto be! | f3 99 d0 49 68 b8 4e 4c 7a 88 a8 e8 d2 7d 25 ae |
| ACF98IFSRmk90AGT | To beornotto be**.** | 1b db 47 a4 3f 76 03 50 b3 7 95 0e 19 1b d0 9f |

Fig. 18. The proposed technique output examples.

### C. Empirical Evaluation

We tested the performance our encryption technique according to the testing rules specified by the national institute for standards and technology–NIST [27]. We specifically prepared the testing data as specified in [26].

*1) Randomness Statistical Tests:* We used the following tests to evaluate the randomness properties of our technique [27].

- *Runs test*: determines whether the number of runs of ones and zeros of various lengths is as expected for a random sequence.
- *Frequency Test* (*Monobit*): determines whether the number of ones and zeros in a sequence are approximately the same as would be expected for a truly random sequence.
- *Discrete Fourier Transform Test* (*Spectral*): detects periodic features (i.e. repetitive patterns that are near each other) in the tested sequence that would indicate a deviation from the assumption of randomness.

*2) Randomness Hypotheses:* We have two hypotheses to test:

- $H_0$ (*Null*): the output of the encryption technique is random.
- $H_1$ (*Alternative*): the output of the encryption technique is not random.

Accepting $H_0$ or $H_1$ depends on a computed value called p-value and a specified value called the significance level $\alpha$. The p-value is computed by the applied statistical test based on an input sequence. The significance level $\alpha$ is specified by the tester (e.g. 0.00001, 0.001, 0.01, 0.05 are typical values for $\alpha$). In particular if p-value $\geq \alpha$, $H_0$ is accepted ($H_1$ is rejected); otherwise $H_0$ is rejected ($H_1$ is accepted).

*3) Test Data:* The testing data were prepared according to [26]. Without losing the generality, we confine the D-Box to be 16×16. This allows for representing each symbol by 8 bits. We used the following sets of data to test our encryption technique.

1) **Key Avalanche Test**. The objective of this data set is to examine the sensitivity of our algorithm to changes in the key.
2) **Plaintext Avalanche Test**. The objective of this data set is to examine the sensitivity of our algorithm to changes in the plaintext.
3) **Plaintext/Ciphertext Correlation**. The objective of this data set is to study the correlation between plaintext–ciphertext pairs.

Firstly, to study the sensitively of our algorithm to the key change, we created and analyzed 456 sequences of size 32,768 bits each. We used a 256-bit (32 bytes) plaintext of all zeros and 456 random keys each of size 128 bits (16 bytes). Each sequence was created by concatenating 128 derived blocks constructed as follows. Each derived block is created by XORing the ciphertext created using the fixed plaintext and the 128-bit key with the ciphertext created using the fixed plaintext and the perturbed random 128-bit key with the $i^{th}$ bit changed, for $1 \leq i \leq 128$.

Secondly, to analyze the sensitivity to the plaintext changes, we created and analyzed 456 sequences of size 32,768 bits each. We used 456 random plaintexts of size 256 bits (32 bytes) and a fixed 128-bit key of all zeros. Each sequence was created by concatenating 128 derived blocks constructed as follows. Each derived block is created by XORing the ciphertext created using the 128-bit key and the 256-bit plaintext with the ciphertext created using the 128-bit key and the perturbed random 256-bit plaintext with the $i^{th}$ bit changed, for $1 \leq i \leq 256$.

Thirdly, to study the correlation of plaintext–ciphertext pairs, we constructed 456 sequences of size 115,712 bits per a sequence. Each sequence is created as follows. Given a random 128-bit key and 452 random plaintext blocks (the

TABLE II. KEY AVALANCHE TEST RESULTS

| Randomness Test | Successes | Failures | Success Rate | Upper limit of CI (0.05) |
|---|---|---|---|---|
| Runs test | 452 | 4 | 99.1% | 36.76 |
| Monobit test | 445 | 11 | 97.6% | 36.76 |
| Spectral test | 424 | 32 | 93.0% | 36.76 |

TABLE III. PLAINTEXT AVALANCHE TEST RESULTS

| Randomness Test | Successes | Failures | Success Rate | Upper limit of CI (0.05) |
|---|---|---|---|---|
| Runs test | 448 | 9 | 98.2% | 36.76 |
| Monobit test | 442 | 14 | 96.9% | 36.766 |
| Spectral test | 418 | **38** | 91.7% | 36.76 |

TABLE IV. PLAINTEXT/CIPHERTEXT CORRELATION TEST RESULTS

| Randomness Test | Successes | Failures | Success Rate | Upper limit of CI (0.05) |
|---|---|---|---|---|
| Runs test | 447 | 9 | 98.0% | 36.76 |
| Monobit test | 443 | 13 | 97.1% | 36.76 |
| Spectral test | 421 | **35** | 92.3% | 36.76 |

block's size is 256 bits), a binary sequence was constructed by concatenating 452 derived blocks. A derived block is created by XORing the plaintext block and its corresponding ciphertext block. Using the 452 (previously selected) plaintext blocks, the process is repeated 455 times (one time for every additional 128-bit key).

During the encryption, we set the number of rounds *t* to 8 rounds for each block because this number gives high performance. (Please see appendix A for more detailed discussion on the number of rounds *t*.)

Tables II, III, and IV show the results of the applied randomness test to the above data sets. The tables show the applied tests, the number of sequences that passed the respective test, the number of failed sequences, and the rate of success. For each test, the significance level was fixed at 0.05, which implies that, ideally, no more than five out of hundred binary sequences may fail the corresponding test. However, in all likelihood, any given data set will deviate from this ideal case. A more realistic interpretation is to use a confidence interval (CI) for the proportion of binary sequences that may fail at the 0.05. The rightmost column shows the maximum number of binary sequences that are expected to fail the corresponding test. For instance, a maximum of 36.76 (or 36) binary sequences are expected to fail each of the three tests.[2]

All the three tables show that the success rate is remarkably high. In fact, 99.1% of the sequences for testing the key avalanche (Table II) passed the Runs test, 97.6% passed the Monobit test, and 93.0% of the sequences passed the Spectral test. Tables III and IV lead to the same conclusions. Additionally, the rightmost column, which specifies the highest number of sequences that are expected to fail the test, shows that number of sequences that failed the corresponding randomness tests is fewer than the expected number at significance level of 0.05. Table III shows, however, that the number of the sequences that failed Spectral test is actually greater than the highest number of sequences that are expected to fail (actual 38, expected 36). Also, in Table IV, the number of the sequences that failed Spectral test is 35, which is very close the maximum number of the sequences that are expected to fail.

[2]The maximum number of binary sequences that are expected to fail at the level of significance $\alpha$ is computed using the following formula [29]: $S(\alpha + 3 \times \sqrt{\frac{\alpha \times (1-\alpha)}{S}})$, where S is the total number of sequences and $\alpha$ is the level of significance.

## X. CONCLUSIONS AND FUTURE WORK

We proposed a full-fledged encryption technique. This technique uses effective operations to manipulate blocks of plaintext and create deep changes to both the structure of the block and the individual symbols. By virtue of these operations, the technique ensures high diffusion, confusion, and avalanche effect. The performance numbers in subsection 9.3 clearly indicate the high avalanche effect. The technique has passed three NIST–recommended randomness tests for evaluating the effeteness of encryption technique. In fact, the least percentage of the sequences that passed these tests was 91.7%.

We have two main objectives for future work. First, we want to replace the built-in random generator, which is used in the key expansion operation, with a more effective random generator. Second, we plan also to apply more randomness tests on a larger set of plaintexts and keys.

## REFERENCES

[1] M. J. Al-Muhammed and R. Abuzitar, $\kappa$–Lookback Random-Based Text Encryption Technique, *Journal of King Saud University-Computer and Information Sciences*, 2017. https://doi.org/10.1016/j.jksuci.2017.10.002

[2] M. J. Al-Muhammed and R. Abuzitar, Dynamic Text Encryption, *International Journal of Security and its Applications* (*IJSIA*), Vol. 11, No. 11, pp. 13–30, November 2017.

[3] J. Daemen and V. Rijmen, The Design of RIJNDAEL: *AESThe Advanced Encryption Standard*, Springer, Berlin, Germany, 2002.

[4] A. Abusukhon, Z. Mohammad, and M.Talib, A Novel Network Security Algorithm based on Encryption Text into a White-Page Image, Proceedings of the World Congress on Engineering and Computer Science, Vol. I WCECS 2016, October 19-21, 2016, San Francisco, USA.

[5] N. A. Azam, A Novel Fuzzy Encryption Technique Based on Multiple Right Translated AES Gray S-Boxes and Phase Embedding, *Security and Communication Networks*, Vol. 2017, 9 pages. https://doi.org/10.1155/2017/5790189

[6] J. Liu, B. Wei, X. Cheng, and X. Wang, An AES S-box to Increase Complexity and Cryptographic Analysis. In *Proceedings of the 19$^{th}$ International Conference on Advanced Information Networking and Applications* (*AINA*'05), pp. 724–728, Taipei, Taiwan, March 2005.

[7] D. Nilesh and M/ Nagle, The New Cryptography Algorithm with High Throughput, 2014 *International Conference on Computer Communication and Informatics* (2014), October 2014.

[8] B. Modi and V. Gupta, A Novel Security Mechanism in Symmetric Cryptography Using MRGA. In: Sa P., Sahoo M., Murugappan M., Wu Y., Majhi B. (eds) *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications. Advances in Intelligent Systems and Computing*, vol 719. Springer, Singapore, 2018

[9] S.-J. Han, the Improved Data Encryption Standard (DES) Algorithm. In *Proceedings of IEEE* $4^{th}$ *International Symposium on Spread Spectrum Techniques and Applications*, pp. 1310–1314, Germany, 1996.

[10] J. Verma and S. Prasad, Security Enhancement in Data Encryption Standard. In: Prasad S.K., Routray S., Khurana R., Sahni S. (eds) *Information Systems, Technology and Management* (*ICISTM*) 2009. Communications in Computer and Information Science, Vol. 31. Springer, Berlin, Heidelberg, 2009.

[11] T. Nie and T. Zhang, A Study of DES and Blowfish Encryption Algorithm, In P*roceedings of IEEE Region* $10^{th}$ *Conference*, Singapore, Jan. 2009.

[12] P. Patil, P. Narayankar, D.G Narayan, S. M. Meena, A Comprehensive Evaluation of Cryptographic Algorithms: DES, 3DES, AES, RSA and Blowfish, Procedia Computer Science, Vol. 78, pp. 617-624, 2016.

[13] R. Anderson, E. Biham and L. Knudsen: Serpent: A Proposal for the Advanced Encryption Standard. Available at http://www.cl.cam.ac.uk/ rja14/Papers/serpent.pdf

[14] C. Burwick, D. Coppersmith, E. D'Avignon, R. Gennaro, S. Halevi, C. Jutla, and N. Zunic, The MARS Encryption Algorithm, IBM, August 1999.

[15] S. A. Y. Hunn, S. Z. binti . Naziri, and N. B. Idris, The Development of Tiny Encryption Algorithm (TEA) Crypto-core for Mobile Systems, 2012 *IEEE International Conference on Electronics Design, Systems and Applications* (*ICEDSA*), Kuala Lumpur, Malaysia, Nov. 2012.

[16] A. Biryukov. Design of a new stream cipherLEX. In M. Robshaw and O. Billet (Eds.): New Stream Cipher Designs, LNCS 4986, Springer–Verlag Berlin Heidelberg, pp. 48–56, 2008.

[17] K. Aoki, T. Ichikawa , M. Kanda , M. Matsui , S. Moriai , J. Nakajima , T. Tokita, Camellia: A 128–Bit Block Cipher Suitable for Multiple Platforms-Design and Analysis, In *Proceedings of the* $7^{th}$ *Annual International Workshop on Selected Areas in Cryptography*, pp.39–56, August 14–15, 2000.

[18] M. F. Mushtaq, S. Jamel, A. H. Disina, Z. A. Pindar, N. S. A. Shakir, and M. M. Deris, A Survey on the Cryptographic Encryption Algorithms, International Journal of Advanced Computer Science and Applications (IJACSA) , Vol. 8, No. 11, pp. 333–344, 2017.

[19] J. Rosenthal, A Polynomial Description of the Rijndael Advanced Encryption Standard, J*ournal of Algebra and Its Applications*, Vol. 2, No. 2, pp. 223–236, 2003.

[20] L. Jingmei, W. Baodian, and W. Xinmei, One AES S-box to Increase Complexity and its Cryptanalysis, *Journal of Systems Engineering and Electronics*, Vol. 18, No. 2, pp. 427–433, 2007.

[21] L.Cui and Y. Cao, A New S-box Structure Named Affine-Power Affine, *International Journal of Innovative Computing, Information and Control*, Vol. 3, No. 3, pp. 751–759, 2007.

[22] M. Khan and N. A. Azam, Right Translated AES Gray S-boxes, *Security and Communication Networks*, Vol. 8, No. 9, pp. 1627–1635, 2015.

[23] C. J. Mitchell, On the Security of 2–Key Triple DES, *IEEE Transactions on Information Theory*, Vol. 62, No. 11, pp. 6260–6267, Nov. 2016.

[24] NIST Special Publication 800–67 Recommendation for the Triple Data Encryption Algorithm (TDEA) Block Cipher Revision 1, Gaithersburg, MD, USA, Jan. 2012.

[25] S. Lucks, Attacking Triple Encryption, In *Proceedings of the* $5^{th}$ *International Workshop Fast Software Encryption* (*FSE*), pp. 239–253, Mar. 1998.

[26] Jr. Juan Soto. Randomness Testing of the AES Candidate Algorithms. http://csrc.nist.gov/archive/aes/round1/r1-rand.pdf, Accessed December 2017.

[27] Rukhin, A., Soto, J., Nechvatal, J., Smid, M., Barker, E., Leigh, S., Levenson, M., Vangel, M., Banks, D., Heckert, A., Dray, J., and Vo, S. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. NIST special publication 800–22, National Institute of Standards and Technology (NIST), Gaithersburg, MD, 2001.

[28] A. Bogdanov, F. Mendel, F. Regazzoni, V. Rijmen, and E. Tischhauser, ALE: AES-Based Lightweight Authenticated Encryption. In: S. Moriai (ed.) FSE 2013. LNCS, Vol. 8424, pp. 447–466. Springer, Heidelberg, 2014.

[29] J. Soto, Randomness Testing of the Advanced Encryption Standard Candidate Algorithms, NIST IR 6390, September 1999.

[30] W. Stallings, *Cryptography and Network Security*: *Principles and Practice*, $7^{th}$ edition, Pearson, 2016.

## APPENDIX A
## ADDITIONAL DATA

### A. Key-Based Numbers

We describe in this section how we generate the integers $I_i$ for reordering the content of D-Box. The process involves two steps.

*1) **STEP I**:* In this step, the encryption key is processed using the key handling procedure proposed by *AES* (advanced encryption standard). The *AES* key handling algorithm takes as input a 4-word (16-byte) key and produces a linear array of 44 words (176 bytes). The key is copied into the first four words of the expanded key. The remainder of the expanded key is filled in four words at a time. Each added word *w*[*i*] depends on the immediately preceding word, *w*[*i*-1], and the word four positions back, *w*[*i*-4]. In three out of four cases, a simple XOR is used. For a word whose position in the w array is a multiple of 4, a more complex function is used. Fig. 19 illustrates the generation of the first eight words of the expanded key, using the symbol *g* to represent the complex function.

The function *g* consists of the following sub-functions:

1) **RotWord** performs a one-byte circular left shift on a word. This means that an input word $[b_0, b_1, b_2, b_3]$ is transformed into $[b_1, b_2, b_3, b_0]$.
2) **SubWord** performs a byte substitution on each byte of its input word, using the static substitution box (or Sbox)—See Fig. 20.
3) The result of steps 1 and 2 is XORed with a round constant, Rcon [*j*].

The round constant is a word in which the three rightmost bytes are always 0. Thus the effect of an XOR of a word with Rcon is to only perform an XOR on the leftmost byte of the word. The round constant is different for each round and is defined as:

```
Rcon[j] = (RC[j],0,0,0),
where RC[j] = 2 . RC[j-1] and RC[1] = 1
```

The multiplication is defined over the field $GF(2^8)$. The values of RC[j] in hexadecimal are defined in Table V.

*2) **STEP II**:* In this step, the last 32 bytes (out of 176) of the sequence generated by Step I are used as a seed for the key-based number generator, which is described in [1]. Fig. 21 shows the algorithmic steps of the generator.

