IJCSIS Vol. 16 No. 3, March 2018
ISSN 1947-5500

International Journal of Computer Science & Information Security

Pennsylvania, USA

Indexed and technically co-sponsored by :

Indexing Service

IJCSIS has been indexed by several world class databases, for more information, please access the following links:

Global Impact Factor
http://globalimpactfactor.com/
Google Scholar
http://scholar.google.com/
CrossRef
http://www.crossref.org/
Microsoft Academic Search
http://academic.research.microsoft.com/
IndexCopernicus
http://journals.indexcopernicus.com/
IET Inspec
http://www.theiet.org/resources/inspec/
EBSCO
http://www.ebscohost.com/
JournalSeek
http://journalseek.net
Ulrich
http://ulrichsweb.serialssolutions.com/
WordCat
http://www.worldcat.org
Academic Journals Database
http://www.journaldatabase.org/
Stanford University Libraries
http://searchworks.stanford.edu/
Harvard Library
http://discovery.lib.harvard.edu/?itemid=|library/m/aleph|012618581
UniSA Library
http://www.library.unisa.edu.au/
ProQuest
http://www.proquest.co.uk

Zeitschriftendatenbank (ZDB)
http://dispatch.opac.d-nb.de/

IJCSIS

ISSN (online): 1947-5500

Please consider to contribute to and/or forward to the appropriate groups the following opportunity to submit and publish original scientific results.

CALL FOR PAPERS
International Journal of Computer Science and Information Security (IJCSIS)
January-December 2018 Issues

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.
See authors guide for manuscript preparation and submission guidelines.

Indexed by Google Scholar, DBLP, CiteSeerX, Directory for Open Access Journal (DOAJ), Bielefeld Academic Search Engine (BASE), SCIRUS, Scopus Database, Cornell University Library, ScientificCommons, ProQuest, EBSCO and more.

Deadline: see web site
Notification: see web site
Revision: see web site
Publication: see web site

Context-aware systems
Networking technologies
Security in network, systems, and applications
Evolutionary computation
Industrial systems
Evolutionary computation
Autonomic and autonomous systems
Bio-technologies
Knowledge data systems
Mobile and distance education
Intelligent techniques, logics and systems
Knowledge processing
Information technologies
Internet and web technologies, IoT
Digital information processing
Cognitive science and knowledge
Agent-based systems
Mobility and multimedia systems
Systems performance
Networking and telecommunications
Software development and deployment
Knowledge virtualization
Systems and networks on the chip
Knowledge for global defense
Information Systems [IS]
IPv6 Today - Technology and deployment
Modeling
Software Engineering
Optimization
Complexity
Natural Language Processing
Speech Synthesis
Data Mining

For more topics, please see web site https://sites.google.com/site/ijcsis/

For more information, please visit the journal website (https://sites.google.com/site/ijcsis/)

Editorial
Message from Editorial Board

*It is our great pleasure to present the **March 2018 issue** (Volume 16 Number 3) of the **International Journal of Computer Science and Information Security** (IJCSIS). High quality research, survey & review articles are proposed from experts in the field, promoting insight and understanding of the state of the art, and trends in computer science and technology. It especially provides a platform for high-caliber academics, practitioners and PhD/Doctoral graduates to publish completed work and latest research outcomes. According to Google Scholar, up to now papers published in IJCSIS have been cited over 10279 times and this journal is experiencing steady and healthy growth. Google statistics shows that IJCSIS has established the first step to be an international and prestigious journal in the field of Computer Science and Information Security. There have been many improvements to the processing of papers; we have also witnessed a significant growth in interest through a higher number of submissions as well as through the breadth and quality of those submissions. IJCSIS is indexed in major academic/scientific databases and important repositories, such as: Google Scholar, Thomson Reuters, ArXiv, CiteSeerX, Cornell's University Library, Ei Compendex, ISI Scopus, DBLP, DOAJ, ProQuest, ResearchGate, LinkedIn, Academia.edu and EBSCO among others.*

A great journal cannot be made great without a dedicated editorial team of editors and reviewers. On behalf of IJCSIS community and the sponsors, we congratulate the authors and thank the reviewers for their outstanding efforts to review and recommend high quality papers for publication. In particular, we would like to thank the international academia and researchers for continued support by citing papers published in IJCSIS. Without their sustained and unselfish commitments, IJCSIS would not have achieved its current premier status, making sure we deliver high-quality content to our readers in a timely fashion.

"We support researchers to succeed by providing high visibility & impact value, prestige and excellence in research publication." We would like to thank you, the authors and readers, the content providers and consumers, who have made this journal the best possible.

*For further questions or other suggestions please do not hesitate to contact us at **ijcsiseditor@gmail.com**.*

A complete list of journals can be found at:
http://sites.google.com/site/ijcsis/

IJCSIS Vol. 16, No. 3, March 2018 Edition

ISSN 1947-5500 © IJCSIS, USA.

Journal Indexed by (among others):

Bibliographic Information
ISSN: 1947-5500
Monthly publication (Regular Special Issues)
Commenced Publication since May 2009

Editorial / Paper Submissions:
IJCSIS Managing Editor
(ijcsiseditor@gmail.com)
Pennsylvania, USA
Tel: +1 412 390 5159

IJCSIS EDITORIAL BOARD

IJCSIS Editorial Board	IJCSIS Guest Editors / Associate Editors
Dr. Shimon K. Modi [Profile] Director of Research BSPA Labs, Purdue University, USA	**Dr Riktesh Srivastava [Profile]** Associate Professor, Information Systems, Skyline University College, Sharjah, PO 1797, UAE
Professor Ying Yang, PhD. [Profile] Computer Science Department, Yale University, USA	**Dr. Jianguo Ding [Profile]** Norwegian University of Science and Technology (NTNU), Norway
Professor Hamid Reza Naji, PhD. [Profile] Department of Computer Enigneering, Shahid Beheshti University, Tehran, Iran	**Dr. Naseer Alquraishi [Profile]** University of Wasit, Iraq
Professor Yong Li, PhD. [Profile] School of Electronic and Information Engineering, Beijing Jiaotong University, P. R. China	**Dr. Kai Cong [Profile]** Intel Corporation, & Computer Science Department, Portland State University, USA
Professor Mokhtar Beldjehem, PhD. [Profile] Sainte-Anne University, Halifax, NS, Canada	**Dr. Omar A. Alzubi [Profile]** Al-Balqa Applied University (BAU), Jordan
Professor Yousef Farhaoui, PhD. Department of Computer Science, Moulay Ismail University, Morocco	**Dr. Jorge A. Ruiz-Vanoye [Profile]** Universidad Autónoma del Estado de Morelos, Mexico
Dr. Alex Pappachen James [Profile] Queensland Micro-nanotechnology center, Griffith University, Australia	**Prof. Ning Xu,** Wuhan University of Technology, China
Professor Sanjay Jasola [Profile] Gautam Buddha University	**Dr . Bilal Alatas [Profile]** Department of Software Engineering, Firat University, Turkey
Dr. Siddhivinayak Kulkarni [Profile] University of Ballarat, Ballarat, Victoria, Australia	**Dr. Ioannis V. Koskosas,** University of Western Macedonia, Greece
Dr. Reza Ebrahimi Atani [Profile] University of Guilan, Iran	**Dr Venu Kuthadi [Profile]** University of Johannesburg, Johannesburg, RSA
Dr. Dong Zhang [Profile] University of Central Florida, USA	**Dr. Zhihan lv [Profile]** Chinese Academy of Science, China
Dr. Vahid Esmaeelzadeh [Profile] Iran University of Science and Technology	**Prof. Ghulam Qasim [Profile]** University of Engineering and Technology, Peshawar, Pakistan
Dr. Jiliang Zhang [Profile] Northeastern University, China	**Prof. Dr. Maqbool Uddin Shaikh [Profile]** Preston University, Islamabad, Pakistan
Dr. Jacek M. Czerniak [Profile] Casimir the Great University in Bydgoszcz, Poland	**Dr. Musa Peker [Profile]** Faculty of Technology, Mugla Sitki Kocman University, Turkey
Dr. Binh P. Nguyen [Profile] National University of Singapore	**Dr. Wencan Luo [Profile]** University of Pittsburgh, US
Professor Seifeidne Kadry [Profile] American University of the Middle East, Kuwait	**Dr. Ijaz Ali Shoukat [Profile]** King Saud University, Saudi Arabia
Dr. Riccardo Colella [Profile] University of Salento, Italy	**Dr. Yilun Shang [Profile]** Tongji University, Shanghai, China
Dr. Sedat Akleylek [Profile] Ondokuz Mayis University, Turkey	**Dr. Sachin Kumar [Profile]** Indian Institute of Technology (IIT) Roorkee

ISSN 1947 5500

TABLE OF CONTENTS

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

18. PaperID 28021851: Removal Based Improved Replication Control and Fault Tolerance Method for Road Traffic Data Management (pp. 141-146)

Dr Annaluri Sreenivasa Rao (1), Dr Attili Venkata Ramana (2), Dr. G. Suresh Reddy (3), Dr C.Sunil Kumar (4)
1) Assistant Professor in IT, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India,
2) Assoc. Prof in CS, Dept. of ECM, Sreenidhi Institute of Science and Technology, Hyderabad, India,
3) Professor, Dept of IT, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India,
4) Professor in ECM, Sreenidhi Institute of Science and Technology, Hyderabad, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

19. PaperID 28021853: AMSFLO: Optimization Based Efficient Approach for Association Rule Mining (pp. 147-154)

Vikram Rajpoot, Research Scholar, MGCGVV Chitrakoot, SATNA, India
Dr. Akhilesh Tiwari, Dept CS/IT, MITS Gwalior, Gwalior, India
Dr. Bharat Misha, Dept of Physical Science, MGCGV Chitrakoot, SATNA, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

20. PaperID 28021854: Decision Making in Uncertainty: A Bayesian Network for Plant Disease Diagnoses (pp. 155-160)

Jawed Naseem, SSITU, SARC, Pakistan Agriculture Research Council, Karachi, Pakistan
S. M. Aqil Burney, College of Computer Science & Information System, IoBM University, Karachi, Pakistan

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

21. PaperID 28021857: The Sharp Increase in Unmasking of Obtrusion into Internet of Things (IoT) IPV6 and IPV6 Low–power Wireless Personal Area Network (6LoWPAN), a Lead Way to Secure Internet of Things Services (pp. 161-167)

Modesta E. Ezema, Computer Science Department, University of Nigeria Nsukka, Enugu, Nigeria
Anthony O. Okwori, Computer Science Department, Federal University Wukari, Wukari, Nigeria
Francis A. Okoye, Computer Engineering Department, Enugu State University of Science and Technology Enugu, Enugu, Nigeria
Christopher C. Ezema, Victoria Island, Lagos Nigeria

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

22. PaperID 28021859: Crowd Conscious Internet of Things Enabled Smart Bus Navigation System (pp. 168-173)

Devi Mukundan, Department of Computer Science and Engineering, R.M.D. Engineering College, Kavaraipettai, Tamil Nadu, India.
Dr. P. Ezhumalai, Department of Computer Science and Engineering, R.M.D. Engineering College, Kavaraipettai, Tamil Nadu, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

23. PaperID 28021870: Knowledge Extraction by Applying Data Mining Technique to Use in Decision Making for Development Policy Making (pp. 174-180)

Md. Humayun Kabir, Dept. of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

24. PaperID 28021871: Parallel Evolutionary Algorithms for Feature Selection in High Dimensional Datasets (pp. 181-188)

Safa Adi, Computer and IT Department, Palestine Polytechnic University, Palestine
Mohammed Aldasht, Computer Engineering Department, Palestine Polytechnic University, Palestine

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

25. PaperID 28021873: Attribute-based Permission Model for Android Smartphones (pp. 189-198)

Abdulrahman A. Muthana (1), Abdulraqeb I. Al-Samei (2)
(1) Faculty of Computer Science and Information Systems, Thamar University, Yemen
(2) Computer Networks, Faculty of Computing & Information Technology, University of Science & Technology, Yemen

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

26. PaperID 28021876: BADEL WEB APPLICATION- An Online Web Portal for Donation and Exchangeable Resources (pp. 199-201)

Ms. Raheemunnisa Begum, Department of Information System, King Khalid University, Abha, K.S.A,
Manal Ali, Yasmin Hasan, Arwa Saad , Shahad Saeed, Atheer Khalid, Malak Saeed , bashayer Mohammed
Department of Information System, King Khalid University, Abha, KSA.

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

27. PaperID 28021877: Hybrid Cluster Based Congestion Aware (HCBCA) Approaches in Wireless Sensor Network (pp. 202-208)

V. Perumal, Department of Computer Science, Erode Arts & Science College (Autonomous), Erode, Tamilnadu, India.
Dr. K. Meenakshi Sundaram, Department of Computer Science, Erode Arts & Science College (Autonomous), Erode, Tamilnadu, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

28. PaperID 28021879: Implementation of OFDM System Using Various Channel Modulation Schemes (pp. 209-215)

Muhammad Zunnurain Hussain, Sr. Lecturer CS & IT Bahria University Lahore Campus

Muhammad Zulkifl Hasan, Sr. Lecturer CS & IT Lahore Garrison University, Lahore

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

29. PaperID 31011831: Information Technology in Animal Health Care System (pp. 216-220)

(1) Omankwu, Obinnaya Chinecherem; (2) Nwagu, Chikezie Kenneth; (3) Inyiama, Hycient
(1) Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
(2) Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria,
(3) Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria.

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

30. PaperID 31121728: Implementation of Secure Cloud data Storage –Data Transaction by Using an Orthogonal Hand Shaking Authentication Mechanism (OHSAM) (pp. 221-227)

Dr. K. Subramanian (#1) & M. Mohamed Sirajudeen (#2)
(#1) Department of Computer Science, H.H Raja College, Pudukottai, Tamil Nadu, India.
(#2) Department of Computer Science, J.J College of Arts and Science, Pudukottai, Tamil Nadu, India

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

31. PaperID 28021844: Analysis of Key Transmission Issues in Optical Wireless Communication for Indoor and Outdoor Applications (pp. 228-233)

Muhammad Waqas Khan, Mehr-e-Munir, Umar Farooq, Amjad Amer
Departmentof Electrical Engineering, Iqra National University, Peshawar, Pakistan

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

32. PaperID 28021880: Sentiment Analysis of Autistic Child by Emotional Advisor (pp. 234-241)

Muhammad Zunnurain Hussain, Senior Lecturer Bahria University Lahore, Campus
Muhammad Zulkifl Hasan, Senior Lecturer, Lahore Garrison University
Dr. Khadeeja Anjum, Demonstrator, Biochemistry Dept. CMH Lahore, Medical College Lahore
Dr. Arha Anwar, Registrar, Oral Medicine Dept. Rashid Latif Dental, College Lahore

Full Text: PDF [Academia.edu | Scopus | Scribd | Archive | ProQuest]

Further Investigation on OXLP: An Optimized Cross-Layers Protocol for Sensor Networks

Ahlam S. Althobaiti, Manal Abdullah
Taif University, King Abdul-Aziz University, Saudi Arabia

Abstract In wireless sensor networks (WSNs), ad hoc networks are WSNs without any prepositioning for the sensor nodes. In modern networks, a WSN is widely distributed to monitor physical or environmental conditions (e.g., temperature, sound).
Exceedingly large amounts of nodes as well as comparatively high node density, lead to scalability of the protocols used in WSNs. However, within a large-scale WSN, the routing process becomes challenging since nodes in this type of network have extremely limited resources for packet storage and routing table updates.
This paper contributes towards evaluating the performance of the Optimized Cross-Layers Protocol (OXLP) developed by the authors, focusing on its scalability. The OXLP protocol improves energy consumption over well-known protocols in the same field. Also, both the packet delivery ratio and packet delay reached a good level compared to other cross-layer based protocols.

Keywords WSNs, Cross Layer, MAC Protocol, Routing Protocol.

1 Introduction

Wireless sensor networks (WSNs) are generally comprised of several sensor nodes which are dispersed either inside or near a geographical location of interest with a view to detect, collect, and distribute data which is related to at least one parameter. In general, network demand for improvement is exponentially expanding with the increase in network dimensions [1]. Unlike traditional networks, WSNs have their own layout and resource limitations. The limitations of the design are dependent on the application and the monitored environment [2].

Cross layer design within a layered architecture is a protocol design which violates the boundaries of layered communication architecture [3]. The need for WSNs with energy-efficient communication generates from the severe limitations of battery-operated sensor nodes. Cross layer design involves layer interaction or architecture violation (e.g., merging layers, violating the layers' Open Systems Interconnections (OSI), forming new interfaces, providing further interdependencies between any two layers).

Robust and scalable protocols for the Internet have simply been designed by joining the layered protocol stack design and the independent layer static interfaces. Yet, in wireless ad-hoc networks, this combination is ineffective [4]. Statistically, the optimal performance for different network

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

parameters, like energy efficiency or delay, can be utilized by the inter-dependencies between different layers.

In general, a MAC protocol not only operates a shared medium's communication traffic, but also allows communication between sensor nodes by creating a fundamental network infrastructure for them. Therefore, it gives nodes the capability to self-organize and attempts to enforce network singularity by ensuring no collisions and errors during communication between the sender and receiver. By using the radio, MAC protocols are capable of efficiently conserving energy. The WSN's design objectives are fulfilled with the help of the MAC protocol since it determines how tasks are performed by nodes (e.g., radio utilization, channel sharing, collision avoidance, extension of lifespan). Therefore, many researchers continue to focus on designing unique solutions for WSN MAC protocols.

When a data transmission request is made, routing occurs. It is the process of selecting paths in a network to determine the best path (i.e., between source and destination). The network layer is used to implement this process on incoming data. In multi-hop based WSNs, the source node cannot directly reach the sink. Therefore, intermediate sensor nodes should relay the source packets to the next step until they reach the sink. However, the implementation of routing tables might be another possible solution. A routing table is defined as the task of the routing algorithm along with the help of the construction / maintenance routing protocol. A routing table covers the lists of node options for any given packet destination.

OXLP (An Optimized Cross-Layers Protocol) [5] is a cross-layer protocol characterized mainly by integrating the functionality of the MAC and network layers with a view towards inclusion of higher layers as well. The OXLP includes features from both the MAC and the network layers; whereas, it significantly reduces the energy consumption of nodes by increasing the sleep periods as much as possible and dealing with collision and control overhead. The OXLP focuses on system performance optimization by proposing a cross-layer protocol at the network/data-link layer for sensor networks. By combining concepts related to routing, access to the medium, and formation of clusters using reduced energy, the OXLP developed a scheme to enhance the lifetime of the network, packet delivery ratio, reduce energy, as well as the delivery delay to the base station (BS). The scheme depended on a collective approach which is supported by the proposed MAC scheme and integrated with an efficient routing protocol.

When designing an efficient WSN protocol, scalability is an essential factor that needs to be considered. Scalable protocols must be able to adapt to various network topologies. This implies that the protocols need to perform efficiently when there is an increase in network size or workload size.

This paper focuses mainly on evaluating the performance of the OXLP protocol with regard to scalability. The performance of the OXLP protocol was evaluated through simulations. Using MATLAB [6], simulator experiments were designed and implemented. The effectiveness of the protocol is demonstrated in terms of packet delivery ratio, network lifetime, delivery delay to the BS, and energy consumption for different traffic loads in the sensor network. The

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

scalability factor for the OXLP was also analyzed.

The rest of this work is organized as follows: Section 2 explains the related work, Section 3 discusses the OXLP protocol in details, Section 4 presents the simulation experiment design for evaluation of the OXLP protocol, and Section 5 concludes the paper.

2 Related Work

Schemes which are schedule-based have several advantages. These schemes are characterized as having minimal collisions, reduced overhearing, an evasion to idle listening, and the provision of a limited end-to-end delay. Since nodes should access the channel during their allocated time, the elevated average queuing delay is considered normal. Yet, these schemes are faced with several important concerns (e.g., overhead and extra traffic, reduced adaptability and scalability, and less throughput). Due to the difficulty of allocating conflict-free time-division multiple access (TDMA) schedules, researchers have focused their attention to MAC protocols based on TDMA [7].

While time division multiplexing was used to base several wireless MAC protocol designs on, large networks need global topology information which some of the designs may not have the scalability for [9] [10].

As proposed in [8], [9] and [10], research in cross layer protocol focuses on the MAC layer since resources can be inefficiently utilized when working with an individual layer. Recent work has combined cross-layer design with TDMA scheduling to lengthen the lifetime of the network.

The research in [11] combined MAC, physical and network layer optimization to compute the interference-free TDMA schedules. This was performed in networks with a relatively small size. The researchers also found a solution to the network lifetime's optimization problem in systems which are cross-layer based. In their research, they also employed the interior point method [12]. While not reusing single frames within the network ensure non-interference, it makes their approach not suitable for WSN's with large sizes because of the substantial end-to-end delay.

To ensure efficient energy schedules, the researchers in [13], proposed combining optimization of the layers with slot reuse. Their proposed model was an optimization of the convex cross-layer. Their model used iteration to allow the network's lifetime to reach its maximum. During every iteration, link schedules evolve until they either achieve a particular energy consumption objective or no optimal solution is reached.

While researchers have proposed various MAC protocols, there is still need for improvement with regard to optimizing system performance, such as optimizing the cross-layers as well as the cross-layers' interactions. Energy consumption can be reduced through the interaction of cross layers, which reduces every single layers' packet overhead. Despite effectively addressing static sensor node performance, current MAC protocol researchers lack sufficient information to compare them with mobile networks. Improving the MAC protocol can enhance not only the reliability of the communication, but also enhance energy consumption.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

3 OXLP: An Optimized Cross-Layer Protocol

This section explains the Optimized Cross-Layers Protocol (OXLP). It is based on both the MAC layer and the network layer; two adjacent layers that ensure the best performance for the sensor network.

3.1. System Model

3.1.1. Energy Model

The OXLP consists of a simple model for the energy consumption of radio electronics (as can be seen in Figure 1). In the model, in transmitting mode, both the radio hardware as well as the power amplifier run on energy that is consumed by transmitter nodes. While in receiving mode, the radio hardware is run by energy that is consumed by receiver nodes.

From Figure 1, let *k (bits)* represent packet size, and E_{elec} *(Joule/bit)* represent the consumed energy required to transmit or receive *k-bit* of data. Let ε_{amp} *(Joule /bit/m²)* represent the power amplifier's energy consumption in the transmitting mode. The energy consumed reaches an adequate level of energy to noise power in receiving mode. Radio dissipation occurs when the source node, *x,* which is *d* far from its destination, transmits a *k-bit* packet, as in Equations 1 and 2:

$$E_{Tx}(k,d) = E_{elec} * k + \varepsilon_{amp} * k * d^2 \quad (2)$$

Equations 3 and 4 express energy that is consumed by the radio in order to receive a *k-bit* packet:

$$E_{Rx}(k) = E_{Rx\_elec}(k) \quad (3)$$

$$E_{Rx}(k) = kE_{elec} \quad (4)$$

During every idle listening interval, the radio's consumption of energy can be expressed as Equation 5:

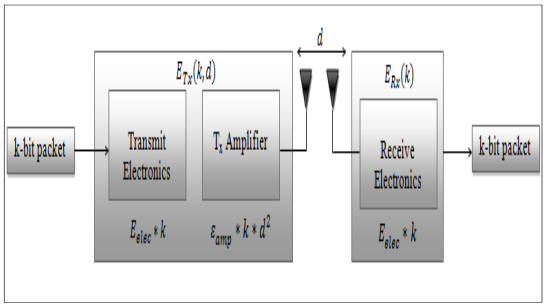
$$E_I(k) = \alpha E_{Rx}(k) \quad (5)$$

Where α is the ratio of the receiving mode's consumption of energy to idle listening interval's consumption of energy.

3.1.2. Overview

Based on joint functionalities of different underlying layers, the OXLP is a protocol which allows integrating the MAC protocol with the routing protocol for energy efficient delivery of data. The network layer utilizes information from the data link layer when the routes establish to efficiently access the medium, as shown in Figure 2. The forwarding process is composed of two phases: the MAC window and the transmission window.

Fig 1. Radio Energy Consumption Model.



Consequently, the total amount of energy consumed in the sensor network might actually be less when using the OXLP than when using direct transmission. To give more clarification, consider the linear sensor network, since the average distance between nodes is *s*.

Consider the energy consumed for the transmission of a single *k-bit* packet from nodes located within *hs* distance from the BS. From Equations 2 and 4, we have:

$$E_{Tx}(k, d = hs) = E_{elec} * k + \varepsilon_{amp} * k * (hs)^2 \quad (6)$$

$$E_{Tx}(k, d = hs) = k\left(E_{elec} + \varepsilon_{amp} h^2 s^2\right) \quad (7)$$

Where *h* is the number of hops, and *s* is the average distance between nodes.

In OXLP, all nodes send one another messages on their way to the cluster head (CH). CHs also send messages to each other

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

on their way to the BS. Therefore, nodes and CHs located within distance, *hs,* from their destination require *h* transmits a distance *s* and *h-1* receives.

$$E_{Tx}(k, d = hs) = hE_{Tx}(k, d = s) + (h-1) * E_{Rx}(k) \quad (8)$$

$$E_{Tx}(k, d = hs) = h\left(E_{elec} * k + \varepsilon_{amp} * k * s^2\right) + (h-1)kE_{elec} \quad (9)$$

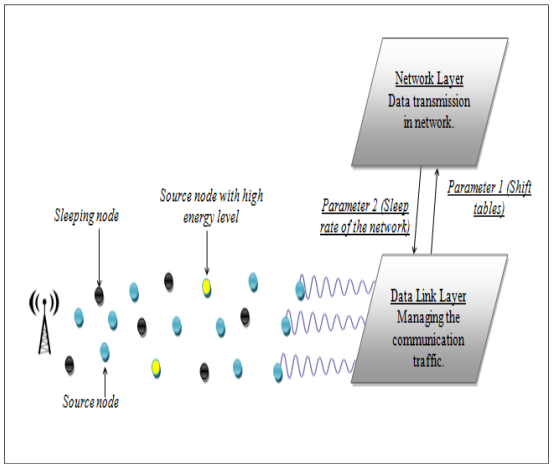
$$E_{Tx}(k, d = hs) = k\left((2h-1)E_{elec} + \varepsilon_{amp} * hs^2\right) \quad (10)$$

Where *h* is the number of hops, and *s* is the average distance between nodes.

3.2. OXLP Protocol

The OXLP process consists of a number of rounds. The MAC window is the beginning of the round, where it organizes the clusters and determines the routing paths. This is followed by the transition window, where data is transferred from the nodes, to the CHs, to the BS.

Fig 2. The Cross-Layer Optimized Framework.



3.2.1. MAC window

The MAC window introduces the core of the OXLP. The basic idea behind the MAC window is to integrate both the MAC and the routing mechanism. This solution allows the simultaneous planning of a proactive routing table and the medium access. The routing table will be maintained by each cluster head, in which each entry contains a destination ID, sender ID and allocated time slot. In doing so, three strong principles are presented:

1. Allocate the time slots in an efficient manner to avoid data collision, while it simultaneously shares the bandwidth resources among several sensor nodes within the entire network, in a fair and efficient manner.

2. In terms of network lifetime, the route of each message intended to the base station is selected in a crucial way.

3. Focus on increasing the sleep periods as much as possible, ensuring efficient awakening and avoid hidden and exposed terminal problems as proposed in [14].

The MAC window has the two following phases:

1. Cluster formation and cluster head selection.

2. Routing path determination and scheduling.

In the following sub-sections, we address the MAC window phases in detail.

3.2.1.1. Cluster Head Selection and Its Cluster Formation.

In OXLP, the CH selection phase apply the same mechanism that is used in the Cluster Status Protocol (CSP) sub-protocol in the admin nodes selection sub-section. as detailed in [14].

3.2.1.2. *Routing Path Determination and Scheduling*.

This phase determines the routing path for both the intra-cluster and inter-cluster communications. The determination of the shortest path from the sensor node to the corresponding CH, and from the CH to the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

BS is the responsibility of the MAC window, and uses Dijkstra's algorithm [15].

These tables can be used to determine appropriate transmission and destination and sleep schedules for all sensor nodes. The transformation from source nodes to BS can be done efficiently in a collision free manner by utilizing such information. To eliminate the need for a routing protocol, the shift tables themselves then serve to inherently form the routes through the sensor network.

3.2.2. Transmission Window

In the transmission window phase, the CH collects information from every sensor node in its cluster, and either directly transmits the information, or uses another CH to transmit it to the BS.

Possible modes for each sensor node in the transmission window are: transmit (T_X), receive (R_X), and sleep (SL). Therefore, every node executes adaptive protocol (AP) to decide the node's current mode (T_X, R_X, or SL). The process is based on the priorities of not only the current node, but also the announced schedules by the MAC window.

4 Simulation-Based Performance Evaluation

This section presents the OXLP's performance evaluation through simulation. MATLAB was used to design and implement the simulation [5] to investigate the OXLP's efficiency.

In the following sub-section, the research evaluates the performance of the OXLP and compares it against both cross-layer based protocols, which are found in the literature such EYES [16] and PLOSA [17], and routing protocols, such as LEACH [18].

4.1. Performance Metrics and Simulation Parameters

The following metrics are used to analyze the following mechanisms:

- **Packet Delivery Ratio (expressed in Percentage):** It is the ratio of packets received by the BS to the total sent.
- **Percentage Sleep Time (expressed in Percentage):** It is the ratio of sleeping slots to the network's total average slots.
- **Average End-to-End Delay (expressed in Milliseconds):** It is the time it takes to transmit a data packet
- **Energy Consumed (measured in Joule):** It is a measure of the rate at which energy is dissipated by sensor nodes in a WSN within a specific time period.
- **Network Lifetime (measured in Seconds):** It is the time since a node first runs energy in a network, until the time the last node (or group of nodes) in the network dies.
- **Control Packet Ratio (expressed in Percentage):** It is the ratio routing control packets sent by the protocol, to the total sent.

The research also assumes that the same amount of energy is needed to send k-bits from point A to point B and vice versa. Table 1 presents a summary of the parameters that were used in the MATLAB simulator. There parameters were chosen for two reasons: firstly, compared to other parameters (e.g., the LEACH), they have a higher metric impact. This allows the proposed method to be compared with other protocols presented in the literature. Secondly, and more importantly, the parameters chosen are common parameters in WSNs evaluations.

The proposed protocols are analyzed in terms of packet delivery ratio, lifetime of the network, delivery delay to the BS, consumed energy, and percentage sleep time, in case of the MAC mechanism for various traffic loads. Where a load is computed by computing the average amount of new packets in every slot,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

and is expressed as a function of λ, the inter-arrival period of messages for a node. By varying the value of λ in the research, the traffic load changes. If λ = 5s, every source node generates a message every 5s. In this research, the value of λ varies from 1 to 5 seconds. The network's channel reaches near full utilization at the highest rate, which is when λ is 1 second, which results in a low bandwidth.

4.2. Simulation Results for OXLP

Figure 3 (a) shows the average packet delivery ratio for the OXLP compared to the EYES protocol, PLOSA protocol and LEACH protocol.

In Figure 3 (b), it is observed that the network with OXLP has the longest lifetime, while that with the LEACH protocol had the shortest lifetime. This result was to be expected due to LEACH's energy consumption.

Figure 3 (c) shows end-to-end delay to the BS for the OXLP compared to the other cross-layer based protocols. The LEACH protocol had the highest delay. That is due to LEACH's limitation, which lies in the route discovery process, as well as the data packet Figure 3 (d) shows that redundant time slot allocations obviously exist in the EYES protocol and the PLOSA protocol, which cause more energy consumption than necessary.

4.3. OXLP Scalability

Scalability is a significant factor in this study and should be highlighted. According to the network growth or the workload, a scalable protocol develops itself to suit the changes in the network size. Mainly, experiments focused on the node density that are based on different performance metrics.

Table 1. Parameters Used in the Simulation.

Parameter	***Value***
Number of sensor nodes	*n= From 20 to 100*
Packet size	*k = 4000 bits*
Network Area	*A= M×M = 100×100*
GW-node Location	*Center BS (50,50) Corner BS (10,10)*
Communication model	*Bi-directional*
Transmitter/Receiver Electronics	E_{elec} *= 50 J/bit*
Initial energy for normal node	E_o *= 0.5 J*
Data aggregation energy	E_{DA} *= 5 nJ/bit/message*
Transmit amplifier	ε_{amp} *= 10 J/bit/m*^{2}

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

To analyze the OXLP's performance in regards to the scalability factor, some of the performance metrics in the OXLP [4] are used. For a WSN to have a longer lifetime, more nodes should be alive since the results are monitored based on parameter performance. However, when analyzing the OXLP, the protocol performance index is reflected by the network's lifetime.

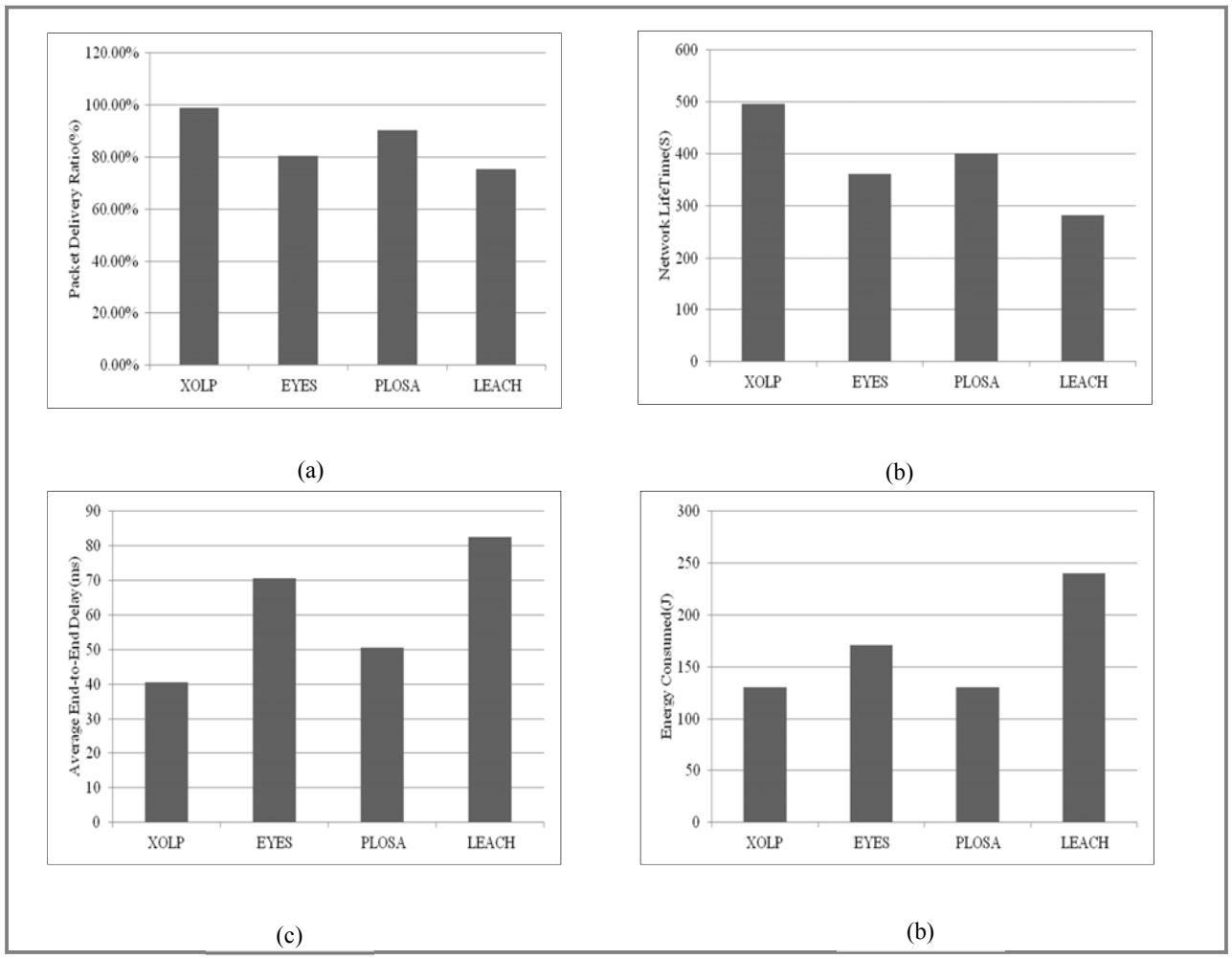
- *Alive Node vs. Network Lifetime*

WSNs demonstrate that the network application is impacted by active and monitor nodes. In addition, they have a limitation in battery-power; knowing when the node reaches a status called dead node, which is when its power level becomes less than the threshold or equal to zero.