The key-based numbers are generated using the steps (2)–(5). As the figure shows, the symbols of the seed are summed

TABLE V. THE VALUES OF RC[J]

| j | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| RC[j] | 01 | 02 | 04 | 08 | 10 | 20 | 40 | 80 | 1B | 36 |

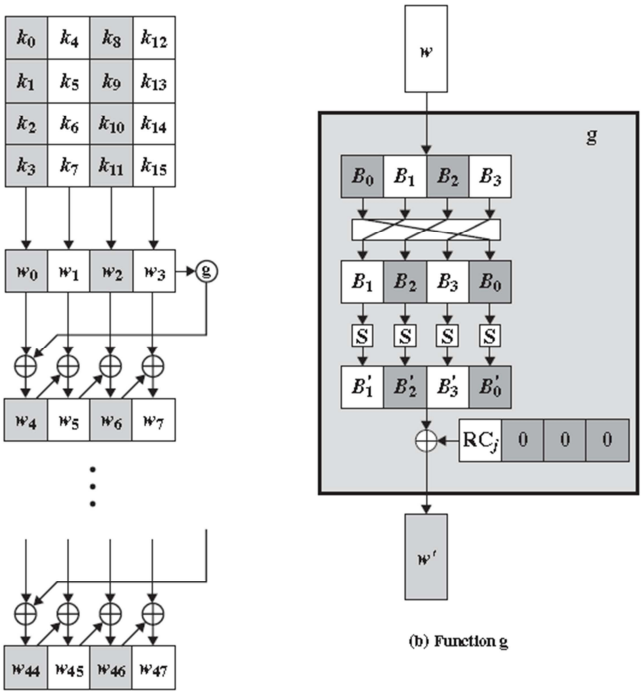



Fig. 19. the AES key handling procedure.

by multiplying the integer value of each seed's symbol $k_i$ by its position in the seed (step 2). The sum is then circularly left shifted *n* positions to yield *Sh* (*n* is the value of the first two symbols in the seed) in step 3 and *Sh* is XORed with the Sum in step 4 to yield *Ls*.

Assuming the number of symbols in the D-Box is *N*, the number *Ls* is therefore adjusted by taking the module *N*. The seed is updated in step (5) by concatenating the seed with the current sum. The seed hence grows after each iteration. If, at any iteration, step (2) results in overflow in the sum, the procedure reduces the seed to 32 symbols by using the middle 32 unicode symbols as a new value for the seed. Steps (2) through (5) repeat until the condition no longer holds. (The condition determines the desired number of values to be generated.)

It is worth mentioning that the numbers created by the generator in Fig. 21 are random. The proof of the randomness properties of the generator in Fig. 21 is beyond the scope of this paper and can be found elsewhere [1].

### B. Number of Rounds (t)

The number of the encryption rounds *t*, applied to each plaintext block, is very important and has a really large impact on the randomness properties of the output (ciphertext). More rounds (large values of *t*) result in higher diffusion and confusion (random output). More rounds, however, increase the time for encrypting a block. Thus, we look for a value for *t* that achieves two objectives: (1) high randomness in the output and (2) short execution time.

| x \ y | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | a | b | c | d | e | f |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 63 | 7c | 77 | 7b | f2 | 6b | 6f | c5 | 30 | 01 | 67 | 2b | fe | d7 | ab | 76 |
| 1 | ca | 82 | c9 | 7d | fa | 59 | 47 | f0 | ad | d4 | a2 | af | 9c | a4 | 72 | c0 |
| 2 | b7 | fd | 93 | 26 | 36 | 3f | f7 | cc | 34 | a5 | e5 | f1 | 71 | d8 | 31 | 15 |
| 3 | 04 | c7 | 23 | c3 | 18 | 96 | 05 | 9a | 07 | 12 | 80 | e2 | eb | 27 | b2 | 75 |
| 4 | 09 | 83 | 2c | 1a | 1b | 6e | 5a | a0 | 52 | 3b | d6 | b3 | 29 | e3 | 2f | 84 |
| 5 | 53 | d1 | 00 | ed | 20 | fc | b1 | 5b | 6a | cb | be | 39 | 4a | 4c | 58 | cf |
| 6 | d0 | ef | aa | fb | 43 | 4d | 33 | 85 | 45 | f9 | 02 | 7f | 50 | 3c | 9f | a8 |
| 7 | 51 | a3 | 40 | 8f | 92 | 9d | 38 | f5 | bc | b6 | da | 21 | 10 | ff | f3 | d2 |
| 8 | cd | 0c | 13 | ec | 5f | 97 | 44 | 17 | c4 | a7 | 7e | 3d | 64 | 5d | 19 | 73 |
| 9 | 60 | 81 | 4f | dc | 22 | 2a | 90 | 88 | 46 | ee | b8 | 14 | de | 5e | 0b | db |
| a | e0 | 32 | 3a | 0a | 49 | 06 | 24 | 5c | c2 | d3 | ac | 62 | 91 | 95 | e4 | 79 |
| b | e7 | c8 | 37 | 6d | 8d | d5 | 4e | a9 | 6c | 56 | f4 | ea | 65 | 7a | ae | 08 |
| c | ba | 78 | 25 | 2e | 1c | a6 | b4 | c6 | e8 | dd | 74 | 1f | 4b | bd | 8b | 8a |
| d | 70 | 3e | b5 | 66 | 48 | 03 | f6 | 0e | 61 | 35 | 57 | b9 | 86 | c1 | 1d | 9e |
| e | e1 | f8 | 98 | 11 | 69 | d9 | 8e | 94 | 9b | 1e | 87 | e9 | ce | 55 | 28 | df |
| f | 8c | a1 | 89 | 0d | bf | e6 | 42 | 68 | 41 | 99 | 2d | 0f | b0 | 54 | bb | 16 |

Fig. 20. the AES SBox.

(1) *Seed* = *key*
**Repeat**
(2) $Sum = \sum_{i=1}^{|seed|} i \times k_i$
(3) *Sh* = *LShift* (*sum*, *n*)
(4) *Ls* = (*Sum* ⊕ *Sh*) % *N*
(5) *Seed* = *concatenate* (*Seed*, *Sum*)
**Until** condition

Fig. 21. the algorithmic steps for generating key-based numbers.

To find the best number of rounds, we conducted many experiments, where we gradually increased the number of rounds from 1 to 10. We used in our experiments the same plaintexts in Section IX. For each value of *t*, we applied our encryption technique to 200 sequences each is of size 32,768 bits. We then subjected the resulting ciphertexts to three randomness tests (Runs, Monobit, and Spectral tests).

Table VI shows the results of the experiments. The results are presented in terms number of rounds *t*, the number of the sequences that passed the corresponding randomness test, and the rate of success. As the numbers show, there is a significant improvement in the success rate as *t* increases from 1 to 8. When *t* is greater than 8, the improvement in the success rate becomes slightly small. For instance, increasing the number of rounds from 8 to 9 slightly improves the rate of success (especially for Spectral), but this slight improvement causes really large increase in the time (see Fig. 22).

Fig. 22 plots the time in milliseconds as a function of the number of rounds. The time linearly increases as the number of rounds increases. This increase in the time appears to be significant. Considering both Table VI and Fig. 22, one can conclude that the amount of the improvement in the rate of success justifies the time increase up to *t*= 8. When *t* is larger than 8, the improvement in the success rate does not really justify the incurred time overhead.[3]

[3]The purpose of presenting the time required for encrypting each sequence is not to show the time performance of our technique. The implementation is only a proof of concept and we did not optimize for time performance. We only mention the time to assert that increasing the number of the rounds increases the success rate, but this increase also entails larger time overhead.

TABLE VI. CHANGES IN THE RANDOMNESS RATE AS THE NUMBER ROUNDS INCREASES

| No. of Sequences | Rounds (t) | Test | Successes | Success Rate (%) |
|---|---|---|---|---|
| 200 | 1 | Runs Test | 100 | 50% |
| | | Monobit | 25 | 22.5% |
| | | Spectral | 0 | 0% |
| 200 | 2 | Runs Test | 119 | 59.5% |
| | | Monobit | 97 | 48.5% |
| | | Spectral | 1 | 0.5% |
| 200 | 3 | Runs Test | 122 | 61% |
| | | Monobit | 96 | 48% |
| | | Spectral | 3 | 1.5% |
| 200 | 4 | Runs Test | 156 | 78% |
| | | Monobit | 133 | 66.5% |
| | | Spectral | 3 | 1.5% |
| 200 | 5 | Runs Test | 183 | 91.5% |
| | | Monobit | 167 | 83.5% |
| | | Spectral | 58 | 29% |
| 200 | 6 | Runs Test | 188 | 94% |
| | | Monobit | 171 | 85.5% |
| | | Spectral | 103 | 51.5% |
| 200 | 7 | Runs Test | 192 | 96% |
| | | Monobit | 182 | 91% |
| | | Spectral | 147 | 73.5% |
| 200 | 8 | Runs Test | 199 | 99.5% |
| | | Monobit | 195 | 97.5% |
| | | Spectral | 186 | 93% |
| 200 | 9 | Runs Test | 200 | 100% |
| | | Monobit | 195 | 97.5% |
| | | Spectral | 189 | 94.5% |
| 200 | 10 | Runs Test | 200 | 100% |
| | | Monobit | 197 | 98.5% |
| | | Spectral | 188 | 94% |

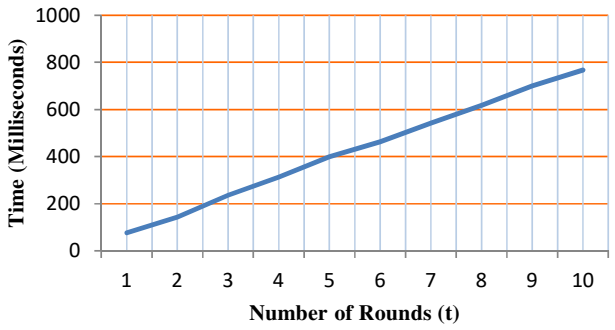



Fig. 22. The time increase as a function of the number of rounds.

# Duplicates Detection Within Incomplete Data Sets Using Blocking and Dynamic Sorting Key Methods

Abdulrazzak Ali[1]
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76300, Ayer Keroh, Melaka, Malaysia

Nurul A. Emran[2]
Computational Intelligence Technologies (CIT) Research Group
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76300, Ayer Keroh, Melaka, Malaysia

Siti A. Asmai[3]
Optimization, Modeling, Analysis,
Simulation and Scheduling (OptiMASS) Research Group
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76300, Ayer Keroh, Melaka, Malaysia

Awsan Thabet[4]
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76300, Ayer Keroh, Melaka, Malaysia

***Abstract*—In database records duplicate detection, blocking method is commonly used to reduce the number of comparisons between the candidate record pairs. The main procedure in this method requires selecting attributes that will be used as sorting keys. Selection accuracy is essential in clustering candidates records that are likely matched in the same block. Nevertheless, the presence of missing values affects the creation of sorting keys and this is particularly undesirable if it involves the attributes that are used as the sorting keys. This is because, consequently, records that are supposed to be included in the duplicate detection procedure will be excluded from being examined. Thus, in this paper, we propose a method that can deal with the impact of missing values by using a dynamic sorting key. Dynamic sorting is an extension of blocking method that essentially works on two functions namely uniqueness calculation function (UF) (to choose unique attributes) and completeness function (CF) (to search for missing values). We experimented a particular blocking method called as sorted neighborhood with a dynamic sorting key on a restaurant data set (that consists of duplicate records) obtained from earlier research in order to evaluate the method's accuracy and speed. Hypothetical missing values were applied to testing data set used in the experiment, where we compare the results of duplicate detection with (and without) dynamic sorting key. The result shows that, even though missing values are present, there is a promising improvement in the partitioning of duplicate records in the same block.**

***Keywords*—Duplicate detection; Incomplete Data Set; Blocking Methods; Sorting key; Attribute Selection**

## I. Introduction

Duplication detection is a crucial process in many data cleaning operations. Nevertheless, this process relies on time-consuming attributes comparison between the records pairs that have been reported as a common bottleneck in duplication detection [1]. To address this problem, the records are normally partitioned into small subsets so that searching for the duplicates is only performed within the small subsets. Common blocking methods that are used to partition the records into blocks (or also called as windows) are standard blocking and sorted neighborhood. These methods work well in big data sets that depend on candidate keys to sort the data set [2]. Blocking methods aims at producing a set of blocks that offers a good balance between the number of detected duplicates and the number of required comparisons. Selection of a suitable blocking method for duplication detection usually depends on domain knowledge [3]. In the context of homogeneous information spaces, these methods typically consider the frequency distribution of the values of attribute names as well as their quality (i.e., the presence of noise or missing values) to derive the most suitable blocking key(s) [4]. In order to obtain an unambiguous sorting order, it is desirable that the sorting keys are unique. In fact, within a data set, candidate key attribute (such as name) can also be used as sorting key rather than using attributes that are commonly defined to partition the data set (such as zip code) [1]. As data uniqueness is a crucial aspect in duplication detection, attributes that contain errors or missing values will hinder the creation of sorting key [5]. In addition, attributes that consists of high repetition ratio such as gender (with only two value states) are less useful as sorting key. Nevertheless, attribute such as surname can be a useful sorting key but it may frequently be reported or keyed incorrectly [4]. The sorting key with uniqueness criteria less repetition ratio will contribute in reducing the unnecessary comparisons of pairs of candidate records in duplicate detection process.

Duplicate detection is usually evaluated in terms of its accuracy. Thus, it relies on how well the blocking method deals with errors. Two kinds of error are of concerned in blocking methods which are:

1) False positive or false acceptance: This error occurs when the actual 'unmatched' records in the candidate pair space are included for comparison. This will cause unnecessary time spent for comparison.
2) (False negative or false rejection: This error occurs when the actual 'matched' records in the candidate pair space are excluded for comparison. This will cause duplicate records to remain in the data set and causes worse consequence than the false positive case.

As described by Christen (2007), the process of dividing the data set into blocks requires the following steps [6]:

1) The data set is sorted based on the blocking key value (BKV) which is created from one or more attributes.

Sorted records rely on the following underlying hypothesis [3]:
The $n$ sorted records, $r_i \leq \ldots \leq r_{i+n}$, may satisfy: $dist(r_i, r_{i+1}) \leq dist(r_i, r_{i+2}) \leq \ldots \leq dist(r_i, r_{i+n})$, where, $dist(r_i, r_j)$, is the distance between records $r_i$ and $r_j$. Most of the blocking methods use the inverted index key where all records that have the same blocking key are placed in the same indexing list.

2) The record identifiers are retrieved from the index data structure for the block and record pairs are created for comparison. Each record is associated with the rest of records in the same block to form the candidate pairs. After the comparison step is complete, a numerical value is assigned.

In the next section, we will present related works in duplicate detection where blocking methods were proposed to improve the efficiency of duplicates detection.

## II. Related Work

There are at least two competing blocking methods used in duplicate detection as stated by Draisbach and Naumann (2010)[1]:

- Standard blocking method that partitions records into separate subsets, for example using "zip code" as the partitioning key.
- Sorted neighborhood method that sorts the data set according to certain key(s), such as "last name", and then slide a window of fixed size across the sorted data and compare pairs only within the window.

### A. Standard Blocking Method

In this method, records are sorted using the block key value and the sorted records are grouped into separate blocks where every record has the same BKV. The key is composed of a set of data set attributes. The comparison is made for all the records within the same block. If the duplicate records detection is required in the data set which has $R$ number of records and $B$ number of blocks, the number of record pairs is $O(\frac{R2}{B})$ for comparison. This method increases the speed of comparison operations because of the limited comparison between records within the same block.

Basically, this method is based on the expectation-maximization (EM) algorithm as proposed by [4] to calculate the probability of matching a pair of records. The probability of matched records pair is defined as class $M$, $P(xi = 1|M)$, and the probability of unmatched records pair is defined as class $U$, $P(xi = 1|U)$ that are calculated from a number of random records. Fig. 1 illustrates a standard blocking technique which was adapted from the original source [1].

Nevertheless, as the records are distributed into equal blocks size and compared within their own block, there is a high possibility for record mismatching (false negative error) that leads to less accurate duplicate detection. Therefore, it is necessary to choose appropriate keys for sorting in order to reduce the error ratio.

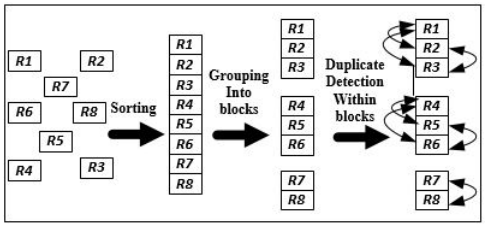



Fig. 1. Standard Blocking Technique(BKV) [1]

### B. Sorted Neighborhood Method

Sorted Neighborhood (SN) (which is also called as windowing) was proposed by Hernandez and Stolfo's (1998) to reduce the time taken for comparison process [7]. As compared to the standard blocking, this method improves more on the records matching ratio. SN requires creating a string of substrings from the selected attributes of the record to form the candidate key. For example, the first three letters of the name field, the first three digits of the identification number and the first three letters of the electronic mail concatenated together are used to create candidate key as shown in Table I.

TABLE I. Example of Creating Candidates Keys

| ID | Name | E-Mail | Candidate keys |
|---|---|---|---|
| 12331232 | Mohammed Ahmed | Moha12@yahoo.com | Moh123Moh |
| 34241212 | Saleh Noor | Salnor@yahoo.com | Sal342Sal |
| 12123234 | Noor Bint Salem | Norsal1@hotmail.com | Noo121Nor |

The records are then sorted based on the candidate keys. Then the comparison of records is performed within the window. After the comparison is finished, the first record is sliding out from the window and the next record is added to the window. If the size of the fixed window is $w$, then any new record enter to the window is compared with the previous one $|w| - 1$ records to find the matching. The process is illustrated by Fig. 2 shows a sorted neighborhood technique adapted from the original source [8].

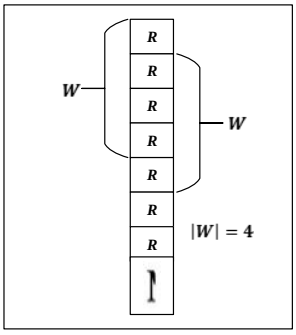



Fig. 2. Sorted Neighborhood (SN) Technique [8]

The accuracy of the SN method depends on the quality of the key and also the window size. With small window size and a high number of duplicates, the problem that can occur is overriding a number of duplicates (false negative

error). However, if the size of the window is too large, unnecessary comparisons will occur. Therefore, the quality of the key in addition to the size of the window is a challenge in SN technique. This problem was resolved by independent execution several times and the records are sorted with a different key in each time with a small window size. The result is derived from the integration of results where this method is called a multi-pass SN as proposed by Hernández and Stolfo (1995) [9].

### C. Duplicate Detection Within Incomplete Data Set

Many methods of duplication detection encountered a problem in compiling and comparing partially complete records. In fact, there are several types of data completeness problem can be found in data sets [10], [11] that can cause partially complete records. The presence of missing values (nulls) is one type data completeness problem that affect duplicate detection. There are several methods proposed to deal with missing values in data sets, such as to ignore records with missing entries, manual data imputation by human and the use an expectation-maximization (EM) imputation algorithm. The first two methods are impractical, while the estimation of missing values in a large number of attributes through EM becomes computationally intractable [12][13][14].

Tamilselvi and Saravanan (2009) presented a framework to detect and eliminate duplicates due to errors and missing data [15]. The proposed framework consists of six steps which are: 1) select the best attributes for duplicate identification (using attribute selection algorithm), 2) form token, 3)cluster the data (using clustering algorithm), 4) compute similarity score, 5) eliminate duplicates and 6) merge. The method of selecting attributes with complete values that do not contain duplicate values to perform tokens matching.

In [14], the authors proposed a method that extends the soft TF-IDF method to address two common cases in the detection of duplicates namely sparsity due to the number of missing entries, and the presence of more than one record duplicate. The proposed method consists of three steps: 1) create similarity scores between the records, 2) clustering the records together in independent groups, and 3)compare the different ways to create similarity scores between records (in addition to a different set of string matching, frequency-inverse methods, and n-gram techniques). The authors pointed out that the method of dealing with missing values can be replaced in the TF-IDF and Jaro-Winkler by performing data imputation with a likely candidate.

## III. Problem Definition

Several important issues need to be considered when record attributes are selected to be used as blocking keys. The first issue is that the quality of the values in these attributes will influence the quality of the generated candidate record pairs.

Ideally, attributes containing the fewest errors, variations, or missing values should be chosen. Any error in an attribute value used to generate a BKV will potentially result in records being inserted into the wrong block, thus leading to missing true matches [12][16].

In SN method, the selected attributes will be used as the blocking key. If the window size is small, then the number of comparisons will be reduced but the false negative value will be increased. If the window size is large then the number of comparisons will be increased with reducing the number of false negatives. For this kind of problem, the calculation of distinct and missing values is very important in the selection of blocking key. A numeric blocking key will not be effective in blocking the record for comparison. Therefore, the identification of measurement type and the type of attributes are important in the selection of blocking key [5]. To avoid miss-sorting due to errors in the attributes that are used to generate the key, again, multi-pass variants of SN produce multiple keys and perform the sorting and windowing multiple times. Like the standard blocking method, the transitive closure is finally calculated. Research has produced many variants of SN method including the one that avoids the choice of keys [15] and a variant for nested XML data [8].

Bigram indexing method is less sensitive to typo errors or some missing information as compared to the previous blocking methods. The features of this method are, first, this method allows each record to be in multiple blocks if needed, and it uses an inverted index [8]. There are three criteria which are very important in attribute selection for data cleaning which are identifying key attributes, classifying attributes (with high distinct value and low missing value) and measurement types of the attributes [17]).

In this section, the impact of the missing values on the blocking key will be elaborated through the following example. Suppose that we have a complete data set that contains employee personal information such as (name, address, city, phone, and sex) as shown in Table II. In this example data set, duplicate records are present in pairs, which are records (1,2), (3,4) and (8,9). In discussing the creation of the sorting key and the effect of missing values on duplicate detection, we consider the following cases:

TABLE II. Personal Information

| no | name | address | city | phone | sex |
|---|---|---|---|---|---|
| 1 | Ahmad | 435 s. la cienegablv. | losangeles | 310/246-1501 | M |
| 2 | Ahmed | 435 s. la cienegablvd. | losangeles | 310-246-1501 | M |
| 3 | Bel | 701 stone canyon rd. | bel air | 310/472-1211 | M |
| 4 | Bell | 701 stone canyon rd. | bel air | 310-472-1211 | M |
| 5 | Brit | 12224 venturablvd. | studio city | 818/762-1221 | M |
| 6 | Brown | 23725 w. malibu rd. | malibu | 310-456-0488 | F |
| 7 | Jim | 9560 dayton way | losangeles | 310/276-0615 | M |
| 8 | Johne | 14016 venturablvd. | sherman oaks | 818/788-3536 | M |
| 9 | Johne | 14016 venturablvd. | sherman oaks | 818-788-3536 | M |
| 10 | Navy | 2709 main st. | losangeles | 310-352-8035 | F |
| 11 | Nicol | 624 s. la breaave. | losangeles | 213-938-1447 | F |
| 12 | Nicolas | 2600 main st. | santamonica | 310/392-9025 | M |

A) **Single sorting key:** In this case, the creation of sorting key depends on one of the record attributes as shown in Table III where the sorting key is the attribute name. The problem, in this case, is sorting key will not appear when the attribute "name" value is missing in a record.

Missing sorting key causes the record to be placed out of the appropriate records block. For example, as "name" value for record 9 is missing, the record becomes the last record after the sorting and it will be excluded from the block.

TABLE III. SINGLE SORTING KEY

| no | name (sortkey) | address | city | phone | sex |
|---|---|---|---|---|---|
| 1 | Ahmad | 435 s. la cienegablv. | losangeles | 310/246-1501 | M |
| 2 | Ahmed | 435 s. la cienegablvd. | losangeles | 310-246-1501 | M |
| 3 | Bel | 701 stone canyon rd. | bel air | 310/472-1211 | M |
| 4 | Bell | 701 stone canyon rd. | bel air | 310-472-1211 | M |
| 5 | Brit | 12224 venturablvd. | studio city | 818/762-1221 | M |
| 6 | Brown | 23725 w. malibu rd. | malibu | 310-456-0488 | F |
| 7 | Jim | 9560 dayton way | losangeles | 310/276-0615 | M |
| 8 | Johne | 14016 venturablvd. | sherman oaks | 818/788-3536 | M |
| 10 | Navy | 2709 main st. | losangeles | 310-352-8035 | F |
| 11 | Nicol | 624 s. la breaave. | losangeles | 213-938-1447 | F |
| 12 | Nicolas | 2600 main st. | santamonica | 310/392-9025 | M |
| 9 | | 14016 venturablvd. | sherman oaks | 818-788-3536 | M |

B) **Multiple sorting key:** In this case, the sorting key consists of the strings of partial attribute values as shown in Table IV. In this table, the sorting key value consists of the first three characters of the attribute "name" and the first three characters of the "city". The sorting key is totally missing when all attribute values used as sorting key are missing; partially missing if at least one attributes value is missing. In both cases, the records with a missing sorting key will be outside of the records block.

TABLE IV. MULTIPLE SORTING KEY

| no | sortkey | name | address | city | phone | sex |
|---|---|---|---|---|---|---|
| 1 | Ahmlos | Ahmad | 435 s. la cienegablv. | losangeles | 310/ 246-1501 | M |
| 4 | Belbel | Bell | 701 stone canyon rd. | bel air | 310- 472-1211 | M |
| 5 | Bristu | Brit | 12224 venturablvd. | studio city | 818/ 762-1221 | M |
| 6 | Bromal | Brown | 23725 w. malibu rd. | malibu | 310- 456-0488 | F |
| 7 | Jimlos | Jim | 9560 dayton way | losangeles | 310/ 276-0615 | M |
| 9 | Johshe | Johne | 14016 venturablvd. | sherman oaks | 818- 788-3536 | M |
| 2 | los | | 435 s. la cienegablvd. | losangeles | 310- 246-1501 | M |
| 10 | Navlos | Navy | 2709 main st. | losangeles | 310- 352-8035 | F |
| 11 | Niclos | Nicol | 624 s. la breaave. | losangeles | 213- 938-1447 | F |
| 12 | Nicsan | Nicolas | 2600 main st. | santamonica | 310/ 392-9025 | M |
| 8 | she | | 14016 venturablvd. | sherman oaks | 818/ 788-3536 | M |
| 3 | | | 701 stone canyon rd. | | 310/ 472-1211 | M |

The purpose of sorting key creation is to partition the data set into smaller subsets so that the number of comparisons can be reduced. To compare the records in the data set, in the traditional way of duplicates detection each record in the data set will be compared with the rest of the records in the data set. If $N$ is the number of records in the data set, the number of comparisons is a matrix $N \times N$. Thus, as there are 12 records in the example, there are $12 \times 12 = 144$ comparisons will be required. To solve this problem, the data set is partitioned by a sorting key into small blocks with a certain length. For example, suppose that $n$ is the length of the block (which also represents the number of records in the block), thus the numbers of comparisons for each block is $\frac{n(n-1)}{2}$. Suppose that $n$ is 4, with 12 records, the data set will be divided into three blocks. The number of comparisons in each block is equal to $\frac{4(4-1)}{2} = 6$. Thus, the total number of comparisons for all blocks is $6 \times 3 = 18$, which is less than the number of comparisons in case if we deal the data set as a single block (where the number of comparisons equals to $\frac{12(12-1)}{2} = 66$).

Missing values can affect the distribution of records on blocks. In our previous example, if the data set is partitioned into blocks of length 4, then in the data set shown in Table III record 9 with a missing value will be is excluded from the comparison process. Similarly, for data set in Table IV, records 2, 3 and 8, will be excluded from the block under measure.

Typically, multiple passes method is used in order to overcome the false negative error as shown in the example. In this way, the records that are not blocked together in one pass will have the potential to be blocked and compared in another pass to avoid from being misclassified. Since two records cannot be matched due to missing values, the variables chosen for the blocking phase should be relatively complete, with only a few missing values. Such blocking strategy will reduce the set of potential matches to a more manageable number. Christen and Goiser (2007) recommend researchers to report the specific steps of their blocking strategy in order to ensure blocking success where in their work, blocking method is used to deal with record linkage problem [18]. To deal with the missing values and their effect in the creation of the sorting key, in the next section, we describe our proposal of blocking with a dynamic sorting key.

## IV. THE PROPOSED METHOD

In this section, we will describe the extension of the traditional method we propose in determining the attributes that are selected in the composition of the sorting key through a dynamic sorting key. In particular, we focus on the problem caused by the presence of missing values in the sorting key attributes. The proposed dynamic sorting key aims to avoid false negative errors by reducing the number of records that are excluded from the blocks due to the missing values. An experiment was conducted in order to evaluate the proposed method.

### A. Attribute Selection

Attributes selection is a pre-requisite process for the dynamic key generation. Attributes selection stage is one important stage upon which duplicates detection relies. In this stage, "appropriate" attributes are selected to create sorting keys in the classification stage. To choose the appropriate attribute there are three criteria to meet: (a) identifying key attributes,

(b) attributes with high distinct value and low missing value and (c) attribute measurement types. In the end, attributes with the highest priority are selected for the further process [19], [15]. Attributes selection process relies on the rate of missing values and repetition values in each attribute and also the threshold value for the acceptance rate.

The attributes selection process passes through three main sequential stages as shown in Algorithm 1. The aim of the attributes selection algorithm is to reduce the time and increase speed in later stages. The algorithm begins by setting the threshold value $T$, where (0<$T$<1). The higher the value of $T$ means the higher the acceptability on the amount of repetition and missing values in the attribute. A good selection of threshold value is very important since the choice of a very low threshold value affects the creation of the sorting key, which in turn affects the detection of duplicates in the data set.

**Algorithm 1** Attributes selection algorithm

**Input:** $N$ Attributes, $n$ number of rows, $T$
**Output:** $S$ Subset of attributes
**Var:** $A$ Attribute set, $i, j$
**begin**

1) $threshold \leftarrow T$
2) Unique $(U)$ value of the attribute $Ai$ if row $i_1^n = Ai + 1$
3) Missing $(C)$ value of the attribute $Ai$ if row $i_1^n = Null$
4) Calculate $AVG$=$avg(U \wedge C)$
5) Compare $(T, AVG)$
6) Rank $(N, AVG)$

**end**

**Compute:**

1) $U_j = n-$ Count(Distinct $A_j$)
2) $C_j = \frac{\sum_{i=0}^{N} completeness_{i,j}}{n}$

- **Uniqueness:** the uniqueness coefficient is calculated to measure the repetition ratio in the attributes values. As it is not possible to rely on attributes that contain duplicate values such as a gender attribute (that contains values only namely male and female) this function is crucial in attribute selection. To calculate the uniqueness coefficient, the uniqueness function (UF) is applied to each attribute in the data set. Figure 3 shows the flow of UF.
- **Completeness:** Figure 4 shows the flow of the algorithm that is called as completeness function (CF) to calculate the number of missing values for the attributes that have a high proportion of uniqueness (based on the threshold value) in the previous stage. To describe CF, suppose that:
  $A$ is a data set.
  $M$ is an array where each element in $M$ is the number of missing values in the attribute in $A$.
  $A_{attr}$ is the attribute identifier in $A$.
  $M_{Key}$ is the attribute identifier in array $M$. The steps taken for CF to measure the missing values are as follow:
  1) Create one dimension array $M$ with a number of elements equal to the number of attributes in $A$. Each element in $M$ consist of key and values, where the key represent the attribute name of $A$ and the value represents the ratio of missing values in the attribute $A_{attr}$ and $M_{Key} = A_{attr}$.

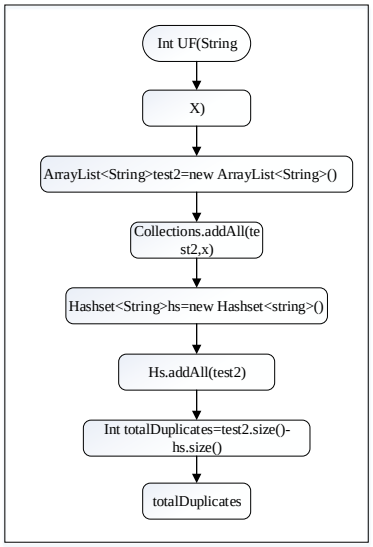



Fig. 3. Uniqueness Calculation Function

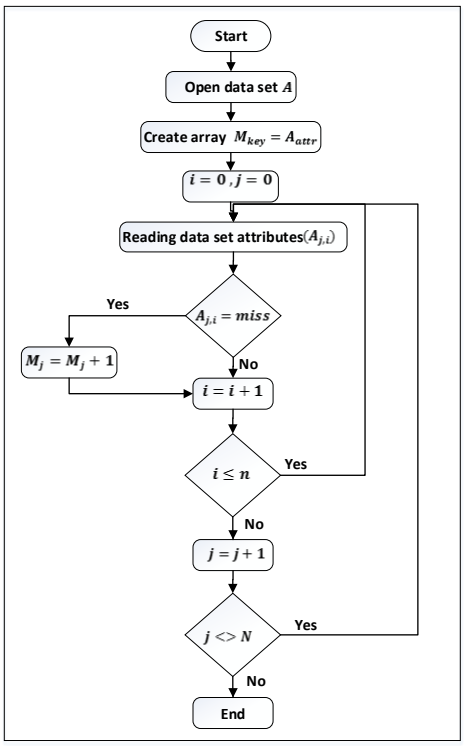



Fig. 4. Completeness Function

2) Check if the attribute which is represented in $A_{j,i}$ contains missing values by comparing the value of $A_{j,i}$ with the missing value representations such as null, blank space or "?" character. Thus, if the value of $A_{j,i}$ equal any representation of missing values, the value of $M_j$ increases by one.