Figure 4, presents the simulation results for the network's lifetime (the first node dies (FND) vs. a live node). It also shows that the network lifetime will decreases when the node density increases. Meanwhile, decreasing the node density from 1,000 to 100 nodes, increases the network's lifetime. Thus, the best network lifetime is reached when the density of the sensor node is smallest.

The OXLP disadvantage is that each node maintains a route structure to each different destination address.

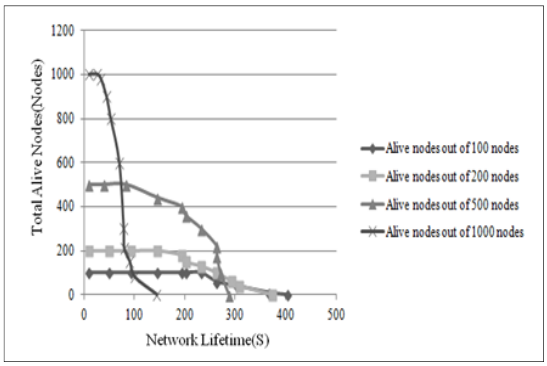
Figure 3. The Simulation Results for OXLP. (a) The Packets Delivery Ratio, (b) The Network Lifetime, (c) The Average End-to-End Delay, (d) The Energy Consumed.



https://sites.google.com/site/ijcsis/
ISSN 1947-5500

It also uses a lot of memory space, which hinders efficiency in large network sizes. It is clear that, with high density networks (1,000 nodes), the network lifetime quickly reaches zero. While with low density networks (100-200 nodes), it takes a long time for the network to die.

Fig 4. Alive Nodes vs. Network Lifetime for Different Node Density.



- *Data vs. Energy*

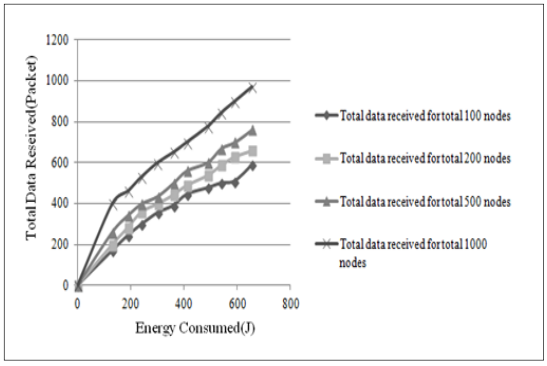
As shown in Figure 5, there is a relation between node density and the BS; where an increase in node density leads to an increase in the data received by the BS.

Moreover, the network, which has a minimum number of nodes, actually dissipates less consumption of energy with an acceptable amount of data that can be received by the BS. As can be seen from Figure 5, when the network has 1,000 nodes, there is more energy consumption with the maximum amount of the information. While, when the network has 100 nodes, less energy is consumed with minimum amount of data. Regardless, in WSNs, the OXLP is a preferred choice in case of increasing the dense network.

4.4. Comparison of WSN Protocols

This section compares results between the proposed cross-layer OXLP approach and other protocols, Table 2 shows that the EYES and PLOSA protocols have been optimized and perform low power consumption to ensure that a node has a several years for its lifetime on a single battery compared to the traditional approaches.

Fig 5. Data vs. Energy for Different Node Density.



The EYES protocol has a lifetime which is at least three times more than that of a SMAC network. This increase is found in dynamic networks. It has a better performance with mobile nodes. Static nodes have passive roles which do not alter. While, on the other hand, mobile nodes are forced to alter their roles as a result of the network changes. As such, it results in a more efficient and even consumption of energy between the nodes, which lengthens the network's lifetime. Therefore, the protocol reserves a standard amount of data for route updates. This space is wasted when the nodes are static.

On the other hand, in the PLOSA protocol, a frame's node access is distributed based on the node's distance to the collector for multi-hop mechanisms. A single frame can be used to complete the forwarding process. Additionally, the PLOSA protocol is

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

able to optimize a device's sleeping period since all of the nodes are able to receive packets to be dispatched only within a certain section of the frame. However, if two nodes send packets in parallel using PLOSA, one of the nodes packets will be delayed, and the node will enter sleep mode. Compared to other modes, a node spends most of its time in sleep mode. Whereby, micro-sensor network locally uses data aggregation in order to decrease the volume of transmitted information to reduce energy consumption as well as inactivity in data transfer. Furthermore, adapting the clusters in a micro-sensor approach depends on which of the nodes are that round's CH (e.g., LEACH). This is a beneficial process since it guarantees communication between nodes and their CHs which require the least transmission power. The LEACH protocol provides the required high performance for severe wireless channel constraints.

The performance of the proposed cross-layer approach was also compared against other cross-layer approaches, as shown in Table 2. Hence, OXLP improves energy conservation which performs high energy-efficiency in WSNs. It provides a longer lifetime for the network. It also uses an optimized MAC protocol based on TDMA, and uses short-dynamic wake-up packets instead of the long preambles. These packets carry the ID for the intended node. Moreover, the proposed method assumes that all nodes sleep while they are not scheduled to be active for sending or receiving data, according to the presented shift table. Hence, the shift table provides the data routing table which enables the nodes in a cluster to communicate based on their scheduled time slot without collision. The OXLP integrates both the MAC and routing mechanisms to create an optimized routing table for data transmission in the network clusters. However, the proposed OXLP increases the sleep states, reduces overhearing and overhead, and avoids the collision problem. It determines the shortest path routes from all sensor nodes to the corresponding CH in an intra-cluster and between CH nodes to the BS node in communications. Moreover, for effective adaptation to occur, network changes should be promptly and efficiently dealt with; the constraint of a node's lifetime and adding new nodes to a network as well as the changing intrusions may modify the

connectivity and topology of the network. The proposed OXLP results in a high delivery rate for data with very low delays, as seen in Table 2. The proposed approach has a limitation regarding finding the shortest path in cases with expanded network scalability. Therefore, the shortest path algorithm used may not apply to large network sizes as well as dynamic networks due to overwhelming additional work.

5. Conclusion

The main goal of researches in the field of WSN is to develop algorithms and protocols that ensure optimal performance; whether they use minimum energy consumption or have the longest network lifetime. Most of the existing solutions are based on a one-layer stack approach. However, recent work has concentrated on utilizing multiple layers for optimizing the network performance.

An Optimized Cross-Layers Protocol (OXLP) is developed to provide an efficient communication method for WSNs. The protocol utilizes adjacent layers (i.e., the MAC layer and the Network layer) to enhance the overall performance of the WSN.

In this work, the performance of the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

OXLP is measured and is determined to have effective protocol scalability. The simulation analysis concentrated on the nodes' energy limitations. The results between comparing node density simulations proved the scalability effect on the lifetime of the sensor.

Table 2. Comparison of WSNs Protocols.

WSNs Protocol	Time Sync Needed	Type	Advantages	Disadvantages
EYES Protocol	No	CSMA, Contention -based	– The nodes are mobile. Therefore, they are forced to alter their roles when dynamic changes occur in the network.	– It provides low efficiency in a static network. For mobile nodes, a standard amount of data is reserved to store route updates. On the other hand, this is wasted space for static nodes.
PLOSA	No	Slotted Aloha	– Energy consumption is limited due to a low packet loss rate. The transceiver of sensor nodes not in use enter low power sleep mode.	– Transmission is delayed when another node sends a packet, and it will enter sleep mode. Using this protocol, nodes spend most of their time in sleep mode. –
LEACH	Yes	TDMA/CSMA	– Clusters are adapted based on which nodes are CHs. This guarantees communication between nodes and the CH that needs the least transmission power. – This protocol provides the	– The possibility of utilizing more transmission power due to the use of fixed clusters as well as the rotation of a cluster's CHs.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

			required high performance for the severe wireless channel constraints by utilizing data aggregation which diminishes energy consumption and inactivity in data transfer.	
OXLP Protocol	Yes	TDMA Wake-up Packet	- It determines the shortest path routes from all sensor nodes to the corresponding CH in an intra-cluster and between CH nodes to the BS node in communications. Moreover, for effective adaptation to occur, network changes should be promptly and efficiently dealt with; the constraint of a node's lifetime and adding new nodes to a network as well as the changing intrusions may modify the connectivity and topology of the network. It increases the sleep states, reduces overhearing and overhead, and avoids the collision problem.	- Like the shortest path scheme, it may have limitations in cases where the network expands as the size of the network increases.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Regarding the simulation results for the OXLP, the routing scheme for cross layers-based protocols that was used in comparison to the OXLP are a functionality-oriented routing algorithm. The performance of these routing algorithms ignores energy consumption in nodes and in information transmission.

The OXLP was found to be energy efficient and increase the network's lifetime. Yet, like shortest path scheme, it does have disadvantages. The OXLP may have limitations in some cases where the network expands as the size of the network increases.

References

[1] TDMA scheduling with maximum throughput and fair rate allocation in wireless sensor networks – IEEE Xplore Document. (n.d.). Retrieved 26 May 2017, from http://ieeexplore.ieee.org/abstract/document/6654739/

[2] Espes, D., Lagrange, X., & Suárez, L. (2015). A cross-layer MAC and routing protocol based on slotted aloha for wireless sensor networks. *Annals of Telecommunications - Annales Des Télécommunications*, *70*(3–4), 159–169. https://doi.org/10.1007/s12243-014-0433-8

[3] Cross layer architectural approaches for Wireless Sensor Networks - IEEE Xplore Document. (n.d.). Retrieved 26 May 2017, from http://ieeexplore.ieee.org/abstract/document/6069374/

[4] Goldsmith, A. J., & Wicker, S. B. (2002). Design challenges for energy-constrained ad hoc wireless networks. *IEEE Wireless Communications*, *9*(4), 8–27. https://doi.org/10.1109/MWC.2002.1028874

[5] Althobaiti, A. S., & Abdullah, M. (2016). OXLP: An optimized cross-layers protocol for wireless sensor networks. In *Communication, Management and Information Technology* (Vols 1–0, pp. 659–670). CRC Press. Retrieved from http://dx.doi.org/10.1201/9781315375083-94

[6] MATLAB - MathWorks. (n.d.). Retrieved 26 May 2017, from https://www.mathworks.com/products/matlab.html

[7] Ye, W., & Heidemann, J. (2004). Medium Access Control in Wireless Sensor Networks. In C. S. Raghavendra, K. M. Sivalingam, & P. T. Znati (Eds.), *Wireless Sensor Networks* (pp. 73–91). Springer US. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4020-7884-2_4

[8] Shi, L., & Fapojuwo, A. O. (2010). TDMA Scheduling with Optimized Energy Efficiency and Minimum Delay in Clustered Wireless Sensor Networks. *IEEE Transactions on Mobile Computing*, *9*(7), 927–940. https://doi.org/10.1109/TMC.2010.42

[9] Ramaswami, R., & Parhi, K. K. (1989). Distributed scheduling of broadcasts in a radio network. In *Technology: Emerging or Converging, IEEE INFOCOM '89. Proceedings of the Eighth Annual Joint Conference of the IEEE Computer and Communications Societies* (pp. 497–504 vol.2). https://doi.org/10.1109/INFCOM.1989.101493

[10] TDMA scheduling algorithms for wireless sensor networks | SpringerLink. (n.d.). Retrieved 26 May 2017, from https://link.springer.com/article/10.1007/s11276-009-0183-0

[11] Cui, S., Madan, R., Goldsmith, A. J., & Lall, S. (2007). Cross-Layer Energy and Delay Optimization in Small-Scale Sensor Networks. *IEEE Transactions on Wireless Communications*, *6*(10), 3688–3699.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

https://doi.org/10.1109/TWC.2007.060072
[12] Boyd, S., & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.
[13] Madan, R., Cui, S., Lall, S., & Goldsmith, N. A. (2006). Cross-Layer Design for Lifetime Maximization in Interference-Limited Wireless Sensor Networks. *IEEE Transactions on Wireless Communications*, *5*(11), 3142–3152. https://doi.org/10.1109/TWC.2006.04770
[14] Energy Efficient with Collision Free MAC Protocol for Wireless Sensor Network | Althobaiti | Journal of Telecommunication, Electronic and Computer Engineering (JTEC). (n.d.). Retrieved 26 May 2017, from http://journal.utem.edu.my/index.php/jtec/article/view/1735
[15] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, *1*(1), 269–271. https://doi.org/10.1007/BF01386390
[16] Hoesel, L. van, Nieberg, T., Wu, J., & Havinga, P. J. M. (2004). Prolonging the lifetime of wireless sensor networks by cross-layer interaction. *IEEE Wireless Communications*, *11*(6), 78–86. https://doi.org/10.1109/MWC.2004.1368900
[17] Espes, D., Lagrange, X., & Suárez, L. (2015). A cross-layer MAC and routing protocol based on slotted aloha for wireless sensor networks. *Annals of Telecommunications - Annales Des Télécommunications*, *70*(3–4), 159–169. https://doi.org/10.1007/s12243-014-0433-8
[18] Heinzelman, W. B., Chandrakasan, A. P., & Balakrishnan, H. (2002). An application-specific protocol architecture for wireless microsensor networks. *IEEE Transactions on Wireless Communications*, *1*(4), 660–670. https://doi.org/10.1109/TWC.2002.804190

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

AN INTELLIGENT APPROACH FOR EFFECTIVE RETRIEVAL OF CONTENT FROM LARGE DATA SETS BASED ON BIVARIATE GENERALIZED GAMMA MIXTURE MODEL

[1]Subash Chandra Chadalavada, [2]Srinivas Yarramalle

[1]Associate Professor, Department of Computer Science & Engineering,

Kakinada Institute of Engineering & Technology, Andhra Pradesh,, India

[2]Professor and Head of the Department, Department of Information Technology,

GITAM Institute of Technology GITAM,Visakhapatnam, Andhra Pradesh, India

E-mail: [1]subash.chadalavada@gmail.com, [2]sriteja.y@gmail.com

ABSTRACT:

The advents in this technological era have resulted into enormous pool of information. This information is stored at multiple places globally, in multiple formats. This article highlights a methodology for extracting the video lectures delivered by experts in the domain of Computer Science by using Generalized Gamma Mixture Model. The feature extraction is based on the DCT transformations. In order to propose the model, the data set is pooled from the YouTube video lectures in the domain of Computer Science. The outputs generated are evaluated using Precision and Recall.

Keywords: *Bivariate Generalized Gamma Mixture Model, DCT Transformations, Feature Extraction, YouTube, Performance Evaluation*

.

1. INTRODUCTION

Today lot of information is available for academic purposes along with the other areas of interest globally in the form of videos. this data is mostly available in heterogeneous form. therefore, identification of relevant and appropriate video source is a challenging task. lot of literature available in this area for retrieving information from the web resources at an optimal speed and minimal accuracy. however, as the data is heterogeneous, acquiring relevant information is exponentially a challenging task. therefore, many models have been proposed to retrieve the relevant information based on the content, features, visual information, text, audio. [1] [2] [3] [4] [5].

As the data is stored from across the globe, it is next to impossible to retrieve the relevant source of information from the data reposited. however, to overcome this challenge, feature vectors play a vital role. lot of feature extraction methodology are proposed in the literature. [6] [7] [8] [9]. along these methodologies, methods based on shape, size, text, content, voice are more significant for the present study. in order to maximize the retrieval accuracy in the present article, we have considered discrete cosine transformation (dct) for extracting features in an efficient manner. the main advantage of dct is that it reduces the dimension and also helps to highlight the appropriate features that is core necessary for the identification of the relevant information to be retrieved. another reason behind the choice of dct is that, it is robust and helps to extract the features more efficiently. dct coefficients, guarantees the maintenance of regularity complexities as it formulates the feature vector by using the orthogonal transformation of the cosine function. another advantage of dct is that itconverts from time domain to frequency domain and helps in extracting the change in speech signals that are core necessary for extracting the lectures. hence, the choice behind the usage of dct is justified. along with dct another feature vector that is considered in this article is linear predictive coding (lpc). the lpc helps to underline the speech signals having very low altitude and hence, in this article, a statistical mixture model based bivariate generalized gamma mixture model is proposed by considering the above two features.

The rest of the paper is organized as follows,. section 2 of the paper highlights about the relevant literature carried out in this area. in section 3 of the article, the bivariate generalized gamma mixture

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

model is presented. section 4 of the paper describes about the data set considered and feature extraction methodology using dct & lpc is proposed in corresponding section 5. the methodology of the proposed model together with experimental results is highlighted in section 6 and the results derived together with performance evaluation are demonstrated in section 7. the paper is concluded with summarization in section 8.

2. RELEVANT LITERATURE

André araujo, jason chaves [2016] presented a research on "large-scale query-by-image video retrieval using bloom filters" they considered the problem of using image queries to retrieve videos from a database. their main contribution is a framework based on bloom filters, which can be used to index long video segments, enabling efficient image-to-video comparisons. they showed that a straightforward application of bloom filters to their problem, using global image descriptors, obtains limited retrieval accuracy. their best-performing scheme adapts the bloom filter framework: the key is to hash discriminative local descriptors into scene-based signatures. the techniques are evaluated by considering different hash functions and score computation methods. large-scale experiments showed that their system achieves high retrieval accuracy and reduced query latency.

Markus mühling[2016]presented a researchon "content-based video retrieval in historical collections of the german broadcasting archive " an automatic video analysis and retrieval system for searching in historical collections of gdr (german democratic republic) television recordings. it consists of video analysis algorithms for shot boundary detection, concept classification, person recognition, text recognition and similarity search.novel algorithms for visual concept classification, similarity search, person recognition and video ocr have been developed to complement human annotations and to support users in finding relevant video shots.

Sungeunhong[2017] examined that "content-based video–music retrieval using soft intra-modal structure constraint" is a new content-based, cross-modal retrieval method for video and music that is implemented through deep neural networks. they train the network via inter-modal ranking loss such that videos and music with similar semantics end up close together in the embedding space.they introduced vm-net, a two-branch deep network that associates videos and music considering interand intra-modal relationships.they showed that inter-modal ranking loss widely used in other cross-modal matching is effective for the cbvmr task.

Cees g.m. snoek [2017] presented "tag-based video retrieval by embedding semantic content in a continuous word space" presented a technique to overcome this gap by using continuous word space representations to explicitly compute query and detector concept similarity.they presented a novel "continuous word space" (cws) video embedding framework for retrieval of unconstrained web videos using tag-based semantic queries. they evaluated the retrieval performance of these three methods on the challenging nist medtest2014 dataset.

Andr´e araujo and bernd girod[2017] presented "large-scale video retrieval using image queries" is about retrieval of videos from large repositories using image queries is important for many applications, such as brand monitoring or content linking. they introduced a new retrieval architecture, where the image query can be compared directly to database videos – significantly improving retrieval scalability, compared to a baseline system that searches the database on a video frame level.they introduced a new comparison technique for fisher vectors, which handles asymmetry of visual information. the basic idea is to carefully select the types of visual information to use in such comparisons, efficiently ignoring clutter that is typical in this case. experimental results demonstrate up to 25% map improvement for two types of asymmetry.

Shishiqiao[2016] presented "deep video code for efficient face video retrieval" to address the problem of face video retrieval.they proposed a novel deep video code (dvc) method which encodes face videos into compact binary codes.they proposed a multi-branch cnn architecture, which takes face videos as inputs and outputs compact binary codes.they owed it to two aspects: first, the integration of frame-level non-linear convolutional feature learning, video-level modeling by temporal feature pooling and hash coding for extracting compact video code. second, the optimization of a smooth upper bound on triplet loss function for hash learning.

Gabriel de oliveira barra,mathias lux and xavier giro-i-nieto [2016] presented "large scale content-based video retrieval with livre" presentslivre, an extension of an existing open source tool for image retrieval to support video indexing.livre consists of three main system components (pre-processing, indexing and retrieval), as well as a scalable and responsive html5 user interface accessible from a web browser.livre supports image-based queries, which are efficiently matched with the extracted

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

frames of the indexed videos.adaptations were done in three main components focusing on the aspects of parsing, indexing and retrieval. besides the implementation they presented an evaluation using a large videodataset with more than 1000 hours of video.

Juca rossetto, st´ephanedupont and metinsezgin [2015] presented "imotion — a content-based video retrieval engine" is a sketch-based video retrieval engine supporting multiple query paradigms.they had presented the imotion system, a content-based video retrieval engine for known-item searches using exemplary images or sketches.since the imotion system was developed to support a wide variety of different kinds of video and implements many diverse features (both low-level and high level) and query paradigms that can be flexibly combined.

M.ravinder and dr.t.venugopal [2016] researched on "content based video indexing and retrieval using key frames discrete wavelet center symmetric local binary patterns (dwcslbp)" is algorithm is applied on a dataset of three hundred and thirty five videos.in which one hundred and forty eight videos are of air-plane type, seventy two videos are of boat type, eighty videos are of car type, and thirty five videos are of war tank type (which are collected from google, bbc, and trecvid 2005).a novel algorithm have been proposed based on discrete wavelet center symmetric local binary patterns, which is useful for content based video indexing and retrieval.they proposed algorithm is applied on a challenging dataset

Klaus schoeffmannet al [2016] presented "content-based retrieval in videos from laparoscopic surgery" to use feature signatures, which can appropriately and concisely describe the content of laparoscopic images, and showed that by using this content descriptor with an appropriate metric, they are able to efficiently perform content-based retrieval in laparoscopic videos. their approaches utilize feature signatures based on low dimensional feature spaces in order to efficiently describe the endoscopic images. they presented different signature-based approaches for content-based video retrieval in recordings from laparoscopic surgery. the signature matching distance allows for video retrieval with high performance already with small-sized feature signatures, which are much faster to compare than larger ones.

Priya singh and sanjeev ghosh [2017] researched on "content based video retrieval using neural network" is based on content fingerprinting and artificial neural network based classification. firstly, the fingerprint extraction algorithm is employed which extracts a fingerprint through the features from the image content of video. these images are represented as temporally informative representative images (tiri). then, the second step is to find the presence of videos in a video database having content similar to that of query video. multi layer feed forward (mlf) neural network that uses back propagation algorithm for training is used for video retrieval. they briefly reviewed the need and significance of video retrieval systems and explain their basic building stages. the first step is feature extraction, here which is extracted using video fingerprinting using tiri-dct algorithm. feature matching is then performed using neural network. for video retrieval a feed forward multilayer network is used and makes use of back propagation algorithm for training.

Shoou-i yu, etal [2015] presented "content-based video search over 1 million videos with 1 core in 1 second." a system which can search 1 million videos with 1 core in less than 1 second while retaining 80% of the performance of a state-of-the-art cbvs system. this potentially opens the door to content-based video search on web-scale video repositories.finally, they proposed system relies on 3 semantics-based features, which enabled them to significantly lower the amount of bytes required to represent each video.

B. m¨unzer ,etal [2017] researched on "when content-based video retrieval and human computation unite: towards effective collaborative video search" took the best from both worlds by combining an advanced content-based retrieval system featuring various query modalities with a straightforward mobile tool that is optimized for fast human perception in a sequential manner.new concept of collaborative video search, which combines the advantages of content-based retrieval and human computation through information exchange about the search status. thus, they conclude that it is more effective to perform a re-ranking based only on explicit input of the expert user who operates the cbvr tool. for now, they only considered a known-item search scenario, but in future work they intended to apply their approach also for ad-hoc search, which allows multiple correct answers.

Lu jiang shoou-i yu, deyumeng, and yi yang [2015], researched on "fast and accurate content-based semantic search in 100m internet videos" a scalable solution for large-scale semantic search in video.they introduced a novel adjustment model that is based on a concise optimization framework with solid interpretations. they also discussed a solution that leverages the text-based inverted index for video retrieval. experimental results validated them efficacy and the efficiency of the proposed method

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

on several datasets. specifically, the experimental results on the challenging trecvid med benchmarks validate the proposed method is of state-of-the-art accuracy.

Aasif ansari and muzammil h mohammed [2015] presented "content based video retrieval systems - methods, techniques, trends and challenges" a complex and wide area of cbvr and cbvr systems has been presented in a comprehensive and simple way.processes at different stages in cbvr systems are described in a systematic way. various querying methods, some of the features like glcm, gabor magnitude, algorithm to obtain similarity like kullback-leibler distance method.using a complete video shot yields better result than that using a key frame representing a shot whereas, system using a query clip is superior than that using a single shot instead. search based on textual information of the video can also be used in cbvr systems.

Narayan Subudhi et al., 2015 exemplify that innumerable methods for image segmentation are available in the literature. Intense research in the recent past has mainly based on the methodologies based on extracting the features and fusion techniques associated with these features. A large amount of these methods have reliance in detecting the core pixels. The authors have also pointed out that the improper selection of the pixels leads to either over classification or miss-classification. The authors have tried to overcome these disadvantages in their present article, by considering only the frontal pixels inside every image. This consideration helps to overcome the difficulties arose out during recognition of boundaries and detection discontinuities.

Christopher Herbon et al., 2014 , the authors in this article have thrown light by considering the colour images for segmentation purpose, in particular in this article, the authors have tried to address the problems and solutions associated in identifying the joints inside the image regions. The authors have considered the concepts of segmentation together with statistical modeling . Split and merge technique is considered and is repetitively applied on to the image and the results of the segmentation process are presented.

R. Loganathan et al., 2013 in their article have addressed about the identification of well beingness inside the hospital. The authors have tried to workout a model that enhances the storage capabilities of storing such images. Each of the diseased patient is considered and the appropriate health picture of the patient is taken into consideration and is segmented to identify the deformities.. The ROI detached are subsequently condensed employing lossless uphold compression method. A novel BOW and Embedded Zero Tree (EZW) is recommended for compression. Experimental outcome elucidate that the method that is counsel enhances the compression ratio.

M. Lalitha et al., 2013 presented the various clustering models towards effective segmentation of photos. The purpose of clustering is to turn into outcome that has noteworthy capable storage and rapid reclamation in different areas. The objective is to deliver a description that is self-sufficient of views. Various clustering methods were also highlighted which are currently available in the state of art for soft computing purposes. This article focused entirely on the basic concept of clustering a photo image, which is taken into consideration. The case studies are highlighted to showcase the importance of each and every clustering model

Hakeem Aejaz Aslam et al., 2013 in their article has projected the k pillar method for segmentation. This segmentation method comprise a means towards the new congregation of agents with well-known resolution films for improving the precision and minimizing the compiling time . The array employ K-means algorithm for picture segmentation, and later this algorithm is further optimized using the Pillar. The results showcase appropriate improvement of clustering results when compared to the existing clustering algorithms.

Sunita p [2013] researched on "image retrieval using co-occurrence matrix &texton co-occurrence matrix for high performance"a new implemented work which is comparison with texton co-occurrence matrix to describe image features. a new class of texture features based on the co-occurrence of grey levels at points. these features are compared with previous types of co-occurrence based features, and experimental results are presented indicating that the new features should be useful for texture. the results demonstrated that it is much more efficient than representative image feature descriptors, such as the auto-correlogram and the texton co-occurrence and the texton co-occurrence matrix.

2.1 Methodolgoies Available

2.1.1 Thresholding Methods

Makes decisions based on information from local pixels and is effective when the intensity levels of the objects fall squarely outside the range of background levels. Because spatial information is ignored, however, the boundaries blurred region can wreak havoc. Boundaries and borders region are closely linked, because there is a strong coordination

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

on the edge of the area. Therefore, edge detection techniques have been used as the basis for the other segmentation technique. The edges identified by edge detection are often disconnected. Limits of a closed segment of the object regions are required to from an image. Discontinuities are bridged if the distance between the two edges is a predetermined threshold.

2.1.2Edge based Methods

These techniques are typically strenuous around the edge detection. Flaw in linking mutually the broken contour lines in the existence of smudge led towards the failure of these models.

2.1.3Region based Methods

In this methodology underlined image is alienated into associated regions by considering the alignment of adjacent pixels at analogous levels of intensity. Contiguous regions are then fused under a decisive factor linking conceivably homogeneity or unevenness of boundaries in the region. More rigorous measure result in division, specified the indistinct borders and luminosity fusion

2.1.4Split and Merge Methods

Split-and-merge Segmentation is based on a separation of the image pixels into a quadrangle tree, occasionally called quad clustering. This process initiate at the origin of the tree in lieu of the complete image. If it is non-uniform (homogeneous), and is alienated into four squares. On the contrary, if square pixels inside the image are uniform, then they are fused into a common group. The segmentation process starts at the root node and continues recursively until no crack or merging is possible.

3. BIVARIATE GENERALIZED GAMMA MIXTURE MODEL

In this article, Bivariate Gamma Mixture Model is chosen for developing the proposed framework. The main advantage of this model is that it can interpret the data more robustly as it considers Bivariate features. The features considered include both text and speech. Another advantage behind the choice of this model is that it can consider the different variants of shape parameters and hence can effectively help to retrieve the relevant lectures more accurately. The Probability Density Function (PDF) of the Bivariate Gamma Mixture Model is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_A^B \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx$$

$$\text{where } A = \log\gamma \text{ and } B = \log(\gamma + 1)$$

4. DATA SET

In order to demonstrate the proposed contribution, we have generated a data set by taking into account the various NPTEL lectures and other lectures available free of cost from the Internet sources. These data sets are pooled together such that it contains a heterogeneous group of lecture material

Figure 1: Dataset considered

5. FEATURE EXTRACTION

In this current article, to extract the features, two models have been considered. Since, this is a Bivariate model, the features extracted using DCT & LPC are considered. The advantage of selecting DCT is already justified in the introductory section of the article. In a DCT based feature extraction, each of the videos is captured and is derived as MxM blocks of fixed sizes. Following the methodology prescribed by Gonzalez and Woods (2002), the DCT coefficients are computed. These coefficients are connected in a zig-zag manner and blocks are formulated such that each block consists of 16 coefficients. These 16 coefficients are necessary to identify the content from each of the video signals. Every block considers these 16 coefficients and thereby these coefficients formulate a sample feature block and finally resulting into N = MxN blocks. This is considered as the training features group. If the video samples considered are N in number, Nx16 blocks of coefficients are thereby generated. These coefficients are considered as one of the features for the Bivariate Generalized Gamma Mixture Model proposed in Section 3 of the paper

5.1 LPC based Feature Extraction

In this sub-section of the article, the feature extraction based on LPC is considered for extraction of features from the speech sample. The main choice behind this selection of LPC vectors is that it helps to recognize the speech sample from audio signals even in the presence of noise more robustly and hence, the LPC signals are considered for the extraction of the low level features.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

6. METHODOLOGY

In order to implement this model, the data set presented in Section 4 of this article is considered. Experimentation is carried out in .Net environment. For experimentation, we have considered various video lectures from Computer Science subjects. In these lectures, the speech coefficients are considered for the identification and along with DCT coefficients. These features are given to the model proposed in Section3 of the article and accordingly the PDFs are generated. Each of the PDFs are considered and based on their maximum likelihood estimates, the relevant information is mapped. The performance evaluation is carried out by Precession and Recall.

The formulas for calculating the same are prescribed below.

$$\text{Precision} = \frac{\text{No of relevant images retrieved}}{\text{Total no of relevant images retrieved}} \quad (1)$$

Precision measures the proportion ofthe total imagesretrieved which arerelevant to the query.

$$\text{Recall} = \frac{\text{Total number of relevant images}}{\text{Number of relevant images retrieved}} \quad (2)$$

7. EXPERIMENTATION

The experimentation is carried out in a dot Net environment and based on the query voice frame, the relevant frames are extracted from the data set The experimentation results after performing the experimentation are presented below.

TABLE 1: EXPERIMENTAL RESULTS

Video Lecture	Retrieved Video Lecture	Precision	Recall
		0.97	0.62

From the above table , it can be clearly noticed that the proposed model is extracting the relevant video frames from the data set more accurately.

8. CONCLUSION

In this article, we present a methodology for effective retrieval of relevant images from the YouTube videos based on Bivariate Generalized Gamma Mixture Model using DCT & LPC. This methodology is found to be of much use in retrieving most relevant video images in case of the large datasets. We used precision and recall for performance evaluations and the result shows that optimal results are generated from the proposed methodology.

REFERENCES:

[1] Neha R. Kasat, Dr.Sudeep D. Thepade, "Novel Content Based Image Classification Method Using LBG Vector Quantization Method with Bayes and Lazy Family Data Mining Classifiers", International Conference on Communication, Computing and Virtulization, 2016.

[2] Anubala, Tajinder Kaur, "Local texton XOR patterns: A new feature descriptor for content based image retrieval", Engineering Science and Technology, an International Journal, 2016

[3] S.K.Vipparthi, S.K. Nagar, "Multi-joint histogram based modelling for image indexing and retrieval", Comput. Electr. Eng., 2014

[4] M.Subrahmanyam, Q.M.Jonathan Wu, R.P.Maheswari, R.Balasubrahmanian, "Modified color motif co-occurance matrix for image indexing and retrieval, Comput. Electr. Eng., 2013

[5] MeenakshiShruti Pal, SushilGarg "Image Retrieval: A Literature Review" International Journal of Advanced Research in Computer Engineering and Technology (IJARCET) Volume 2, Issue 6, June 2013.

[6] Z.Tang, X.Zhang, X. Dai, J.Yang, T.Wu, "Roburst image hash function using local color features", Int. J. Electroni. Commu., 2013

[7] Y.Rui, T.S.Huang, "Image retrieval: current techniques, promising directions and issues", J. Vis. Commu. Image Representation, 1999

[8] Vaishali D. Dhale, A. R. Mahajan, Uma Thakur, "A Survey of feature Extraction Methods for Image Retrieval," International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 10, October 2012 ISSN: 2277 128X.

[9] Hui Hui Wang, Dzulkifli Mohamad, N.A. Ismail "Approaches, Challenges and Future Direction of Image Retrieval," Journal of

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Computing, Volume 2, Issue 6, June 2010, ISSN 2151-9617.

[10] A. Araujo, J. Chaves, H. Lakshman, R. Angst, and B. Girod, "LargeScale Query-by-Image Video Retrieval Using Bloom Filters," arXiv, vol. 1604.07939, 2016.

[11] M¨uhling, M., Meister, M., Korfhage, N., Wehling, J., H¨orth, A., Ewerth, R., Freisleben, B.: Content-based video retrieval in historical collections of the german broadcasting archive. In: 20th Int. Conf. on Theory and Practice of Digital Libraries. pp. 67–78 (2016)

[12] Cees G.M. Snoek "Tag-based Video Retrieval by Embedding Semantic Content in a Continuous Word Space" IEEE Winter Conference on Applications of Computer Vision: WACV 2016: Lake Placid, New York, USA,7-10 March 2016

[13] Andr´e Araujo and Bernd Girod "Large-Scale Video Retrieval Using Image Queries" IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, 2017

[14] Gabriel de Oliveira Barra ,Mathias Lux and Xavier Giro-i-Nieto "Large Scale Content-Based Video Retrieval with LIvRE" IEEE 2016

[15] Luca Rossetto, St´ephaneDupont and MetinSezgin"IMOTION — A Content-Based Video Retrieval Engine" Springer International Publishing Switzerland, 2015.