- **Aggregation and ranking:** In the final step, we calculate the average of the uniqueness and the completeness scores for each selected attribute. The result ($AVG$) is compared with the threshold value ($T$) that was set at the beginning. Attributes with a high average of duplicates values and missing values are excluded from the candidate list. This means the result fails to fulfill $T$. Attributes that are in the candidate list are ranked according to their average scores.

### B. Dynamic Sorting Key

Dynamic sorting key creation depends on the attribute selection stage, that determines the high-rank attributes. In this method, three high-ranking attributes are selected to form a sorting key. In addition, the sub-string size is determined from the first three characters of each selected attribute where the sub-strings are grouped together to form a single string representing the sorting key after removing the blanks in the values of the attributes (if any). Figure 5 shows the steps to create the dynamic sorting key. A new attribute (called as

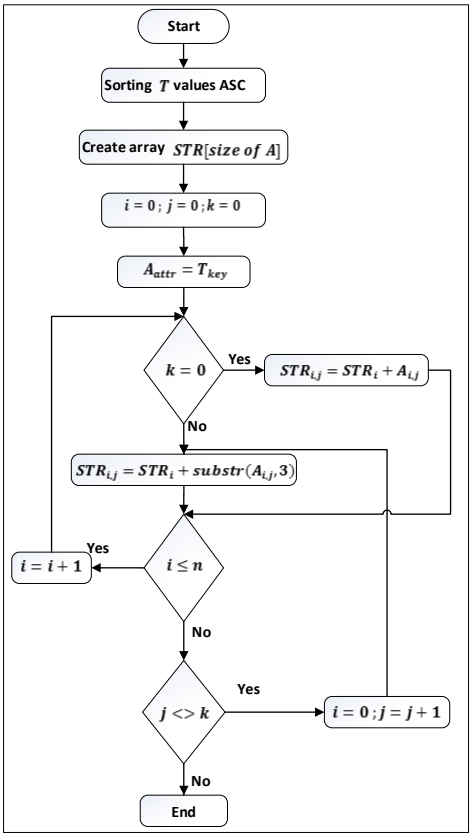



Fig. 5. Dynamic Sorting Key

ksort) is added to the data set to store the values of sorting keys that were created by dynamic sorting key function. After the dynamic sorting key is generated, the data set is sorted based on its values. The sorting key is used to partition the data set into small blocks to reduce the comparisons between the records pairs. In our method, we use DuDe toolkit (as proposed by Draisbach and Naumann (2010) [20]) to test the performance of duplicates records detection within incomplete data set with dynamic sorting key. In Section V, we will present the results of duplicate detection with the proposed method.

### C. Data Set

To test our method, we used the restaurant data set that provided by Hasso-Plattner-Institute (HPI)[1] which has been frequently used in duplicate detection research. The restaurant data were extracted from the RIDDLE repository. This data set consists of real 864 restaurant records, which was taken from the Fodor's and Zagat's restaurant guides that contains 112 duplicates. The data set comprises the names and addresses of restaurants. We have used restaurant data set that are configured by Duplicate detection toolkit (DuDe). Restaurant data set underwent a series of changes as reported in [20]. These changes are:

- changed file format from "arff" to "csv".
- removed header information.
- inserted row with column names.
- added a unique identifier for each record.
- deduplicated information in data sets.

Several additional changes have been made to the restaurant data set to prepare it for the experiment. The changes in the so-called data preparation step are:

- Partitioned the comma-separated value (CSV) file into attribute headings. The data set was originally separated by a semicolon.
- Eliminated blank spaces between the values for all attributes.
- Applied conceptual of arbitrary pattern to add missing values to the data set by deleting values from the data set attributes randomly (following the method in [21]).

### D. Experiment Configuration

To show the effect the missing values on the blocking method, we apply our method by using DuDe toolkit that has been used to detect duplicate records. DuDe is an open source tool that is opened for modifications and additions on the functions code for research purposes. We add attributes selection functions to select the proper attributes before the dynamic sorting key can be created. Figure 6 shows the experiment flow. In the experiment, a java class was created to read restaurants data sets that were stored in CSV format. Data preparation step as described earlier was performed before the attribute selection algorithm can be applied. To generate the dynamic sorting key, the UF was applied on the restaurant data set to get the uniqueness factor for each attribute. The CF was applied against the attributes with the highest uniqueness score (low repeated values) before the dynamic sorting key can be created. Figure 7 shows the values of dynamic keys (Ksort) created from the first three characters of the attributes values "Phone", "Address" and "Class" (that exhibit low repeated values and low missing values).

In the proposed method, we modified the configuration of SN algorithm that is used in DuDe toolkit. The modification

[1]https://hpi.de/naumann/projects/data-quality-and-cleansing/dude-duplicate-detection.html#c114678

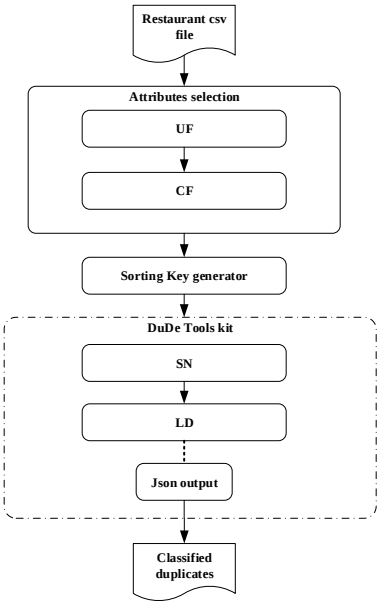



Fig. 6. Experiment Flowchart

| B | D | E | F | G | H | I |
|---|---|---|---|---|---|---|
| Ksort | name | ddress | city | phone | type | class |
| 3104350 | | 435s.lacienegablv | losangeles | 310/246-150 | american | 0] |
| 2136706 | | 6703melroseave. | losangeles | 213/857-003 | californian | 6] |
| 3102378 | | 23725w.maliburd | malibu | 310/456-048 | californian | 8] |
| 404103675 | 103west | 103w.pacesferryr | atlanta | 404-233-599 | continental | 675] |
| 21220m172 | 20mott | 20mottst.between | newyork | 212/964-038 | asian | 172] |
| 21221w23 | 21club | 21w.52ndst. | newyork | 212/582-720 | american | 23] |
| 21221w23 | 21club | 21w.52ndst. | newyorkcit | 212-582-720 | american(nev | 23] |
| 415222453 | 2223 | 2223marketst. | sanfrancisc | 415/431-069 | american | 453] |
| 2129jo173 | 9jonesstre | 9jonesst. | newyork | 212/989-122 | american | 173] |
| 404163379 | abbey | 163poncedeleona | atlanta | 404/876-853 | international | 379] |
| 40423574 | abruzzi | 2355peachtreerd. | atlanta | 404/261-818 | italian | 74] |
| 40423574 | abruzzi | 2355peachtreerd. | atlanta | 404-261-818 | italian | 74] |
| 415172454 | acquarello | 1722sacramentos | sanfrancisc | 415/567-543 | italian | 454] |
| 310293112 | adrianosri | 2930beverlyglenc | losangeles | 310/475-980 | italian | 112] |
| 212700174 | adrienne | 7005thave.at55th | newyork | 212/903-391 | french | 174] |
| 212764602 | afghankeb | 764ninthave. | newyorkcit | 212-307-161 | afghan | 602] |

Fig. 7. Sorting Key

allows SN algorithm to pair and to sort the records based on dynamic sorting key where window size was set to 10. Listing 1 shows the configuration of the SN algorithm.

Listing 1. Configuration of the SN algorithm

```
//Attributes Selection
//Missing values compensate
//Dynamic sorting key(Ksort) generator
new SN algorithm with window size 10 and
in-memory processing enabled
algorithm=new SortedNeighborhood
Method(Ksort,10);
algorithm.enableInMemoryProcessing();
algorithm.addDataExtractor(extractor);
```

The SN algorithm returns the pairs to be classified by the comparator that uses Levenshtein Distance (LD) for the comparison string $ksort$ that was created in the previous step with similarity measure equals to 0.9. Listing 2 shows the configuration of the comparator.

Listing 2. Configuration of the comparator

```
levComp = new LevenshteinDistance
Comparator("ksort");
```

As DuDe is used to classifying duplicate pairs, the post processor is configured to compute the transitive closure and to write the final output of the pairs that are classified as a duplicate in Json file.

## V. Results, Analysis and Discussions

As described in the previous section, the experiment produces two sets of result namely DuDe with dynamic sorting key's (denoted as Dude') and the original Dude's (denoted as Dude). The results show that using Dude', the number of comparison pairs for the restaurant data set was 7,740 pairs and the number of duplicate records detected was 104. With Dude, lower number of pairs of comparisons and duplicate records was yielded, with 7,731 and 92 respectively. Figure 8 shows the results of the experiments in Dude's toolkit interface.


Fig. 8. Experiment Results

According to the results of the experiments, out of 112 actual duplicates, Dude' managed to detect 92% of it, with the ratio of false negative equals to 7.14%. This error rate is low and it suggests that that Dude' offers more accuracy than Dude with 82% of duplicates detected and 17.86% of false

negatives. In terms of detection speed, Dude' has demonstrated a lower run time of 351 ms as compared to DuDe with 583 ms. Figure 9 illustrates the results.

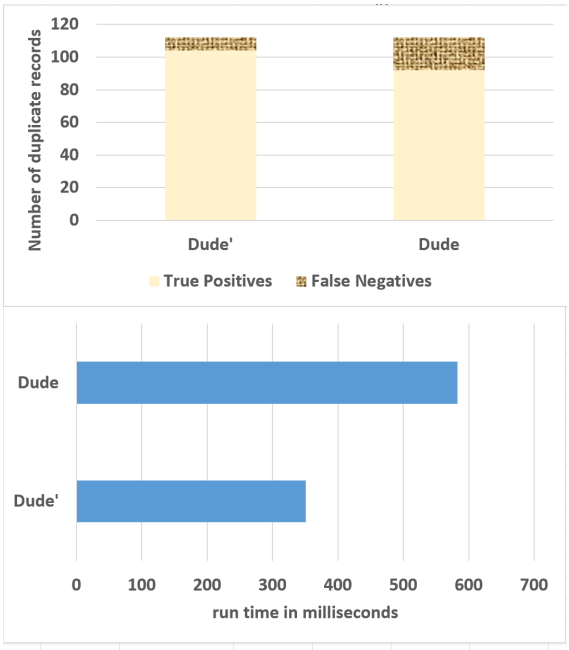



Fig. 9. Comparison of Accuracy between DuDe' and DuDe

## VI. Conclusion

As a conclusion, detecting duplicates within a data set is a challenge especially when missing values are present. In this paper, we focus to answer the question of how duplicate records can be accurately detected within incomplete data sets. In search of the answer, we proposed an extension to SN blocking method that adds dynamic sorting key algorithm within the attribute selection step. To evaluate the proposed method, we conducted an experiment where a real restaurant data set and DuDe toolkit were used. The results show that not only the number of false negatives duplicates can be reduced, the proposed method is also faster than the duplicate detection that was performed without the proposed extension. Nevertheless, whether the proposed method behaves well or not in duplicate detection using other blocking methods such as bigram indexing and canopy clustering with TFIDF (Term Frequency/Inverse Document Frequency) is an open problem for future work.

## Acknowledgment

We would like to thank Hasso-Plattner-Institute (HPI) for the data set and duplicate detection tool (DuDe) used this research. We also acknowledge financial assistance received from the Universiti Teknikal Malaysia Melaka (UTeM) in publishing this work.

## References

[1] U. Draisbach and F. Naumann, "A generalization of blocking and windowing algorithms for duplicate detection," *Data and Knowledge Engineering (ICDKE), 2011 International Conference on. IEEE*, pp. 18–24, 2011.

[2] B. Carlo and M. Scannapieca, *Data Quality*. ACM Computing Classification, 2006.

[3] S. Yan, D. Lee, M.-Y. Kan, and L. C. Giles, "Adaptive Sorted Neighborhood Methods for Efficient Record Linkage," *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 185–194, 2007. [Online]. Available: http://doi.acm.org/10.1145/1255175.1255213

[4] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 414–420, 1989.

[5] J. T. J and V. Saravanan, "Token-based method of blocking records for large data warehouse," *Advances in Information Mining*, vol. 2, no. 2, pp. 5–10, 2010.

[6] P. Christen, "Improving data linkage and deduplication quality through nearest-neighbour based blocking," *Discovery*, 2007.

[7] M. A. Hernández and S. J. Stolfo, "Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem," *Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 9–37, 1998.

[8] J. Shin, "Comparative Study on Blocking Methods in Record Linkage," Ph.D. dissertation, Oklahoma State University, 2009.

[9] M. A. Hernández and S. J. Stolfo, "The Merge / Purge Problem for Large Databases," *ACM SIGMOD Record*, pp. 127–138, 1995.

[10] N. Emran, S. Embury, and P. Missier, "Model-driven component generation for families of completeness," in *6th International Workshop on Quality in Databases and Management of Uncertain Data, Very Large Databases (VLDB)*, 2008.

[11] N. A. Emran, "Data completeness measures," in *Advances in Intelligent Systems and Computing*, vol. 355, 2015, pp. 117–130.

[12] T. D. Pigott, "A Review of Methods for Missing Data," *Educational Research and Evaluation*, vol. 7, no. 4, pp. 353–383, 2001.

[13] N. J. Horton and K. P. Kleinman, "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models," *The American Statistician*, 2007.

[14] Y. Van Gennip, B. Hunter, A. Ma, D. Moyer, R. de Vera, and A. L. Bertozzi, "Unsupervised record matching with noisy and incomplete data," 2018.

[15] J. Tamilselvi and V. Saravanan, "Detection and elimination of duplicate data using token-based method for a data warehouse: A clustering based approach," *International Journal of Dynamics of Fluids*, vol. 5, no. 2, pp. 145–164, 2009.

[16] G. Papadakis, E. Ioannou, T. Palpanas, C. Niederé, and W. Nejdl, "A blocking framework for entity resolution in highly heterogeneous information spaces," *IEEE Transactions On Knowledge and Data Engineering*, 2013.

[17] J. Jebamalar Tamilselvi and C. Brilly Gifta, "Handling Duplicate Data in Data Warehouse for Data Mining," *International Journal of Computer Applications*, vol. 15, no. 4, pp. 7–15, 2011.

[18] P. Christen and K. Goiser, "Quality and Complexity Measures for Data Linkage and Deduplication," *Springer*, pp. 127–151, 2007.

[19] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Technique (The Morgan Kaufmann Series in Data Management Systems)*. Elsevier, 2011.

[20] U. Draisbach and F. Naumann, "DuDe: The Duplicate Detection Toolkit," *Proceedings of the International Workshop on Quality in Databases (QDB)*, 2010.

[21] C. K. Enders, *Applied Missing Data Analysis*. The Guildford Press, 2010.

# An Agent Cellular Residential Mobility Model : From Functional and Conceptual View

Elarbi Elalaouy*, Khadija Rhoulami, Moulay Driss Rahmani
LRIT associated unit to CNRST (URAC°29), Faculty of Science, Mohammed V University in Rabat
4 Av.Ibn Battouta B.P. 1014 RP, 10006 Rabat, Morocco

***Abstract*—Residential mobility is of great challenge to sustainable cities. Developing computer models based simulation could be a powerful tool to support informing urban decisions especially with the fact that half of the world's population now lives in cities. The present paper presents our detailed model of residential mobility which use an alliance Multi-agent systems and Cellular automata (MAS-CA) approach. Conventional Urban modelling approaches will be presented firstly with a distinct light sheded on the alliance CA-MAS approach. The model will be then exposed in its two functional and conceptual views. At the end, results of a scenario of growth population is simulated and discussed. Results of the simulation shows significant conformity with the underlying model hypothesis.**

***Keywords—Residential mobility; land use change; computer model; multi-agent systems; cellular automata***

## I. Introduction

Residential mobility is an urban phenomenon that has been the subject of research in several disciplines, starting with sociology, demography, geography, planning and urban planning [1] [2] [3] [4]. It is characterized by residential movements, both individual movements (decision to move and housing choice) and aggregated ones (population flows) of an urban population. Residential mobility is mainly concerned with intra-urban mobility taking place within an urban agglomeration, without taking into account the mobility over longer distances (example of interurban mobility).

Studies of residential mobility have among others the following objectives : 1) understanding individual and aggregate residential choices of city dwellers; 2) studying spatial structures resulting from the residential choices of the population; 3) anticipating and predicting residential dynamics; 4) testing the possible evolutions in the face of modifications of the infrastructure such as insertion of building zones, industrial zones, etc; 5) meeting needs of the population in terms of housing and transport infrastructure.

Conventional studies of residential mobility carried out by state institutions (which are statistical in nature) are important [5]. They make it possible to quantify the phenomenon of residential mobility, to clarify its causes and present effects. However, the decisions and urban actions derived from these studies can only be gradual or progressive as they are a reaction to a present urban reality. The current state of the question shows that conventional studies cannot alone achieve objectives (3-4-5) since these objectives require predictive and simulation capabilities that are deficient to conventional studies.