[16] M.Ravinder and Dr.T.Venugopal "Content Based Video Indexing and Retrieval Using Key Frames Discrete Wavelet Center Symmetric Local Binary Patterns (DWCSLBP)" International Journal of Computer Science and Information Security (IJCSIS) ,May 2016

[17] Priya Singh and Sanjeev Ghosh "Content Based Video Retrieval using Neural Network" International Journal of Electronics, Electrical and Computational System 2017

[18] Shoou-I Yu [etal] "Content-Based Video Search over 1 Million Videos with 1 Core in 1 Second." ICMR'15, 2015, Shanghai, China

[19] B. M¨unzer [etal] "When Content-Based Video Retrieval And Human Computation Unite: Towards Effective Collaborative Video Search" IEEE,2017

[20] Lu Jiang Shoou-I Yu, DeyuMeng, and Yi Yang "Fast and Accurate Content-based Semantic Search in 100M Internet Videos" 2015

[21] Aasif Ansari and Muzammil H Mohammed "Content based Video Retrieval Systems - Methods, Techniques, Trends and Challenges" International Journal of Computer Applications (0975 – 8887) Volume 112 – No. 7, February 2015

[22] Sunita.P "Image Retrieval Using Co-Occurrence Matrix &Texton Co-Occurrence Matrix For High Performance" International Journal of Advances in Engineering & Technology, Jan. 2013. ©IJAET.

[23] K. Schoeffmann, C. Beecks, M. Lux, M. S. Uysal, and T. Seidl, "Content-based retrieval in videos from laparoscopic surgery," in Proceedings of the SPIE Medical Imaging Conference, 2016, pp. 97 861V–97 861V–10.

[24] ShishiQiao "Deep Video Code for Efficient Face Video Retrieval" inAsian Conference on Computer Vision Workshop (ACCV workshop), Springer (2016).

[25] S. Hong, W. Im, and H. S. Yang, "Content-based Video-Music Retrieval using Soft Intra-modal Structure Constraint," CoRR, vol. abs/1704.06761, 2017. A. EL Allaoui, M. Merzougui, M. Nasri, M. EL Hitmy and H. Ouariachi. Evolutionary Image Segmentation By Pixel Classification Application To Medical Images. IJCIIS International Journal of Computational Intelligence and Information Security, ISSN: 1837-7823, Vol. 2, No. 3 pp. 12- 24. March 2011.

[26] A. EL Allaoui, M. Nasri, M. Merzougui , M. EL Hitmy et B. Bouali. Medical Image Segmentation By Region Evolutionary Approach. The 3rd International Conference on Multimedia Computing and Systems ICMCS'12 Tangier, CDROM, 10-12 Mai 2012.

[27] P. Arbeláez, M. Maire, C. Fowlkes and J. Malik.Contour Detection and Hierarchical Image Segmentation IEEE TPAMI, Vol. 33, No. 5, pp. 898-916, May 2011.

[28] Jinping Liu , WeihuaGui, Qing Chen, Zhaohui Tang, and Chunhua Yang. "An unsupervised method for flotation froth image segmentation evaluation base on image graylevel distribution." In Control Conference (CCC), 2013 32nd Chinese, pp. 4018-4022. IEEE, 2013.

[29] Badri Narayan Subudhi, Ishan Patwa, Ashish Ghosh, and Sung-Bae Cho. "Edge Preserving Region Growing for Aerial Color Image Segmentation." In Intelligent Computing, Communication and Devices, pp. 481-488. Springer India, 2015.

[30] Christopher Herbon, Klaus Tönnies, and Bernd Stock. "Detection and segmentation of clustered objects by using iterative classification, segmentation, and Gaussian mixture models and application to wood log

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

detection." In Pattern Recognition, pp. 354-364. Springer International Publishing, 2014.

Subash Chandra Chadalavada, received M.Sc. in Physics with Electronics specialization from Andhra University and M.Tech. (CSE) from JNTU Kakinada is working as Associate Professor in Department of Computer Science Engineering, Kakinada Institute of Engineering and Technology, Korangi. He is an active member of CSI & ISTE. He has 10 years of teaching experience. There are a few of publications both national and International Conferences / Journals to his credit. His area of interest includes Information Security, Cloud Computing, Computational Photography and other advances in Computer Applications

Dr. Srinivas Yerramalle was awarded M.Tech (Computer Science & Technology) from Andhra University in 1999. He was awarded Ph.D in Computer Science with Specialization in Image Processing from Acharya Nagarjuna Univerisity in 2008. He Published Several Research papers in National & International Journals/ Conferences. He Guided 9 Students for Ph.D. Presently, working as Professor and Head of the Department, Department of IT, GITAM University, Visakhapatnam.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Optimization of Resource Allocation Strategy Using Modified PSO in Cloud Environment

Neeraj Kumar Pandey[1], N.K.Joshi[2]

1,2 Uttaranchal University, Dehradun Uttarakhand INDIA

ABSTRACT:

Resource scheduling is a most important functioning area for the cloud manager and challenge as well. It plays very vital role to maintain the scalability in the cloud resources and 'on demand' availability of cloud. The challenges arise because the Cloud Service Provider (CSP) has to pretend to have infinite resource while he has limited amount of resource. Resource allocation in cloud computing means managing resources in such a way that every demand (task) must be fulfilled along with considering the parameter like throughput, cost, make span, availability, utilization of resource, time and reliability. The Modified Resource Allocation Mutation PSO (MRAMPSO) strategy based on the resource scheduling and allocation of cloud is proposed. In this paper MRAMPSO schedules the task with help of Extended Multi Queue (EMQ) by considering the resource availability and reschedule the task that fails to allocate. This approach is compared with slandered PSO and Longest Cloudlet to Fastest Processor (LCFP) algorithm to show that MRAMPSO can save execution time, makes span, transmission cost and round trip time.

Keyword: CSP (Cloud Service Provider), VM (Virtual Machine), PSO (Particle Swarm Optimization)

1. INTRODUCTION:

The National Institute of Standards and Technology (NIST) defines the cloud as "a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [1]. Cloud computing creates pool of infrastructure that connects different computing components and provides them as a service so it is called XaaS i.e anything as a service. Here X can be replaced by memory, storage, network, Operating system, Application, Security etc. These services can be demanded any time from the CSP in any amount so the scheduling of resources is the major issue for the cloud manager. Scheduling algorithms plays vital role to manage the resources which cloud be available any time the cloud user demands. Scheduling is a NP hard problem So the solution is found in the heuristic algorithms. Scheduling ensures the optimum usage of the available resources along with the concern of the other changed parameters of services. PSO is a heuristic algorithm is used to solve the scheduling problem and many other NP hard problems. The purpose of the paper to enhance the performance of the resource scheduling problem in the cloud environment. The existing algorithm tries to solve the problem with the parameters like make span, time, cost, resource utilization and task scheduling. The proposed work is based on the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

heuristic strategy using Modified Resource Allocation Mutation particle swarm optimization (MRAMPSO). This strategy will attain the optimized resource scheduling with the task scheduled by Extended Multi Queue Scheduling (EMQS). The remaining paper is organized as follows: section 2 will describe related work. Section 3 will describe the proposed model of the optimization technique. Section 4 presents the formulation of task scheduling problem. Section 5 proposes EMQS strategy of task scheduling model. Section 6 will represent the proposed MRAMPSO. Section 7 will show the experimental result. Section 8 will have conclusion and future scope.

2. RELATED WORK

There is wide range of research done by different researcher to solve the resource allocation problem in cloud environment. Every solution tries to optimize the existing solution to considering some tedious parameter so that improvement can be proposed and this serious issue of CSP can be properly addressed. Most of the researcher improves the parameter like cost, speed, resource scheduling and reliability, make span and availability.

In the paper [2] a new optimized model of task scheduling is proposed which used the Particle swarm Optimization (PSO) to solve the scheduling of tasks with some heuristic way. The small position value rule to minimize the cost of provisioning the existing resource a PSO strategy is used. The experiments show that the PSO executes faster than other two strategies. So PSO prove better in scheduling problems in cloud environment. This paper does not focus on the efficiency of the scheduling model and SLA.

In paper [3] set of task-service pair is represented as a candidate set. Each particle will learn from each feasible pair of different dimension. The position building technique guarantees every position is reasonable. This scheme significantly minimizes the search space and improves the algorithm performance. The new algorithm produces ultimate performance on job scheduling–resource allocation schemes in cloud environment. There is no discussion of cost in this paper. It could cover this area also.

In paper [4] the proposed algorithm improves QoS parameter preferred by user. The paper focuses on the scheduling workflow. The experimental results significantly improve the CPU utilization. The energy efficiency of the workflow is not discussed in the paper which can be opted as a future improvement.

In the paper [5] heuristics is used to reduce the execution cost of application workflow in the cloud environment. In the paper aggregate cost of execution by differing the correspondence cost among resource and the execution cost of compute resources is acquired. The result is compared with “Best Resource Selection” (BRS) and found that PSO based task mapping takes three times less cost than BRS algorithms. The different cost parameters are calculated but no reliability and SLA conditions are discussed in the paper.

The second side of the proposed model is based on the task scheduling strategy. There a huge number of strategies, their optimization and some traditional models of task scheduling have been proposed by different researcher. Some key research models which inspire the proposed model are discussed below.

Paper [6] proposes priority based task scheduling model used Analytical Hierarchy Process (AHP) and multi-attribute decision making models to evaluate the priority of the task. It covers the most important attribute of any job to categorise them. The improvement in the cost performance is not measured in the paper which is a major concern of the resource scheduling.

Paper [7] a parallel scheduler is being proposed i.e. Naphele scheduler. It evaluates the maximum time allowed to any process to get executed (Critical Time). If any job reached to critical time and failed to execute completely them it

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

is thrown to the waiting queue. Naphele scheduler allows the parallel processing of jobs which is highly desirable for the optimistic resource allocation strategy. The strategy discussed in the paper works in the non-pre-emptive mode of selecting jobs which may not perform in the priority and efficiency criteria of job execution.

Paper [8] proposed a decentralized architecture of energy efficient resource allocation policy of cloud. The energy consumption of any data centre is depending upon the efficiency of the resource allocation algorithms because it influences the time consumed by any VM to complete the task. The performance of the resource allocation strategy must be measured before the evaluation of energy efficiency in the cloud.

Paper [9] focuses on the major concern of the client and the CSP i.e. cost. The execution time of task plays a vital role in the cost of execution. So in this paper the combination of resource cost and the execution time of any resource calculate the total cost of execution of a task. The system can be improved by robustness checking and reliability factor optimization.

Paper [10] proposes an economic model of the task scheduling with the help of bio-inspired algorithm, an intelligent combinatorial double auction based dynamic resource allocation technique. The price prediction system is proposed for dynamic pricing for set of task. The entire task is keeps the SLA in mind while making polices based on cost. The extension can include the cost of execution and reliability of proposed model.

Paper [11] discussed a different type of algorithm in nature i.e heuristic algorithms. This algorithm is considered due to NP hard nature of the resource allocation problem. The inference of proposed model of solution is being taken from the same nature of algorithm. The mathematical model proved the problem with some restriction which can be improved in order to extend the work.

3. PROPOSED STRATEGIC MODEL

The motivation of this model is allocation of resources to the Virtual Machine (VM) in the efficient way by applying the two way scheduling one is for the tasks which are recently arrived and other is resources are scheduled before the submission. The proposed model is in figure 1 have five different phases.

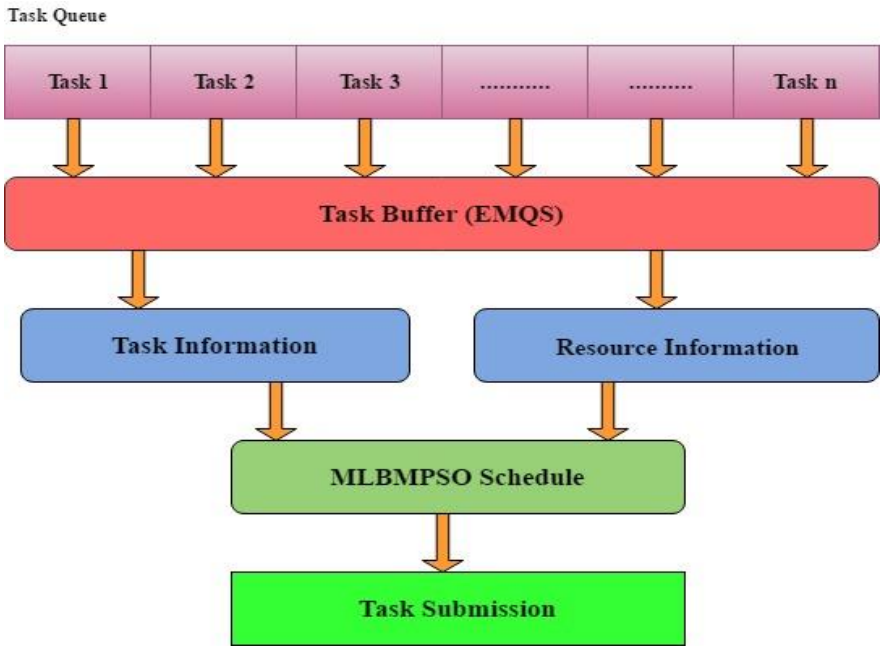


Fig.1: Proposed Model Structure

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Task Buffer

Task buffer receives the task from the queue containing the task arrived to the CSP cloud manager. The selected tasks have gone through the Extended Multi Queue Scheduling Algorithm (EMQS) which selects the task as per the resources by the dynamic task selector and identify the resource to execute. Figure 2 depicts the extended multi queue dynamic job selector which is used to select the job demanding for resources. In the dynamic job selector the jobs are divided into the two queues initially one is for the jobs having different priority levels and another is which have same priority. The priority queue is used to pick the job which is at high urgency (Priority). They are allocated the resource first and then the dynamic selector comes to the queue of jobs which do not have any priority. The Dynamic Selector then come to the non-priority queue and divides queue on the basis of their burst time i.e. the small bust time and large burst time (the category range can be decided by administrator, DA will dynamically select the job from each queue one by one so that each queue must be served eventually and no job of any queue will wait for undefined time. here the proposed model considers the following criteria to divide the queues:

1. 70 % jobs are stored in first queue (Small Jobs)
2. 30% jobs are stored in second queue (Big Jobs)
3. Priority based jobs are stored in the third queue

The priority based multi queue job scheduling is applied with the help of extended priority based task scheduling in cloud. It is helps the resource allocator to allocate the resource in economic way so that maximum profit can be made without violating SLA.

The following strategy is applied for selecting the job by the Dynamic Job Selector [12].

1. For all Ts(Task Submitted)
2 Find the priority of submitted task.
3. Maintain the ready queue based on newly arrived job
4. for all newly arrived job
If (priority of new jobTn> Priority of job in execution Te)
Te=Tn
5 Run the job queue using priority scheduling
6. Else
7. Divide the jobs according their burst time and use traditional Scheduling

The above strategy selects one eligible job for allocating resource. This strategy works along with traditional algorithm like CBA, Priority Scheduling and FCFS. By implementation of above strategy performs better than the traditional algorithm.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

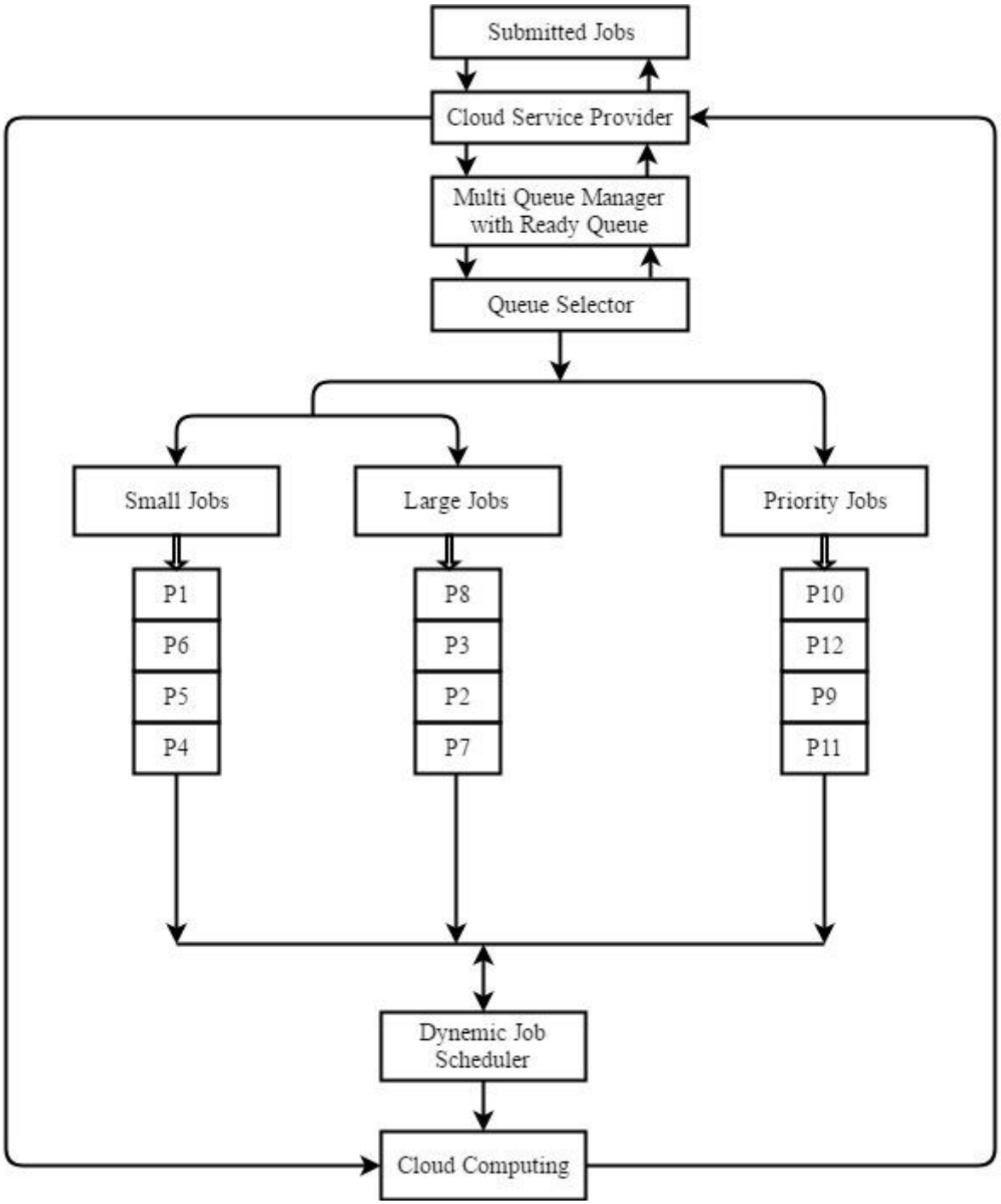


Fig 2. Extended Dynemic Multi Queue Job Selector

Task Information

This phase collect the necessary information about submitted task like Expected Execution Time(EET), Resource Required(RR) to execute the task, Round Trip Time (RTT), Expected Transmission Time(ETT). This information will help the scheduler to manage the execution of the task along with the specified parameter so that the optimal solution can be produced.

Resource Information

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

This phase is responsible for the collecting information of available resources which is necessary for the optimal provisioning. The resources contain the information of host, data centre and VMs. Multiple VMs can be generated by each host which can be assigned to different task. The information list contains VM list, memory availability, band width and mips of each VM (every host can have more than one VM so every VM can have different speed mips). This information is passed to the next phase of the model for further processing.

MRAMPSO

Modified Resource allocation Mutation PSO is used to allocate the resources to the task that are provided by the EMQS. The first problem is which task must be allocated what resource. Second no task should be left unallocated and no more than one VM should be allocated to any task. Solving these two problems we will increase reliability and decrease the task execution cost.

4. FORMULATION OF TASK SCHEDULING PROBLEM

There are various task arrived on the real times (may be thousands) basis so cloud manager have to assign these task to VM. Figure 2 simply depicts the assignment of the task to the corresponding VM or more than one VM. But this scenario always creates a problem because there are n VMs which are allocated to more than one task and vice versa. PSO will select the optimal distribution of task to VMs to achieve the objective. This strategy will reduce the expected processing time of task i to VM j i.e.

(Processing Time) $EET_{ij} = length_i/mips(VM_j)$ (1)

(Expected Transmission Time) $ETT_{ij} = file\_size_i / bandwidth.$ (2)

(Expected Round Trip Time) $ERTT_{ij} = (ETRT_{ij} + latency) + (EET_{ij} + latency)$ (3)

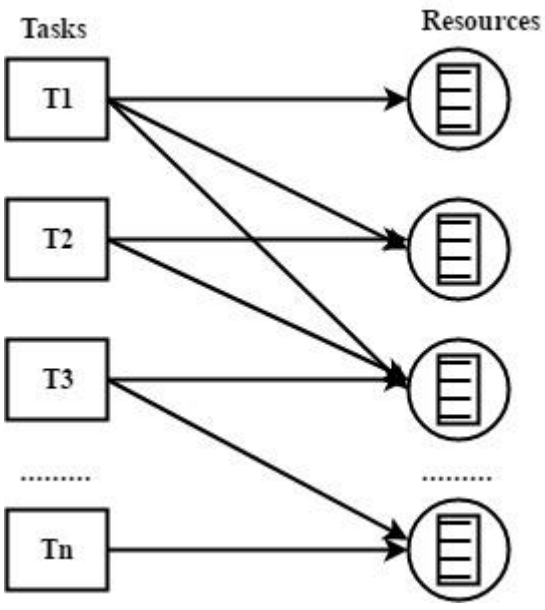


Fig.3: Tasks mapping to Resources

The length i represent the number of task and the speed of VM represents in mips (number of instruction executed per second) in equation (1). ETT is calculated in the size of the file to be executed per bandwidth of the network in equation (2). The EET and ETT affect the ERTT along with the latency of these two parameters in equation (3).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

5. EXTENDED MULTI QUEUE SCHEDULING (EMQS)

The proposed algorithm selects the task from the pool of jobs then divides the task into two parts one is based on priority and other is a pool on task of equal priority. Then the equal priority job is again divided into the Expected Execution Time (EET) of the task. A dynamic scheduler is set to pick the job from the different queue. It picks the job one by one from each queue so that no queue will remain untouched. So in the proposed model, task is selected from the queue by using EMQS strategy resulting the optimum selection of jobs to assign the available resource. The task information and resource information is collected from the selected job by EMQS. The algorithm used by EMQS is as follows [12]:

Extended Multi-Queue Scheduling

---

1. For all T_s(Task Submitted)

2. Find the priority of submitted task.

3. Maintain the ready queue based on newly arrived Task

4. For all newly arrived task

If (priority of new task T_n> Priority of task in execution T_e)

$T_e=T_n$

5. Run the task queue using priority scheduling

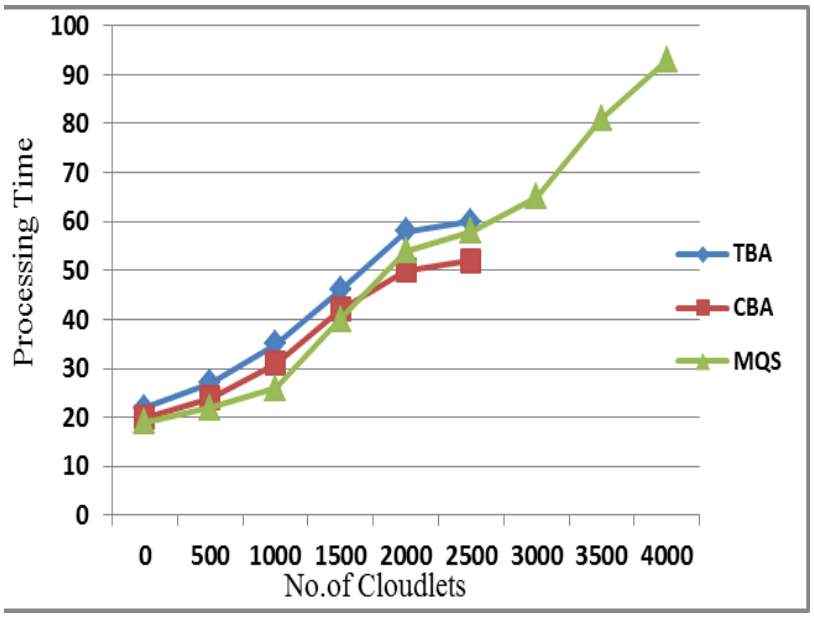
6. Else

7. Divide the tasks according their EET and use Scheduling strategy

This algorithm gives the optimum selection of tasks when simulated with the cloudsim. The following graph simply represents the comparison between traditional task scheduling strategies and EMQS. Thus we found the optimum result when compared with the other scheduling algorithm. So this strategy can be opted in the selection of the task from the newly arrived arrives job pool. The proposed models adopt it for the further processing in order to optimize the resource scheduling strategy in the cloud environment.

Result of the EMQS is represented along with the comparisons with the traditional algorithm. It shows better results when number of cloudlets are increased.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500



Graph 1: Comparison of the EMQS with existing methods

So this strategy can be adopted and amalgamated with PSO to get desired result in resource scheduling.

6. MODIFIED RESOURCE ALLOCATION MUTATION PSO (MRAMPSO)

PSO is a population based meta-heuristic search algorithm based on the simulation of the social behaviour of birds within the flock and fish school proposed by Kennedy and Eberhart [13,14]. This algorithm is famous for the its effectiveness and the simplicity to solve the Broad range problem (NP hard) like resource scheduling problem and task allocation problem. In this strategy every participating particle is act like a solution at their individual position. Then the velocity vector of particle changes and the position of every particle are recalculated. Then we find the new position of every particle. This process continues until the optimized solution is not found. The following algorithm will describe the MRAMPSO strategy which ensures the execution of each task by appropriate VM with the lowest cost and high reliability [15].

*Load of VM_i = (resource of VM_i/Total_resources) *N (number of tasks)* *(4)*

Equation (4) is used to determine the load of the VM after each iteration of executing task and the assigning the task to VM. It helps to update the load of VM every time so that the VM can be managed at the time of allocation. So by this way the actual load is calculated and the overloading or under loading of VM can be prevented which is the one aspect of the proposed strategy. In the MRAMPSO algorithm the best PSO is applied before the distribution of the VM to any task so that individual best position of the particle is already achieved. Then the VM is managed by using following algorithm and the load is also being checked every time after assigning the VM. Because we are sorting the task as per their EET and the load of the VM so that lowest loaded VM can be assigned [16].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Modified Resource Allocation Mutation PSO Algorithm

1. Find the best existing PSO solution
2. For ∀ T_i € T do
   Determine the Task Queue (Task waiting for resources)
   Determine the task which is wrongly allocated (allocated to many VM)
   End For
3. For ∀ VM_i € VM do
   Determine the current task allocated to VM_i (VM_i Load)
   Determine the actual current Load of VM_i
   End for
4. Sort the VM according to current Load
5. Sort the Task based on resource required
6. For ∀ sorted (VM_i) € VM(available) do
   ∀ sorted task (T_i)
   If (real load>current load VM_i)
   Allocate $VM_i \rightarrow T_i$
   Increase the load of VM → VM++
   Else
   Break; // check for another Vm and sort ascending load and exit
   End if
   End for
7. End

Above strategy will be used to allocate VM to the task selected by dynamic selector.

7. SIMULATION RESULT AND EVALUATION

Cloudsim is used to implement the proposed MRAMPSO algorithm. The result of the algorithm is compared with longest VM longest cloudlet algorithm [9], mutation PSO without considering the standard PSO and load balancing algorithm. The evaluation is being done considering the parameters i.e. average cost, average make span, average execution time, average round trip time [9].These parameters are considered in comparison with the mutation PSO, Longest Cloudlet to Fastest Processor with the MRAMPSO algorithm. The graph1, graph 2 and graph 3 shows comparison of different parameters of different algorithm with proposed strategy and their effect.

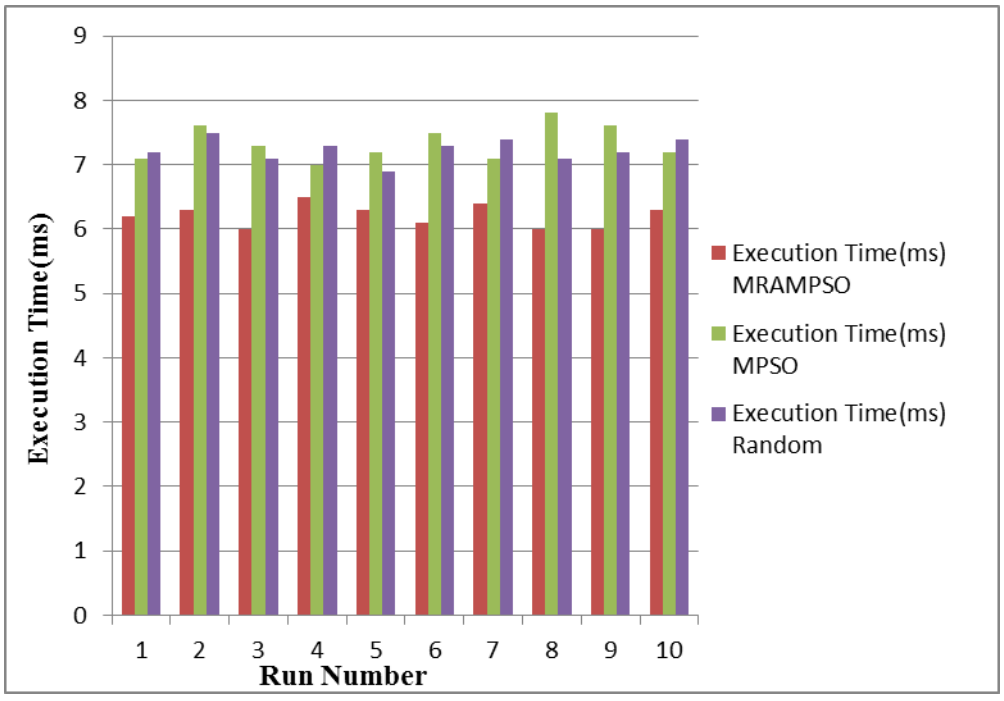
The following table displays the data set on which simulation is performed and result is measured. In this phase the parameters like Task Length, No. of Tasks, size of file, data centre size and number of hosts is considered. Graph 3 displays the comparison and performance of MRAMPSO and random algorithm between the parameters like ET and EET.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 1: Resource Parameters

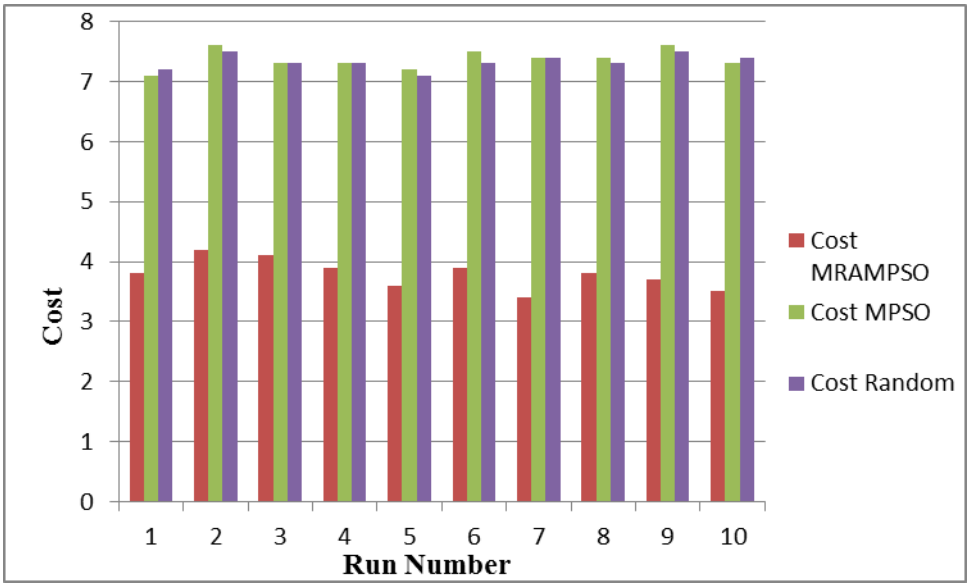
Parameter	Value	Parameter	Value	Parameter	Value
Task Length	1000-3000	No of VM	100	Data Center Size	8
number of task	2000	Speed(MIPS)	1000-2000	No of Host	4-8
File_Size	1-500	RAM	256-2048		
Output_Size	1-500	BW	500-1000		

Graph 2 represents the expected execution time of every task by dividing the execution time with the speed of processor. The result is measured with 10 different run and the EET is collected with respect to MRAMPSO, MPSO and random algorithm. The data set for execution is same as table 1. So the MRAMPSO is performing better than the other two algorithms.



Graph 2: Average ET based on EET

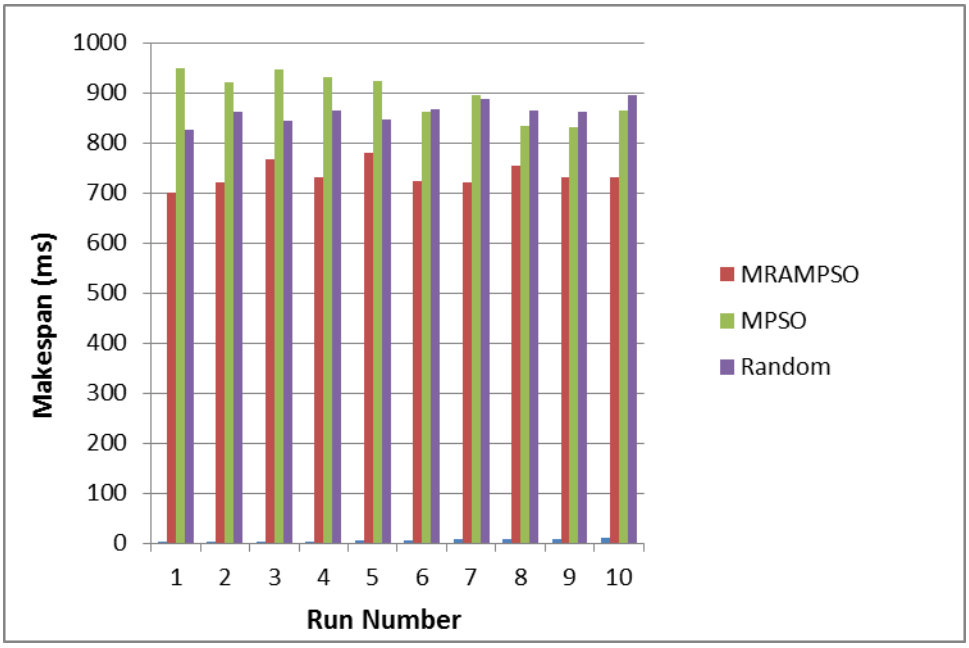
https://sites.google.com/site/ijcsis/
ISSN 1947-5500



Graph 3: Average Cost based on ETT

The next parameter is cost of execution which is very vital issue in the SLA and an important factor for the cloud business. So through simulation and considering the parameter of table 1 the graph 3 shows the estimated cost of the MRAMPSO, MPSO and random with comparisons. So from this experiment we can see the level of diminishing cost by MRAMPSO in comparison to other two strategies.

It impacts a lot in the overall performance of the algorithm because we can increase the profit by minimizing the execution cost



Graph 4: Make Span based on EET

In the above result the graph 4 depicts that as we increase the run number the make span time of MRAMPSO decreases in comparison to MPSO and random methods. The performance of MRAMPSO is measured with the parameter i.e execution time, cost and make span and it performs better than traditional strategies.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

8. CONCLUSIONS AND FUTURE SCOPE

In this paper a resource scheduling algorithm MRAMPSO is used to and minimize the makes span, minimize the execution time (ET), minimize expected RTT are proposed and implemented using java based open source simulator cloudsim. The process starts when the tasks arrive in the job pool (where newly arrived jobs are stored, ready for execution and waiting for execution). The EMQ strategy is used to select the task as per their priority. It makes two types of queue one for priority job and other for the task having same priority. The tasks having same priority are divided according to their ETT. So that task of every queue must be executed in every iteration and no tasks have to wait infinitely. This job selection mechanism is done by dynamic task scheduler. After submission of the job by EMQ the task is arrive to the MRAMPSO strategy (as represented to figure 1). Then MRAMPSO makes the resource available to each job given by the dynamic scheduler. The MRAMPSO strategy apply on the all job do that no job left unallocated and no more than one VM is allocated to any job. So by considering this strategy modified resource allocation mutation PSO gives the optimized result as compared to random algorithm and slandered mutation algorithm. In this paper we simulated the MRAMPSO by considering the parameter i.e. execution time (ET), average transmission time (ETT) and cost of the execution. These result of simulator are then compared to the existing random algorithm and mutation PSO, it gives optimized result than existing algorithm.

The future scope of this hybrid strategy can be if other parameter like reliability and robustness can be considered for current strategy and the existing algorithm. Then many other aspect of the resource allocation strategy cam be covered and it can help the cloud manager to maintain the scalability property of the cloud in future. The scope can be dividing large problem into small sub problem and distribute them to high speed VM so that cost and EET can be reduced.

9. REFERENCES

[1] P. Mell, T. Grance “The NIST Definition of Cloud Computing” Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899-8930 September 2011.

[2] Guo L,Zhao S,Shen S, Jiang C, "Task Scheduling Optimization in Cloud Computing Based on Heuristic Algorithm," Journal of Networks (7) 2012, pp. 547-553.

[3] Z. Wu, Z. Ni, Lichuan Gu, Xiao Liu “A Revised Discrete Particle Swarm Optimizat-ion for Cloud Workflow Scheduling” IEEE International Conference on Computational Intelligence and Security (2010) pp 184-188.

[4]Varalakshmi P, Ramaswamy A, Balasubramanian A,Vijaykumar P,"An Optimal Workflow Based Scheduling and Resource Allocation in Cloud",Advances in Computing and Communications, First International Conference, ACC 2011, pp. 411-420.

[5] S.Pandey, L. Wu, S. Guru, R. Buyya “A Particle Swarm Optimization-based Heuristic for Scheduling Workflow Applications in Cloud Computing Environments” 24th IEEE International Conference on Advanced Information Networking and Applications 2010 pp.400-407.

[6] Swachil J. Patel, Upendra R. Bhoi “Improved priority based job scheduling algorithm in cloud computing using iterative method” IEEE Fourth International Conference on Advances in Computing and Communications 2014 pp.199-202.

[7]G. Upadhye,T. Dange “Cloud resource allocation as nonpreemptive approach” IEEE 2nd International Conference on Current Trends in Engineering and Technology(ICCTET)2014 pp.352-356.

[8]A. Beloglazov and R. Buyya “Energy Efficient Resource Management in Virtualized Cloud Data Centers” 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing 2010 pp. 826-831.

[9] A. Vichare, Zenia P. Gomes, Noella Fernandes and Flavin Cardoza “Cloud computing using OCRP and virtual machines for dynamic allocation of resources” IEEE International Conference on Technologies for Sustainable Development (ICTSD-2015), Feb. 04 – 06, 2015, Mumbai, India.

[10] X. Wang, Xueyi Wang, Hao Che, Keqin Li, Fellow,Min Huang, and Chengxi Gao “An Intelligent Economic Approach for Dynamic Resource Allocation in Cloud Services” IEEE Transactions On Cloud Computing, Vol. 3, No. 3, July-September 2015 pp.275-289.

[11] S. Khatua, P.K. Sur, R.K. Das and N. Mukherjee, “Heuristic-Based Resource Reservation Strategies for Public Cloud,”, IEEE Trans. Cloud Comput., vol. 4, no. 4, pp. 392– 401, Oct.–Dec. 2016.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[12] N. K. Pandey, S.Chaudhary, N.K.Joshi "Extended Multi Queue Job Scheduling in Cloud " International Journal of Computer Science and Information Security ISSN 1947-5500 Volume 15 No. 11 November 2017 pp.1-8.

[13] A. Al-maamari and F. A. Omara "Task Scheduling Using PSO Algorithm in Cloud Computing Environments" International Journal of Grid Distribution Computing Vol. 8, No.5, (2015), pp.245-256.

[14] Eberhart R , Kennedy J, "A New Optimizer Using Particle Swarm Theory", 6th International Symposium Micro Machine and Human Science, Nagoya;1995,p. 39-43.

[15]A.I.Awad,N.A.El-Hefnawy, H.M.Abdel_kader "Enhanced Particle Swarm Optimization For Task Scheduling In Cloud Computing Environments" International Conference on Communication, Management and Information Technology (ICCMIT 2015) pp. 920-929.

[16] V. A. Chalack,S. N. Razavi,S.J. Gudakahriz "Resource Allocation in Cloud Environment Using Approaches Based Particle Swarm Optimization" International Journal of Computer Applications Technology and Research Volume 6–Issue 2, 87-90, 2017, ISSN:-2319–8656 pp.87-90.

A Survey of deep learning algorithms for malware detection

Ankur Singh Bist
PHD Scholar Department of Computer Science and Engineering, SVU, India

Abstract: *Deep learning algorithms have drawn the attention of researchers working in the field of computer vision, speech recognition, malware detection, pattern recognition and natural language processing. In this paper, we present an overview of deep learning techniques like Convolutional neural network, deep belief network, Autoencoder, Restricted Boltzmann machine and recurrent neural network. With this, current work of deep learning algorithms on malware detection is shown with the help of literature survey. Suggestions for future research are given with full justification. We also showed the experimental analysis in order to show the importance of deep learning techniques.*
Keywords: *Deep belief network, Autoencoder, Restricted Boltzmann machine and Convolutional neural network.*

1. Introduction

Machine learning techniques have been adopted in various fields like pattern recognition, computer vision and speech recognition. Machine learning has brought so many changes in our life which includes variety of applications ranging from intelligent games to self driving systems. Due to advancements in hardware during last decade, deep learning has become active area of research. Malware detection is the core part of computer security. The main purpose of malware detection is to identify malicious activities caused by malwares. It is a big task to design an algorithm that can detect all kinds of malware with perfect accuracy in a reasonable amount of time. Malware detection requires an automated technique which demands minimal human intervention. It is due to increasing volume of malicious codes and their mutants. Signature based detection technique is quite popular but mutants of existing malware can conceal their behaviour in intelligent manner hence signature based detection is not suitable for zero day malwares [12-13]. In order to trace aberrant activity of zero day malwares machine learning techniques are used under static, dynamic and hybrid detection category.

The purpose of this article is to present a timely review of deep learning techniques in the field of malware detection. It is aimed to give the readers an introduction to different deep learning techniques as well as latest modified architectures of deep networks. The rest of paper is organised as follows. In section 2, different deep learning techniques with their recent variations are reviewed. Section 3 and 4 gives structure of experimental analysis and conclusion and future work respectively.

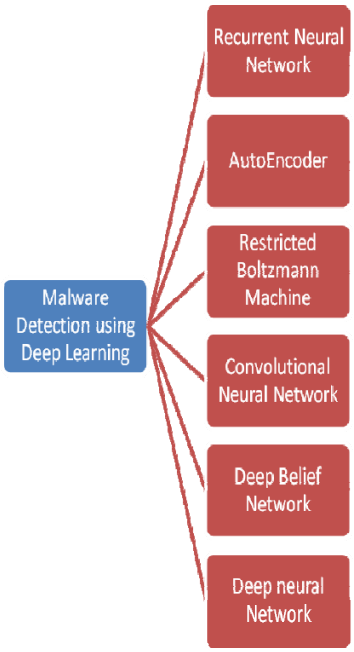


Figure1. Block diagram of malware detection using deep learning

2. Deep Learning Algorithms

The idea of deep learning evolved from neural networks. Neural networks become very popular because of its utility in practical scenarios. Other popular machine learning methods generally used for malware research are SVM, Random forest, Naive Bayes, Multilayer Perceptron, KNN, Ada Boost and Decision Tree(J48) but for dealing big data deep

networks are good [44-47]. Malware samples are growing with very fast pace so deep networks are now becoming popular in antimalware research. Figure1 gives basic the basic terminology of deep learning algorithms. The following describes basic deep learning algorithms.

2.1 Restricted Boltzmann Machine (RBM)

RBMs are very popular in deep learning networks due to their simple architecture as compared to other popular models. RBM contains two layers where first layer denotes stochastic visible units and other layer denotes stochastic observable units. A bias unit is alsc there whose states remains on and the purpose o this layer is to tune different inherent properties.

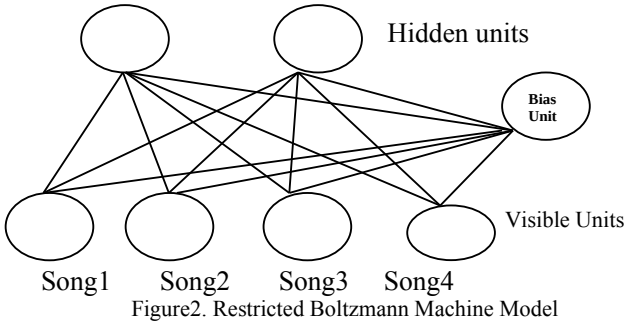


Figure2. Restricted Boltzmann Machine Model

Suppose we have a group of four hindi songs and we asked user to give input which ones they desire to listen. If the purpose involves the learning of two latent units then the RBM will look like as shown in Figure2. In order to evaluate state activation, firstly activation energy is calculated.

$$a_i = \sum_j w_{ij} x_j$$

Where w_{ij} is the weight of the connection between i and j, and x_j is either 0 or 1. Suppose r_i =α (a_i), where α (y)=1/(1+exp(-x)) is the logistic function after that on unit i with probability p_i , and off with probability 1- p_i. Second important question in this domain is to understand the learning pattern. For each iteration, select training dataset, compute activation energy.

$$a_k = \sum_t w_{tk} x_t$$

Then set x_k to 1 with probability α (a_k) and to 0 with probability 1- α (a_k) and for all edges e_{ik} calculate Positive(e_{ik})= x_i* x_k. Calculate Negative(e_{ik})= x_i* x_k for all edges. Weight updating algorithm can be written as-

w_{ik}=w_{ik}+ λ*(Positive(e_{ik})-Negative(e_{ik})), where λ is the learning rate, finally iterate these steps for all training samples until we may obtain error below certain threshold value.

2.2 Deep Belief Networks (DBN)

Deep belief networks belong to the category of deep learning algorithms. DBN generally use the concept of stacking Boltzmann machines. In the deep learning algorithms successful training was big issue and DBNs belongs to deep architecture that could be trained successfully. The main point of the model is that new evidences are taken as prior belief and rearranged to generate posterior and finally convergence towards approximation of facts takes place.

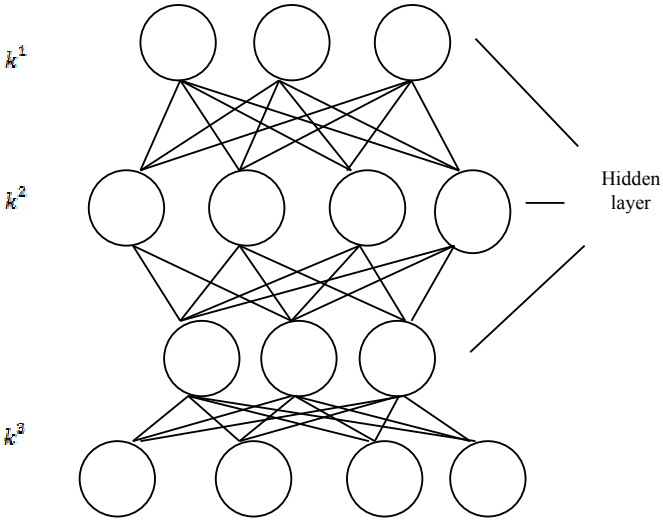


Figure3. DBN Architecture

The learning of weight w is done through restricted Boltzmann machine. Probability of a visible vector can be defined as-

$$P(v) = \sum_j P(k|w) P(v|k,w)$$

Learning of w involves P(v|k,w) same but replace P(k|w) by refined design of the aggregated posterior distribution. Deep belief network generally utilizes a logistic function of input that is weighted in order to find out the probability that a binary latent variable has a value of 1 in the top-down generation or bottom up approach, other category of variables can be used based upon variety of situations.

2.3 Autoencoder

An Autoencoder belongs to the category of neural network that tends to reconstruct from its input. If the vector (1, 1, 1, 1, 0, 0, 1)is given to Autoencoder then the Autoencoder will give (1, 1, 1, 1, 0, 0, 1) as output. The crucial part is hidden layer, for example if one has inputs in 7 dimensions and uses 3 neurons in hidden layer then Autoencoder will take 7 features and encode them in 3 features so that it can give rise to the seven dimension output. We move from (1,1, 1, 1, 0, 0, 1) to (x, y, z) and from (x, y, z) to (1, 1, 1, 1, 0, 0, 1). Training happens in such

a manner that reconstruction error reaches to its minimum level. We can take an example.

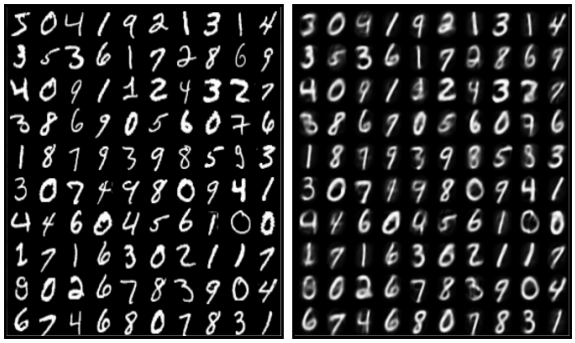


Figure4. 28x28 MNIST image

Let us take the example of MNIST dataset, it contains handwritten digit in 28x28 image format so total inputs become 28x28=784 then the second task is to select the hidden neurons, then we made training with 28 hidden neurons. Finally it performs the desired task as shown in Figure4. The performance of autoencoders can be improved by evaluating optimal values for hyperparameters.

2.4 Convolution Neural Network

The biologically inspired variants of multilayer perceptrons are convolutional neural network as shown in Figure5.

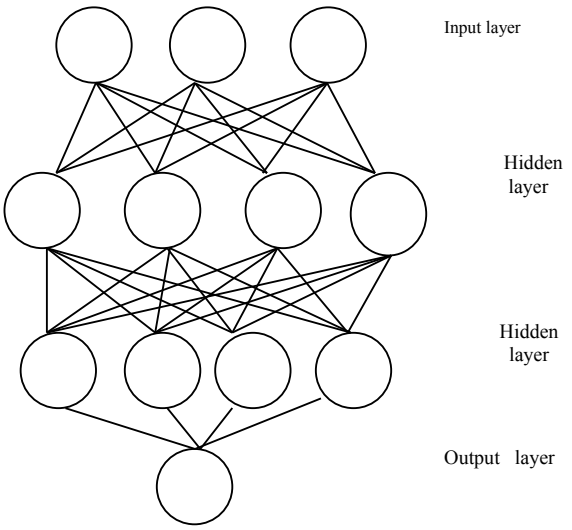


Figure5. DBN Architecture

There are various steps involved in CNN as shown in Figure6. The first layer which is responsible for receiving input vector is called convolution filter. It is the process where the model labels the input vector by referring to what it has taken in the past. The output obtained from this layer transferred to next layer. In order to reduce the sensitivity of the filters to noise, smoothening of inputs from Convolutional layer can be done. The activation layer performs the task of controlling signal flowing from one layer to another layer. In last layers, neurons are fully connected to previous layers. In general it can be stated as:

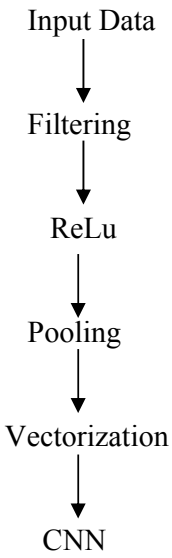


Figure6. Steps involved in CNN

2.5 Recurrent neural network

Recurrent neural networks are deep learning models having good computational power. A RNN has looping structure that makes information to be carried in neurons while scanning the input.

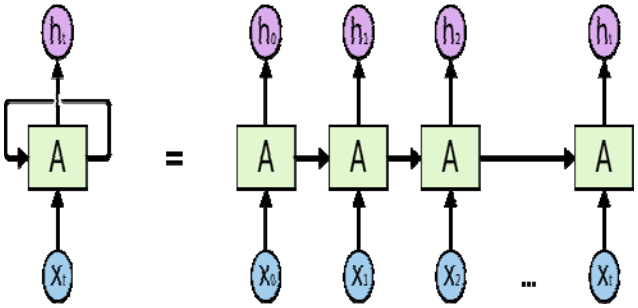


Figure7. RNN architecture

In the above figure X_t belongs to input and h_t belongs to output. The target is to use h_t as output and compare the same with test data. Generally multilayer perceptron can be used for any function approximation then it appears that there is no need of RNN. There are various problems related to time series where RNN will perform better that can store information from long span of time but there exist problem of gradient vanishing problem at the same time.

2.6 Variations of deep learning algorithms

In past few years deep learning equation has gained lot of popularity. Research from various fields used deep learning algorithms to fulfil their requirements.

2.6.1 Advancements in RBM

Currently RBMs are being used for variety of tasks including feature learning, dimensionality

reduction, collaborative filtering etc. Selection of parameters is very crucial for learning algorithms. To avoid this problem Bengio proposed Discriminative Restricted Boltzmann Machines (DRBM). Conditional restricted Boltzmann machines (CRBM) [1] resolved the basic problem of multi-label classification as shown by basic RBM. Using the concept of discriminative learning algorithm, Elfwing proposed the concept of DRBM [2] focusing on important feature i.e. temperature, temperature based restricted Boltzmann was proposed by Li et al. [3].

2.6.2 Advancements in DBN

To mitigate the problem of learning scalability deep convex networks are introduced [4]. Further enhancement in performance can be done by tuning process. DBN in combination with back propagation neural network is also used for designing automatic diagnosis system [5]. DBN in combination of softmax classifier is used for image retrieval [6]. Proposed model performs better than other previous approaches like CBIR (Content based image retrieval), shape base algorithm etc. CDBNs (Convolutional deep belief network) are introduced to enhance the scope of deep belief networks [7].

2.6.3 Advancements in Autoencoders

Denoising Autoencoders (DAE) are introduced to increase the robustness [8, 9]. To solve real time problems k-sparse AE are introduced [10]. Separable deep encoder [11] was designed to deal with zero day noise problems. To enhance the performance of regularised autoencoders, authors [14] proposed contractive autoencoders.

2.6.4 Advancements in CNN

In order to improve the efficiency authors [15] designed recursive Convolutional network (RCN). Feature extraction and feature learning are very important classification process. Jarrett et al. [16] and Masci et al. [17] developed convolution with Autoencoder and stacked convolution Autoencoder. Convolutional restricted Boltzmann machine (CRBM) [18] and CDBNs [19] are widely popular. To train large amount of data a new version of CNN with fast fourier transform [20] has been proposed. Some advance versions of CNN [21] also have been launched to solve various problems like speech recognition and image recognition.

2.7 Deep learning on malware

Lot of articles have been written for malware detection using deep learning. Alom et al. [22] used RBM based DBN on NSL-KDD dataset [23-24] and attained 97.5% accuracy. Li et al. [25] used methods like support vector machine, decision tree etc. for malware detection. Feature extraction was one of the major problems as identified by authors. They used Autoencoder and DBN on KDDCup1999 dataset [26]. Proposed model found satisfactory and better than past model. Tao et al. [27] focused on data fusion algorithms. SVM, J48 and BPNN are used for the task of classification then the authors applied deep Autoencoder algorithm that was far better than other methods for big network traffic classification. Niyaz et al. [28] proposed signature and anomaly based detection technique. Machine learning techniques like artificial neural network, support vector machine, Naive Bayes, self organising map are used to solve the desired purpose. Autoencoder and softmax regression were also used for the research. Deep learning algorithm performed better in various senses like accuracy, precision, recall and F-measure.

Salama et al. [29] proposed a hybrid approach where svm is used in combination with DBN which includes three important steps: preprocessing, DBN feature reduction and classification. NSL-KDD dataset was taken for analysis and authors obtained satisfactory results. Kim et al. [30] proposed architecture by merging the long short term memory and RNN for analysing the intrusion. KDDcup1999 was taken as testing dataset and obtained 98.8% accuracy with false alarm rate of 10%.

An intrusion detection system was designed in software defined network based on NSL-KDD dataset by Tank et al. [31]. Experiments showed that accuracy rate was 91.7%. Yuan et al. [32] implemented an online android malware detection technique in order to identify malicious app and achieved 96.76% accuracy using DBN. Proposed model performs better than C4.5, logistic regression, svm, naive bayes and multilayer perceptron. Kolosnjaji et al. designed deep neural network to process system call sequences and obtained 85.6% precision and 89.4% recall value.

3. Experimental Results

Figure 2 describes the proposed system for malware detection using deep learning. Our goal is to show the effectiveness of deep learning techniques for malware detection. Cuckoo Sandbox [37] is used as virtual machine to execute the entire experiment. Past literature ([38], [39], [40], [41], [42]) show that API calls can be used as important feature for malware classification. API calls are mapped with numerical values so that it can be used as a proper input for different classifiers. By mapping data to certain numeric values we found large vectors.

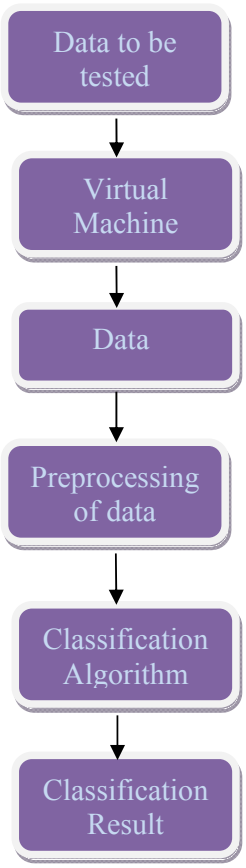
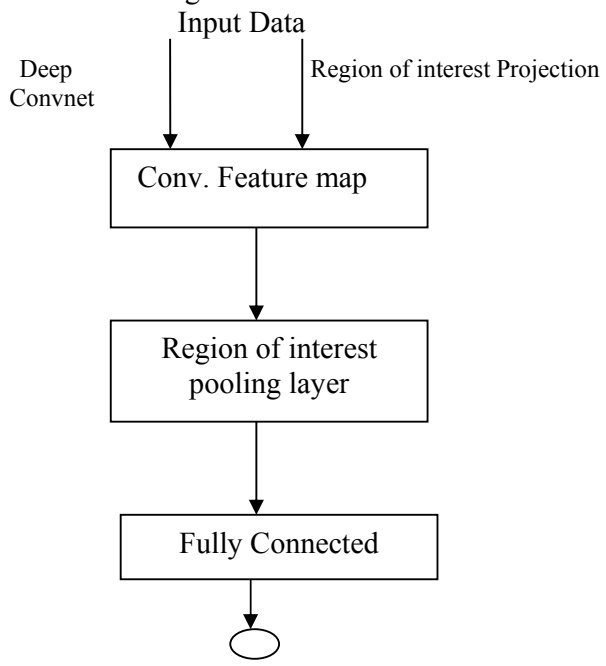


Figure8. Architecture of proposed method

In order to speed up the classification task, we normalized the data values between -1 and 1. Scaling is done using following equation where μ_i is mean, σ_i is variance and x_i is feature:

$$\frac{x_i - \mu_i}{\sigma_i}$$

We are interested to find out the impact of deep learning techniques for malware detection. Support Vector Machine (SVM), K-Nearest Neighbour, J48 Decision Tree and Fast R-CNN are used as classification algorithms.



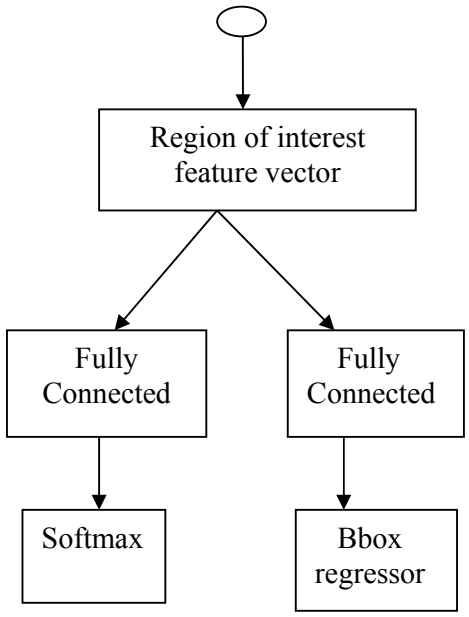


Figure9. Model of Fast RNN

Figure9 explains the model of fast RNN [33]. Fast RNN process input with many Convolutional and max pooling layer to create Convolutional feature map. Purpose of region of interest is to extract feature vector, it executes by dividing the h * w window into a grid of H * W sub windows and the network is completely connected with softmax and bbox regressor. Model produces softmax probabilities and per class bounding box regression offset. The model is trained end to end with multitask loss. Finally, Fast R-CNN can be concluded as an efficient model that trains nine times faster than R-CNN and three times faster than SPPnet. It runs 200 times faster than R-CNN and ten times faster than SPPnet [43].

Table 1. Accuracy values for malware dataset.

Classifier	Sequence length	Accuracy
SVM	11	81.23
KNN	11	84.54
Decision Tree (J48)	12	89.74
RNN	20	97.89
Fast R-CNN	21	98.66

4. Conclusion and Future Work

Deep learning is an extension of machine learning. In this paper we presented a survey of latest deep learning techniques. Deep learning techniques have wide range of applications in pattern recognition, speech recognition etc. Restricted Boltzmann machine, Deep belief Networks, Autoencoders, Convolutional Neural

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Networks and Recurrent Neural Networks are discussed with proper examples. Recent variations in these models are also discussed. In this analysis we took 458 malware samples and 500 benign files for analysis. API sequence is used as feature for classification. Conventional machine learning and deep learning models are used for classification where Fast R-CNN performs better than all techniques taken for analysis as shown in Table1.

As lots of advancements are going on in the field of hardware resources, it will lead to better use of deep learning techniques for real time applications. As far as our experimental analysis is concerned we will take more advance learning models for classification. Secondly, the development a deep learning model that can classify malicious data with less training samples is one of the important questions in malware research. Third point is to design a deep learning method that can work properly for imbalance dataset. Fourth point is the use of advance pre-processing technique for malware dataset. Fourth point is the intelligent used of optimization techniques with deep learning algorithms. Fifth point is the selection of features in dataset. More number of appropriate features may lead to better accuracy values.

As discussed in section 2.6 there are lots of variants of Restricted Boltzmann machine, Deep belief Networks, Autoencoders, Convolutional Neural Networks and Recurrent Neural Networks. Conditional restricted Boltzmann machines (CRBM), Seperable deep encoders, Stacked Convolutional Autoencoders, Convolutional Restricted Boltzmann Machines have shown promising results for various problems of pattern recognition and speech processing. Now it will be interesting to evaluate the impact of these recently evolved techniques in the field of malware detection.

References

[1] Mnih, Volodymyr, Hugo Larochelle, and Geoffrey E. Hinton. "Conditional restricted boltzmann machines for structured output prediction."*arXiv:1202.3748* (2012).

[2] Elfwing, Stefan, Eiji Uchibe, and Kenji Doya. "Expected energy-based restricted Boltzmann machine for classification." *Neural networks* 64 (2015): 29-38.

[3] Li, Guoqi, et al. "Temperature based restricted boltzmann machines." *Scientific reports* 6 (2016).

[4] Deng, Li, and Dong Yu. "Deep convex net: A scalable architecture for speech pattern classification." *Twelfth Annual Conference of the International Speech Communication Association*. 2011.

[5] Abdel-Zaher, Ahmed M., and Ayman M. Eldeib. "Breast cancer classification using deep belief networks." *Expert Systems with Applications* 46 (2016): 139-144.

[6] Liao, Bin, et al. "An image retrieval method for binary images based on DBN and softmax classifier." *IETE Technical Review*32.4 (2015): 294-303.

[7] Arel, Itamar, Derek C. Rose, and Thomas P. Karnowski. "Deep machine learning-a new frontier in artificial intelligence research [research frontier]." *IEEE computational intelligence magazine* 5.4 (2010): 13-18.

[8] Vincent, Pascal. "A connection between score matching and denoising autoencoders." *Neural computation* 23.7 (2011): 1661-1674.

[9] Vincent, Pascal, et al. "Extracting and composing robust features with denoising autoencoders." *Proceedings of the 25th international conference on Machine learning*. ACM, 2008.

[10] Makhzani, Alireza, and Brendan Frey. "K-sparse autoencoders." *arXiv preprint arXiv:1312.5663* (2013).

[11] Sun, Meng, Xiongwei Zhang, and Thomas Fang Zheng. "Unseen noise estimation using separable deep auto encoder for speech enhancement." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.1 (2016): 93-104.

[12] Bist, Ankur Singh. "Detection of metamorphic viruses: A survey." *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*. IEEE, 2014.

[13] Bist, Ankur Singh, and Sunita Jalal. "Identification of metamorphic viruses." *Advance Computing Conference (IACC), 2014 IEEE International*. IEEE, 2014.

[14] Rifai, Salah, et al. "Contractive auto-encoders: Explicit invariance during feature extraction." *Proceedings of the 28th International Conference on International Conference on Machine Learning*. Omnipress, 2011.

[15] Eigen, David, et al. "Understanding deep architectures using a recursive convolutional network." *preprint arXiv:1312.1847* (2013).

[16] Jarrett, Kevin, Koray Kavukcuoglu, and Yann LeCun. "What is the best multi-stage architecture for object recognition?." *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009.

[17] Masci, Jonathan, et al. "Stacked convolutional auto-encoders for hierarchical feature extraction." *Artificial Neural Networks and Machine Learning–ICANN 2011* (2011): 52-59.

[18] Desjardins, Guillaume, and Yoshua Bengio. "Empirical evaluation of convolutional RBMs for-vision." *DIRO, Université de Montréal* (2008): 1-13.

[19] Krizhevsky, Alex, and G. Hinton. "Convolutional deep belief networks on cifar-10." *Unpublished manuscript* 40 (2010).

[20] Mathieu, Michael, Mikael Henaff, and Yann LeCun. "Fast training of convolutional networks through ffts." *arXiv preprint arXiv:1312.5851* (2013).

[21] Sainath, Tara N., et al. "Learning filter banks within a deep neural network framework." *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013.

[22] Alom, Md Zahangir, VenkataRamesh Bontupalli, and Tarek M. Taha. "Intrusion detection using deep belief networks." *Aerospace and Electronics Conference (NAECON), 2015 National*. IEEE, 2015.

[23] Kim, Sang Kyun, Peter Leonard McMahon, and Kunle Olukotun. "A large-scale architecture for restricted boltzmann machines." *Field-Programmable Custom Computing Machines (FCCM), 2010 18th IEEE Annual International Symposium on*. IEEE, 2010.

[24] Kayac, H. G. "k, AN Zincir-Heywood, and MI Heywood," Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets,"." *Third Annual Conference on Privacy, Security and Trust, St. Andrews, New Brunswick, Canada*. 2005.

[25] Li, Yuancheng, Rong Ma, and Runhai Jiao. "A hybrid malicious code detection method based on deep learning." *methods* 9.5 (2015).

[26] Tavallaee, Mahbod, et al. "A detailed analysis of the KDD CUP 99 data set." *Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on*. IEEE, 2009.

[27] Tao, Xiaoling, et al. "A Big Network Traffic Data Fusion Approach Based on Fisher and Deep Auto-Encoder." *Information* 7.2 (2016): 20.

[28] Tavallaee, Mahbod, et al. "A detailed analysis of the KDD CUP 99 data set." *Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on*. IEEE, 2009.

[29] Salama, Mostafa, et al. "Hybrid intelligent intrusion detection scheme." *Soft computing in industrial applications* (2011): 293-303.

[30] Sak, Haşim, Andrew Senior, and Françoise Beaufays. "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." *Fifteenth Annual Conference of the International Speech Communication Association*. 2014.

[31] Tang, Tuan A., et al. "Deep Learning Approach for Network Intrusion Detection in Software Defined Networking." *Wireless Networks and Mobile Communications (WINCOM), 2016 International Conference on*. IEEE, 2016.

[32] Yuan, Zhenlong, Yongqiang Lu, and Yibo Xue. "Droiddetector: android malware characterization and detection using deep learning." *Tsinghua Science and Technology* 21.1 (2016): 114-123.

[33] Girshick, Ross. "Fast r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2015.

[34] http://vxer.org/, Last visited: 26 January 2018.

[35] https://keras.io/, Last visited: 26 January 2018.

[36] https://www.tensorflow.org/, Last visited: 26 January 2018.

[37] Qiao, Yong, et al. "Analyzing malware by abstracting the frequent itemsets in API call sequences." *Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on*. IEEE, 2013.

[38] Wu, Songyang, et al. "Effective detection of android malware based on the usage of data flow APIs and machine learning." *Information and Software Technology* 75 (2016): 17-25.

[39] Salehi, Zahra, Ashkan Sami, and Mahboobe Ghiasi. "MAAR: Robust features to detect malicious activity based on API calls, their arguments and return values." *Engineering Applications of Artificial Intelligence* 59 (2017): 93-102.

[40] Calleja, Alejandro, et al. "Picking on the family: Disrupting android malware triage by forcing misclassification." *Expert Systems with Applications* 95 (2018): 113-126.

[41] Lin, Chih-Hung, Hsing-Kuo Pao, and Jian-Wei Liao. "Efficient dynamic malware analysis using virtual time control mechanics." *Computers & Security* (2017). [42] Aafer, Yousra, Wenliang Du, and Heng Yin. "Droidapiminer: Mining api-level features for robust malware detection in android." *International Conference on Security and Privacy in Communication Systems*. Springer, Cham, 2013.

[42] Zhou, Yajin, and Xuxian Jiang. "Dissecting android malware: Characterization and evolution." *Security and Privacy (SP), 2012 IEEE Symposium on*. IEEE, 2012.

[43] https://github.com/rbgirshick/fast-rcnn, Last visited: 26 January, 2018.

[44] Makandar, Aziz, and Anita Patrot. "Malware image analysis and classification using support vector machine." *Int. J. Adv. Trends Comput. Sci. Eng* 4.5 (2015): 01-03.

[45] Ham, Hyo-Sik, et al. "Linear SVM-based android malware detection." *Frontier and innovation in future computing and communications*. Springer, Dordrecht, 2014. 575-585.

[46] Alam, Mohammed S., and Son T. Vuong. "Random forest classification for detecting android malware." *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing*. IEEE, 2013.

[47] CAI, Zhi-ling, and Qing-shan JIANG. "Android Malware Detection Framework using Protected API Methods with Random Forest on Spark." *Artificial Intelligence* 10 (2016): 9789813206823_0019.

Prevention of Cross-Site Scripting using Hash Technique

Dr.Muna M.T.Jawhar
Dept. of Software Engineering
College of Computer Sc. & Math, University of Mosul.
Mosul, Iraq.