Issues of environmental and sustainable development of cities, advances in computer science have allowed the creation of a new movement of modeling and simulation of urban dynamics. This new urban dynamics modeling and simulation movement focuses on the development of computer simulation models that harness data on population and urban infrastructure, predict mutual dynamics of population and infrastructure, and analyze results of the simulation in order to understanding urban systems and support with decision making.

In this article, we will present firstly four main urban modelling approaches that have been used in the literature; we shed then the light on the alliance CA-MAS approach. Secondly, we will outline our developed residential mobility model from a functional and conceptual view. Finally, results are then presented to highlighting usefulness of this CA-MAS modeling approach for supporting decisions making and urban planning.

## II. Urban Modelling Approaches

Modeling and simulation of urban dynamics, including residential mobility, has been marked by the use of several approaches. We present four main modeling approaches.

### A. The Regional Approach

The regional approach is an approach originated from economics. This approach considers the city as divided into several administrative areas. Each zone includes data on infrastructure (land-use distributions, housing categories, housing prices, etc.) and on population (socio-economic and cultural group distributions, etc.), mainly coming from statistics on these administrative areas [1]. These areas are the basic units of the regional models that attempt to manage the flows between zones. The objects exchanged can be population, commodity, capital, etc. Zones have weights that depend on their scales.

### B. The Statistical Approach

The statistical approach is based on the analysis and exploitation of general population censuses and household surveys [5].. This approach attempts to measure urban phenomena (such as residential mobility, urban growth, urbanization, etc.), describe the evolution of these phenomena at a spatio-temporal scales, and prepare cross-tabulations for example of behaviors of household mobility and housing characteristics (location, housing type, occupancy status) with those of individuals and households (age, occupation, income level, etc.)

### C. Cellular Automata Approach

The cellular automata approach is a bottom-up approach. It divides the urban space into a grid of cells. Each cell has a state among a finite set of system states. future states depend on transition rules based on a local spatio-temporal neighborhood. The system is homogeneous in the sense that the set of states is the same for each cell and the same transition rules are applied to each cell. Time advances in discrete steps and the application of transition rules on all cells can be synchronous or asynchronous [6].

Conventional CA models use neighborhood-based rules with local scope. New applications of CA models are becoming more and more free of this specificity. They use rules whose neighborhood is defined by a distance to the relevant cell [7]. For example, in a city, when looking for a new home, the exploration of a household can be done in the radius of its current home. This radius can cover gradually a district, a town or the entire city.

CA models use either IF-THEN transition rules or probabilistic transition rules. The IF-THEN approach consists of using simple rules implemented as conditions to model the dynamics of the system [8]. This approach, according to our knowledge, is the least used in contemporary time in urban modeling. The probabilistic approach, on the other hand, is becoming increasingly important in modeling spatial dynamics. The Markov chain models incorporated within CA models are an example of models that introduce probabilistic transition rules [9]. The authors of [10] reported that an approach combining CA and Markov chains benefiting from the potentials of both methods is a preferred approach. The cellular automaton represents the spatial distribution of the land use of an urban space, and above it, the Markov chains predict the total rate of land use change from the past changes.

Cellular modeling techniques are the core of many spatial dynamics models. The use of these techniques for geographic process modeling dates back to the nineties. Advanced CA models currently exist for modeling and simulating spatial dynamics, evacuation systems, river basin dynamics, forest fire spreading [11] [12] [13] [14]. In particular, these cellular techniques have also contributed to the modeling and simulation of residential mobility. Among the first theoretical works that simulated the residential movements of individuals in an urban space are the works of James M. Sadoka and the works of Thomas Schelling as reported in [6]. The two authors designed a cellular space inhabited by a population belonging to two social groups and then tried following their own models of transition rules to simulate manually the residential displacements of individuals over cellular space.

Both models are very simple in construction but it inspired researchers in modeling and simulation of residential mobility [1] [2] [15]. These latter models inherit Sadoka and Schelling's idea of conception, which allows on one side to place each individual entity, which can be either an individual or a household, in a cell that is used as housing, and on the other hand, to use the transition rules as a proxy to trigger the decision-making of individual entities. For example, in the work [2] a model of residential mobility simulation based on cellular automaton was designed. The population is organized into individual households each of which occupies a cell of the automaton which represents the inhabited dwelling. The transition rules of the CA model make it possible to govern households in their housing change decisions. Emerging residential dynamics are generated from the interaction of changes in the social structure of households and the positive or negative influence of the neighborhood. This example sheds light on how the transition rules can be used as proxy between the cellular automaton that represents the studied city and the population that lives there.

Modeling efforts, mainly using the cellular automaton approach, for the modeling of the resulting spatial dynamics in addition to the decisions of the urban actors, succeed as the work [2] demonstrates. However, this cellular method succeeds only if the unit of analysis is in harmony with the used cellular arrangement, the decisions of the urban actors are at the same level of organization and the local neighborhood affects the different actors with a simple and uniform way . When the actors are not related to the locations in an intrinsic manner of the cells of the cellular automaton, the neighborhood relations of the conventional cellular automaton cannot reflect the neighborhood relations of the urban actors. The solution to this problem depends on techniques that use non-uniform transition rules and that can dynamically change the configurations of cell connections. Because these techniques far exceed the capabilities of conventional CA models, conventional CA methods cannot be widely adapted to modeling urban dynamics. A model of urban dynamics may require multiple urban actors arranged in space, agents arranged at different scales of organizations (e.g., population, urban managers), or agents controlling certain parts of urban space.

In summary, CA techniques have clearly demonstrated their ability to model the spatial aspects of urban dynamics, yet they face challenges when considering the decisions of urban actors, especially when these actors are working at different organizations scales. To overcome these challenges, it is necessary to implement complex and multiscale transition rule sets that apply to different groups of cells. Even if this solution is possible, the deviations from generic CA will be to the detriment of giving up the potential benefits of generic CA. In particular, In order to ensure reuse, and extensibility in other applications and domains through modeling and simulation platforms, it is necessary that the form of the CA preserves as many features of generic models as possible.

### D. Multi-Agent Systems Approach

While the cellular automata approach focuses on the representation of space and its transition, the multi-agent approach is interested in the representation of actors so-called agents, their actions and interactions in the system. Agents are the fundamental elements in multi-agent models. Several characteristics define the agents : they are autonomous ; they share an environment through communication and interaction ; and they make decisions that bind behavior to the environment. Agents have been used to represent various entities such as atoms in physics, cells in biology, animals, robots, humans and organizations. Autonomy means that agents act and interact with their total will, even when these agents are dominated from an external entity, these agents do not perform actions unless they have chosen them. Wooldridge defines intelligent

agents as being able to act flexibly [16]. This imply that agents are objective-oriented in their actions and that they are able to interact with other agents and a common environment, which is everything outside the agents. In an urban modeling context, a shared environment can be an urban space where the actions of each agent can affect those of others. It can also be a land market through which agents buy and sell housing and land.

Agents must perform actions that result from a thought process that links their goals to the environment through their behavior. This thinking process, also known as the cognition model, can range from simple stimulus-response decisions to the point where agents become proactive and make decisions. An autonomous agent needs, at a minimum, to perceive his environment and react accordingly to the perceived knowledge. The reaction can be considered as a cognitive model, in the sense that it is the response of agents to change. Some models of cognition go beyond reaction, they consider agents as possessors of rationality and perfect analytical capacity that allow them to solve the complex problems of mathematical optimization. These models trying to make the agents perfectly rational are contradicted by the experimental proofs. It is an open question if these models of perfect rationality are appropriate to multi-agent urban models, knowing the interdependencies and interrelations in its systems. The difficult recognition of the complex environment in which human decision-making occurs has led to shift to a new variant of agent models with reduced rationality. Instead of implementing a solution that completely predicts the future states of the system they are part of, agents use heuristic classification and prediction methods. Bayesian networks, neural networks and genetic algorithms are examples of methods that have been used to design agent cognition models in some urban systems [17] [18] [19].

The applications of the agent approach for designing complex systems are not to be demonstrated, however, residential mobility applications designed solely based on the agent approach are not up to the required height since the agent approach is not originally interested in modeling spatial environments. On the other hand, there is some modeling work in residential modeling that has opted for the use of the agent approach. These residential mobility agent models, do not integrate urban space as environment, they do not have spatial representation as reported by [20]. To represent the agent environment, these agent models incorporate land market models. The work of [21] illustrates this kind of model through a model of multi-agent simulation of displacements of the population in a mono-centric urban city. The model does not incorporate the urban space in which the population moves, it uses a reduced grid of two dimensions, which makes it possible by differentiating housing units by their distance to the center of the city, as well as by amenities by location.

These models rarely encountered in literature, described a research track that begins to be forgotten, in contemporary times, for two main reasons. The first reason is that these models do not include spatial information of the city such as equipment units, hospitals, schools etc. The second reason, even if they are agent based models, they do not deal with individual residential behaviors including mobility decision, and housing choice decision. This research, which uses the agent approach in residential mobility modeling, shows, on the one hand, that the agent approach is useful for the design of urban actors, in particular households, but on the other hand, they illuminate well that this agent approach does not have the means to represent the urban space, place where interactions of human agents happen.

## III. Proposed Residential Mobility Model

### A. The Alliance Approach of CA&MAS

The alliance approach of CA-MAS is a recent approach for modeling urban dynamics that are characterized by both societal and spatial dynamics. Despite its novelty, it has in its repertoire various applications from urban modeling as it is an emerging approach of two approaches each having a great maturity in the areas of modeling and simulation of complex systems. Torrens reports that this new approach is useful for the development of planning support tools, with the ability to simulate individual households and built-up areas in a realistic, dynamic and flexible way [22]. It are the limitations of the previously presented approaches that are causing the movement towards this new modeling and simulation approach. The lack of dynamic representation in urban simulations, the lack of detail in spatio-socio-economic representations and the lack of inclusion of bottom-up processes in complex systems are the major disadvantages of conventional regional models [22].

Regional models operate at a very high level. In fact, when heterogeneity and interaction at a local level are important in urban dynamics, regional models will have a very limited capacity to explain these dynamics. Statistical models provide a sophisticated view of multi-dimensional decision indicators. However, these models condense the information into estimated parameters that represent the average effects of the available data. Indeed, these models can be useful for the projection of dynamics and spatial interactions only for phenomena guaranteeing stationarity and uniformity in space and time. Although cellular models offer greater flexibility for the representation of spatio-temporal dynamics, they have limited means to represent urban actors. Cellular models do not fail to model urban actors only when they are arranged in the same way of cells of the cellular automaton. None of the previous approaches can effectively represent the different decisions of urban actors that are characterized by autonomy, heterogeneity and decentralization except multi-agent approach. But the latter does not have the means to represent the urban space where interaction of human agents happen.

The alliance MAS-CA approach addresses limitations of each of the previous approaches. In particular, this alliance approach has a great capacity for modeling the urban dynamics resulting from the interactions of human and spatial entities. The approach benefits from the complementary advantages of both MAS and CA approaches. The MAS approach can represent urban actors, their behaviors and their interactions with each other and with their environment. The CA approach can represent spatial units like houses, apartments, schools, etc.. The MAS-CA approach benefits also from the theoretical underpinnings of both approaches. In the SMA Framework, spatial dynamics, the development of built-up areas and the creation of vacant housing are considered emerging dynamics

of the actions of urban actors, particularly the population living in the urban area; In the CA framework, spatial dynamics are considered as resulting dynamics of statistical concepts of transition rules (eg., Markov chain concepts) applied to a series of cellular maps of the studied region. These two theoretical foundations make it possible to see the city in these two societal and spatial aspects and can bring more representation to the intrinsic processes that govern urban areas.

The new alliance approach MAS&CA creates a new motivation for the modeling and simulation of residential mobility. Although urban models, particularly those of residential mobility, are not simple in design and development, researchers are increasingly relying on previous approaches and are increasingly involved in this new framework of the alliance of MAS-CA approaches. Among the most important models, we could find [1] [20] [4]. These residential mobility models are generally based on similar theoretical frameworks (with of course, specific features of each model) derived mainly from advances in the social and environmental sciences. These theoretical frameworks represent residential dynamics according to a bottom-up approach called the "stress-resistance" approach, which sees residential dynamics as the result of the residential actions of different households in the population. In its standard version (see Figure 1), households take two basic steps, the first stage concerns the decision to leave the dwelling, the second stage concerns the decision to reside in a new location [1]. In the first stage, households estimate the "stress" of moving by comparing the current stress to the desired residential situation; if the stress is high enough, they decide to change housing. In the second stage, those who are willing to move estimate the "resistance" to relocate by comparing available alternatives to their current home, and then decide to either move to one of the alternatives or stay in their current home. He continues to say that this approach has several names "dissatisfaction", "utility" and "residential dissonance". The urban system, in these models of residential mobility, includes urban space and population particularly individual households which perform residential actions namely the mobility decision and the housing choice decision.

These residential mobility models, based on the stress-resistance approach, are more or less characterized by the following modules: 1) a module for representing urban infrastructure; 2) a spatial dynamics module; 3) a household representation module 4) a households mobility module 5) a housing choice module 6) and a population projection module.

### *B. Context and Hypothesis*

The residential mobility simulation model that we propose is part of this new trend of combining CA-MAS approaches for the modeling and simulation of urban systems. This model represents the urban space by a two-dimensional cellular automaton. Thus, urban space consist of a set of juxtaposed uniform spatial units. The model represents the households constituting the population of the urban space by mobile agents having the mobility capacity in the urban space. Household agents use socio-economic data from household surveys. Agents make decisions independently.

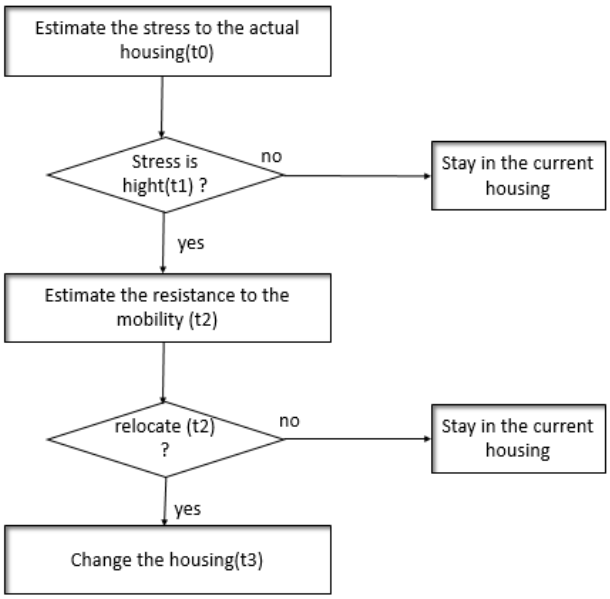



Fig. 1. The stress-resistance approach, $t_1 < t_2 < t_3$ indicates consecutive moments of the temporal sequence of a household related to its residential actions (adapted from [1])

Each agent decides for himself whether he wants to change housing or not. If this decision is positive, he starts looking for a new home that suits him. These agents are cognitive and thus have a reasoning ability about their mobility and housing choices. In fact, each agent has a mobility decision function designed using a probabilistic model based on mathematical concepts of Bayesian networks. It also has a function of choosing the best housing from a list of suitable housing to his desire. Thus, an agent explores the urban space starting with the nearest neighborhood. In case of need, the householder increases his field of research until finding suitable housing.

The model also governs the dynamics of the urban space through a complementary layer of Markov chain and Cellular automaton. The transition rules of the cellular automaton are then governed according to a probabilistic model using mathematical concepts of Markov chain. The model also governs the growth of the population through a population projection sub-model. On an annual basis, a new generation of households is generated from the previous generation using demographic indicators such as birth rate, member departure rate, marriage rate, divorce rate and expectancy of life.

### *C. The Contribution of the Model*

The objective of our model is the simulation of residential mobility and the associated urban growth over a horizon of ten-years. The city being the place where the two phenomena run, is the result of residential choices at the individual level of households. The urban system as a whole consist of an urban space, a set of households, a set of dwellings and a time dimension. Urban space including housing is approached by a grid of cellular automaton. The households making up the population are approached by a multi-agent system.

The model of residential mobility we are presenting is an extension of two-research work [23] [24]. The first research work focus on the functional aspect of the proposed residential

mobility model, and describe its mathematical equations. The second research work focus on the conceptual aspects of the proposed residential mobility model, which consists of UML 2.0 based models. The objectif of the UML 2.0 modeling is to boost the development phase of residential mobility Framework. This research work bring these two aspects together in order to show the proposed residential mobility model as entirely as possible. In addition, the present work describe, from an algorithmic view, the four functional models, which represent the core of the proposed residential mobility model.

## IV. Detailed Description of the Model

### A. The Functional View

The RM model consists mainly of four mathematical models as below. This new description of the functional view has origin in the research work [23] which describe in detail the mathematical models. Figure 2 depicts the architecture of the RM model and its components.

- **The mobility decision model:** which uses a classification technique to find out whether each given household with given characteristics has a propensity for changing housing or not. The propensity for mobility noted $P^t(relocation|K)$ , where t is the time, K the vector of household characteristics, is calculated as well detailed in [23]. Figure 5 shows the corresponding algorithm**.**
- **The Housing Choice Model:** allows each household to explore the neighborhood gradually (rank 1, rank 2, etc.), look for potential dwellings, and ultimately choose the most suitable dwelling to the household. The equation denoted $P_{C_{ij}}$ is used to calculate the attraction potential of a cell $C_{ij}$ . Figure 7 shows the corresponding algorithm.
- **Spatial prediction model:** which makes it possible to predict spatial dynamics, this one uses either spatial prediction techniques or manual configurations that can be parameterized in the RM Framework. In our model, the Markovian equation that governs the spatial transition is denoted $P\left(C_{ij}^{t+1} = \mathrm{e}\right)$ , it designates the probability that the spatial unit $C_{ij}$ transit to the state e at time t + 1. The Figure 4 shows the corresponding algorithm
- **The population projection model:** which allows by means of demographic rates such as the rate of birth, divorce, marriage, departure etc, to project the next population from a given population. This model could use projection technique or manual configurations introduced in the Framework. The Figure 6 shows the corresponding algorithm.

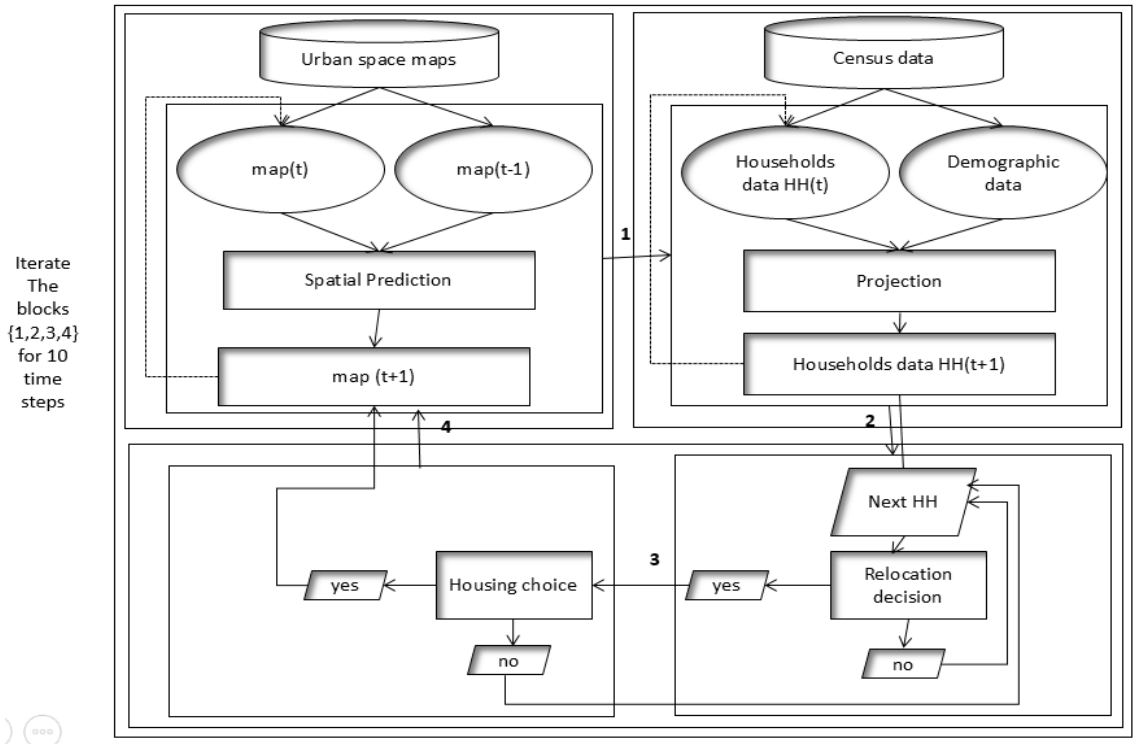



Fig. 2. Architecture of the residential mobility model gathering the above described sub-models.

### B. The Conceptual View

The model consists mainly of four conceptual models designed using the UML 2.0 modeling language. This new description of the conceptual view has its origin in the modeling work [24] which describe in detail these conceptual models. The Figure 3 expose a light weighted combination of the four models in one model.

- **The model of the urban space:** represents the urban space by means of a cellular automaton approach. The

urban space is therfore considered as a set of juxtaposed spatial cells having states representing the soil occupancy.

- **The population model:** represents the population of households and their characteristics, in the simulation framework; the population are approached by a multi-agent system**.**
- **The model of population dynamics:** allows archiving of population generations over a calendar of years. This model allow on one hand to trace the demographic evolution of the population and on the other hand observe the movements of households in the urban space.
- **The model of urban space dynamics:** which traces the evolution of urban space during a calendar of years of simulation. This traceability will thus make it possible to observe the evolution of built-up areas in the studied urban space.

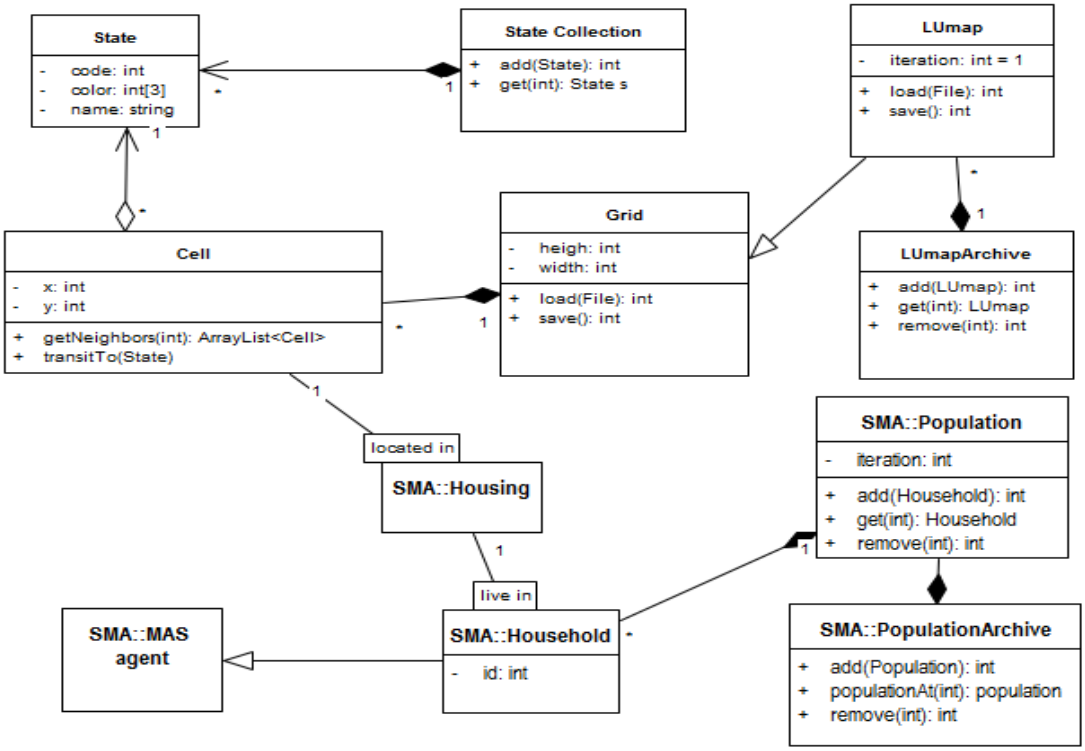



Fig. 3. The dynamic model CA-MAS of residential mobility

### C. Data of the Model

The model uses three types of input data, with which begin the simulation of residential mobility, namely maps of urban space, household survey data, and initialization parameters:

- **Household data:** represent socio-economic data of all households in the urban area. Households are classified according to their household type (Family With Children, Family Without Children, Single Family and Single Person). Households have also characteristics such as occupancy status, age of householdholder and average income. Households are also classified as owners and renters according to their occupancy status.
- **Maps of urban space:** represent the land use of urban space. The model requires two reference maps at time $t_0$ and $t_1$. It is from these two maps that the map at time $t_2$ is predicted, and so on. The preparation the urban space map require different spatial data layers of the studied area in particular (road and rail infrastructure, urban planning, building areas etc).

Maps of urban space are designed according to a predefined format and use static and dynamic states of land use : road or rail network, river or lakes, green space or forest, commercial building and equipment building. The elementary dynamic states are owned single house, rented apartment, rented single house, rented apartment, and building area

- **Initialization parameters:** represent demographic rates and other simulation parameters. Demographic rates, such as birth rate, departure rate, divorce rate, marriage rate and life expectancy govern the life cycle of households. In other words, the transition, for example, from single-person households to a Family Without Child households is governed by the marriage rate, and so on. There are other parameters which control the simulation such as the total duration of the simulation, the spatial resolution of the cellular automaton and the environmental quality of the housing type.

**Algorithm: spatial prediction of time t + 1**
Inputs: t : integer / * time * /
* Get <Dynamic Cell> List at time t
* For each Cell Cij
- Calculate transition probability $P\left(C_{ij}^{t+1} = \mathrm{e}\right)$ to all dynamic cell
- Choice of state e having the highest probability $P\left(C_{ij}^{t+1} = \mathrm{e}\right)$
- Calculate potential of attraction
- If <Cij> is Houssing && potentiel >= 1.0
  - Cij transit to the current state e
- Else if <Cij> is "Planned building area" && <Cji> near to Housings
  - <Cji> transit to e state

Fig. 4. Spatial prediction algorithm

**Algorithm : Mobility of households at t time**
Inputs: t: integer / * time t * /
* Get <Household> agents List at t time
* Get <House> List at t time
* Get <Apartment> List at t time
* For each agent
If Desir = 'LivingAHouse'
- Find a suitable House list
- get the best <House>
- get action of the agent
- If action = 'BuyAHouse' && OccupancyStatus = 'Owner'
  - Agent migrate to the new House
- If action = 'RentAHouse' && OccupancyStatus = 'Renter'
  - Agent migrate to the new House

If Desir = 'LivingAnApartment'
- Find suitable apartment list
- Get best <apartment>
- get action of the agent
- If action = 'BuyAnApartement' && OccupancyStatus = 'Owner'
  - Agent migrate to the new apartment
- If action = 'RentAnApartement' && OccupancyStatus = 'Renter'
  - Agent migrate to the new <apartment>

Fig. 5. Algorithm of Mobility of household agent a t time

$$\mathbf{Effectif_{PS}^{t+1}} \mathbf{+= Effectif_{PS}^{t}} * (\mathbf{1} - \mathbf{T_{mariage}}) \quad (1)$$

$$\mathbf{Effectif_{PS}^{t+1}} \mathbf{+= (Effectif_{FAE}^{t} + Effectif_{FM}^{t})} * \mathbf{T_{départ}} \quad (2)$$

$$\mathbf{Effectif_{FSE}^{t+1}} \mathbf{+= \frac{1}{2}} * \mathbf{Effectif_{PS}^{t}} * \mathbf{T_{mariage}} \quad (3)$$

$$\mathbf{Effectif_{FSE}^{t+1}} \mathbf{+= Effectif_{FSE}^{t}} * (\mathbf{1} - \mathbf{T_{divorce}}) \quad (4)$$

$$\mathbf{Effectif_{PS}^{t+1}} \mathbf{+= 2} * \mathbf{Effectif_{FSE}^{t}} * \mathbf{T_{divorce}} \quad (5)$$

$$\mathbf{Effectif_{FAE}^{t+1}} \mathbf{+= Effectif_{FAE}^{t}} * (\mathbf{1} - \mathbf{T_{divorce}}) \quad (6)$$

$$\mathbf{Effectif_{PS}^{t+1}} \mathbf{+= Effectif_{FAE}^{t}} * \mathbf{T_{divorce}} \quad (7)$$

$$\mathbf{Effectif_{FM}^{t+1}} \mathbf{+= Effectif_{FAE}^{t}} * \mathbf{T_{divorce}} \quad (8)$$

$$\mathbf{Effectif_{FAE}^{t+1}} \mathbf{+= Effectif_{FSE}^{t}} * \mathbf{T_{naissance}} \quad (9)$$

Fig. 6. Algorithmic equations of the population projection model

**Algorithm : Choice of Suitable Housing list a t time**
Input : t:entier /*time*/
* If agent occupy a Housing
For each Exploration {rank 1, rank 2, rank 7, town}
- Get <Housing> list a t time
- For each Housing
  - Calculate score = RelocationPropensity*CurrentHousingQuality
  - If AttractionPotential > score
    - Add Housing to the list of Housings checking desirs of the agent
- If Housing List Size > 0
  - Return the Housing List
- Else got the next rank

* If agent don't occupy a Housing
Exploration of the entire town
- Get <Housing> list at t time
- For each Housing
  - Calculate score = RelocationPropensity*CurrentHousingQuality
  - If AttractionPotential > score
    - Add Housing to the list of Housings checking desirs of the agent
- If Housing List Size > 0
  - Return the Housing List
- Else got the next rank

Fig. 7. Algorithm of choice of suitable housing at t time.

## V. Results and Discussion

The simulation starts by parametring the number of households, housings and maps that represents initial data of the studied area. Real data of a Salé City could not be acquired entirely, so we will expose here some results of our model of residential mobility we have developed by simulating a growth population scenario that uses a small synthetic City. The number of housings and households is respectively 216 and 160; birth rate, departure rate, divorce rate, marriage rate and life expectancy have respectively 0.7, 0.3, 0.5, 0.6, and 70. A number of housings are vacant at the simulation beginning.

Households have features which form their distinct profiles and which subsequently affect their residential choices notably residential mobility and housing choice. Population evolves yearly in terms of household's number, using the demographic projection model based on statistical data. Calibration of household data had performed with a distribution of housing types by household categories similarly to [20].

From the maps at Figure 8, the initial transition matrix and the probability distribution are computed. We then compute transition matrix and probability distribution at next stage based on Markov chain principles , and use it subsequently to predict yearly land use change in terms of built up areas as done by the spatial prediction model.

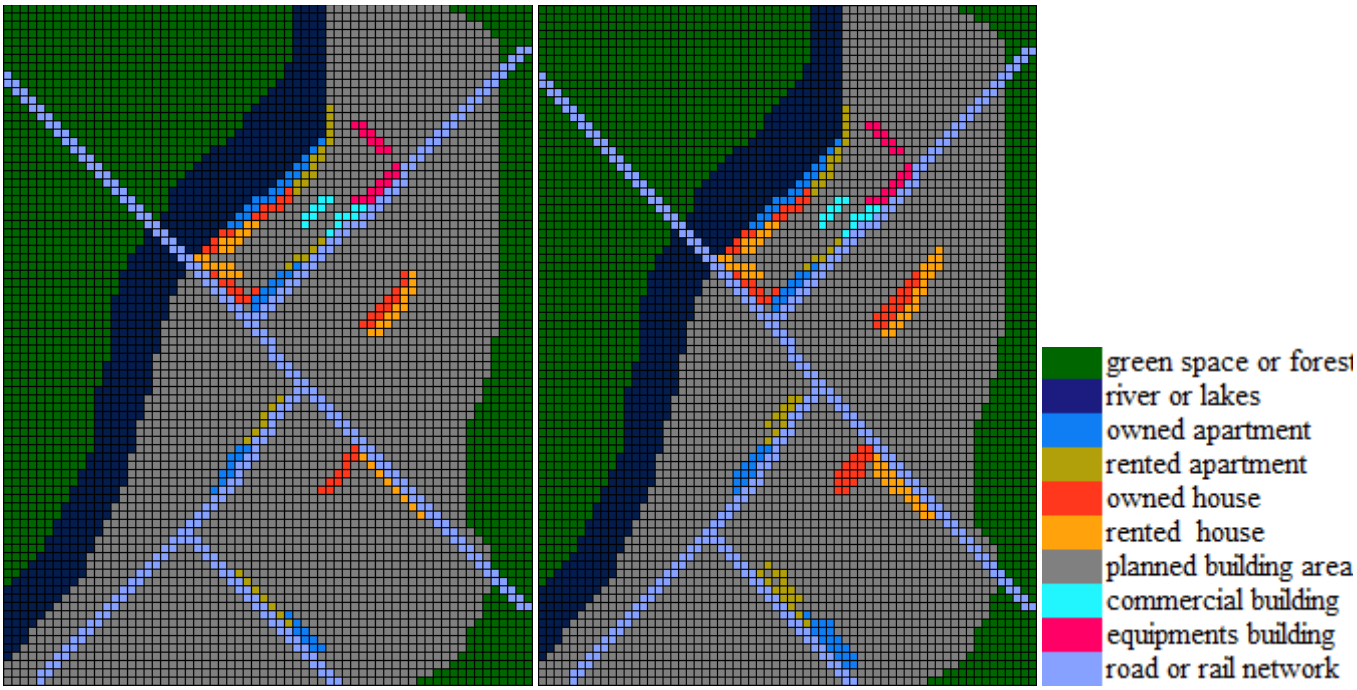
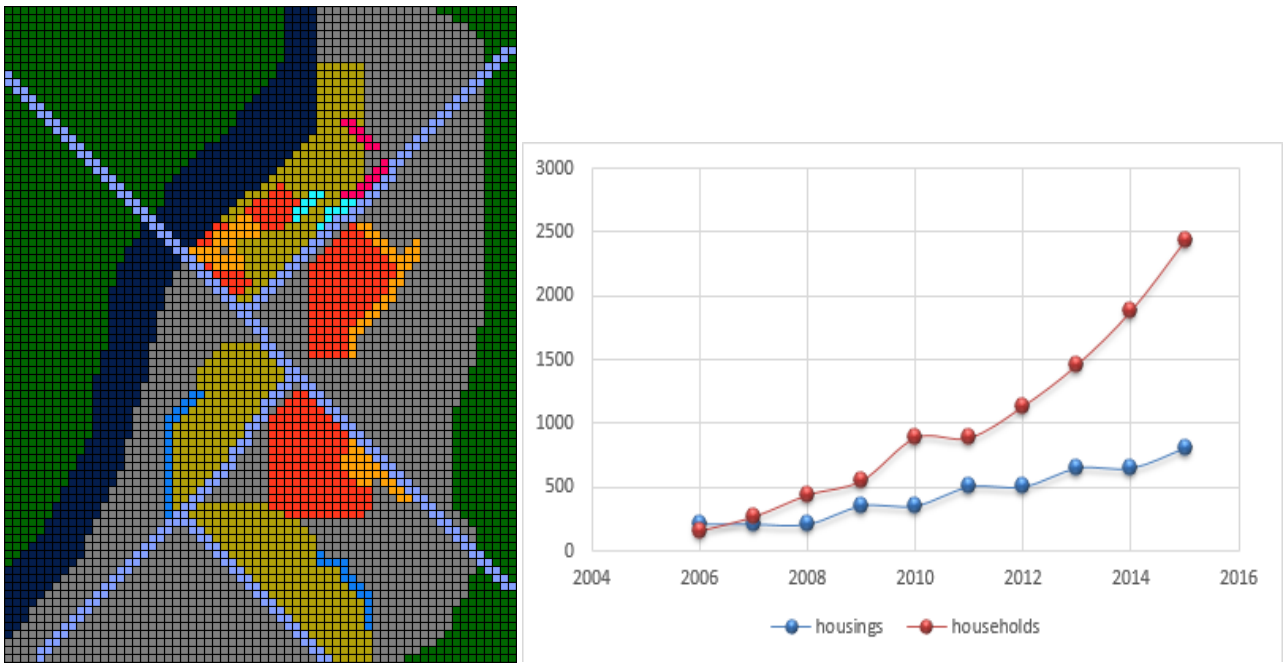