Abstract— **Cookies are the mechanisms that maintain an authentication state between the user and web application. Therefore cookies are the possible targets for the attackers. Cross Site Scripting (XSS) attack is one of such attacks against the web applications in which a user has to compromise its browser's resources (e.g. cookies). In this paper, a novel technique of SHA_512 Hash Technique is introduced whose aim is to make cookies worthless for the attackers. The work done in HTTP protocol with windows10.**

Keywords-component; ***Cookies, HTTP protocol, Cross-Site Scripting Attacks, Hash function.***

I. INTRODUCTION

Normally, users in client side through web browsers request the resources from the server side through web server of the web application, and the web server respond with the resources through HTTP protocol [1] in which no sessions are retained [2]. Therefore, web applications generally use cookies to provide a mechanism for creating stateful HTTP sessions. Cookies are often used to store the session [3] for the web applications that require authentication. Since the cookies can both identify and authenticate the users [4], this makes the cookies a very interesting target for the attackers. Now-a-days, Cross-Site Scripting (XSS) attack is a common vulnerability which is being exploited in modern web applications through the injection of advanced HTML tags and Java Script functions. A weak input validation on the web application causes the stealing of cookies from the web browser's database. In many cases, the attacker who can obtain the valid cookies from XSS attack can directly hijack the user's session.

Cross-Site Scripting attack continuously leads the most wide spread web application vulnerabilities lists (e.g. OWASP [5] etc.). XSS are broadly classified into two main attacks which are Persistent and Non-Persistent Attacks [6] [7]. Persistent attack (also called as stored attack) holes exist when an attacker post the malicious code on the vulnerable web application's repository. As a result, if the stored malicious code gets executed by the victim's browser, then stored attack gets exploited on the victim's web browser. Secondly non-persistent attack (also called as reflected attack) means that the vulnerable malicious code is not persistently stored on a web server but it is immediately displayed by the vulnerable web application back to the victim's web browser. If so, then the malicious code gets executed on the victim's web browser and finally, victim's browser has to compromise its resources (e.g. cookies). The rest of the paper is organized as follows. Section II discusses the related works, and section III mention background of cookies and the XSS attack, detection and prevention of XSS attacks. Section IV discusses our proposed technique. Section V discusses our proposed technique. Finally we conclude and brief the future work in last section.

II. RELATED WORK

A.S. Christensen, A. Moller, and M.I. Schwartzbach suggested the study of static string analysis for imperative languages. They have shown usefulness of string analysis for analyzing reflective code in Java programs and checking for errors in dynamically generated SQL queries. They used finite state automata (FSAs) as a target language representation to analyze Java. Methods from computational linguistics were also applied to generate good Finite State Automata approximation of CFGs [8].

Y.W Huang and others used counterexample traces to minimize the number of sanitization routines inserted and to identify the reason of errors that enhance the precision of both code instrumentation and error reports [9]. Variables representing current trust level were assigned states which further were used in verifying the legal information flow in a web application. Now in order to verify the correctness of all safety states of program Abstract Interpretation, Bounded Model Checking technique was used[10].

In [11] the authors provide an approach to address security risks by using proven security library known as Enterprise Security API from Open Web Application Security Project. They also provided an assessment of the approach against the existing way of handling cross site scripting vulnerabilities.

In [11], a Webmail XSS fuzzer called L-WMxD (Lexical based Webmail XSS Discoverer), which works on a lexical based mutation engine is an active defence system to discover XSS before the webmail application is online for service. The researchers have run the L-WMxD on over 26 real-world Webmail applications and found vulnerabilities in 21 Webmail services, including some of the most widely used Yahoo-Mail.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

In [12] authors conduct a thorough analysis of the current state-of-the-art in browser-based XSS filtering and uncover a set of conceptual shortcomings, that allow efficient creation of filter evasions, especially in the case of DOM-based XSS. To validate their findings, they reported on practical experiments using a set of 1,602 real-world vulnerabilities, achieving a rate of 73% successful filter bypasses. Motivated by their findings, they proposed an alternative filter design for DOM-based XSS, that utilizes runtime taint tracking and taint-aware parsers to stop the parsing of attacker controlled syntactic content. To examine the efficiency and feasibility of their approach, they presented a practical implementation based on the open source browser Chromium.

III. CROSS-SITE SCRIPTING

Cross-site scripting (XSS) attack is one of the most common vulnerabilities in web applications. It considered as one of the top 10 web application vulnerabilities of 2013 by the Open Web Application Security Project (OWASP) [13]. According Cenzic Application Vulnerability Trends Report (2013) Cross Site Scripting represents 26% of the total population respectively [14] and considers as top most first attack. Two recent incidents highlighted the severity of xss vulnerability are Apple Developer Site (July 18, 2013) and Ubuntu Forums (July 14 and July 20, 2013) [15]. Cross Site Scripting attack carried out using HTML, JavaScript, VBScript, ActiveX, Flash, and other client-side languages. A weak input validation on the web application leads Cross Site Scripting attacks to gather data from account hijacking, changing of user settings, cookie theft. Detection or prevention of XSS is a topic of active research in the industry and academia. To achieve those purposes, automatic tools and security system have been implemented, but none of them are complete or accurate enough to guarantee an absolute level of security on web application. One of the important reasons of this shortcoming is that there is lack of common and complete methodology for the evaluation either in terms of performance or needed source code modification which in an overhead for an existing system. A mechanism which will easily deployable and provide a good performance to detect and prevent the Cross-site scripting (XSS) attack is essential one.

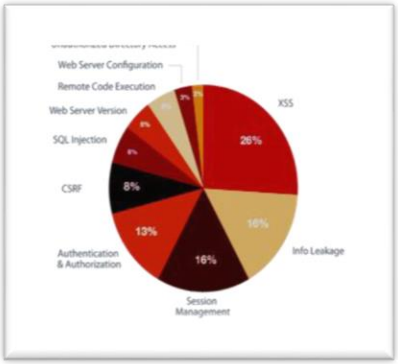


Figure 1. Web Application Security Vulnerability Population (2013)

Types of Cross-Site Scripting Attack

There are three distinct types of XSS attacks: non persistent, persistent, and DOM-based.

A. *Persistent XSS Attack:*

In the persistent XSS an attacker can inject the malicious code into the page persistently and that means the code will be stored in the target servers as an html text, such as in a database, in a comment field, messages posted on forums, etc., and this code will be stored in the page which will show to the user victim later on. If the user victim goes to the page which is embedded with XSS attacking code, the code will execute on the user victim's browser, which, in turn sends the user's sensitive information from his site to the attacker's site. The persistent XSS attack is also known as stored XSS attack. Compared with "REFLECTED XSS", this type of XSS does more serious harm. If the "STORED XSS" vulnerability is successfully exploited by hackers, it will persistently attack the users until administrator remove this vulnerability. The following Figure 2. shows architecture of exploiting the persistent XSS attack by a malicious attacker.

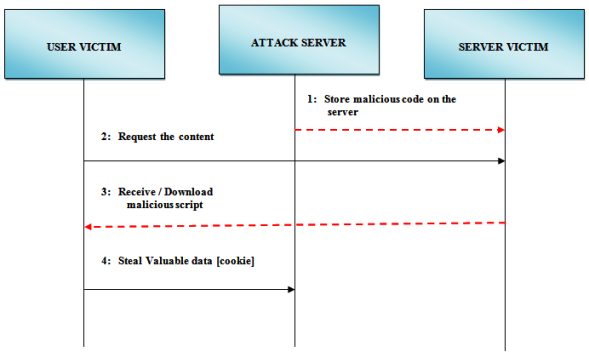


Figure 2. Architecture of Exploiting the Persistent XSS Attack

B. *Non-persistent XSS Attack:*

Non-persistent cross-site scripting vulnerability is the common type of XSS attacks. The attack code is not persistently stored, but, instead, it is immediately reflected back to the user. It is also known as reflected XSS attack. In this type the injected code is sent back to the user victim off the server, such as in an error message, search result, or any other response that includes some or all of the input sent to the server as part of the request. To do this, the attacker sends a link to the user victim (*e.g.*, by email). If the user victim clicks on the link, the vulnerable web application displays the requested web page with the information passed to it in this link. This information contains the malicious code which is now part of the web page that is sent back to the web browser of the user, where it is executed. The following Figure 3 is an architecture which shows the sequence of steps of exploiting the reflected XSS vulnerability by a malicious attacker.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

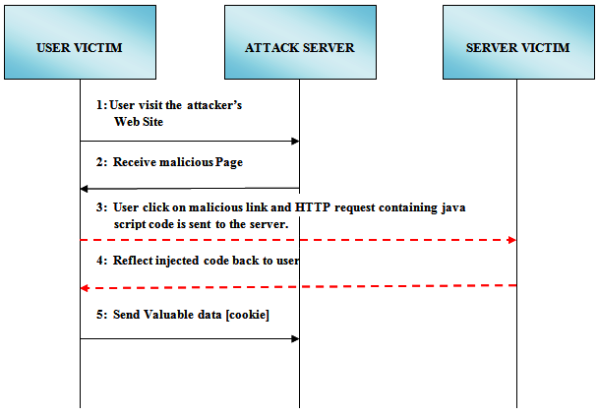


Figure 3. Architecture of Exploiting the Non-persistent XSS Attack

C. *DOM-based XSS:*

DOM-based cross-site scripting attacks are performed by modifying the DOM "environment" in the client side instead of sending any malicious code to server. DOM is short for Document Object Model and it is a platform and language neutral interface. DOM allows the scripting to change the HTML or XML document, the HTML or XML document can be modified by a hacker's scripting or program. Therefore, DOM-based XSS uses DOMs vulnerability to make the XSS come true. This type of XSS vulnerability is totally different from the reflected or stored XSS attack and it does not inject malicious code into a page. So, it is the problem of the insecure DOM object which can be controlled by the client side in the web page or application. For this reason, hackers can let the attack payload execute in the DOM environment to attack the Victim side. The following Figure 4 is an architecture which shows the sequence of steps of exploiting the reflected XSS vulnerability by a malicious attacker.

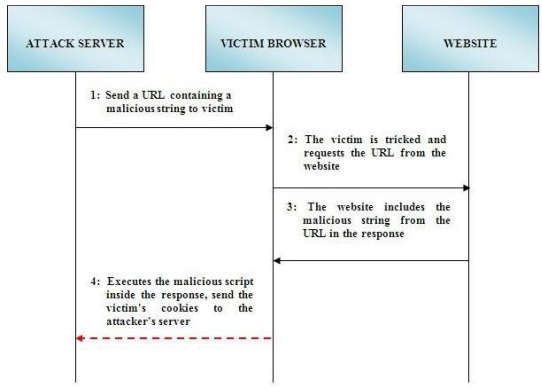


Figure 4. Architecture of Exploiting the DOM-based XSS Attack

IV. What is a Cookie?

Cookies are the best current way to identify users and allow persistent sessions[16]. Cookies are small repositories of data that are stored within your web browser by a web server. Cookies were first developed by Netscape but now are supported by all major browsers[17]. They are rife with security concerns, and some of them can even track your online activity. Whenever you visit a web site, the cookie stored in your browser serves as a type of ID card. Each additional time you login or request resources from the same web server, the cookie saved in your browser sends its stored data to the web server. This allows web site administrators, and even Internet marketers, to see which of their pages are getting the most hits, how long users stay on each page, which links they click on, and a wealth of other information.

And believe it or not, cookies are extremely prevalent these days. Have you ever purchased something from Amazon.com? If so, then you've used cookies before, whether you knew it or not. It's quite common for online ecommerce sites to use cookies to record and store personal information you've entered, which products you've searched for, which items are in your online shopping cart, and other information so it doesn't need to be tediously reentered each time you want to make a purchase.

Furthermore, cookies are used to make a website more personal. Many sites offer preference options to let you customize the look, feel, and experience of any given web service. Once you revisit the site or resource, you'll find that all your preferences were preserved. Though cookies make browsing the web a lot more convenient, they do have a lot of security drawbacks, as we'll discuss next.

V. Types of Cookies and Security Problems

We can classify cookies broadly into two types: session cookies and persistent cookies. A session cookie is a temporary cookie that keeps track of settings and preferences as a user navigates a site. A session cookie is deleted when the user exits the browser. Persistent cookies can live longer; they are stored on disk and survive browser exits and computer restarts. Persistent cookies often are used to retain a configuration profile or login name for a site that a user visits periodically. The only difference between session cookies and persistent cookies is when they expire [18].

There are two different versions of cookie specifications in use: Version 0 cookies (sometimes called "Netscape cookies"), and Version 1 ("RFC 2965") cookies. Version 1 cookies are a less widely used extension of Version 0 cookies [15].

The cookie contain the following information:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE I. THE COOKIE CONTAIN THE FOLLOWING INFORMATION

Information	Value	Optional or original
Cookie attribute	NAME=VALUE.	Original
Expires	contain the time and date.	Optional
Domain	A browser sends the cookie only to server hostnames in the specified domain like "acme.com"	Optional
Path	This attribute lets you assign cookies to particular documents on a server	Optional

even though your browser has ways to manage cookies, some are nearly impossible to delete. The problem is that special types of cookies aren't stored within your browser, so even if you opt for a different web browser (Firefox, Chrome, etc.), the cookie will still be active. And many of these types of cookies are much larger than the average 4KB HTTP cookies – some of them ranging to 100KB or even 1MB. If you attempt to delete the cookie but notice that it keeps coming back every time you restart your browser, you've discovered a zombie cookie and may need special security software to remove it.

VI. PROPOSED METHOD

In this section, we present a novel procedure whose main objective is to make the cookies useless for the attackers. This approach is easily implemented on the web server without any changes required on the web browser. In this paper the web server will produce a hash of value of name attribute and domain in the cookie and send this hash value to the browser, so the browser will keep the hash value of cookie in its database rather than the original value. Now each time, if the browser wants to reconnect as a part of active connection, the browser has to include the hash cookie value into its corresponding request so that the web server will also rewrite this hash cookie value to the original value, which is generated by the web server. Rewriting of hash value to original value is necessary to be done at the server side, so that the user at the browser side will get authenticated by the web server. As the browser stores the hash value of cookies, so even the XSS attack can steal the cookies from browser's database, the cookies cannot be used later to hijack or take off the user's session.

We have conducted the experiments on version 0 cookies in which three attributes (name, domain) are specified for the identifying the cookies uniquely. First we must bring the cookies by capture it. We used Web cookies sniffer for capture the cookies as in the figure (5), it is a program to capture cookies in real time from network is a Windows utility, a new application by Nirsoft that captures all cookies saved on your computer by websites via browsers and applications, and then provide you with all information about the saved cookie[19].

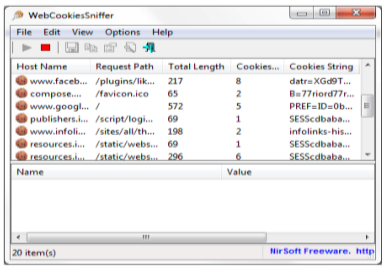


Figure 5: capture program

The cookies stored in files, then we take the name and domain to generate the hash value, all the other attributes will remain same.

We take this cookies, name and path

hp@microsoft
microsoft.com/

We have used the SHA Hashing function for generate hashing value. Following are some of the steps which are used to explain the SHA-512 Hashing Technique[18][20]:-

- Append padding bits.
- Append length.
- Initialize hash buffer.
- Process the message in 1024 bit (128 word) which forms the heart of the algorithm.
- Output the final state value as the resulting hash.

The hash output for the previous cookies is :

fe6eb4f1e09bb3ca5cea07ba9f5f5a8aa462a1adb1634104fbe60d64ab1d33dfd08f0344eeeea5033efb066d7ba23bd84d428e3b78964fcc7c2a96b7037968ba

Figure 6: the output of SHA-512

The user on the Clint side submits the user-id and password to the web server of the web application with HTTP protocol.

- The web server submits the corresponding information from the browser and generates a cookie.
- Now the web server will dynamically generate the hash of value of the name attribute in the cookie and store both these values (original as well as hash value) in the form of a table on the server side.

TABLE II: THE COOKIE CONTAIN THE FOLLOWING INFORMATION

Cookies	Hash value
hp@microsoft microsoft.com/	fe6eb4f1e09bb3ca5cea07ba9f5f5a8aa462a1adb1634104fbe60d64ab1d33dfd08f0344eeeea5033efb066d7ba23bd84d428e3b78964fcc7c2a96b7037968ba

- Subsequently, the web server will send the hash value of the name attribute in the cookie to the web browser.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- The web browser will store this hash value into its repository.

Since the cookies at the browser's database now are not valid for the web applications. Therefore XSS attack will not be able to impersonate the user using stolen cookies which are converted into its hash form. Now if the browser wants to reconnect to the web server as a part of the active connection, it has to include cookie (hash value) with its corresponding request to the web server. The web server will use the information in the table to rewrite back the values of name attribute in the cookie (sent by the web browser) to the original value generated by the web server as shown in the next Figure.

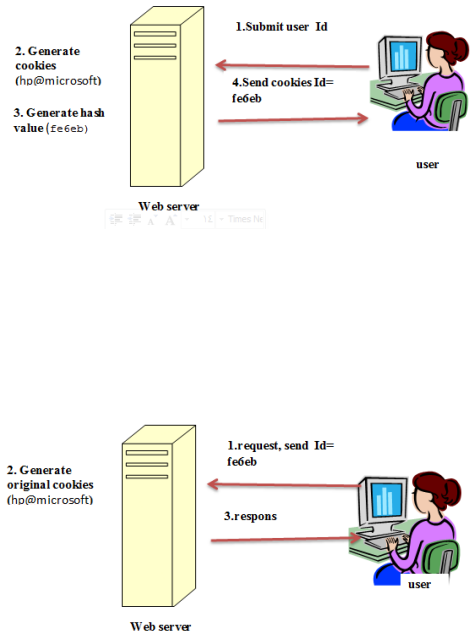


Figure 7: communication between user and web server

VII. Conclusion and Future Work

This paper has presented the SHA-152 Hash function technique, whose main purpose is to make the cookies worthless for the attackers even if the attacker successfully exploits the vulnerabilities of victim's web browser. This technique has been implemented on the web server and the results showed that our technique worked well with the Version 0 cookies on the modern web browsers with HTTP protocol.
Currently we are working on how our proposed technique works with the Version 1 cookies on the real world websites. In future, we would also like to develop an analysis on the Clint.

References

[1] D. Gourley, B. Totty, M. Sayer, S. Reddy, and A. Aggarwal, HTTP The Definitive Guide, 1st ed., O'Reilly Media, US, 2002.

[2] Behrouz A. Forouzan, TCP/IP protocol suite, third edition, 2008.

[3] D. Kristol, "HTTP State Management Mechanism,", in Internet Society, 2000. Available: http:// www.ietf.org/rfc/rfc2965.txt.

[4] "Cross Site Scripting Techniques and mitigation,", GovCertUK, revision 1.0, October 2009. Available: www.govcertuk.gov.uk.

[5] Open Web Application Security Project, "The ten most critical web application security vulnerabilities", 2007, www.owasp.org/index.php/OWASP_Top_Ten_Project.

[6] J. Garcia-Alfaro and G. Navarro-Arribas, "Prevention of Cross-Site Scripting Attacks on Current Web Applications,", Available: http://hacks-galore.org/guille/pubs/is-otm-07.pdf.

[7] S. Saha, "Consideration Points: Detecting Cross-Site Scripting," (IJCSIS) International Journal of Computer Science and Information Security,Vol. 4, No. 1 & 2, 2009.

[8] A. S. Christensen, A. Møller and M. I. Schwartzbach, "Precise analysis of string expression", In proceedings of the 10th international static analysis symposium, LNCS, Springer-Verlag, vol. 2694, pp. 1-18.

[9] Y. W Huang, F. Yu, C. Hang, C. H. Tsai, D. Lee and S. Y. Kuo, "Verifying Web Application using Bounded Model Checking," In Proceedings of the International Conference on Dependable Systems and Networks, (2004).

[10] Bhanu Prakash Gopularam1 and N. Nalini2," Cross Site Scripting Security Vulnerabilities and Risk Mitigation Using Enterprise Security API for Web-Apps on Public Infrastructure", Elsevier Publications 2013.

[11] Zhushou Tang, Haojin Zhu, Zhenfu Cao, Shuai Zhao, L-WMxD: Lexical Based Webmail XSS Discoverer, IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 976-981, 2011.

[12] Ben Stock, Sebastian Lekies and Tobias Mueller ,"Precise client-side protection against DOM-based Cross-Site Scripting", This paper is included in the Proceedings of the 23rd USENIX Security Symposium. August 20–22, 2014.

[13] https://www.owasp.org/index.php/Category:OWASP_Top_Ten_Project.

[14] H. Chen and M. V. Gundy, "Using randomization to enforce information flow tracking and thwart crosssite scripting attacks," In Proceedings of the 16th Annual Network and Distributed System Security Symposium (NDSS), (2009).

[15] D. Gourley, B. Totty, M. Sayer, S. Reddy, and A. Aggarwal, HTTP The Definitive Guide, 1st ed., O'Reilly Media, US, 2002.

[16] http://www.thewindowsclub.com/capture-cookies-save-computer-webcookiessniffer

[17] William Stallings, "Cryptography and network security, principles and practices", fifth edition, 2008.

[18] Aumasson J.P, Meier W., phan R. and Henzen L., "the hash function black", 2014.

[19] V. Nithya1, S. Lakshmana Pandian2 and C. Malarvizhi3, " A Survey on Detection and Prevention of Cross-Site Scripting Attack", International Journal of Security and Its Applications Vol. 9, No. 3 (2015), pp. 139-152

[20] https://isis.poly.edu/~jcappos/papers/tr-cse-2013-02.pdf.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Non-Uniform Gap Distribution Library Sort

Neetu Faujdar, Rakhi Mittal
Amity University Uttar Pradesh, Department of CSE, Noida, India
Email: neetu.faujdar@gamil.com, rakhimittal1108@gmail.com

***Abstract*—**The current study is focused towards a sorting algorithm referred as Library Sort. Previous scientific limitations have revealed many utilities of sorting algorithms. Each sorting algorithm varies according to their strength and weaknesses. The Bender et al proposed the library sort with uniform gap distribution (LUGD) i.e. there is equal number of gap provided after each element. But what happens if there are many elements that belong to the same place in the array and there is only one gap after that element. To overcome the problems of LUGD, this paper focuses on the implementation of library sort algorithm along with non-uniform gap distribution (LNGD). The proposed algorithm is investigated using the concept of mean and median. The LNGD is tested using the four types of test cases which are random, reverse sorted, nearly sorted and sorted dataset. The experimental result of proposed algorithm is compared to the LUGD and LNGD proved to provide better results in all the aspects of execution time like re-balancing and insertion. Improvement has been achieved that ranges from 8% to 90%.

***Index Terms*—** Sorting, Insertion Sort, Library Sort, LNGD.

I. INTRODUCTION

Michael A. Bender proposed the library sort which can be defined as follows [2], it is a sorting algorithm that uses insertion sort, but with gaps in the array to accelerate subsequent insertions [3]. Library sort is also called the gapped insertion sort. Suppose, if a librarian wants to keep his books alphabetically on a long shelf starting with the alphabet '*A*' and continuing to the right till the alphabet '*Z*' and there is no gap between the books, and if the librarian has some more books that belong to section '*B*', then he has to search the correct place in section '*B*'. Once he finds the correct place in section '*B*', he will have to move every book from the middle of section '*B*' to the end of the section '*Z*' in order to make space for the new book. This is called insertion sort. However, if the librarian leaves a gap after every letter book, then there will be a space after section '*B*' letter books and then he will only have to move a few books to make space for the new books, this is the primary principle of library sort. Thus the library sort is an extension of insertion sort. The author achieves the $O(\log n)$ insertions with high probability using the evenly distributed gap, and the algorithm runs $O(n\log n)$ time complexity with high probability [7]. The time complexity of library sort $O(n\log n)$ is better than the insertion sort, which is $O(n^2)$ [3]. In library sort author used the uniform gap distribution after each element and gap is denoted by the 'ε'. The detailed analysis of the library sort has been done in the paper [1]. By execution time analysis, author found that, execution time is inversely proportional to the value of epsilon in most of the cases. At some point the value of epsilon (ε) reached at the saturation point due to the presence of extra spaces. Later, these extra spaces are used to insert the data. By space complexity analysis, found that the space complexity of the library sort algorithm increases linearly that is, when the value of epsilon increases, the memory consumption also increased in the same proportion. By execution time analysis of re-balancing, found that increase the re-balancing factor '*a*' from 2 to 4 then the execution time of library sort algorithm will also increase as it moves towards the traditional insertion sort. So, to find out the best result of library sort algorithm, the value of epsilon should be optimized and re-balancing factor should be minimized or ideally equal to 2. This is what about the library sort using the uniform gap distribution. The application of leaving gaps for insertions in a data structure is used by [8]. This idea has found recent application in external memory and cache-oblivious algorithms in the packed memory structure of Bender, Demaine and Farach-Colton and later is used in [9-10].

In this paper, a new approach proposed that is library sort with non-uniform gap distribution (LNGD). And new approach results are analyzed using the four types of test cases which are random, nearly sorted, reverse sorted and sorted test cases. The paper also compares the LUGD with the LNGD in all the aspects of execution time like re-balancing and insertion.

The remaining section of this paper is organized as follows: Section 2 contains the detailed description of LNGD algorithm. The execution time comparison of LUGD and LNGD has been shown in section 3. The testing and comparison based on re-balancing is described in section 4. The performance analysis in all aspects has been shown in section 5. In section 6 concluded the presented work along with its future scope.

II. RELATED WORK

In this section, briefly surveyed related work based on insertion and library sort. Tarundeep *et al* presented a new sorting algorithm EIS (Enhanced Insertion Sort). It is based on the hybrid sort. The suggested algorithm achieved $O(n)$ time complexity which is compared $O(n^2)$ of insertion sort. In this the effectiveness of the algorithm is also proved. The hybrid based sort is analyzed, implemented, tested and compared with other major sorting algorithms [14].

Partha *et al* shows that how to improve the performance of insertion sort. He suggested a new approach which compared to the original version of

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

insertion sort. The new approach is also compared with bubble sort. The experimental results have shown that proposed approach performs better in worst case [15].

Michael *et al* proposed the library sort algorithm which is also called gapped insertion sort. It is the enhanced of the insertion sort. The authors showed that library sort has insertion time $O(\log n)$ with high probability. The total running time is $O(n \log n)$ with high probability [2].

Franky *et al* investigated the improvement of worst case running time of insertion sort. So the author presented the rotated library sort. The suggested approach has $O(\sqrt{n} \log n)$ operation per insertion and the worst case running time is $O(n^{1.5} \log n)$ [16].

Neetu *et al* overcomes some issues of library sort. The detailed experimental analysis of library sort is done by the author. The performance of library sort is measured using the dataset [1].

III. Problem Statement

Library sort is an improvement of insertion sort with $O(n \log n)$ time complexity. Library sort is also called gapped insertion sort or say that insertion sort with gaps. The Bender et al proposed the library sort with uniform gap distribution i.e. there is equal number of gap provided after each element. But what happens if there are many elements that belong to the same place in the array and there is only one gap after that element.

So to overcome this problem in this paper, library sort with non-uniform gap distribution is proposed i.e. element can have non-uniform gap. The suggested algorithm is investigated using the concept of mean and median.

IV. Proposed Library Sort Algorithm with Non-Uniform Distribution (LNGD)

The LNGD algorithm consists of three steps. The first two steps will be the same as the LUGD algorithm [1], but the third step will be different.

Step1. Binary Search with blanks: In library sort insert a number in the space where it belongs and to find that binary search is used. The array '*S*' is sorted but with gaps. As in computer, gaps of memory will hold some value and this value is fixed to sentinel value that is '-1'. Due to this reason, directly cannot use the binary search for sorting. So, modify the binary search. After finding the mid element, if it comes out to be '-1' then move linearly left and right until get a non-zero value. These values are named as $m1$ and $m2$. Based on these values, define values for new low, high and mid used in a binary search. Another difference in the binary search is that, it is not only searches the element in the list, but also reports the correct position where we have to insert the number [1]. Working of step 1 is illustrated with the help of an example.

***Example*:** In the following array '-1' shows the gaps in the array. The array position is starts from 0 up to 9. Now let search an element say 5.

1	-1	3	-1	5	-1	7	-1	9

low=0
high=9
mid=(0+9)/2 = 4 = *S*[4]

Here *S*[4] = 5 got the element and terminate the search.

1	-1	3	-1	-1	-1	7	-1	9	-1

In this array, do not have element 5 but we are going to search it.

Here also low =0

High=9

Mid =(0+9) /2 =4 = *S*[4]

S[4]= *S*[mid]= -1

In this case, find $m1$ and $m2$ as a mid which are represented by *S*[$m1$] and *S*[$m2$] greater than '-1' in both the direction limiting to low and high respectively. Here the value of $m1$=*S*[2]=3 and the value of $m2$=*S*[6]=7. According to $m1$ and $m2$ values, update the low and high to perform binary search.

Step2. *Insertion*: Library sort is also known by the name gapped insertion sort. The presence of gap for inserting any new number at a particular position removes the task of shifting the elements up to the next gap [1]. Working of step 2 is explained with the help of an example.

***Example*:** Insert the elements in the manner of 2^i in the array. i.e. in the power of 2. This is stored in *S*[*i*]. *S*[*i*] =pow (2, *i*) where '*i*' is the pass number i.e *i*= 0, 1, 2, 3... if *i*=0 then *S*1= 2^0 =1. Now search the position of the insertion element in the array and add the element at position returned by the search function. Next time *i*=1 then *S*1=2^1=2, and *S*[*i*] = pow(2, *i*-1) to pow(2, *i*) i.e. the value of *S*1 is 1 to 2 and so on for all values of '*i*'.

Step3. *Re-balancing*: Re-balancing is done after inserting 2^i elements where *i*=1, 2, 3, 4... and the spaces are added when re-balancing is called. In the previous approach, the gaps were uniform in nature. In the proposed technique, non-uniform gap distribution is given based on the property of insertion sort. This property tells that more updates should be done in the beginning of an array for generating more gaps. Gaps are generated using the equation (1).

$$Ratio = n * (\mu / \sigma) / 2 \qquad (1)$$

Here μ is mean and σ is standard deviation.

$$ee = 2 * n / ratio \qquad (2)$$

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Initially have *e*+*ee* gaps, but '*ee*' is decreased when parsed the number equal to the ratio.

ALGORITHM: LNGD Re-balancing

Input: List of elements and re-balancing factor *e*.
Output: List with non uniform gaps.

```
Compute μ and σ
Ratio ← n*( μ/σ)/2
ee ← 2*n/ratio
        while(l < n) do
                if(j% ratio ==0 && j>0 && e+ee>0
        )then
                ee--
                endif
                if(S[j] ! = -1) then
                        reba[i] = S[j]
                        i++
                        j++
                        l++
                        for k=0 to ee+e do
                                reba[i] = -1
                                i++
                        endfor
                else
                        j++
                endif
                for k = 0 to i do
                        S[k] = reba[k]
                endfor
        endwhile

end
```

V. Execution Time Comparison of LUGD and LNGD

LUGD and LNGD algorithms have been tested on a dataset [T10I4D100K (.gz)] [11-13] by increasing the value of the gap (ε). The dataset contains the 1010228 items. Data set contains four cases [26-30].

(1) Random with repeated data (Random data)
(2) Reverse sorted with repeated Data (Reverse sorted data)
(3) Sorted with repeated data (Sorted data)
(4) Nearly sorted with repeated data (Nearly sorted data)

Table 1 shows the execution time of LUGD and LNGD algorithms in microseconds using the above mentioned cases.

The performance of the LUGD and LNGD are compared with random data, nearly sorted data, reverse sorted and sorted data. The execution time in micro-seconds is presented in Table 1. The Results are presented for different value of 'ε'. Epsilon (ε) is the minimum number of gaps between the two elements. The execution time comparison of LUGD and LNGD algorithms has also been shown in Fig. 1 to 4. In all Fig. 1 to 4, the *X*-axis represents the different value of gap and the *Y*-axis represents the execution time in microseconds.

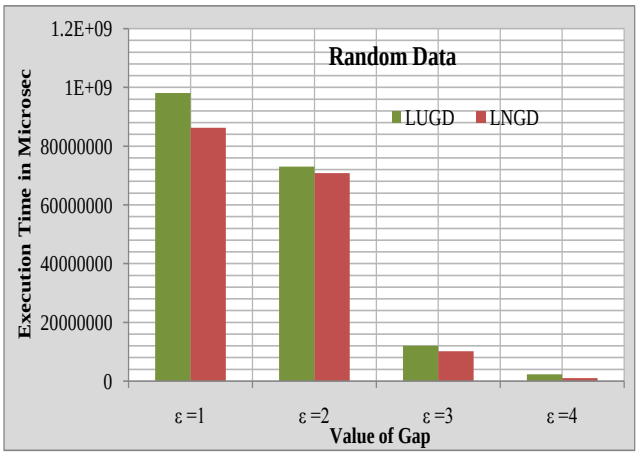


Fig. 1. Execution time comparison of LUGD and LNGD

TABLE 1. Execution Time of LUGD and LNGD Algorithms in *microsecond* based on Gap Values

Time in Microseconds								
Data Set	**Random**		**Nearly Sorted**		**Reverse Sorted**		**Sorted**	
Value of ε	**LUGD**	**LNGD**	**LUGD**	**LNGD**	**LUGD**	**LNGD**	**LUGD**	**LNGD**
ε =1	981267433	862909204	864558882	306063385	1.451E+09	1.329E+09	861929937	313078205
ε =2	729981576	708580455	620115904	230939335	1.065E+09	1.022E+09	609647355	234697961
ε =3	119727535	101921406	358670053	185759986	278810310	125152235	356489846	195120953
ε =4	23003046	10557332	117188830	107729204	263693774	116417058	116590140	106897060

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Fig. 1 shows the comparison of LUGD and LNGD for different values of the gap. It can be seen from the graph that the LNGD has outperformed LUGD. The maximum improvement in execution time by LNGD is 36.7% for the value of ε=4.

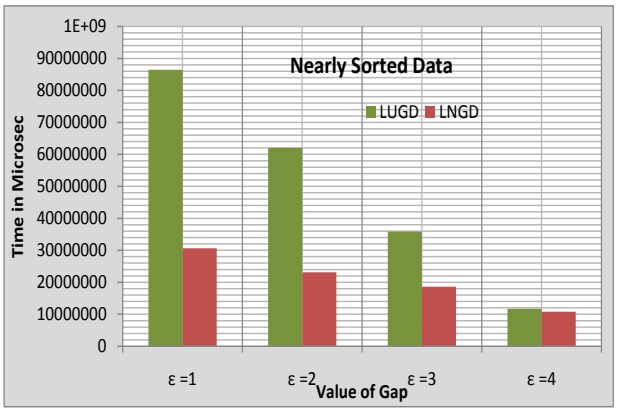


Fig. 2. Execution time comparison of LUGD and LNGD

Fig. 2 describes the execution time of the two algorithms LUGD and LNGD on the nearly sorted data. Major improvement found in the case of ε=1. Observed that the improvement in execution time by LNGD is 64.59% at ε =1. With observations, found that execution time is inversely proportional to the value of 'ε'. The execution time is calculated as 8% in the case of ε=4.

Fig. 3. The execution time comparison between LUGD and LNGD using reverse sorted data. In the case of reverse sorted data the trend for execution time is reversed. It is nearly 8% for the ε =1, and it further decreases for ε =2, ε =3 and ε =4 upto 55%. The same has been shown in Fig. 3.

Fig. 4 describes the execution time of both the algorithms on the sorted data, the improvement can be seen from the ε=1 to ε =4. It is maximum at ε =1 that is 63.67% and minimum at ε =4 that is 8%.