Fig. 8. City maps 2006 and 2007 (from left to right)




(a) (b)

Fig. 9. (a) Expected 2015 built-up areas, (b) housings and hiuseholds evolution

After development of built-up areas, households (including those generated by the population projection model) begin their residential choices. New generated Households search for vacant housings using the housing choice model that allows them to explore gradually housings in their neighborhoods. Other households verify their residential satisfaction, if they are not satisfied with their actual housing; they use the housing choice model to look for new residence. Thus, some households move to their new housing, while others do not because they could not find housings that correspond to their intention.

The model iterate yearly over a calendar of ten years. But lower duration could be also specified. Figure 9 expose the predicted built-up areas and the evolution of households and housings numbers. At the initial year of the simulation, we note that number of households is less than housings. Since the second year, in contrast to housings, household's number grow quickly. This result show a gap between population growth and constructed houses, which means that the annual built-up development rate is too slow regarding population growth rate. Such result, besides others ones that could be extracted by analyzing simulation outputs, could support decisions makers, to prepare effective urban decisions ( e.g., around housing needs, urban regeneration, households' energy consumption etc).

## VI. Conclusion and Perspectives

The proposed model could be used to give insights into the dynamics of residential mobility and the linked land use dynamic. We used an interesting alliance CA-MAS approach for designing and implementing the proposed model. In the Framework of the model, the City is considered as an emergent and auto-organized system where the global consequences are results of a set of individual behavior of the smallest units notably households and housings. The proposed model, as being presented, is exposed in functional and conceptual aspects in order to show it as entirely as possible. We presented also, the core of the proposed residential mobility model from an algorithmic view, which could bring also sufficient information on the implementation phase.

The evaluation of the model have taken place as it has developed. Two levels of evaluation were conducted: human evaluation focused on observing predicted land use maps; and an operational evaluation that focused on observing the input outputs of the model's internal variables as well as the progress of the simulation between the different time steps. The two-level evaluation demonstrated a significant degree of consistency between model assumptions and simulation results.

A third level of evaluation, of operational type, could be envisaged when acquiring all real data of an effective City. This type of evaluation could use confusion matrix and kappa index and play a support for Framework comparaison.

## References

[1] I. Benenson, «Agent-based modeling: From individual residential choice to urban residential dynamics,» Spatially integrated social science: Examples in best practice, vol. 42, pp. 67-95, 2004.

[2] V. Dabbaghian, P. Jackson, V. Spicer et K. Wuschke, «A cellular automata model on residential migration in response to neighborhood social dynamics,» Mathematical and Computer Modelling, vol. 52, pp. 1752-1762, 2010.

[3] V. Gaube et A. Remesch, «Impact of urban planning on household's residential decisions: An agent-based simulation model for Vienna,» Environmental modelling \& software, vol. 45, pp. 92-103, 2013.

[4] V. Gaube, A. Remesch et B. Smetschka, «Planning, Residential Decisions and Energy Use in Vienna,» chez Social Ecology, Springer, 2016, pp. 489-503.

[5] S. Fol, Y. Miot et C. Vignal, MOBILITÉS RÉSIDENTIELLES, TERRITOIRES ET POLITIQUES PUBLIQUES, Presses universitaires du Septentrion, 2014.

[6] R. Hegselmann et A. Flache, «Understanding complex social dynamics: A plea for cellular automata based modelling,» Journal of Artificial Societies and Social Simulation, vol. 1, p. 1, 1998.

[7] K. R. E. Elalaouy et M. D. Rahmani, «Cellular Automata: From Theoretical Concepts to Urban Modeling a New Methodology and Perspective,» Journal of Engineering and Applied Sciences,, vol. 13, pp. 4260-4266, 2018.

[8] E. Dubos-Paillard, Y. Guermond et P. Langlois, «Analyse de l'évolution urbaine par automate cellulaire. Le modèle SpaCelle,» L'espace g{\'e}ographique, vol. 32, pp. 357-378, 2003.

[9] H. S. Moghadam et M. Helbich, «Spatiotemporal urbanization processes in the megacity of Mumbai, India: A Markov chains-cellular automata urban growth model,» Applied Geography, vol. 40, pp. 140-149, 2013.

[10] Y. Du, W. Wen, F. Cao et M. Ji, «A case-based reasoning approach for land use change prediction,» Expert Systems with Applications, vol. 37, pp. 5745-5750, 2010.

[11] J. P. Antoni, V. Judge, G. Vuidel et O. Klein, «Constraint Cellular Automata for Urban Development Simulation: An Application to the Strasbourg-Kehl Cross-Border Area,» chez Geomatic Approaches for Modeling Land Change Scenarios, M. T. Camacho Olmedo, M. Paegelow, J. Mas et F. Escobar, Éds., Cham, : Springer International Publishing, 2018, pp. 293-306.

[12] J. Ji, L. Lu, Z. Jin, S. Wei et L. Ni, «A cellular automata model for high-density crowd evacuation using triangle grids,» Physica A: Statistical Mechanics and its Applications, 2018.

[13] B. K. Pandey et D. Khare, «Analyzing and modeling of a large river basin dynamics applying integrated cellular automata and Markov model,» Environmental Earth Sciences, vol. 76, p. 779, 2017.

[14] Z. Zheng, W. Huang, S. Li et Y. Zeng, «Forest fire spread simulating model using cellular automaton with extreme learning machine,» Ecological Modelling, vol. 348, pp. 33-43, 2017.

[15] S. Lauf, D. Haase et B. Kleinschmit, «Land-use scenario modelling based on human decisions--Combining system dynamics and cellular automata,» 2012.

[16] M. Wooldridge, An introduction to multiagent systems, John Wiley & Sons, 2009.

[17] V. Kocabas et S. Dragicevic, «Bayesian networks and agent-based modeling approach for urban land-use and population density change: a BNAS model,» Journal of geographical systems, vol. 15, pp. 403-426, 2013.

[18] J. Barth{\'e}lemy et T. Carletti, «An adaptive agent-based approach to traffic simulation,» Transportation research procedia, vol. 25, pp. 1238-1248, 2017.

[19] M. Durak, N. Durak, E. D. Goodman et R. Till, «Optimizing an agent-based traffic evacuation model using genetic algorithms,» chez Winter Simulation Conference (WSC), 2015, 2015.

[20] Dagmar Haase et R. Seppelt, «Modeling and simulating residential mobility in a shrinking city using an agent-based approach,» Environmental Modelling \& Software, vol. 25, pp. 1225-1240, 2010.

[21] H. Wang et C.-J. Chang, «Simulation of Housing Market Dynamics: Amenity Distribution and Housing Vacancy,» chez Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World, Piscataway, 2013.

[22] P. M. Torrens, «Cellular automata and multi-agent systems as planning support tools,» chez Planning support systems in practice, Springer, 2003, pp. 205-222.

[23] E. Elalaouy, K. Rhoulami et M. D. Rahmani, «Towards an agent based model for simulating residential mobility and urban expansion,» Proceedings of the Mediterranean Conference on Information \& Communication Technologies. Springer International Publishing, 2016. p. 343-351., 2015.

[24] E. Elalaouy, K. Rhoulami et M. D. Rahmani, «A Novel Modeling based Agent Cellular Automata for Advanced Residential Mobility Applications,» INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, vol. 8, pp. 337-343, 2017.

# Urdu Sentiment Analysis

Khairullah Khan[1], Atta Ur Rahman[3], Aurangzeb Khan[4], Ashraf Ullah Khan[6], Bibi Saqia[7]
Department of Computer Science
University of Science & Technology Bannu
Bannu, Pakistan

Wahab Khan[2]
Department of Computer Science and Software Engineering
International Islamic University, Islamabad

Asfandyar Khan[5]
Institute of Business and Management Sciences University of Agriculture Peshawar, Pakistan

***Abstract*—Internet is the most significant source of getting up thoughts, surveys for a product, and reviews for any type of service or activity. A Bulky amount of reviews are produced on daily basis on the cyberspace about online products and objects. For example, many individuals share their remarks, reviews and feelings in their own language utilizing social media networks such as twitter and so on. Considering their colossal Quantity and size, it is exceedingly knotty to look at with and interpret specified surveys. Sentiment Analysis (SA) aims at extracting people's opinion, felling and thought from their reviews in social websites. SA has recently gained significant consideration, however the vast majority of the resources and frameworks constructed so far are tailored to English as well as English like Western languages. The requirement for designing frameworks for different dialects is expanding, particularly as blogging and micro-blogging sites are becoming popular. This paper presents a comprehensive review of approaches of Urdu sentiment analysis and outlines of relevant gaps in the literature.**

***Keywords—Urdu; sentiment analysis; social media; survey***

## I. Introduction

In the near past about a decade, galore social networks are introduced while the already existed networks such as Twitter, Facebook and Instagram etc have blew up their presence on cyber space [1, 2]. The mentioned social networks have billions of users and yields a vast quantity of digital data, comprised of images, text, audio and video etc [2]. As indicated by [3], the evaluated measure of data on the web will be around 40 thousand Exabytes, or 40 trillion gigabytes, in 2020.

Online user comments are regularly utilized by individuals needing to know about different customer reviews about product or services of interest [4, 5]. These comments are frequently given in a free-text format by online business or social sites [4-6]. With the guide of this rich source of data individuals can certainly make a better decision [5, 6]. It is important to collect potential reviews in every field for good decision making and future prediction. Governments, Societies and organisations have an awesome enthusiasm for catching and collecting what individuals consider about a specific topic [1]. What the bigger portion of folks conceive Donald Trump, Facebook's procuring of WhatsApp and the fresh iPhone 10 are indispensable questions in the current age [1, 5].

Sentiment analysis can be used to solve such inquiries [1, 5]. Because SA is the process of investigating and extracting of individuals' remarks, audits and sentiments about a particular topic, an item, a news alert and a mobile application etc [1, 2, 5, 7]. The advantages of performing SA are countless [7]. It can help in measuring the public's opinion on controversial issues in a more precise, extensive and moderate form than open surveys [7]. It can likewise enable organizations to tailor their services, administration, product etc. to their clients' needs and in this way increment their benefits [7].

Abundant literature is available on strategies, difficulties, and utilization of SA for English language [1]. However, very little research has been led in SA of the Urdu language [8]. The SA designed for English language cannot be utilized for Urdu language having different script and morphological structure [8, 9]. Urdu has linguistic features that are diverse as well as contradicting to the English dialect [8, 9].

Like some other languages, Urdu online resource are in addition to getting popular with as individuals would like to partake impressions and convey opinions in their native languages [8, 9]. From Literature study it is clear that techniques employed for counterpart languages cannot be adopted for handling Urdu language issues [8, 9]. The main objective of this study is to report and investigate Urdu SA techniques and challenges.

The structure of the paper is organized as: section 2 describes SA in detail, in section 3 the recent trends and approaches of Urdu SA are provided, in section 4 literature review is provided, in section 5 Urdu SA challenges are described in detail, section 6 highlights SA techniques while in section 7 conclusion is provided.

## II. Sentiment Analysis

Sentiments are vital to every human activity since they are critical influencers of our practice. At whatever point we have to draw a conclusion, we need to experience other people' feelings. In reality, organizations and businesses need to know about consumer and public opinion about their services and products. Individual clients additionally need to know the opinion of existing consumers of an item before obtaining it, and others' opinions about political candidates in election campaign. When an individual required opinion about someone, he inquired about him from his friends and family members. The executive members of an organization conduct

external survey and organize polls in order to know public opinions to make a strategic decision.

Obtaining public and consumer opinions have long been an immense business itself for marketing, public relations, and political campaign.

With the development of web-based social networking (e.g. Twitter, Facebook, and YouTube etc.) on the Web are progressively utilizing the contents available on the social media to make a suitable strategic decision. These days, if one needs to purchase a product, one is no more restricted to review people opinion on the Web about that item. For an organization, it is not obligatory to carry on surveys, public opinion poll, and center groupings called for to get together masses' impression, because there is a lot of such kind of data openly accessible on the Web.

However, coming up and checking out opinion sources on the WWW and rectifying the data arrested in them persists an imposing errand in view of the proliferation of a different site. Each site commonly contains a large volume of sentiment message that isn't generally effortlessly decrypted in long in long sites and discussion postings. A moderate human reader will experience issues recognizing authoritative sites and pulling out and summing up the opinions in them. Automated SA frameworks are thus needed. SA application have spread to practically every possible domain, such as customer services, consumer product, financial and health care services to political campaign etc [10].

## III. Urdu Sentiment Analysis

This section describes the overview Urdu SA. Urdu language has unique feature and therefore some extra steps are needed for SA process. For example, Urdu script is written from right-to-left. The shape of alpha bits changes with change of its position in words. Urdu has different stop words. Some common grammatical mistakes are raised while dealing with stop words. For example کہ is confused with کے , پہ with پے and other rhyming words but having different meanings.

In Urdu writings the usage of space is not consistent and usually leads to either space insertion or space omission problem in words. The Space omission problem e.g. the Urdu word " انکا " which is in reality a compounding of cardinal words, but the system processes it as an individual word. The space Insertion problem e.g. the word "عقلمند " (Aqalmand, Intelligent) s in reality a single word but when tokenized , will be handled as compound word i.e "عقل" and "مند" which is addressed by a two-stage system.

The literature shows various SA techniques that have some common steps. The most commonly used steps are given in figure 1.

The first step is pre-processing. In this step a series of sub steps are carried out e.g. Noise Removal, Sentence Boundary Detection, Words Tokenization, and in some cases Part of Speech Tagging. Hence the related problems are needed to be handled in pre-processing step.

The next commonly used step is polarity identification i.e. for a word, sentence or text it to determine that it is positive, negative or neutral.

In order to prepare different models or algorithm successfully, it's essential to provide right training data to the models and furthermore the data must be sufficiently vast to prepare the model effectively [8]. The Urdu dataset can be collected from the commonly used Urdu News portal such as BBC Urdu, AAJ News, Abb Takk News, ARY News, *Geo* News, Samaa News, Dunya News, Dawn News, Express News, jang News etc

In the next phase tokenization of Urdu text is performed. Tokenization is the way toward separating the given content into units called tokens [8, 9]. The tokens might be words, numbers or punctuation marks. Tokenization does this assignment by finding word boundaries [8, 9]. When tokens are created from the sentences then it is passed to the polarity identification stage [8, 9]. Polarity of each word is determined by comparing it with sentiment lexicon [8]. Polarities of each word are determined as: Positive=1, Negative= -1, and Neutral=0 [8, 9].