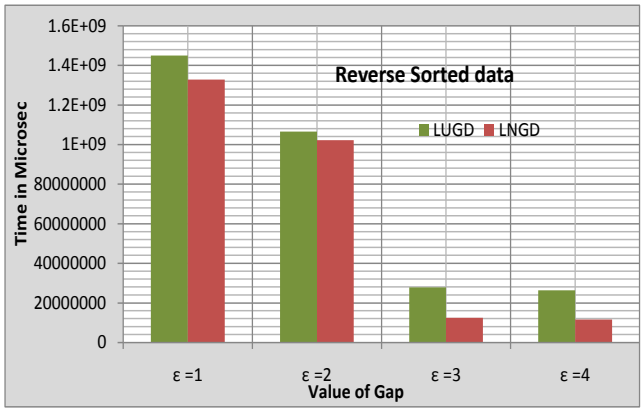


Fig. 3. Execution time comparison of LUGD and LNGD

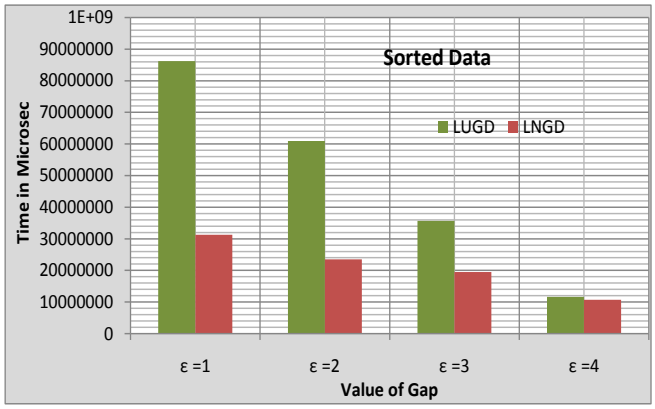


Fig. 4. Execution time comparison of LUGD and LNGD

VI. Rebalancing Based Comparison of LUGD and LNGD

Re-balancing is used after inserting a^i element, this increases the size of the array. The size of the array will depend on 'ε'. To do this process, require an auxiliary array of the same size so as to make a duplicate copy with gaps. Re-balancing is necessary after inserting a^i element, and re-balancing calculated till a^i where 'a'= 2, 3, 4 and i = 0, 1, 2, 3, 4….. with the value of gaps 'ε' = 1, 2, 3, 4. Now showing that how to calculate the re-balancing of LUGD and LNGD with the help of an example.

1. **Example of re-balancing using LUGD algorithm**

(A) For example, when ε=1, then how re-balancing will be performed if a=2.

$2^i = 2^0, 2^1, 2^2, 2^3, 2^4$ ……

=1, 2, 4, 8, 16 ……

1.1 Re-balance for 2^0 =1

1	-1

1.2 Re-balance for 2^1=2

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

1	2

After re-balancing this array is

1	-1	2	-1

1.3 Re-balance for 2^2=4

1	2	3	4

After re-balancing this array is as follows:

1	-1	2	-1	3	-1	4	-1

In this manner, re-balance can done of the array in the power of 2^i.

(B) Example when ε=1, then how re-balancing will be performed if a=3.

$3^i = 3^0, 3^1, 3^2, 3^3, 3^4$ ……
$= 1, 3, 9, 27$……

1.1 Re-balance for 3^0 =1

1	-1

1.2 Re-balance for 3^1=3

In the above array only one space is empty. This shows that only one element can be inserted. On the other hand, according to re-balancing factor 3^1=3 require two spaces in the array. In this situation, need to shift the data to make space for the new element. In this way the performance of the algorithm degrades as having the larger number of swapping to generate the spaces which is same as that in the case of traditional insertion sort.

2. Example of re-balancing using LNGD algorithm

(A) For example, how re-balancing will be performed if a=2.

$2^i = 2^0, 2^1, 2^2, 2^3, 2^4$ ……
$=1, 2, 4, 8, 16$ ……

In the proposed algorithm used two parameters '*ee*' and 'ratio' which is defined prior in the algorithm along with the value of gaps. To understand this concept, consider an example; say the list to be sorted is 1, 2, 3, and 4. The average and standard deviation are calculated first. The mean and standard deviation are calculated to 2.5 and 1.2 respectively. The ratio and '*ee*' is calculated using equation (1) and (2).

Ratio = 3.

ee = 8/3=2 as integer

The total gaps are 1+2 = 3

2.1 Re-balance for 2^0 =1

1	-1	-1	-1

2.2 Re-balance for 2^1=2

After re-balancing the array is described below.

1	-1	-1	-1	2	-1	-1

In this case initially *j*=1 that means have *e*+*ee* gaps that is equal to 3. At second iteration '*j*' is equal to 2, now have the condition that is *j*=ratio so we decrement the value of '*ee*' by 1. Initially we have 3 gaps, then 2 gaps.

After re-balancing, the array is:

2.3 Re-balance for 2^2=4

1	2	3	4

After re-balancing this array is as follows:

1	-1	-1	-1	2	-1	-1	3	-1	-1	4	-1

Initially have 3 gaps for *j*=1.

- For *j*=2, *j*%ratio is equal to zero, therefore '*ee*' is decremented by 1.
- For *j*=3, the value remains unchanged to 2 gaps
- For *j*=4, again the value is decremented by 1 so there is only single gap.

(B) Example for how re-balancing will be performed if a=3.

$3^i = 3^0, 3^1, 3^2, 3^3, 3^4$ ……
$= 1, 3, 9, 27$……

2.4 Re-balance for 3^0 =1

After re-balancing, the array is described as:

1	-1	-1	-1

2.5 Re-balance for 3^1=3

1	2	3	4

After re-balancing, the array is as follows:

1	-1	-1	-1	2	-1	-1	3	-1	-1

The spaces are calculated using the equation (1). In the similar manner, re-balancing of the array for the remaining value of the '*a*' is held till the re-balancing is not possible. The reason for re-balancing is not being possible in the array according to requirement spaces is not possible. In this way performance of algorithm degrades as require the larger number of swaps to generate the spaces which is same as that in the case of traditional insertion sort.

Table 2 describes the execution time of the LUGD and LNGD algorithm using different type of data set that are

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

random data, nearly sorted data, reverse sorted data and sorted data. Along with the different dataset value, the table also describes the value of 'ε' and re-balancing factor 'a'. The re-balancing comparison of LUGD and LNGD algorithms is shown in Fig. 5 to 8. From fig. 5 to 8, the *X*-axis represents the value of gap (ε) and re balancing factor (a) and *Y*-axis represents the execution time in microseconds.

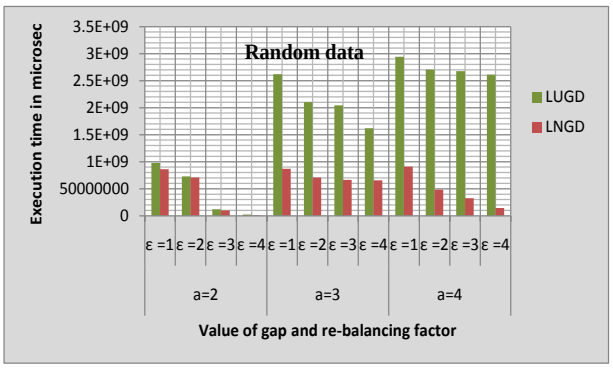


Fig. 5. Rebalancing time comparison of random data

Fig. 5 describes the plot at random data for the different values of 'ε' and re-balancing factor 'a'. It is observed from Fig. 5, as if increase the re-balance factor in the case of LUGD the execution time also increases significantly, but in the case of LNGD the improvement of execution time achieved upto 94% in comparison to LUGD.

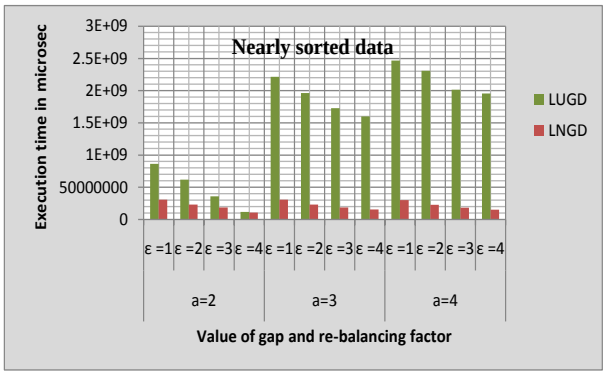


Fig. 6. Rebalancing time comparison of nearly sorted data

Fig. 6 shows the comparison of LUGD with LNGD at different gaps and re-balancing factors. Again the improvement is upto 92% at a=4 and ε=4.

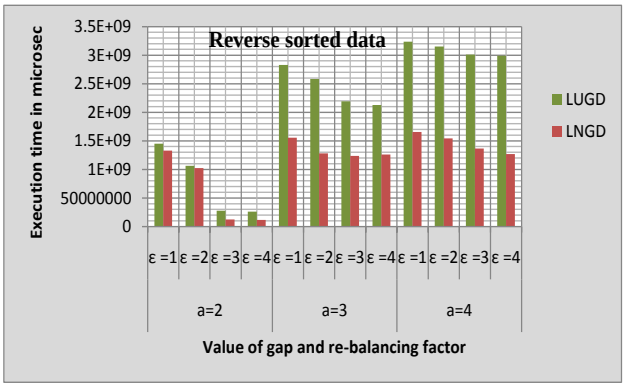


Fig.7. Rebalancing time comparison of reverse sorted data

Fig. 7 shows the comparison of execution time of reverse sorted data at the different values of 'ε' and re-balancing factor 'a'. Initially, in this case results are improved by 8%, but maximum upto 57% at ε=4 and a=4.

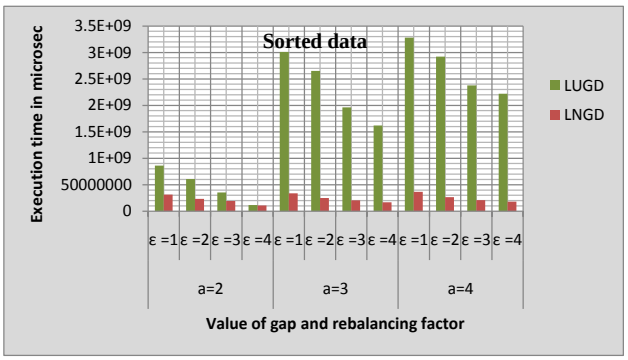


Fig.8. Rebalancing time comparison of sorted data

Fig. 8 represents the result on the sorted data with the different value of gaps and re-balancing factor. The result shows that maximum improvement achieved is upto 91% in comparison to that of LUGD.

TABLE 2. TIME TAKEN BY LUGD AND LNGD ALGORITHM IN MICROSECOND FOR DIFFERENT VALUE OF RE-BALANCING AND GAP

Type of Dataset									
Re-balancing	**Value of ε**	**Random**		**Nearly**		**Reverse**		**Sorted**	
		LUGD	**LNGD**	**LUGD**	**LNGD**	**LUGD**	**LNGD**	**LUGD**	**LNGD**
2	ε =1	981267433	862909204	864558882	306063385	1450636163	1328993502	861929937	313078205
	ε =2	729981576	708580455	620115904	230939335	1065247938	1022310950	609647355	234697961
	ε =3	119727535	106921406	358670053	185759986	278810310	125152235	356489846	195120953
	ε =4	23003046	14557332	117188830	107729204	263693774	116417058	116590140	106897060
3	ε =1	2622591059	869209660	2214715182	308010395	2832112301	1556949795	3011802732	339671533

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

	ε =2	2103580421	709709631	1964645906	231895871	2585747568	1280383725	2651992181	249815851
	ε =3	2043974421	666999939	1728175857	185620741	2195021514	1239442185	1962122927	206018101
	ε =4	1620914312	657130080	1600879365	155075390	2130261056	1263332585	1620374625	170410859
4	ε =1	2942693856	912631839	2467933298	300698051	3239333534	1656255915	3281368964	367329723
	ε =2	2705332601	484314092	2510103530	227562280	3154811065	1545638611	2923182920	266428823
	ε =3	2676681610	327850683	2613423098	183893005	3013676930	1366501241	2378347887	210265375
	ε =4	2611656774	146342570	2157740458	153786055	2993363707	1270043884	2222906193	181557846

VII. Conclusion and Future Work

The proposed approach of LNGD proved to be a better algorithm in comparison to that of LUGD. Improvement has been achieved that ranges from 8% to 90%. The improvement of 90% has been found in the cases where the LUGD was performing poorer. The performance of LNGD is better for different values of re balancing factor which was not achieved in the case LUGD. The LNGD and LUGD both algorithms are implemented in C language. The program of both algorithms is designed at Borland C++ 5.02 compiler and executed on the Intel core I5 processor-3230 M CPU @ 2.60 GHz machine, and the programs runs at 2.2 GHz clock speed.

In the future, investigation can be done on the locality of data in more details. This will help not only in allocating the spaces accurately, but may also reduce the extra spaces which have been allocated and will act as an overhead both on the space and execution time of the program.

References

[1] Faujdar Neetu, Satya Prakash Ghrera, "A Detailed Experimental Analysis of Library Sort Algorithm," *12th IEEE India International Conference (INDICON)*, pp. 1-6, 2015.

[2] Bender, Michael A., Martin Farach-Colton, "Insertion sort is *O*(*n*log*n*)," *Theory of Computing Systems*, 39 (3), pp. 391-397, 2006.

[3] Faujdar Neetu, Satya Prakash Ghrera, "Performance Evaluation of Merge and Quick Sort using GPU Computing with CUDA," *International Journal of Applied Engineering Research*, 10(18), pp. 39315-39319, 2015.

[4] Janko, Wolfgang, "A list insertion sort for keys with arbitrary key distribution," *ACM Transactions on Mathematical Software (TOMS)*, 2(2), pp. 143-153, 1976.

[5] Estivill-Castro, Vladmir, and Derick Wood, "A survey of adaptive sorting algorithms," *ACM Computing Surveys (CSUR)*, 24(4), pp. 441-476, 1992.

[6] Pardo, Luis Trab, "Stable sorting and merging with optimal space and time bounds," *SIAM Journal on Computing*, 6(2), pp. 351-372, 1977.

[7] Thomas, Nathan, "A framework for adaptive algorithm selection in STAPL," *Proceedings of the tenth ACM SIGPLAN symposium on Principles and practice of parallel programming ACM*, pp. 277-288, 2005.

[8] Itai, Alon, Alan G. Konheim, and Michael Rodeh, "A sparse table implementation of priority queues," *Springer Berlin Heidelberg*, pp. 417-431, 1981.

[9] Bender, Michael A et al, "A locality-preserving cache-oblivious dynamic dictionary," *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*, pp.1-22, 2002.

[10] Brodal, Gerth Stølting, et al, "Cache oblivious search trees via binary trees of small height," *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1-20, 2002.

[11] Frequent Itemset Mining Implementations Repository,http://fimi.cs.helsinki.fi accessed on 10/11/2014.

[12] Zubair Khan, Neetu Faujdar, "Modified BitApriori Algorithm: An Intelligent Approach for Mining Frequent Item-Set," *Proc. Of Int. Conf. on Advance in Signal Processing and Communication*, pp. 813-819, 2013.

[13] Faujdar Neetu, and Satya Prakash Ghrera, "Analysis and Testing of Sorting Algorithms on a Standard Dataset, "*Fifth International Conference on Communication Systems and Network Technologies (CSNT)*, pp.962-967, 2015.

[14] Sodhi, Tarundeep Singh, Surmeet Kaur, and Snehdeep Kaur, "Enhanced insertion sort algorithm," *International journal of Computer applications*, 64(21), pp. 35-39, 2013.

[15] Dutta, Partha Sarathi, "An Approach to Improve the Performance of Insertion Sort Algorithm," *International Journal of Computer Science & Engineering Technology (IJCSET)*, 4(5), pp. 503-505, 2013.

[16] Lam, Franky, and Raymond K. Wong, "Rotated library sort," *Proceedings of the Nineteenth Computing: The Australasian Theory Symposium Australian Computer Society, Inc*, 14(1), pp. 21-26, 2013.

[17] Shen, Xipeng, and Chen Ding, "Adaptive data partition for sorting using probability distribution," *International Conference on Parallel Processing, ICPP*, 2004.

[18] A Kaushik, SS Reddy, L Umesh, BKY Devi, N Santana, N Rakesh, "Oral and salivary changes among renal patients undergoing hemodialysis: A cross-sectional study," Indian journal of nephrology, volume 23, pp. 2013/3.

[19] Sandeep Pratap Singh, Shiv Shankar P Shukla, Nitin Rakesh, Vipin Tyagi," Problem reduction in online payment system using hybrid model," arXiv preprint arXiv:1109.0689, pp. 2011/9/4.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A Comparative Study of Educational Data Mining Techniques for skill-based Predicting Student Performance

Mustafa M. Zain Eddin[1], Nabila A. Khodeir[1], Heba A. Elnemr[2]
[1] Dep. Of Informatics research, Electronics Research Institute, Giza, Egypt
[2] Computers and Systems Dep., Electronics Research Institute, Giza, Egypt
mustafa@eri.sci.eg
nkhodeir@eri.sci.eg
heba@eri.sci.eg

Abstract— **Prediction of student performance has become an essential issue for improving the educational system. However, this has turned to be a challenging task due to the huge quantity of data in the educational environment. Educational data mining is an emerging field that aims to develop techniques to manipulate and explore the sizable educational data. Classification is one of the primary approaches of the educational data mining methods that is the most widely used for predicting student performance and characteristics. In this work, three linear classification techniques; logistic regression, support vector machines (SVM), and stochastic gradient descent (SGD), and three nonlinear classification methods; decision tree, random forest and adaptive boosting (AdaBoost) are explored and evaluated on a dataset of ASSISTment system. A k-fold cross validation method is used to evaluate the implemented techniques. The results demonstrate that decision tree algorithm outperforms the other techniques, with an average accuracy of 0.7254, an average sensitivity of 0.8036 and an average specificity of 0.901. Furthermore, the importance of the utilized features is obtained and the system performance is computed using the most significant features. The results reveal that the best performance is reached using the first 80 important features with accuracy, sensitivity and specificity of 0.7252, 0.8042 and 0.9016, respectively.**

I. INTRODUCTION

Intelligent Tutoring Systems (ITSs) are computer-assisted tutoring systems that permit personalization of the system interactions. ITSs consider a student model as an input to different tutoring systems components to adapt their behaviours accordingly. Student model is a representation of the student features [1] such as knowledge [2, 3], performance, interests [4], goals [5, 6] and individual traits [7].

Diverse techniques are utilized in student modelling process. They are generally classified into two categories: cognitive science and data mining approaches [8]. Cognitive Science approaches modelled how humans learn based on domain modelling and expert systems [9]. Model-tracing (MT) and constraint-based modelling (CBM) are the two common techniques in cognitive science approach. Student modelling in MT is represented in terms of rules that represent the procedural knowledge of the domain. The model reasons for student knowledge through tracing student execution of the defined domain rules [10]. In CBM, student modelling is based on recognizing the student errors in terms of violations of defined domain constraints. Domain constraints are used to represent not only the procedural knowledge as MT but also the declarative knowledge [11].

On the other hand, educational data mining student modelling approaches are based on the generated data through interactions between the system and a large number of students. For example, the student answers to the presented questions and different assistance types that are requested by the student. Reasonable data mining techniques are used as a tool to understand and model the student properties of interest using such data. The focus of student models that are based on cognitive science is modelling of the student knowledge. On the other hand, data mining techniques model different features such as student performance, student affect, and student learning [12, 13].

Generally, the student model is used to predict the student response to tutorial actions and to assess the student or the ITS. Data mining-based student models can achieve such tasks [14].

Different data mining techniques are designed to achieve predictive tasks. Based on the values of different independent variables, the value of a target dependent variable is predicted [12, 14]. In ITSs, the most popular predicted variables are the correctness of a student's response to a question.

The focus point of this paper is to discuss the implementation of educational data mining techniques for predicting student performance. In this study, we appraise and compare some of the most widespread classification algorithms that tackle this problem. Six data mining techniques: logistic regression, support vector machine (SVM), stochastic gradient descent (SGD), decision tree, random forest and adaptive boosting (AdaBoost), are investigated and assessed. Furthermore, the feature importance method is accomplished to reduce the feature space as well as selecting the significant features to enhance the system performance.

The paper is organized as follows: A short overview of educational data mining techniques is presented in section II,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

while section III reviews some related work on using classification techniques in students' performance prediction. The procedure for the comparative study is exhibited in section IV. The conclusion and future work are provided in section V.

II. EDUCATIONAL DATA MINING TECHNIQUES

Educational data mining is a procedure applied to distill valuable information and patterns from a massive educational database [15]. The valuable information and patterns may be exploited for predicting students' performance. Accordingly, this can help the educationalists in affording an effectual teaching methodology.

Typically, in educational data mining technique, predictive modelling is applied to portend the student performance. Several tasks are commonly used to build the predictive model, including classification, categorization and regression.

Yet, classification is considered the most widespread method to predict students' performance. Several classification techniques can be applied to portend the students' performance. Among these algorithms are the logistic regression, SGD, SVM, decision tree, random forest and AdaBoost classifier.

Classification aims to take an input vector x and assign it to one of K discrete classes C_k where $k = 1, 2, \dots, K$. Typically, the classes are separated and therefore each input is assigned to one and only one class. In other words, the input space is thereby divided into decision regions whose boundaries are called decision surfaces. In linear models for classification, the decision surfaces are linear functions of the input vector.

Linear prediction methods, such as logistic regression and SVM, have been widely used in statistics and machine learning. In this paper, we will focus on logistic regression [16, 17, 18], SVM [19] and SGD [20].

On the other hand, the decision tree has the ability to capture the non-linearity in the data by dividing the space into smaller sub-spaces based on the taken decision. In this work, the decision tree technique is explored for predicting the student performance [21, 22]. Ensembles of decision trees, random forest and AdaBoost, which combine the predictions of several base estimators, are also considered. [23, 24]. A brief description of the utilized data mining methods is presented in the next section

A. Logistic Regression

Regression aims to estimate the conditional expectations of continuous variables using input-output relationships. Logistic regression is used for categorical or binary outcomes. Therefore logistic regression is a linear model for classification rather than regression. It is extensively utilized owing mainly to its simplicity. Logistic regression investigates the relation between a categorical dependent variable and several independent variables, and predicts the probability of incidence of an event by fitting the data into a logistic curve. Generally, logistic regression is categorized into two models: binary logistic regression and multinomial logistic regression. The binary logistic regression is normally used if the dependent variable is dichotomous, whether the independent variables are continuous or categorical. Whereas, the multinomial logistic regression are used if the dependent variable is not dichotomous and contains more than two classes. [16, 17, 18].

B. Support Vector Machine

SVM is a supervised learning method that analyses the training data and builds a model, which in turn assigns the unlabelled data to the appropriate class. For a binary class learning task, the SVM aims to obtain the superior classification function to discriminate between the two classes elements in the training data. In this work, we have adopted linear SVM, where a linear classification function is used to create a separating hyperplane that goes along the middle path separating the two classes. Outwardly, there exist infinitely several hyperplanes that may detach the training data. The SVM technique attempts to find the hyperplanes that maximize the gap between both classes so as to achieve the testing classification without perpetrating misclassification errors. Maximizing the margin to get the largest possible distance between the separating hyperplane and the instances on either side of it is proven to reduce an upper bound on the expected generalization error [19].

C. Stochastic Gradient Descent Machine

SGD is a linear classifier beneath convex loss functions like SVM and Logistic Regression. It is considered an effective approach to discriminative learning. Despite the availability of this approach in the machine learning society since long time, it has lately got more attention owing to its potency with large-scale and sparse data mining problems. It has been successfully exploited in issues related to natural language processing and text classification.

Generally, gradient descent is deemed the best method used insomuch that the parameters cannot be determined analytically and need to be searched for through an optimization algorithm [20]. Therefore, SGD is considered an optimization algorithm used to find the values of parameters of a function (f) that minimizes a cost function (cost). The aim of the algorithm is to get model parameters that reduce the error of the model on the training dataset. That is achieved by changing the model that moves it through a gradient or slope of errors down on the way to a minimum error value. This provides the algorithm with its name of “gradient descent".

D. Decision Trees

Decision tree is a supervised learning scheme based on a non-parametric model and utilized for classification and regression. Decision tree constructs a model to portend the target variable value through learning plain decision rules inferred from the data features [21].

Decision tree algorithms start with a set of cases and create a tree data structure that can be utilized to classify new cases. Each case is defined by a set of features which can have numeric or symbolic values. A label representing the name of a class is associated with each training case [22]. Decision

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Tree is characterized by simplicity to understand and to interpret. Decision Tree is able to handle both numerical and categorical data. Other techniques are usually specialised in analysing datasets that have only one type of variable. On the other hand, decision Tree model can create over-complex trees that do not generalise the data well (overfitting). Setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree is necessary to avoid this problem.

Small deviations in the data may generate a completely distinct decision tree, which causes instability to the tree. This problem is alleviated by adopting an ensemble of the decision tree.

E. Ensemble Methods

To enhance generalizability robustness over a single classifier, ensemble methods are used which combine the predictions of several base estimators built with a learning algorithm.

Principally, the ensemble methods are categorized into the averaging methods and the boosting methods. In the averaging methods, several estimators are created independently and then their average prediction is computed. Random forests classifier is an example of the average methods. Conversely, the base estimators in the boosting methods are constructed sequentially and one attempts to decrease the bias of the joined estimator. AdaBoost is an example of boosting methods.

The ensemble methods target to combine several feeble models to generate a powerful ensemble. Generally, the joined estimator is usually superior to any of a single base estimator as its variance is reduced [23, 24].

1) Random forests

Random Forest algorithm is an example of averaging ensemble methods. Random Forests is more robust than decision trees and able to model large feature spaces. Random Forests is a bagged classifier linking a collection of decision tree classifiers which constitute a forest of trees [23]. The varied set of classifiers is created by introducing randomness in the classifier construction. The prediction of the ensemble is given as the averaged prediction of the discrete classifiers. In random forests, each tree in the ensemble is grown on a different bootstrap sample that containing randomly drawn instances with replacement from the original training sample. In addition, random forest uses random feature selection where at each node of the decision tree t, m features are nominated at random out of the M features and the best split selected out of this m. When splitting a node during the building of the tree, the split that is selected is no longer the best split among all features. Alternatively, the selected split is the best split between a random subset of features. Accordingly, the forest bias typically increases to some extent regarding the bias of an individual non-random tree. However, averaging is usually more than compensating for the increase in bias that gives an overall better model

2) Adaptive Boosting (AdaBoost)

Boosting is a general ensemble method that produces a strong classifier from a number of weak classifiers. It is based on building a model from the training data, then creating a second model that tries to correct the errors from the first model. Models are added up to the training set is predicted perfectly or a maximum number of models are added. AdaBoost was the first really successful boosting algorithm established for binary classification. It can be used in combination with many other types of learning algorithms to increase performance. The output of the other learning algorithms is combined into a weighted sum that denotes the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favour of those instances misclassified by preceding classifiers [24].

For each consecutive iteration, the weights of sample data are singly modified and then, the learning process is reapplied to the reweighted sample. The incorrectly predicted training examples induced by the boosted model at the previous step possess increased weights, while the correctly predicted ones hold decreased weights. Usually, this permits decreasing the variance within the model.

III. RELATED WORK

Predicting student performance in solving problems is the focus of a number of literature. Different educational data mining techniques, such as decision trees [25], artificial neural networks [26], matrix factorization [27], collaborative filters [28] and probabilistic graphical models [29], have been applied to develop prediction algorithms. These classifiers can be used to identify weak students and thus assist the students develop their learning activities.

Pereira et al, utilizes decision tree classifiers to predict the student marks based on previous semester marks and internal grades [25]. The accuracy of the classifiers was computed and it was shown that the decision tree classifier CHAID has the highest accuracy followed by C4.5.

Thai et al, compared the accuracy of decision tree and Bayesian Network algorithms for predicting the academic performance of undergraduate and postgraduate students at two very different academic based on their grades and demographic information. They concluded that the decision tree was more accurate than the Bayesian Network [30].

Xu et al. developed an ongoing prediction of the students' performance using ensemble learning technique. Exponentially Weighted Average Forecaster (EWAF) is used as a building block algorithm to enable progressive prediction of students' performance [31].

Feng and Heffernan utilized skill model for predicting student performance in a large scale test Massachusetts Comprehensive Assessment System (MCAS). Skill model is a matrix that relates questions to the needed skills to solve the problem. Based on the assessed skills of the students, the model of the performance of the predicted MCAS test score is measured. They used Mixed-effects logistic regression model

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

to predict the student response based on student response data through time, and skills parameter. Two different grain-sized skill models were tried. Skill model that has large grain-sized of skills gave more accurate prediction [32].

Ostrow et al. implemented partial credit models within ASSISTments system to predict student performance. Partial credit scoring is defined based on penalties for hints and attempts. The maximum likelihood probabilities for the next problem correctness within each test fold are used as predicted values. Implementing partial credit scoring improves prediction of student performance within adaptive tutoring systems [33].

Wang et al. introduced the Opportunity Count Model (OCM) and investigated the significance of considering OC in student models [34]. The OCM built separate models for differing OCs by using random forest to determine fluctuations in the importance of student performance details across a dataset stratified by OC.

IV. Dataset and Method

The ASSISTment system is based on Intelligent Tutoring System technology and is delivered through the web. The main feature of ASSISTments is that they offer instructional assistance in the process of assessing students. ASSISTments utilizes amount and type of the assistance that students receive as a way to assess the student knowledge. In addition, the questions and related needed skills to be solved are defined in [33]. As shown in table I, The dataset contained performance details for 96,331 transaction log logged by 2,889 unique students for 978 problems spanning 106 unique Skill Builders and 269 of these problems are multi-skilled problems. Multi-skilled problems are problems that need more than one skill to be solved. Such problems are called original problems.

TABLE I
Used Dataset statistics

Students	Questions	Original Questions	Logs
2889	978	269	96331

ASSISTment system provides a number of original questions and associated scaffolding questions. The original questions typically have the same text as in MCAS test whereas the scaffolding questions were created by content experts to train students who fail to answer the original question. The process starts when the student submits an incorrect answer, then the student is not allowed to try the original question further, but instead must then answer a sequence of scaffolding questions that are presented one at a time. Students work through the scaffolding questions, probably with hints, until he finally gets the problem correct.

Scaffolding questions allow tracking the learning of individual skills where each question is related and tagged by the needed skills to be solved. That is used to express the skill model. The dataset of our experiment has 103 skills that are defined for all used questions.

The next sections indicate the reason of for selecting dataset features. The conducting of the comparative study of applying various classification techniques on the addressed dataset is considered also for predicting the student performance.

A. Dataset Features

Our model works in predicting either the student will answer a question or not based on a set of features. These features are the student-id who is accessing the system, the skills of the question if the question is a multi-skilled question or single-skill and either the student tried to solve this question before or not (is_done).

The system based on this information can learn and predict the student's behavior against different questions even a newly added question, as the question itself is an independent factor in our learning progress, but we care about the skills. So whatever the question is even the questions not included in the learning process or added later, we can predict student's behavior only based on the skills of a question.

B. Comparative Analysis

This work presents a comprehensive comparative study of the applicability of the data mining techniques for predicting student performance. Several data mining techniques, are investigated and evaluated, logistic regression, SVM and SGD as linear prediction methods as well as decision tree, random forest and AdaBoost classifier as non-linear prediction methods

C. Evaluation methods

The evaluation stage is an essential step for selecting the appropriate classifier for a given data. In this work, we adopted several methods to achieve the evaluation task; mean squared error (MSE), accuracy, sensitivity and specificity. MSE measures the average of the squares of the errors or deviations. The model is initially fit on a training dataset, which is a set of examples used to fit the classifier parameters. The resulting model is run with the training dataset and produces a result, which is then compared with the target, for each input vector in the training dataset, to measure the MSE for the training data. Then, the fitted model is used to predict the responses for the observations in another part of the dataset that called the validation dataset. The obtained results are compared with each input vector in the validating dataset to compute the Testing MSE. The MSE is obtained as follows

$$MSE(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i, \hat{y}_i)^2 \qquad (1)$$

where $n_{samples}$ is the number of predictions, $\hat{y}$ is the vector of observed values, and yˆ is the vector of predicted values.

Estimation of the accuracy, sensitivity and specificity is based on four terms, namely:

True positives (TP): positive instances that are correctly predicted.

True negatives (TN): negative instances that are correctly predicted.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

False positives (FP): positive instances that are incorrectly predicted.
False negatives (FN): negative instances that are incorrectly predicted.

The accuracy is an empirical rate of correct prediction, the sensitivity is the ability to correctly classify the output to a particular class, and the specificity is the ability to predict that the outputs of other classes are not part of a stated class. These performance measures are computed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (3)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

A k-fold cross-validation is performed to assess the overall performance of the implemented classifier techniques. The dataset is split randomly into k consecutive subsets called folds of approximately the same size. The model is trained and tested k times. In this paper, 10-fold cross-validation is applied.

D. *Result and Discussion*

TABLE II
THE AVERAGE TRAINING AND TESTING MSE FOR DIFFERENT CLASSIFIERS

Classifier	Training MSE	Testing MSE
Linear SVM	0.5069	0.5079
SGD Classifier	0.4922	0.4929
Logistic Regression	0.3085	0.3089
Decision Tree	0.1158	0.2746
Random Forest Classifier	0.1263	0.2717
AdaBoost Classifier	0.281	0.2817

In the light of the training and testing MSE illustrated in figure 1 and table II, we can see that the linear SVM has the worst training and testing MSE (0.507 & 0.508, respectively), which means that the linear SVM has high variance and high bias. Decision tree classifier, on the other hand, achieves the least MSE for the training phase (0.116), while random forest classifier attains the least MSE for the testing phase (0.272) with an insignificant difference of the decision tree (0.275).

Fig. 1. The average Training and Testing MSE for the different classifiers

Furthermore, the accuracy, sensitivity and specificity for the different classifiers are recorded in Table II and displayed in figure 2, figure 3 and figure 4, respectively. It can be observed that the worst accuracy value (0.49) is made using the linear SVM classifier. Additionally the poorest specificity (~0.0) is produced when applying the linear SVM and SGD techniques. Conversely, both have the highest sensitivity, ~1.0. While the minimum sensitivity (0.66) is reached using the logistic regression method. Regarding the accuracy, the decision tree and random forest classifiers realize the best performance (~0.73) with a trivial difference. On the other hand, the specificity accomplishes its maximum value (0.9) using the decision tree.

TABLE III
AVERAGE ACCURACY, SENSITIVITY AND SPECIFICITY FOR THE DIFFERENT CLASSIFIERS

Classifier	Accuracy	Sensitivity	Specificity
Linear SVM	0.4921	0.9975	0.0042
SGD classifier	0.5071	1	0
Logistic Regression	0.6911	0.6641	0.7485
Decision tree	0.7254	0.8036	0.901
Random Forest classifier	0.7283	0.8378	0.8562
AdaBoost classifier	0.7183	0.671	0.764

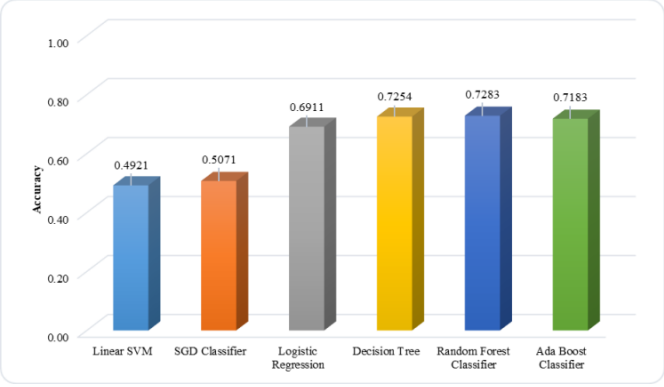


Fig. 2. Accuracy of the different classifiers

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

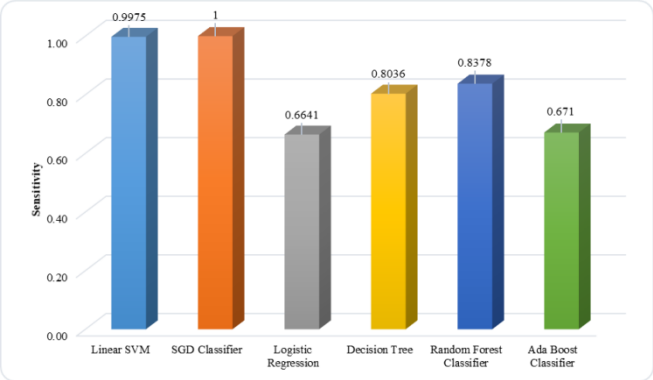


Fig. 3. Sensitivity of the different classifiers

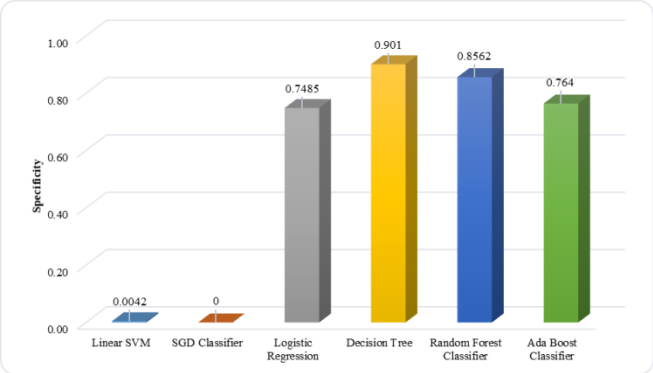


Fig. 4. Specificity of the different classifiers

The perfect sensitivity and deficient specificity of the linear SVM and SGD, signifies that both classifiers are good for capturing actual true cases, but they fail to catch false positives. Furthermore, the results indicate that the decision tree and random forest classifiers have roughly the best training MSE, testing MSE, accuracy and specificity as well as a high sensitivity. Nevertheless, their sensitivity is less than that of linear SVM and SGD classifiers. However, the SVM and SGD Classifiers have very small accuracy and poor specificity, as well as very high training and testing MSE. Additionally, despite that the random forest classifier has a comparable performance to decision tree approach, it disburses a long time for training large datasets. Therefore, it may be emphasized that the decision tree classifier is the best techniques for predicting new data that hadn't been fitted before.

E. *Feature Selection*

Feature selection is the process of selecting the most significant features that can be used for model construction. Furthermore, feature selection techniques delivers a way for reducing the computation time as well as improving the prediction performance. In order to reduce the feature domain, the feature importance for the triumphed classifier (decision tree) is detected using the Gini coefficient.

The Gini importance reveals the frequency of selecting a specific feature for a split and the potential of this feature for discrimination during the classification problem. As stated this criterion, the features are ranked and nominated preceding to the classification procedure. Every time a split of a node is made on variable, the Gini impurity criterion for the two descendant nodes is less than the parent node. Adding up the Gini decreases for each individual variable over all trees in the forest, gives a fast variable importance that is often very consistent with the permutation importance measure. The Gini importance principle demonstrated robustness versus noise and efficiency in selecting beneficial features.

Each feature has an importance in detecting the student's behavior against questions, the table below shows the most important features in predicting student's behavior. The higher ratio indicates the more important feature.

As shown in Fig. 5, the most important features for our model is the student-id, then whether the student fulfilled this question or not, without caring about his previous answer was wrong or not, thereafter, if this question is multi_skilled or not.

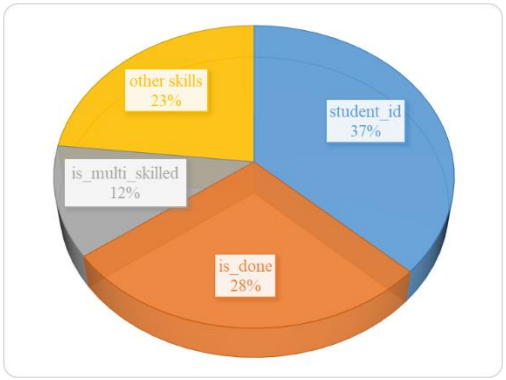


Fig. 5. Different features important for decision trees classification

The impact of the feature selection on the student performance prediction is investigated. The features are first rated descending according to their importance. Then, the effect of selecting different number of the first ranked features on predicting the student performance is recorded. Fig. 6 portrays the effect of the number of features on the system performance. The results reveal that the best performance is achieved using the initial 80 important features. Despite the slight improvement in the prediction accuracy (0.7252), sensitivity (0.8042) and specificity (0.9016), the feature space is reduced significantly.

Fig. 6. The system performance using different number of features

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

V. Conclusion

Students' performance prediction is typically valuable to assist the instructors and learners refining their teaching and learning process. This paper discusses the performance of different data mining techniques in predicting the student performance. Logistic regression, SVM, SGD, decision-Trees, and AdaBoost classifier are investigated and their classification performance is compared. The obtained results unveiled that data mining tools may broadly be employed by education institutions to predict the students' performance. The uppermost classification accuracy (0.7254), sensitivity (0.8036) and specificity (0.901) are produced using decision tree scheme. Feature selection technique based on Gini coefficient is performed to reduce the feature space and improve the system performance. The superior performance is attained using the first 80 important features.

For future work more extended datasets may be utilized. Furthermore, hybrid classification techniques can be applied for predicting the students' performance.

VI. Acknowledgment

We used the 'Assistments Math 2006-2007 (5046 Students)' dataset accessed via DataShop [35].

References

[1] J. Beck, M. Stern, and E. Haugsjaa, *Applications of AI in Education*, ACM Crossroads, 1996

[2] P. Brusilovsky, E. Schwarz, and G. Weber: ELM-ART: An intelligent tutoring system on World Wide Webí. Third International Conference on Intelligent Tutoring Systems, ITS-96, 1996, Montreal.

[3] Z. Pardos, N. Heffernan, Modeling individualization in a Bayesian networks implementation of Knowledge Tracing. In De Bra, P., Kobsa, A., Chin, D. (eds.) UMAP2010 Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization, LNCS vol. 6075, pp. 255-266,. 2010, Springer, Heidelberg

[4] M. Feng, and N. Heffernan, Assessing Students' Performance: Item Difficulty Parameter vs. Skill Learning Tracking. Paper presented at the National Council on Educational Measurement 2007 Annual Conference, Chicago.

[5] R. Roll, V, Baker, and K. Koedinger, What goals do students have when choosing the actions they perform? In Proceedings of the 6th International Conference on Cognitive Modelling, (pp. 380-381), 2004,. Pittsburgh, PA.

[6] M. Mhlenbrock, and O. Scheue, Using action analysis in active math to estimate student motivation. In Proceedings of the 13th Annual Workshop of the SIG Adaptivity and User Modelling in Interactive Systems (ABIS 2005), pp. 56-60, 2005, Saarbrcken Germany.

[7] C. Carmona, G. Castillo, and E. Millan, Designing a dynamic network for modelling students' learning style. In Proceedings of the 8th IEEE International Conference on Advanced Learning Technologies, pp. 346-350, 2008. Santander, Cantabria.

[8] Y. Gong,. Student modeling in Intelligent Tutoring Systems. Worcester Polytechnic Institute, 2014

[9] N. Khodeir, H. Elazhary and N. Wanas :Rule-based Cognitive Modeling and Model Tracing for Symbolization in a Math Story Problem Tutor. International Journal of Emerging Technologies in Learning 12 (4). 2017.

[10] J. Anderson, A Corbett, Cognitive tutors: Lessons learned., Journal of the Learning Sciences 4(2): 167-207, 1995.

[11] S. Ohlsson, Learning from performance errors." Psychological Review 103: 241-262. 1996

[12] W. Hämäläinen, M. Vinni,. Comparison of machine learning methods for intelligent tutoring systems. In international conference in intelligent tutoring systems, Taiwan, 525-534. 2006.

[13] C. Romero , S. Ventura, "Educational data mining: A survey from 1995 to 2005", Expert Systems with Applications 33 (2007) 135–146

[14] Tan, P., M. Steinbach, et al., Eds. Introduction to Data Mining., Addison-Wesley, 2005.

[15] A. Shahiria, W.Husaina, and N. Rashida, A Review on Predicting Student's Performance using Data Mining Techniques", Procedia Computer Science 72 414 – 422, 2015 .

[16] Nasrabadi, Nasser M, Pattern Recognition and Machine Learning, Journal of electronic imaging, 16(4), 2007.

[17] M. Schmidt, N. Le Roux, and F. Bach, Minimizing Finite Sums with the Stochastic Average Gradient. 162(1,2), p 83—112, 2017

[18] A. Defazio, F. Bach, S. Lacoste-Julien: SAGA, A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives, Advances in neural information processing systems, p1646—1654, 2014

[19] C. Burges, A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2: 1:47, 1998.

[20] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms" - In Proceedings of ICML '04. 2004.

[21] T. Hastie, R. Tibshirani and J. Friedman. Elements of Statistical Learning, Springer, 2009.

[22] Quinlan, J. Ross, C4. 5: programs for machine learning, Morgan Kaufmann, 1993.

[23] A. Prinzie and P. Van den, Random multiclass classification: Generalizing random forests to random mnl and random nb, International Conference on Database and Expert Systems Applications, p. 349--358, 2007.

[24] J. Zhu, H. Zou, S. Rosset, T. Hastie, Multi-class AdaBoost , 2009.

[25] BA Pereira, A. Pai, C Fernandes. A Comparative Analysis of Decision Tree Algorithms for Predicting Student's Performance - International Journal of Engineering Science and Computing, April, 2017

[26] W., Y.-h., and Liao, H.-C. Data mining for adaptive learning in a tesl-based e-learning system. Expert Systems with Applications 38(6), p 6480–6485, . 2011.

[27] N. Thai-Nghe, L.; Drumond, L. and T. Horvath, T.;

[28] L. Schmidt-Thieme, Multi-relational factorization models for predicting student performance. In Proc. of the KDD Workshop on Knowledge Discovery in Educational Data. Citeseer. 2011.

[29] A., Toscher, and M. Jahrer, Collaborative filtering applied to educational data mining. KDD cup. 2010

[30] Z. Pardos, and N. Heffernan, Using hmms and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset. Journal of Machine Learning Research W & CP. 2010.

[31] N. Thai, Janecek, P. Haddawy, A comparative analysis of techniques for predicting academic performance. In Proc. of 37th Conf. on ASEE/IEEE Frontiers in Education, 2007

[32] J. Xu, Y. Han, D. Marcu, M. van der Schaar. Progressive Prediction of Student Performance in College Programs. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17) p. 1604—1610, 2017.

[33] M. Feng and N. Heffernan, Assessing Students' Performance Longitudinally: Item Difficulty Parameter vs. Skill Learning Tracking

[34] K. Ostrow, C. Donnelly, and N. Heffernan, "Optimizing partial credit algorithms to predict student performance," Proc. 8th International Conference on Educational Data Mining, pp.404–407, Madrid, Spain, 2015

[35] Y Wang, K Ostrow, S Adjei, N Heffernan, The Opportunity Count Model: A Flexible Approach to Modeling Student Performance. Proceedings of the Third ACM Conference on Learning @ Scale Pages p 113-116, 2016.

[36] K. Koedinger, R. Baker, K, Cunningham, A. Skogsholm, B. Leber, J. Stamper, A Data Repository for the EDM community: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) Handbook of Educational Data Mining. Boca Raton, FL: CRC Press, 2010.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Illumination Invariant Hand Gesture Classification against Complex Background using Combinational Features.

Anjali Patil
Electronics Engineering
DKTE TEI, Ichalkaranji
Maharashtra, India
anjalirpatil@gmail.com

Dr. Mrs. S. Subbaraman
Professor, Electronics Department
Walchand College of Engineering, Sangli.
Maharashtra, India
s.subbaraman@gmail.com

***Abstract*: Hand gesture classification is popularly used in wide applications like Human-Machine Interface, Virtual Reality, Sign Language Recognition, Animations etc. The classification accuracy of static gestures depends on the technique used to extract the features as well as the classifier used in the system. To achieve the invariance to illumination against complex background, experimentation has been carried out to generate a feature vector based on skin color detection by fusing the Fourier descriptors of the image with its geometrical features. Such feature vectors are then used in Neural Network environment implementing Back Propagation algorithm to classify the hand gestures. The set of images for the hand gestures used in the proposed research work are collected from the standard databases viz. Sebastien Marcel Database, Cambridge Hand Gesture Data set and NUS Hand Posture dataset. An average classification accuracy of 95.25% has been observed which is on par with that reported in the literature by the earlier researchers.**

***Index Terms:* Back-propagation, Combinational Features, Fourier Descriptor, Neural Network, Skin color, Static hand gesture**

I. INTRODUCTION

Hand gesture recognition plays an important role in the areas covering the applications from virtual reality to sign language recognition. The images captured for hand gestures fall into two categories viz. glove based images and non-glove based images. Hand gestures recognition also is correspondingly classified as glove based recognition and non-glove based i.e. vision based recognition.

In glove based approach, users have to wear cumbersome wires which may hinder the ease and naturalness with which the user interacts with computers or machines. The awkwardness in using gloves and other devices can be overcome by using vision based systems that means video based interactive systems. This technique uses cameras and computer vision techniques to recognize the gestures in a much simpler way. [1] [2]. Vision based approaches are further classified as 3D model (which is exact representation of shape but is computationally expensive) and appearance based 2D model which is projection of 3-D object onto 2-D plane and is economical computationally. This paper focuses on appearance based methods for recognition of hand postures.

As shown in Figure 1, after capturing the image of hand gesture, segmentation is done based on the skin color. In the skin color detection process the RGB color model is first transformed to appropriate color space and a skin classifier is used to find a pixel is skin pixel or non skin pixel. Skin color is the low level features extraction technique which is robust to scale, geometric transformations, occlusions etc. By the skin classification the region of interest is observed which then is used to find the boundary of the hand. After extracting the hand contour, the Fourier Descriptors (FDs) are extracted and combined with the geometrical features. The feature vectors, thus formed, are given to artificial neural network used as a classifier to classify the hand gestures.

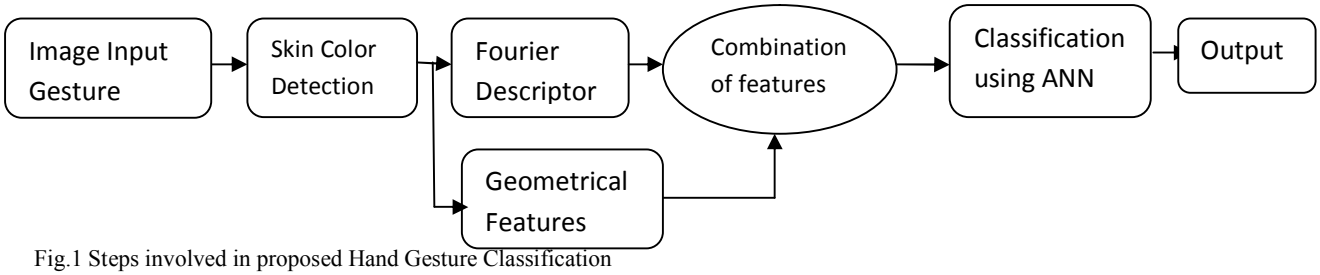


Fig.1 Steps involved in proposed Hand Gesture Classification

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The detailed implementation is explained in the successive sessions. The main objective of this paper is to present the contribution of the work done in direction to classify the hand gestures with the help of skin color correctly from images captured under different illumination conditions. The system which will be robust against variation in illumination and hence can be called as illumination invariant.

The rest of this paper is organized as follows: Section 2 presents the literature review on illumination normalization and skin color detection. Experimental work is discussed in Section 3. Detailed results are presented in Section 4 followed by conclusions and future scope in Section 5.

II. Related work

Detection of skin color in an image is sensitive to several factors such as illumination conditions, camera characteristics, background, shadows, motions besides person dependent characteristics such as gender, ethnicity, makeup etc. A good skin color detector must be robust against illumination variations and must be able to cope up with the great variability of skin color between ethnic groups. Another challenge in detecting human skin color is the fact that the objects in the real world which may be in the background of the image can have skin tone colors, for example, leather, wood, skin-colored clothing, hairs etc. The systems not taking care of this aspect may have false detections. The purpose of the research work is to identify and classify the hand gestures with this type of uncontrolled environment.

Image is represented in different color spaces including RGB, normalized RGB, HSV, YCbCr, YUV, YIQ, etc. Color spaces efficiently separating the chromaticity from the luminance components of color are typically considered preferable (Luma-Chroma model). This is due to the fact that by employing chromaticity-dependent components of color only, some degree of robustness to illumination changes can be achieved. Different Skin color models with comparison of their performance have been presented by Terrillon et.al. in [3].

The detection and segmentation of skin pixels using HSV and YCbCr color space has been explained by Khamar et. al. in [4] wherein an approach to discriminate color and intensity information under uneven illumination conditions is highlighted. The threshold based on histograms of the Hue, Saturation and Value (HSV) has been to classify the pixels into skin or non-skin category. The typical values of threshold applied to the chrominance components followed the limits as 150 <Cr<200 && 100<Cb<150. Chromacity clustering using k means of YCbCr color space to segment the hand against the uneven illumination and complex background has been implemented in [5] by Zhang Qiu et.al . The different experiments performed on the Jochen Triesch Static Hand Posture Database II were reported with comparison in terms of time consumed.

Bahare Jalilian et.al. detected regions of face and hands in complex background and non-uniform illumination in [6]. The steps involved in their approach were skin color detection based on YCbCr color space, application of single Gaussian model followed by Bayes rule and morphological operations. Recognition accuracy for images with complex background reported was 95%.
YCbCr color space was used in [7] by Hsiang et.al. to detect hand contour based on skin color against the complex background. Convex hull was calculated and the angle between finger spacing and the finger tip positions were derived to classify the hand gesture. The accuracy of the recognition rate reported was more than 95.1%.

HSV based skin color detection was implemented by Nasser Dardas et.al in [8], The method has been reported to have real time performance and is robust against rotations, scaling and lighting conditions. Additionally it can tolerate occlusion well. The thresholding proposed was H between 0^o to 20^o and S between 75 and 190. The segmenting resulted in giving the hand contour which was subsequently compared with the templates of the contours of the hand postures. Four gestures were tested by the authors which indicated an average accuracy of more than 90%.

HSV based hand skin color segmentation was used by Zhi-hua et.al in [9]. They presented an efficient and effective method for hand gesture recognition. The hand region is detected using HSV color model whrein they applied the thresholds as 315, 94, and 37 on H, S, V respectively through the background subtraction method. After hand detection, segmentation was carried out to separate out palm and fingers. Fingers and thumb were counted to recognize the gesture. The total classification accuracy of 1300 images tested by them has been reported was 96.69%. However the system failed to work satisfactorily in case of complex background.

Wei Ren Tan et.al [10] proposed a novel human skin detection approach that combined a smoothened 2-D histogram and Gaussian model, for automatic human skin detection in color image(s). In their approach, an eye detector was used to refine the skin model for a specific person. This approach drastically reduced the computational costs as no training was required, and it improved the accuracy of skin detection to 90.39% despite wide variation in ethnicity and illumination.

Log Chromaticity Color Space (LCCS) was proposed in [11] by Bishesh Khanal et.al. which gave illumination invariant representation of image. LCCS resulted into an overall classification rate (CR) of about 85%. A better CR (90.45%) was obtained when LCCS was calculated as against only luminance. In [12] , Yong Luo et.al. removed illumination component by subtracting the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

mean estimation from the original image. To make the standardization of the overall gray values of the different face images, ration matrix and modulus mean was calculated and used as features. The reported recognition rate using PCA was 92% for Yale B+ face database and using LDA 94.28%. Hsu et. al. addressed the issue of illumination changes, by first normalizing the image using the geometric mean followed by a natural log of the normalized image.[13]. The false rejection and false acceptance ratios reported by them were as low as 0.47% and 0% respectively.

Mohmed Alshekhali et.al. in [14] proposed the technique for detection of hand and determination of its center, tracking the hands trajectory and analyzing the variations in the hand locations, and finally recognizing the gesture. Their technique resulted in overcoming the background complexity and gave satisfactory results for the camera located up to 2.5 meters from the object of interest. Experimental results indicate that this technique could recognize 12 gestures with more than 94% recognition accuracy.

Extensive literature review reveals that the Luminance-Chrominance color model can be used to detect the skin color which provides robustness against illumination variation. Chroma (Chrominance) sampling is the key for color based segmentation in real time environment. YCbCr found to be promising for complex background while HSV indicates its robustness against the variation in the intensity of illumination while capturing the images. In order to achieve the benefits of both YCbCr and HSV, an approach based on the combination/fusion of two viz. YUV (variant of YCbCr) and HSV color space is proposed in this paper to detect the skin color. YUV color space which was initially coded for PAL analog video, now is also used in the CCIR 601 standard for digital video. The detailed implementation of this fusion and the results thereof are discussed in section III.

III. Experimental Work

As discussed in section II the first clue to segment the hand from the image is skin color. For this purpose Luminance-Chrominance color model is used. Pure color space (chrominance value) is used to model the skin color; for instance UV space in YUV and SV space in HSV color space. But under varying illumination conditions, the skin color of the hands from different databases, either different persons or even same person, may vary. The sample images of hand gestures captured under varying illumination conditions used in this paper are shown in the Figure 2. These are available online for research purpose and are from Sebestien Marcel database (Figure 2.a) [15], Cambridge Hand Gesture database (Figure 2.b) [16] and NUS Hand Posture database II (Figure 2.c) [17].

To reduce the effects of illumination variation effects, a normalized color space is used. Normalization is achieved by combining YUV and HSV color spaces. For this firstly the RGB image is converted into the YUV and HSV color spaces using (1) to (6). This separates the luminance and chrominance components from the image. Separation of the chrominance approximates the "chromaticity" of skin (or, in essence, its absorption spectrum) rather than its apparent color value thereby increasing the robustness against variation in illumination. In this process, typically the luminance component is eliminated to remove the effect of shadows, variations in illumination etc.

Fig. 2: Images of Hand Gestures with Variation in Illumination a) 'Five' from Sebestien Marcel database [15] b) 'Flat' from Cambridge Hand Gesture database [16] and c) 'B' from NUS Hand Posture database II [17].

YUV is an orthogonal color space in which the color is represented with statistically independent components. The luminance (Y) component is computed as a weighted sum of RGB values while the chrominance (U and V) components are computed by subtracting the luminance component from B and R values respectively. Mathematically this conversion is given by the following equations:

$$Y = R + 2G + \frac{B}{4} \quad (1)$$

$$U = R - G \quad (2)$$

$$V = B - G \quad (3)$$

HSV is invariant to dull surfaces and lighting. HSV approximates the way humans perceive and interpret color. The research shows that the luminance may vary due to ambient lighting conditions and is not reliable measure to detect the skin pixels. Saturation and Value V (brightness) can be used in order to minimize the influence of shadow

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

and uneven lighting. Conversion from RGB to HSV color space is done using following equations:

$$H=\frac{\cos^{-1}\frac{1}{2}[(R-G)+(R-B)]}{\sqrt{((R-G)^2)+(R-B)(G-B)}} \quad (4)$$

$$S=1-\frac{\min(R,G,B)}{R+G+B} \quad (5)$$

$$V=\frac{1}{3}(R+G+B) \quad (6)$$

The same algorithm mentioned in [4] is used for human skin detection from YUV and HSV color spaces. Histogram is used for deciding the threshold level for discriminating the skin and non skin pixels. The output will be image with only skin pixels. The largest blob is detected as hand. Arm removal algorithm is implemented to segment the palm for further processing.

After segmenting the hand using skin color, the boundary of the hand is detected. The object is generally described by its boundary in a meaningful manner. Since each boundary is composition of collection of all connected curves, the concentration is upon the description of connected curves. In hand gesture recognition, the techniques which provides unique features that are used primarily for shape representation as well as its time complexity is less, is chosen so that the recognition of static hand gestures can be done in real time. It is also expected that the technique used should be invariant to translation, rotation, and scaling.

Different methods in the literature include the use of eccentricity, scale space and Fourier descriptors for shape detection. 2-D Fourier transformation is extensively used for shape representation and analysis. The details of the literature review describing the use of Fourier descriptors for 2-D shape detection and hand shape detection and its implementation can be found in [18]. The coefficients calculated by applying Fourier transform on the input image forms the Fourier descriptors of the shape. These descriptors generally represent the shape in a frequency domain. The global features of the shape are given by the low frequency descriptors and finer details of the shape are given by the higher frequency descriptors. The number of coefficients obtained after transformation are generally large, some of them are sufficient to properly define the overall features of the shape. High frequency descriptors that are generally used to provide the finer details of the shapes are not used for discrimination of the shape, so they can be ignored. By doing this, the dimensions of the Fourier descriptors used for capturing shapes are significantly reduced and the size of feature vector is also reduced.

As shape is connected object and is described using a closed contour that can be represented as a collection of the pixel coordinates in x and y direction. The coordinates can be considered to be sampling values. Suppose that the boundary of a particular shape has P pixels numbered from 0 to P - 1. The p^{th} pixel along boundary of the contour has position (x_p, y_p). The contour can be described using two parametric equations:

$$x(p)= x_p$$
$$y(p)= y_p \quad (7)$$

The Cartesian coordinates of the boundary pixel is not considered as Cartesian coordinates instead they are converted to the complex plane by using the following equation:

$$s(p)= x(p)+ iy(p) \quad (8)$$

The above equation means that the x-axis is treated as real axis and y-axis as imaginary axis of a sequence of complex numbers. Although the interpretation of the sequence was recast, the nature of the boundary itself was not changed. Of course this representation has one great advantage: It reduces a 2-D to 1-D problem. The Discrete Fourier Transform of this function is taken and frequency spectra are obtained. Discrete Fourier transform of s(p) is given by

$$a(u)=\frac{1}{P}\sum_{k=0}^{p-1}s(p)e^{-j2\pi up/P} \quad (9)$$

Where u= 0, 1,2, ….P-1.

The complex coefficients a(u) are called the Fourier descriptors of the boundary. The inverse Fourier transform of these coefficients restores s(P) and given by the following equation:

$$S(P)=\frac{1}{P}\sum_{u=0}^{p-1}a(u)e^{-j2\pi p/P} \quad (10)$$

where p=0, 1, 2, ….P-1

To increase the robustness of the system the geometrical features like eccentricity, aspect ratio of the area and perimeter of the closed contour are also calculated from the properties of the region of the hand contour. The feature vector is formed combining the skin color based shape features and geometrical features. The complete algorithm of feature vector formation and classification is represented in the following algorithm

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Algorithm:

a) Read RGB image
b) Convert RGB to HSV and YUV
c) Apply skin detector algorithm based on the threshold on S and U.
d) Perform the morphological operations.
e) Find the largest blob
f) Detect the palm by using arm removal algorithm.
g) Extract the boundary co-ordinates of the contour
h) Apply Fast Fourier Transform and calculate Fourier descriptors
i) Calculate the geometrical properties of the blob
j) Combine the features (Skin color + Fourier + Geometrical Features.) to form the feature vector.
k) Repeat the procedure for all the images in the training and testing database.
l) Train the Backpropogation neural network to classify the gestures.
m) Test the network and find out the accuracy.

IV. Results and Discussion

As mentioned in Section III, the performance of the system is tested using three different datasets with details as given below.

1. Sebestien Marcel dataset consists of total 6 postures viz. A, B, C, Point, Five and V of 10 persons in 3 different backgrounds (light, dark and complex).
2. Cambridge Hand Gesture consists of 900 image sequences of 9 gesture classes. Each class has 100 image sequences performed by 2 subjects, captured under 5 different illuminations and 10 arbitrary motions. The 9 classes are defined by three primitive hand shapes and three primitive motions. For the experimentation we are focusing on the hand shapes in different illumination conditions.
3. NUS Hand Posture database consists of the postures by 40 subjects, with different ethnicities against different complex backgrounds. The database used in this experimentation consists of 4 hand postures repeated five times by each of the subjects. Hand posture images of size 160x120. 100 images are used for training and 100 for testing.

The skin detector is first applied to extract skin regions in the images from the three databases using fusion of HSV and YUV color space and applying the threshold. The results of the skin detector algorithm on the images of three sets are presented in Fig. 4,5 and 6. Hand posture shown in these figures are number 'Five' from Sebastian Marcel dataset II, 'B' from NUS and 'Flat' from Cambridge hand gesture dataset. The purpose of presenting the same hand shape for all the database is to show that the proposed system works better for complex background and illuminations conditions. The fig. 4 shows that the algorithm and works quiet better for Sebastian Marcel dataset. Fig.5 represents empirical results that show the detection of the hand region is not up to the mark for Cambridge hand gesture database with the 5th illumination conditions as can be seen from the Fig. 2b. Fig. 6 interprets the result of the skin detection on the NUS dataset. After detecting the skin, morphological operations were performed to get the closed contour of the hand. As explained in the algorithm in section III, the Fourier descriptors were chosen as features and hence were calculated from the closed contour. The descriptors were then normalized by nullifying the 0^{th} Fourier descriptor to get the invariance to the translation. Scale invariance was obtained by dividing all Fourier descriptors by the magnitude of the 1^{st} Fourier descriptor. Rotation invariance is achieved by only considering the magnitude of the Fourier coefficients.

The feature vector was formed by considering 20 coefficients of Fourier descriptors (which are invariant to scale, rotation and translation) and two geometrical features viz. area to perimeter aspect ratio and eccentricity, thus making a total of 22 features. Geometrical features were calculated from the closed contour of the segmented hand. The feature vectors thus formed were then used to train and test the multilayer feed forward neural network to classify the hand gesture. For learning the network, Back propagation algorithm with Levenberg-Marquardt algorithm has been used to train the network. The activation function used is "Sigmoid". Fig. 3 shows the architecture of the NN used in this experiment.

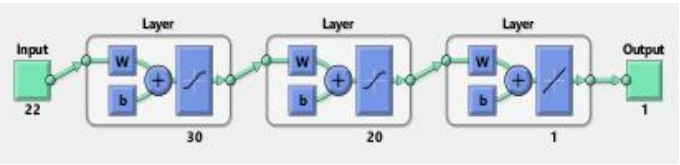


Fig. 3. Neural Network Architecture

The same experiment is tested for NUS Hand Posture dataset I which consists of 10 classes of postures with 24 samples of each. As there is uniform background the classification accuracy observed is 100%.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

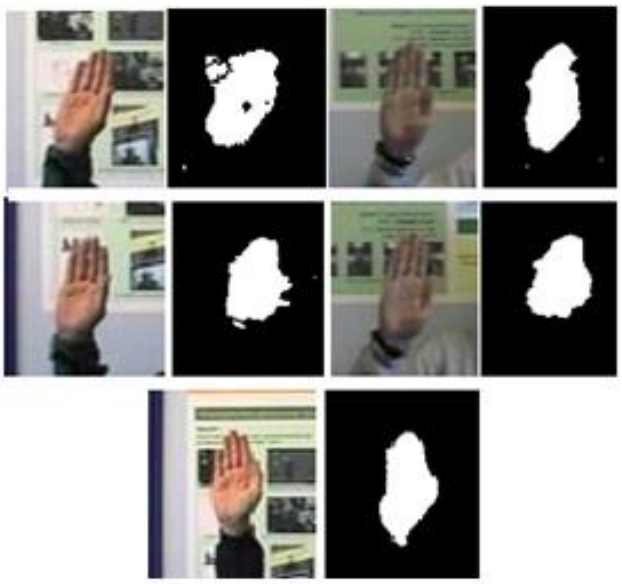


Fig. 4. Results of skin detector -Sebastian Marcel dataset II.

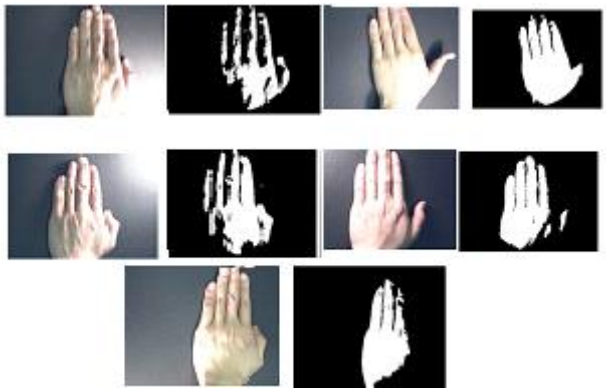


Fig. 5. Results of skin detector for posture 'flat' from the Cambridge dataset

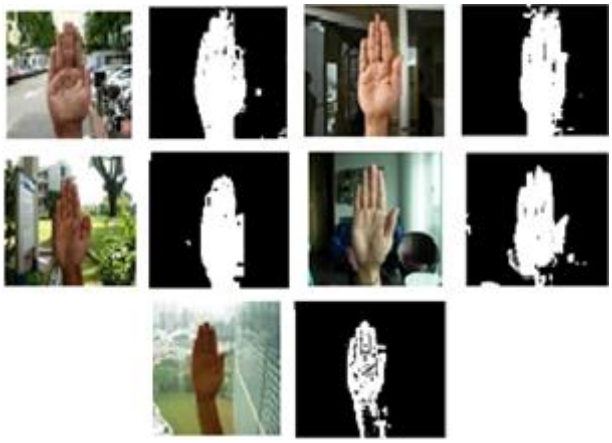


Fig . 6. Result of skin detector for posture 'B' from NUS Hand Posture Dataset II.

The results of the proposed work are presented in the following tables. For the experimentation six postures 'A', 'B', 'C', 'point', 'V', 'Five' from Sebastien Marcel database has been used. Table 1 describes the individual accuracy for each of these postures. The average accuracy achieved is 96%.

TABLE 1: CLASSIFICATION ACCURACY FOR SEBESTIEN MARCEL DATABASE

Static gesture	No. of gesture samples	Correct	Incorrect	Classification Accuracy %
A	100	96	04	96
B	100	94	6	94
C	100	95	5	95
Point	100	98	2	98
V	100	96	4	96
Five	100	97	3	97
Avearage Classification Accuracy				96

The proposed work is compared with the existing state of art techniques on the same benchmark dataset in Table 2 which expose that the experiment conducted in this paper is comparable with those of existing techniques.

TABLE 2: COMPARISON WITH EXISTING STATE OF ART TECHNIQUES FOR SEBASTIAN MARCEL DATABASE

Paper No.	Features	Classifier	Accuracy (%)
[19]	Modified Cesnsus Transform	AdaBoost	81.25
[20]	Haar like features	AdaBoost	90.0
[21]	Haar wavelets	Penalty score	94.89
[22]	Scale space features	AdaBoost	93.8
[23]	Bag of features	Support Vector Machine	96.23
[24]	Normalized Moment of Inertia (NMI) and Hu invariant moments	Support Vector Machine	96.9
Proposed method	Skin color and Fourier Descriptor	Artificial Neural Network	96

Four hand gestures 'A', 'B', 'C', 'D' are used for the experiments from the NUS Hand Posture dataset. 100 samples for each posture are used for training and 100 are used for testing. The results of the experiment are presented in Table 3. The average accuracy of 95.25% is achieved.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE 3. CLASSIFICATION ACCURACY FOR NUS HAND POSTURE DATABASE II

Static gesture	No. of gesture samples	Correct	Incorrect	Classification Accuracy %
A	100	96	4	94
B	100	94	6	94
C	100	97	3	97
D	100	96	4	96
Avearage Classification Accuracy				95.25

The results obtained through this experimentation are compared with the state of art techniques. The comparison reveals that the proposed method is better than the existing methods. The details of this are given in the table 4.