When single polarities are assigned to each word then the combine polarity of the sentence is calculated [8, 9]. For example if a specific sentence has two positive words and one negative word, overall polarity would be determined as + 1(+2-1), consequently declaring it as a positive comment [8, 9]. The output demonstrates that the remark/feeling of the reviewer had a positive, negative or neutral conclusion towards the news, items or products [8, 9].

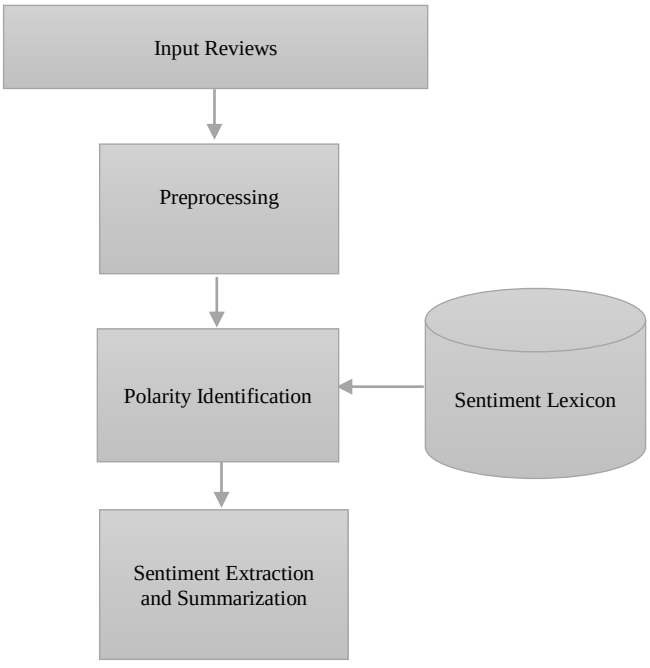



Fig. 1. Common Steps of Sentiment Analysis.

TABLE I. SUMMARY OF EXITING WORK ON URDU SENTIMENT ANALYSIS

| Authors | Year | Task | Model/Approch | Polarity | Data scope | Data set/source | Language |
|---|---|---|---|---|---|---|---|
| Afraz Z. Syed | 2010 | Lexicon Based SA | Classification | Pos/Neg | Urdu Web Forums | Movies and Products reviews | Urdu |
| Afraz Z. Syed | 2011 | Adjectival Phrases as the Sentiment Carriers | Classification | Pos/Neg | Urdu Web Forums | movies and electronic appliances | Urdu |
| Faiza Hahim | 2011 | Lexicon Based SA | Classification | Pos/Neg/Neutral | Urdu News Headlines | Product and Movie reviews | Urdu |
| Smruthi Mukund | 2011 | identify Opinion Entities | SVM/Kernels Method | N/A | Urdu News Headlines | BBC Urdu News Portal | Urdu |
| Smruthi Mukund | 2012 | Analyzing Urdu Social Media for Sentiments | SVM | Pos/Neg | Newswire data | cricket and movies | Urdu |
| Syed Afraz Z | 2014 | Identification and Extraction of Appraisal Expressions | Classification | Pos/Neg | Urdu Web Forums | Movies and Products reviews | Urdu |
| Misbah Daud | 2015 | Opinion Mining System | Machine Learning | Pos/Neg/Neutral | Roman Urdu | 1620 comments | Roman Urdu |
| S. Abbas Ali | 2016 | Salience Analysis of NEWS Corpus | Heuristic Approach | Pos/Neg | whole News except the heading | Urdu News Corpus | Urdu |
| Muhammad Bilal | 2016 | Sentiment classification | Classification: Models used are Naïve Bay, Decision Tree (DT) and KNN | Pos/Neg | Roman-Urdu and English | The model performance was evaluated on dataset of 150 positive and 150 negative reviews | Roman Urdu |
| A. Nazir | 2017 | Opinion Extraction | lexicon-based approach | Pos/Neg | Urdu Web Forums | 100,000-tagged words downloaded from http://www.cle.org.pk | Urdu |

## IV. LITERATURE REVIEW

There are galore practical application and sweetening on SA approaches that were advised over the latest couple of years. This study expects to give a more intensive look on these improvements and to summarize and classify a few articles displayed in this field as indicated by the different SA systems. The author has collected different articles which represent critical upgrades to the SA systems utilized for Urdu language. This paper covers a wide variety of SA fields published in the last few years for Urdu language processing. They are arranged by the objective of the article showing the algorithm, data sets and information utilized as a part of their work. This study can be helpful for new comer scientists in this field as it covers the most well-known SA procedures and applications in a single research paper. It talks about additionally new related fields in SA which have pulled in the analysts of late and their comparing articles.

Websites such as IEEE Explore, Springer Linker, Science Direct, ACM Portal and Googol Search Engine were utilized as a base for this exploration [11]. Diverse keyword terms such as ('Urdu ‘and ' Sentiment Analysis') and ('Urdu' and 'opinion mining') etc were searched for exploration. The articles displayed in this survey are outlined in Table 1.

## V. CHALLENGES OF SENTIMENT ANALYSIS IN URDU LANGUAGE

**Datasets:** There are rear datasets and corpora available to apply for SA in Urdu language. Data are generally collected from social networks and online forums, and newspapers.

**Lexicon:** In lexicon based SA we require a sentiment annotated lexicon of Urdu words constructed from a huge amount of text. The most suitable and easy way of text collection is online resources such as web blogs, social media sites, online news or electronic journals etc. Although Urdu is a very rich language its available resources on the internet are limited [12]. Furthermore, most of the data are Urdu websites are available in graphics/image format and hence are not easily retrievable. There are few publicly available corpus for Urdu sentiments and few lexicons have been created so far [9] and most of them are not openly available.

**Opinion** Target and Opinion Words Detection

For opinion mining and SA identification of opinion target is an important task […]. Noun has been most popularly employed for opinion target detection while adjectives have been employed as opinion words. In Urdu language we have problem e.g in the sentence

اسلام محبت اور امن کا درس دیتا ہے

The words اسلام, محبت and امن are nouns where اسلام is opinion target about which positive opinion is represented through the words محبت and امن. Although adjective is not available in this sentence it is still opinionated. Therefore, the opinion mining system must be able to classify nouns as opinion target and opinion words based on the sentence structure and opinion lexicon [..].

Another serious issue in Urdu language is that in day to day activities, online blogs and even in newspaper we observe English words while talking in Urdu. Similarly, Roman Urdu is most popularly used in most of the online platforms such Facebook, Twitter etc. In Such case traditional system fails to detect opinion words opinion targets from Urdu text. Hence the system should have additional preprocessing step such as to detect, clean and convert in normal text.

**Detecting opinion spam**

The issue of opinion spam is also observed in Urdu SA. Opinion spam is fake sentiment utilized for misguiding the users. This is typically done in organizations to advance or business to promote.

**Resolving co-reference**

Co-reference resolution also observed in Urdu SA, it is indistinct when the sentiments refer to multiple opinions.

**Feature extraction**

Urdu text is usually unstructured which make morphological analysers and POS taggers extremely difficult for Urdu language processing.

**Segmentation:** Segmentation issue can be further classified as a) Space-inclusion, b) Space-deletion issues. For example a single word can have a space in it, such as, " خوب صورت "(khoob surat, beautiful). On the other hands, space between two distinctive words can be deleted such as, "دستگیر"(dastgeer, benefactor) [9].

## VI. Sentiment Classification Techniques

The core sentiment classification approaches can be placed in three broad categories e.g. machine learning approach, lexicon based approach and hybrid approaches See Figure 2 [13].

The term "Machine Learning" is identified with enabling computer to learn training it first on pre labelled data that enables it to arrive at a prediction about test data that it may be given over later time [13].

In various research analysis the data sets are annotated manually, while some annotate it automatically by utilizing online lexicons such as SentiWordNet etc [13]. There are not many resources available for Urdu SA. Subsequently, the majority of the past research annotated the data manually. In SA, supervised ML is performed by giving the computer a set of phrases (features) and their polarity, giving the computer the capacity to predict the polarity of the unseen text [13, 14]. Unsupervised ML is utilized when it is hard to discover a labeled feature.

Supervised ML classifier can be divided into four principle classes, such as decision tree, linear, rule base and probabilistic classifiers [13]. Where linear classifier can be further classified into Support Vector Machine (SVM) and Neural Network (NN). Probabilistic classifier utilizes the probability of specific term/phrase to predict the polarity of unseen text. It can be further divided into Naive Bayes (NB), Bayesian Networks (BN) and Maximum Entropy (ME) classifier [13].

The lexicon based strategy is typically utilized when the data are unlabelled [1]. Lexicon are utilized to label the data and to predict the polarity of each word. In the Urdu dialect, only a few lexicon are available [1]. A few researcher have made Urdu lexicon, however, the greater part of these lexicons are not openly accessible [1].

The hybrid approach utilizes both lexicon and machine learning-based techniques [1]. This approach is more predominant in the current literature and have a higher efficiency as compared to lexicon based and machine learning techniques alone [1]. The lexicon scores are typically utilized as a features in the classifier [1].

In dictionary based approach a set of sentiment words is collected manually with their orientation [1]. This set is developed by searching in the WordNet or thesaurus for their equivalent words and antonyms. The recently discovered words are added to the seed list then the following emphasis begins [1]. The iterative process stops when no new words are found. After the access is finished, manual investigation can be performed to evacuate or correct errors [1].

The Corpus-based approach tackles the issue of finding opinion words with context specific orientation [1]. Its strategies rely upon syntactic examples or examples that happen together along with a seed list of opinion words to find other opinion words in a vast corpus [1]. Corpus based approach can be further classified into Statistical and Semantic approach. In a statistical approach the polarity of a word can be identified by finding the occurrence frequencies of the word in a corpus. Semantic technique gives similar sentiment values to semantically close words [1]. For example WordNet gives various types of semantic connections between words used to find sentiment polarities [1].

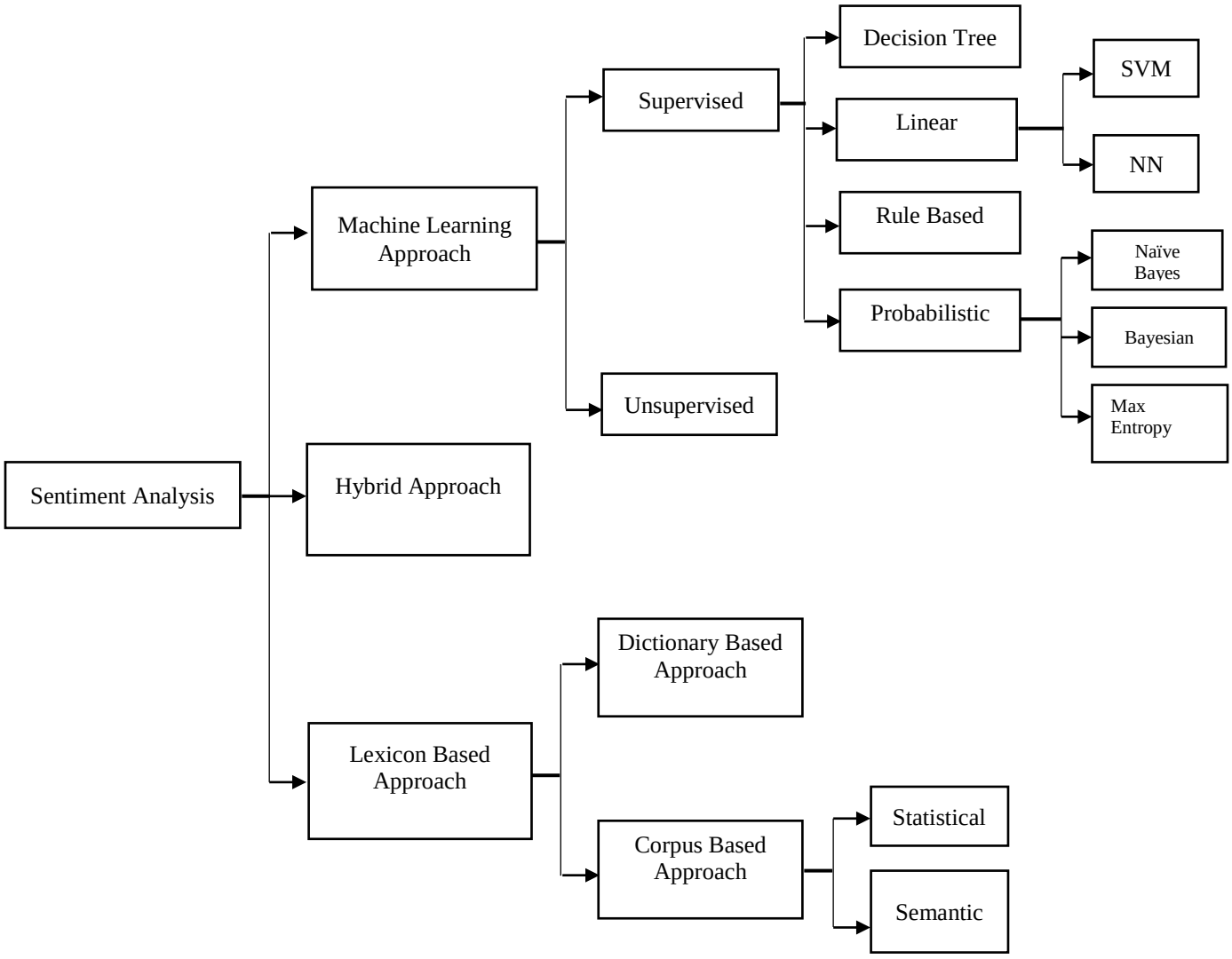



Fig. 2. Sentiment Classification Techniques.

## VII. Conclusion

It is evident from this study that a lot of potential strategies and methodologies are available but still little work is done on Urdu sentiments analysis. Table 1 summarizes all related articles to Urdu SA published up to date. Masses of clients share their opinions via web-based networking media, making it a significant platform for tracking and exploring public opinions. Online networking is one of the greatest stages where huge texts are published each day which makes it a perfect source for catching people opinions on different topics such as products, services and celebrities etc. The primary objective of this paper is to give an outline of most recent updates in SA and classification techniques utilized for Urdu language. Various improvements are possible in the field of Urdu SA, such as utilizing different data sources from multiple Urdu web forums. Another critical expansion is to exploit effective techniques that increase the use of online market and social media to extract sentiments from user reviews or comments for the purpose of improving products and services.

## References

[1] M. El-Masri, N. Altrabsheh, and H. Mansour, "Successes and challenges of Arabic sentiment analysis research: a literature review," Social Network Analysis and Mining, vol. 7, p. 54, 2017.

[2] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "Arabic Language Sentiment Analysis on Health Services," arXiv preprint arXiv:1702.03197, 2017.

[3] J. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," IDC iView: IDC Analyze the future, vol. 2007, pp. 1-16, 2012.

[4] T. Shivaprasad and J. Shetty, "Sentiment analysis of product reviews: A review," in Inventive Communication and Computational Technologies (ICICCT), 2017 International Conference on, 2017, pp. 298-301.

[5] M. E. Basiri, A. R. Naghsh-Nilchi, and N. Ghassem-Aghaee, "A framework for sentiment analysis in persian," Open Transactions on Information Processing, vol. 1, pp. 1-14, 2014.

[6] M. Korayem, D. J. Crandall, and M. Abdul-Mageed, "Subjectivity and Sentiment Analysis of Arabic: A Survey," in AMLTA, 2012, pp. 128-139.

[7] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on, 2013, pp. 1-6.

[8] A. Z. Syed, M. Aslam, and A. M. Martinez-Enriquez, "Lexicon based sentiment analysis of Urdu text using SentiUnits," in Mexican International Conference on Artificial Intelligence, 2010, pp. 32-43.

[9] Z. U. Rehman and I. S. Bajwa, "Lexicon-based sentiment analysis for Urdu language," in Innovative Computing Technology (INTECH), 2016 Sixth International Conference on, 2016, pp. 497-501.

[10] B. Liu, "Sentiment analysis and opinion mining," Synthesis lectures on human language technologies, vol. 5, pp. 1-167, 2012.

[11] K. Khan, B. B. Baharudin, and A. Khan, "Mining opinion from text documents: A survey," in Digital Ecosystems and Technologies, 2009. DEST'09. 3rd IEEE International Conference on, 2009, pp. 217-222.

[12] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "Urdu Named Entity Dataset for urdu Named Enity Recognition Task," in 6th International Conference on Language & Technology 2016, pp. 51-55.

[13] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 5, pp. 1093-1113, 2014.

[14] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "A survey on the state-of-the-art machine learning models in the context of NLP," Kuwait journal of Science, vol. 43, pp. 66-84, 2016.