TABLE 4. COMPARISON WITH EXISTING STATE OF ART TECHNIQUES FOR NUS HAND POSTURE DATABASE II

Paper No.	Features	Classifier	Accuracy
[25]	Shape based and texture based features	GentleBoost	75.71
[26]	Viola jones	Real Time Deformable Detetctor	90.66
[27]	NUS standard model features (SMFs)	Fuzzy Rule Classifier Support Vector Machine	93.33 92.50
[28]	Shape texture color	Support Vector Machine	94.36
Proposed method	Skin color and Fourier Descriptor	Artificial Neural Network	95.25

Three primitive hand shapes 'flat', 'Spread' and 'V' from Cambridge Hand gesture database are used for testing the proposed algorithm. The results of the proposed work are presented in the Table 5. The experiment is carried out for each set of database and reported in the table. The average accuracy is 93.67% .

The results obtained through this experimentation are compared with the state of art techniques and reported in Table 6.

V. CONCLUSION AND FUTURE SCOPE

The paper proposed a system for hand segmentation and classification. The main component of the system is to track the hand based on skin color under different illumination conditions and with complex background. Fusion of HSV and YUV color space to detect the skin color gave the invariance to the illumination even in the complex background. The closed contour of the segmented hand is used to detect the shape of the hand gesture. Fourier descriptors are calculated as shape descriptors. To improve the robustness for the shape detection, the geometrical features are added in the feature vector. The feature vector thus achieved by combining the shape features and geometrical features are given to the artificial neural network for classification. The average classification accuracy of 95.25% is achieved for all the three databases.

The hand postures in the databases have the different viewing angle. So the classification accuracy can be further increased by extracting the view invariant features from the images. This lays a direction for further reseach in this area.

TABLE 5. CLASSIFICATION ACCURACY FOR CAMBRIDGE HAND GESTURE DATABASE

Static gesture	**No. of gesture samples**	**Set 1**	**Set2**	**Set 3**	**Set 4**	**Set 5**
Flat	100	93	96	96	92	94
Spread	100	94	95	93	93	93
V	100	95	96	95	92	94
Avearage Classification Accuracy		94	95.67	94.67	92.33	93.67

TABLE 6. COMPARISON WITH EXISTING STATE OF ART TECHNIQUES FOR CAMBRIDGE HAND GESTURE DATABASE

Paper No.	Features	Classifier	Accuracy (%)
[29]	PCA on Motion gradient orientation	Sparse Bayesian Classifier	80
[30]	Canonical Correlation Analysis (CCA) + SIFT	Support Vector Machine	85
[31]	Concatenated HOG	Kernel Diceriminant analysis with RBF kernel	91.1
[32]	Fourier Descriptors (Static postures 4 shapes)	Support Vector Machine	92.5
Proposed method	Skin color and Fourier Descriptor	ANN	94.50

REFERENCES

[1]. Vladimir I. et.al. 'Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review', IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, July 1997.

[2]. Sushmita Mitra, Tinku Acharya, 'Gesture Recognition: A Survey' IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, Vol. 37, No. 3, MAY 2007.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[3]. Terrillon, J., Shirazi, M. N., Fukamachi, H., Akamatsu, S, 'Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images', IEEE International Conference on Face and Gesture Recognition, pp. 54-61 (2000)
[4]. Khamar Basha Shaika et.al 'Comparative Study of Skin Color Detection and Segmentation in HSV and YCbCr Color Space' , 3rd International Elsevier Conference on Recent Trends in Computing 2015
[5]. Zhang Qiu-yu, Lu Jun-chi, et.al, "Hand Gesture Segmentation Method Based on YCbCr Color Space and K-Means Clustering", International Journal of Signal Processing, Image Processing and Pattern RecognitionVol. 8, No. 5 (2015), pp. 105-1 16
[6]. Bahare Jalilian, Abdolah Chalechale , 'Face and Hand Shape Segmentation Using Statistical Skin Detection for Sign Language Recognition', Computer Science and Information Technology .DOI: 10.13189/csit.2013.010305
[7]. Hsiang-Yueh. Lai 1, Han-Jheng. Lai, "Real-Time Dynamic Hand Gesture Recognition", IEEE DOI 10.1109/IS3C.2014.177 Published in: Computer, Consumer and Control (IS3C), 2014 International Symposium
[8]. Nasser H. Dardas and Emil M. Petriu," Hand Gesture Detection and Recognition Using Principal Component", IEEE International Conferenc Computational Intelligence for Measurement Systems and Applications (CIMSA), 2011
[9]. Zhi-hua Chen,1 Jung-Tae Kim,1 Jianning Lian "Real-Time Hand Gesture Recognition Using Finger Segmentation" The Scientific World Journal Volume 2014 (2014), Article ID 267872, http://dx.doi.org/10.1155/2014/267872
[10]. Wei Ren Tan, et.al., "A Fusion Approach for Efficient Human Skin Detection", IEEE transaction on Industrial informatics Vol8(1) PP138-147
[11]. Bishesh Khanal, et.al "Efficient Skin Detection Under Severe Illumination Changes and Shadows". ICIRA 2011 - 4th International Conference on Intelligent Rob otics and Applications, Dec 2011, Aachen, Germany. pp.1-8, 2011.
[12]. Luo et.al. "Illumination Normalization Method Based on Mean Estimation for A Robust Face Recognition" Research article.
[13]. Hsu, R., Ab del-Mottaleb, M., Jain, A. K.: Face detecting in color images. IEEE Trans. on Pattern Analysis and Machine Intelligence 24, 696-706 (2002)
[14]. Mohamed Alsheakhali, Ahmed Skaik "Hand Gesture Recognition System", International Conference on Information and Communication Systems (ICICS 2011)
[15]. S. Marcel, "Hand posture recognition in a body-face centered space," in Proc. Conf. Human Factors Comput. Syst. (CHI), 1999, pp. 302Processing, 2005, vol. 2, pp. 954 – 957
[16]. T-K. Kim et.al, " Tensor Canonical Correlation Analysis for Action Classification ",In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, 2007.
[17]. Pramod Kumar Pisharady, Prahlad Vadakkepat, Ai Poh Loh, "Attention Based Detection and Recognition of Hand Postures Against Complex Backgrounds", International Journal of Computer Vision, vol.101, no.3, pp.403-419, February, 2013
[18]. Mrs.A. R.Patil, Dr. S. Subbaraman, "Static Hand Gesture Detection and Classification using Contour based Fourier Descriptor", 9th World Congress International conference on Science and Engineering Technology at GOA. Jan. 2016 http ://iferp.in/digital_object_identifier/WCASETG9718
[19]. Agn`es Just et.al.,' Hand Posture Classification and Recognition using the Modified Census Transform', 7th IEEE International Conference on Automatic Face and Gesture Recognition (FGR'06) 0-7695-2503-2/© 2006 IEEE
[20]. Q. Chen, N. Georganas, and E. Petriu, "Real-time vision-based hand gesture recognition using Haar-like features," in Proc. IEEE IMTC, 2007,pp. 1–6.
[21]. Y. Fang, K. Wang, J. Cheng, and H. Lu, "A real-time hand gesture recognition method," in Proc. IEEE Int. Conf. ultimedia Expo, 2007, pp. 995–998.
[22]. W. Chung, X. Wu, and Y. Xu, "A real time hand gesture recognition based on Haar wavelet representation," in Proc. IEEE Int. Conf. Robot. Biomimetics, 2009, pp. 336–341
[23]. Nasser H. Dardas and Nicolas D. Georganas, "Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques". IEEE Transactions on Instrumentation and Measurement, vol.60, no.11, November 2011
[24]. Y. Ren and C. Gu, "Real-time hand gesture recognition based on vision," in Proc. Edutainment, 2010, pp. 468–475
[25]. Serre, T., et.al.(2007), 'Robust object recognition with cortex-like mechanisms', IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(3), 411–426
[26]. Uriel Haile Hernandez-Belmonte and Victor Ayala-Ramirez ,'Real-Time Hand Posture Recognition for Human-Robot Interaction Tasks', Sensors 2016, 16, 36; doi:10.3390/s16010036 www.mdpi.com/journal/sensors
[27]. P. Pramod Kumar et.al,' Hand Posture And Face Recognition Using A Fuzzy- Rough Approach', International Journal of Humanoid Robotics Vol . 7, N o. 3 (2010) 331–356 World Scientific Publishing Company DOI: 10.1142/S0219843610002180
[28]. Pramod Kumar Pisharady et.al, 'Attention Based Detection and Recognition of Hand Postures Against Complex Backgrounds',Journal Computer Vision DOI 10.1007/s11263-012-0560-5 published by springerlink
[29]. Shu-Fai Wong and Roberto Cipolla, ' Real-time Interpretation of Hand Motions using a Sparse Bayesian Classifier on Motion Gradient Orientation Images', BMVC 2005 doi:10.5244/C.19.41
[30]. T.-K. Kim and R. Cipolla. Canonical correlation analysis of video volume tensors for action categorization and detection. IEEE Trans. Pattern Analysis and Machine Intelligence, 31(8):1415–1428, 2009.
[31]. Masoud Faraki Mehrtash T. Harandi Fatih Porikli , 'Image Set Classification by Symmetric Positive Semi-Definite Matrices'
[32]. Heba M.Gamal et.al, ' Hand Gesture Recognition using Fourier Descriptors', 978-1-4799-008 274 0-0/13/$31.00 ©2013 IEEE

Authors' Profiles

Mrs.Anjali R Patil have completed BE (2002)and ME(2009) from Shivaji University, Kolhapur. Currently she is Pursuing PhD in Electronics Engineering from Shivaji University Kolhapur. Her research interest includes image processing, Pattern recognition, Soft computing etc.

Dr. Mrs. Shaila Subbaraman, Ph. D. from I.I.T., Bombay (1999) and M. Tech. from I.I.Sc., Bangalore (1975) has a vast experience in industry in the capacity of R & D engineer and Quality Assurance Manager in the field of manufacturing semiconductor devices and ICs. She also has more than 27 years of teaching experience at both UG and PG level for the courses in Electronics Engineering. Her specialization is in Micro-electronics and VLSI Design. She has more than fifty publications to her credit. She retired as Dean Academics of autonomous Walchand College of Engineering, Sangli and currently she is working as Professor (PG) in the same college. Additionally she works as a NBA expert for evaluating engineering programs in India in accordance with Washington Accord. Recently she has been felicitated by "Pillars of Hindustani Society" award instituted by Chamber of Commerce, Mumbai for contribution to Higher Education in Western Maharashtra.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

COIRS: Cost Optimized Identity based Ring Signature with Forward Secrecy in Cloud Computing

Muthi Reddy P [#1], Rekha Rangappa Dasar [#2], Tanuja R [#3], Manjula S H [#4], Venugopal K R [#5]

[#1,2,3,4,5] *Department of Computer Science and Engineering,*
University Visvesvaraya College of Engineering, Bangalore University,
Bengaluru-560001, Karnataka, India

[1] muthireddy2016@gmail.com

***Abstract*—In recent days Cloud computing is a rising technique which offers data sharing with more *efficient*, *effective* and *economical* approaches between group members. To create an authentic and anonymous data sharing, *IDentity based Ring Signature (ID-RS)* is one of the promising technique between the groups. Ring signature scheme permits the manager or data owner to authenticate into the system in anonymous manner. In conventional Public Key Infrastructure (PKI) data sharing scheme contains certificate authentication process, which is a bottleneck because of its high cost. To avoid this problem, we proposed Cost Optimized Identity based Ring Signature with forward secrecy (COIRS) scheme. This scheme helps to remove the traditional certificate verification process. Only once the user needs to be verified by the manager giving his public details. The cost and time required for this process is comparatively less than traditional public key infrastructure. If the secret key holder has been compromised, all early generated signatures remains valid (*Forward Secrecy*). This paper discuss about how to optimize the time and cost when sharing the files to the cloud. We provide a protection from collision attack, which means revoked users will not get the original documents. In general better *efficiency* and *secrecy* can be provided for group sharing by applying above approaches.**

***Index Terms*—Anonymity, Authenticity, Forward secrecy, Group sharing, Ring signature**

I. INTRODUCTION

Cloud computing is an Internet based technology because of its widespread and popular use. It enables both users and enterprises to keep their information in cloud storage and allows resource sharing [1], [2], [3], [4]. Cloud computing is widely used because of its two main applications, which are as follows: i) Vast amount of information storage: Cloud storage allows the users to store the files on users request. Cloud storage provides benefit to store huge amount of storage facility. ii) Allows users to easily share their data: Cloud computing technology provides another facility that is to easily share files to the public and to the individual. It allows sharing of data through a third party which becomes more economically useful. Privacy of both the data and group members identities are most significant notion in cloud computing. Consider a *Smart Grid* example as shown in fig. 1, users in smart grid may get their data usage file without any encrypted format and they get encouraged to share their private information with others. Consider an example, if the user wants to upload their files to the cloud platform like Microsoft Azure, from that gathered copy of energy data files several statistical copies are created. Anyone could match the data files about energy consumption with others. This may lead to critical problems to energy usage while accessing, analyzing and responding back to the cloud. Because of its openness, deployment of data sharing took place in a standalone background, it is open to several secrecy problems [5], [6], [7]. There are many secrecy criteria to be reached in order to achieve data *efficiency* and *secrecy*, i.e.,

i) *Authenticity of Data:* The signed data usage file would be confusing in the example of smart grid, if that data file is copied by the adversaries. At the same time this type of problems can be solved by using some cryptographic techniques such as digital signatures, hash functions, encryption or decryption techniques or message authentication techniques. User might face other issues in smart grid system like *anonymity* and *efficiency*.

ii) *Data Anonymity:* The signed energy usage file is enclosed with huge amount of information of consumers, sharing in the smart grid is processed in fine grained fashion.Then the signed energy file anyone can copy the information of consumers from the system. The copied information may be of electrical utilities used for a particular time etc., therefore, it is not easily possible to hold the anonymity condition of consumers.

iii) *Data Efficiency:* The smart grid (it is an electric grid

consisting a variety of operational, vitality measures, smart apparatuses, sustainable power source assets, smart meters) for data sharing system contains a large number of users, to save the consumption of energy from such smart grid systems. A realistic system must decrease its communication cost and computation as less as possible or else it would lead to energy wastage, this is against to the aim of smart grid. To overcome above metrics and provide more secure in data sharing COIRS model is introduced and it reduces group accessing time and cost of the files. We dedicate this paper to examining essential goals for understanding the three properties as described above.

1) Data Authenticity
2) Anonymity
3) Efficiency

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

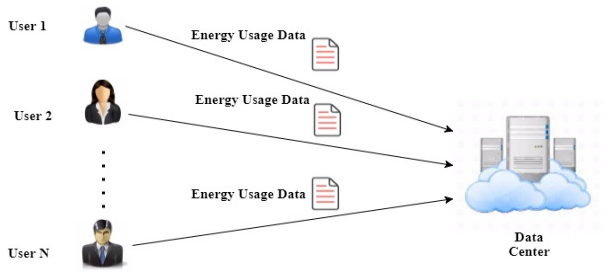


Fig. 1: File Data Sharing in Smart Grid.

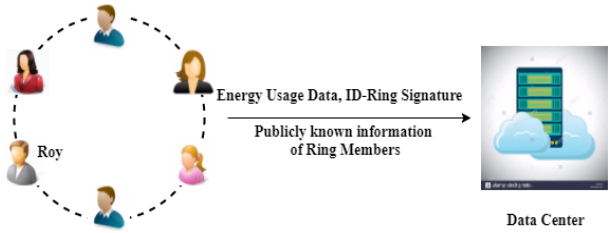


Fig. 2: Identity Based Ring Signature.

Instead of those secrecy issues there are other secrecy tools, such as availability (even under network attacks, service is being provided at an acceptable level) and access control. We discussed how our COIRS model is used in identity based cryptosystem and advantages in big data system in next part.

A. *Identity based cryptosystem*

Shamir [8] has introduced the first IDentity-based cryptosystem. It removes the necessity for proving the validity of Public Key (PKey) certificates, the maintenance in conventional public key infrastructure is both cost and time consuming. By collecting the publicly known users unique identity like address or email-id for the public key of the user is calculated. In ID-based Cryptosystem, private keys can be generated by a private key generator and later master-secret for users is calculated. Identity-based cryptosystem scheme removes the necessity of certicate validation, which is a part of traditional PKI and links an implicit PKey to all members inside the system. In ID-based signature, one does not require to validate the certicates first which is a contradiction to the conventional public key infrastructure. The removal of such certificate verification makes the entire verification process more effective. This would definitely lead to a major save in both computation and communication cost when huge number of consumers are involved (smart-grid). Here we assign some cost value to particular file to optimize the overall cost required for the process. One constant cost value for the file is assigned. The file size increases then their cost value varies. RS is a group focused signature along with secrecy assurance on signer. The client can sign secretly in the interest of a group's individual choice, while individuals are absolutely ignorant of signature generated using their identity information. Verifier can check that a data has been signed by one of the individuals of the group. However the real character of the signer isn't being shared [9]. RS could be utilized for the application of whistle blowing [10] and an anonymous authentication for groups [11]. Numerous different applications which don't need group development stage however require signer secrecy.

B. *An advantage in big data system*

Because of its normal structure, ID-based framework has a positive advantage in Big Data. RS in ID-based framework has an imperative favorable position over its partner in ordinary open key framework, for the most part in the huge information diagnostic scheme. Consider an event including 20,000 individuals in the group, the signature verifier of a traditional PKI based framework should approve all 20,000 certificates first,then one can take out the actual message verification process along with the signature. Unlike traditional PKI, in ID-based RS just the ring client's information along with the message and signature sets are required. Subsequently, we would be able to eliminate the expensive certicate validation process, which spares a lot of calculation time and execution time. As the quantity of clients in the ring builds, sparing will be more basic if a more elevated amount of secrecy is needed. As outlined in fig. 2, ID-based RS plot is more preferable, where huge number of individuals are involved with the framework like smart grid framework is as following:

i) The vitality information proprietor (say, Roy), first make a ring or group by choosing an group of clients. This stage just requires public information of the users, similar to changeless or private locations, and Roy does not require the relationship between any ring individuals.

ii) Roy uploads his private details of electronic utilization, along with a group signature and the identity details of all group individuals.

iii) by approving the produced ring signature, one can be ensured that the information or message is certainly conveyed by a legitimate occupant, meanwhile we cannot find out actual signer of the group. Anonymity of the message provider is guaranteed along with the data or message authenticity. At the same time the verification process is highly efficient because it does not include any certificate verification method.

By adding more users in the ring one can achieve a higher level protection, but the possibility of key disclosure might increase. Key exploration is the real disadvantage of ordinary advanced signatures. Assume the SKey of an user is compromised, every single past signature of that client becomes valueless: future signatures are rejected and also already issued signatures can't be trusted. It doesn't resolve the issue of forgeability for past produced signatures.

C. *Motivation*

1) Key Exposure: The idea of forward secrecy is proposed to protect the legitimacy of past signatures regardless of the possibility that the present SKey holder is compromised.

2) Big Data Key Exposure : The exposure of key in a RS scheme is more serious issue: suppose a user's private key is uncovered by any unauthorized user then user can develop a legitimate ring signatures of any records in the interest of that ring. Indeed, even more in worst condition, the ring can be characterized by his own choice. Even one can not recognize whether a ring signature is created preceding the

key introduction or by which client. Subsequently, forward secrecy is a necessary prerequisite for all frameworks to share information.

D. Contribution

A creative idea called ID-based forward secure ring signature is introduced which is an essential key for developing a COIRS framework. It gives a formal clarification on ID-based forward secure RS; we exhibit a solid outline of COIRS scheme, no past ID-based RS scheme had the property of forward secrecy, we demonstrate the secrecy of the proposed method under the standard RSA model assumption. Use of COIRS plan contains:

1) The disposal of expensive certicate verification process makes it versatile and particularly reasonable for enormous information explanatory condition.
2) The secret key is small in size.
3) Exponentiation is done in key update process.
4) We are calculating the energy usage required by the data owner to upload files to the cloud and downloading energy for the data centre for providing files to the clients.
5) We are determining the cost required by the owner to upload the files and data centre to download the files requested by the clients.

Organization: In section II, we give related work on forward secrecy to provide authentication access and cost optimization. In section III, we describe architecture of COIRS model. In section IV, we discuss mathematical model of COIRS scheme. In section V, we deal on experimental analysis. We concluded our model in section VI.

II. RELATED WORK

Liu *et al.,* [12] proposed a novel that can completely maintain fine-grained update request and authorized auditing by providing a proper examination for feasible forms of fine-grained data updates. Based on the above idea enhancement is made, that can significantly diminish communication expenses for verifying small updates, and significantly reduce the overhead for big-data applications. Yang *et al.,* [13] studied first outlined an evaluating structure for distributed storage frameworks and proposed an efficient and protection safeguarding inspecting convention. Then, they stretched out evaluating algorithms to help the information dynamic operations, which is efficient and provably secure. The examination and re-enhancement comes about in demonstration that proposed evaluating conventions are secure and efficient, particularly it reducing the calculation cost. Nabeel *et al.,* [14] proposed a vital issue in broad daylight mists by which to specifically share reports in view of fine-grained Access Based Control Policy Scheme (ACPS). An approach is to scramble records fulfilling diverse strategies with various keys utilizing an open key cryptosystem, for example, property based encryption, as well as intermediary re-encryption [15].

Dai *et al.,* [16] studied inventions to decrease vitality utilization by server farms considering the position of virtual machines onto the servers in the server farm astutely. This discuss as a number of programming issue, demonstrate it in NP-hard, at that point investigate two eager guess calculations, least vitality virtual machine and least correspondence virtual machine planning calculation, to learn the vitality while fulfilling the inhabitant's administration level agreements. Bera *et al*., [17] demonstrates the quick paced improvement of energy frameworks that requires keen networks to encourage continuous control and checking with bidirectional correspondence and power flows. To concentrate on dependable, efficient, secured and financially survey on control administration prerequisites. Li *et al*., [18] worked on despite the fact that, it recommended that a half breed cloud may spare cost contrasted and assembling an intense private cloud, extensive leasing expense and correspondence cost are still presented in such a world view. The most effective method to improve such operational cost ends up plainly one noteworthy worry for the SaaS suppliers to receive the crossover cloud figuring world view. Yang *et al.,* [19] presented novel strategies in light of compiler code investigation that viably lessen the exchanged information measure by exchanging just the basic store objects and the stack outlines really referenced in the server. The tests display that the decreased size decidedly impacts the exchange time itself as well as the general adequacy of execution offloading and eventually, enhances the execution of versatile distributed computing altogether as far as execution time and vitality utilization is concerned.

Yao *et al*., [20] built a novel structure named cost optimization for internet content multihoming. COMIC progressively adjusts end-clients' heaps among server farms and CDNs in order to limit the substance benefit cost. To guarantee superior for content conveying, content diministration uses an innovation known as substance multihoming: substance are produced from numerous geologically appropriated server farms and conveyed by different dispersed substance circulation systems. The power costs for server farms and the utilization costs for CDNs are real supporters of the substance benefit cost. As power costs change crosswise over server farms and use costs fluctuate crosswise over CDNs, planning server farms and CDNs has a huge outcome for advancing substance benefit cost.

Trombetta *et al*., [21] recommended three conventions tackling this issue on concealment based, speculation based k-mysterious and secret databases. The conventions depend on surely understood cryptographic presumptions, and we give hypothetical investigations to evidence their soundness and test results to represent their productivity. Zhou *et al.,* [22] proposed a plan that enables an association to store information safely in an open cloud while keeping up the touchy data identified with the association's structure in a private cloud. Clients of open distributed computing do not know where their information is put away. They have a misguided judgment of losing their information.

Amelie *et al.,* [23] studied about the difficulties of controlling administration rates and applying the N-strategy to improve operational cost inside an execution. The cost work has been created in which the expenses of energy utilization, framework clock and server start-up are altogether mulled over. Yu *et al.,* [24] have built up an efficient id-based threshold ring signature scheme. Edge ring signature empowers any

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

group of *t* substances immediately recruiting discretionary *n-t* elements to create an openly undeniable t-out-of-n edge signature for the benefit of the entire gathering of the *n* elements, while the genuine underwriters stay a mysterious. Bellare *et al.,* [25] studied a forward secure digital signature conspires, it is a digital signature diagram in which open key is settled yet mystery signature key is refreshed at consistent interims in order to give a forward secrecy appropriately, bargaining of the mystery key does not enable the oppose to produce the signatures relating to the past. This can be helpful to alleviate the harm caused by key introduction without requiring the dissemination of keys [26], [27].

III. COIRS MODEL

In this section, we are discussing the mathematical assumption, secrecy model and designed architecture of COIRS secrecy model. The different notations for efficiency comparison is explained in table I.

A. Mathematical assumption

a) Denition: Let $M = uv$, where *u* and *v* are two b-bit prime numbers where $u = 2u' + 1$ and $v = 2v' + 1$ for some primes u', v'. Let *r* be a prime, r $> 2^{\ell}$ for a some constant parameter ℓ, where $gcd(r, \phi(M)) = 1$. Let *x* is a random element in Z_M^*. We say that an algorithm *A* resolve the RSA dilemma if it accept an input the tuple *(M, r, x)* and outputs an element *z* such that z^r = *x mod M*.

B. Secrecy model

Cost Optimized Id-based Ring Signature (COIRS) scheme is a part of Probabilistic Polynomial Time (PPT) algorithms. This PPT contains the following operations:

1) *Setup*:
   - Input ← (1^{γ}, *Prm*, *MSGG*, *S*).
   - Results ← PKG generates Master Secret key *(MSkey)* and parameter list *Prm*.
2) *Extract*:
   - Input ← *Prm*, an identity $ID_i \in \{0,1\}^*$, 1^{γ}, *MSkey*.
   - Results ← Users Secret Key $(SKey_{i,0}) \in K$ such that the secret key is valid for time $t = 0$. When we say identity ID_i corresponds to user secret key $SKey_{i,0}$ or vice versa, we mean the pair *(*ID_i*,* $SKey_{i,0}$*)* is an input-output pair of *Extract* with respect to *Prm* and *MSkey*.
3) *Update*:
   - Input ← $SKey_{i,t}$ for a time period *t*.
   - Results ← New user Secret Key $SKey_{i,t+1}$ for the time period *t+1*.
4) *Sign*:
   - Input ← Parameter list *Prm*, *t*, group size *n* of length polynomial in γ, a set *L* = $ID_i \in \{0,1\}^* | i \in [1,n]$ of *n* user identities, $MSg \in MSGG$ and $SKey_{\pi,t} \in K, \pi \in [1,n]$ for time *t*
   - Results ← signature $\alpha \in S$.
5) *Verify*:
   - Input ← parameter list *Prm*, *t*, group size *n* of length polynomial inγ, a set *L* = $ID_i \in \{0,1\}^* | i \in [1,n]$ of *n* user identities, $MSg \in MSGG$ and a signature $\alpha \in S$.
   - Results ← generated signature $\alpha \in S$ is valid or invalid.

a) Correctness: A (1, n) COIRS scheme should satisfy the verification on correctnesssignatures signed by honest signer are veried to be invalid with negligible probability.

C. Architecture of COIRS scheme

The architecture of Cost Optimized Identity based Ring Signature with forward secrecy (COIRS) scheme is illustrated in fig. 3. The architecture mainly consists of four components:

1) User
2) Admin
3) Private Key Generator (PKG)
4) Public Cloud

a) User: User is the one who wants to share their personal information to others or they wish to keep secret or confidential data hidden from unauthorized persons. In COIRS scheme, user registers to a cloud by filling all his details. Admin or manager of the particular group grants the authorization permission to users to perform the desired upload/download operations. By agreeing terms and conditions of the registered cloud, user can perform the upload and download the operations. After logging in to the particular group by getting OTP to user *email id* which is entered while registering at the first time. The user becomes a group member in addition the user has rights to perform the tasks. For every task of a group signature is generated by a particular user on behalf of the group to maintain secrecy and forward secrecy to avoid unauthorized access.

b) Admin: Admin gives access to the registered users before performing the tasks. Admin then collects all registered user's public details and uploads his information with users details to maintain the users log records. Admin will keep the information about file details of all the user details, accessing details etc.

c) Private Key Generator (PKG): It generates the private keys for all registered users and these key will be vary every time while performing new task. PKG sets up the group's average time, to calculate the average time required by the group to upload and download the files.

d) Public Cloud: Public cloud is the cloud infrastructure where any user can access the information from the cloud. Here there are several cloud service providers like *Microsoft Azure*, *Dropbox*, $Google^+$, *Amazon*, etc., these service providers provide the services to requesting users by using some algorithms to maintain privacy and secrecy of the data.

IV. MATHEMATICAL MODEL OF COIRS SCHEME

In this section, we are going to give the description and analysis of our COIRS scheme.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

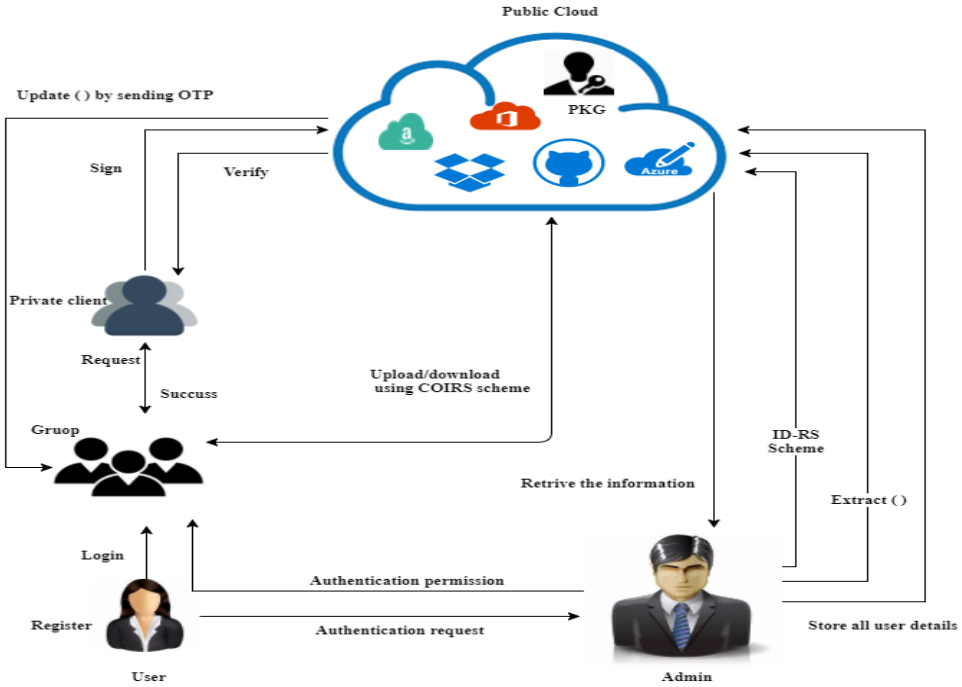


Fig. 3: Architecture of COIRS scheme.

A. The Design

Assume that the user private key and group member identities are valid up to T periods and do the time period intervals as public ans set the message space MSGG = (0, 1)*

- *Setup*: Let γ is a secrecy parameter using as input to a setup phase, the *PKG* generates two random b-bit prime numbers *u* and *v* such that $u = 2u' + 1$ and $v = 2v' + 1$ for some primes u', v'. It computes *M=uv*. For fixed parameter ℓ, it selects a random prime number *r* such that $2^{\ell} \leq r \leq 2^{\ell+1}$ and $gcd(r, \phi(M)) = 1$. It selects two hash functions $HF_1 : 0,1^* \rightarrow Z_N^*$ and $HF_2 : 0,1^* \rightarrow 0,1^{\ell}$. The public parameters *Prm* are $(b, \ell, r, M, HF_1, HF_2)$ and the *MSkey* is u,v.
- *Extract*: The PKG generates the user secret key for user *i*, with user's identities $ID_i \in 0,1^*$ requests for a secret key at time period *t* (integer), where $0 \leq t \leq T$. $SKey_{i,t}$ = $[HF_1(ID_i)]^{1/r^{(T+1-t)}}$ mod M.
- *Update*: $SKey_{i,t}$ as a input for a time period t, if t < T the user updates the secret key as $SKey_{i,t+1}$ = $SKey_{i,t}^{r}$ mod N. Or else, the algorithm yields $\perp$ means the secret key has expired.
- *Sign*: To sign a message $MSg \in (0,1)^*$ in time period *t* where $0 \leq t \leq T$, on behalf of a ring of identities L= $ID_1, ..., ID_n$ a user with identity $ID_\pi \in L$ and secret key $SKey_{i,t}$:
  1) For all $i \in 1, ..., n$, choose random $A_i \in Z_M^*$ and compute $R_i = A_i^{e^{(T+1-t)}}$ mod M and $h_i = HF_2(L, m, t, ID_i, R_i)$
  2) Choose random $A_\pi \in Z_M^*$ and compute $R_\pi = A_\pi^{e^{(T+1-t)}}$ mod M * $\Pi_{i=1,i\neq\pi}^{n} HF_1(ID_i)^{-h_i}$ mod M and $HF_\pi = HF_2(L, MSg, tID_i, R_\pi)$
  3) Compute $SKey_{i,t}^{h_\pi} * \Pi_{i=1}^{n} A_i$ mod M.
  4) Output the signature for the list of identities *L*, the message MSg and the time period *t* as $\gamma = (R_1 ... R_n, h_1 ... h_n, s)$.
- *Verify*: To verify a signature α for a message MSg, a list of identities *L* and the time period *t*, check whether $h_i = HF_2(L, m, t, ID_i, R_i)$ for i = 1,..., M and $\gamma^{e^{(T+1-t)}} = \Pi_{i=1}^{n}(R_i HF_1(ID_i)^{h_i})$ mod M. We will get valid output if all equality's satisfied. Otherwise the result will be invalid.

B. Correctness

We are checking whether our secrete key is valid or not by considering the equations on left hand side with the right hand side. The secrete key verification becomes succuss then LHS=RHS.

$\gamma^{e^{(T+1-t)}} = \Pi_{i=1}^{n}(R_i HF_1(ID_i)^{h_i})$ mod M

LHS=$\gamma^{e^{(T+1-t)}}$

= $((SKey_{\pi,t})^{h_\pi} * \Pi_{i=1}^{n} A_i$ mod M) $^{e^{(T+1-t)}}$

=$((HF_1(ID_\pi)^{1/r^{(T+1-t)}})^{h_\pi} * \Pi_{i=1}^{n} A_i$ mod M) $^{e^{(T+1-t)}}$

=$(HF_1(ID_\pi)^{h_\pi} * \Pi_{i=1}^{n}(A_i)^{r^{(T+1-t)}}$ mod M)

RHS=$\Pi_{i=1}^{n}(R_i * HF_1(ID_i)^{h_i})$ mod M

= $(\Pi_{i=1,i\neq\pi}^{n}(R_i * HF_1(ID_i)^{h_i})) * (R_\pi * HF_1(ID_\pi)^{h_\pi})$ mod M

= $(\Pi_{i=1,i\neq\pi}^{n}(A_i^{r^{(T+1-t)}} * HF_1(ID_i)^{h_i})) * (A_\pi^{r^{(T+1-t)}} * \Pi_{i=1,i\neq\pi}^{n} HF_1(ID_i)^{-h_i} * HF_1(ID_\pi)^{h_\pi})$ mod M

= $(\Pi_{i=1}^{n}(A_i)^{r^{(T+1-t)}}) * HF_1(ID_\pi)^{h_\pi}$ mod M

=LHS

Therefore, LHS= RHS.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE I: Notations for efficiency comparison

Notation	Definition
Pkey	*Public Key*
1^{γ}	*Security Parameter*
M	*Group Size*
Prm	*Public System Parameter*
L	*List of Identities of all Users*
MSkey	*Master Secret Key*
MSg	*Message*
K	*User Secret Key Space*
S	*Signature Space*
MSSG	*Message Space*
ID	*Identity of User*
G	*Cyclic Bilinear Group*
SKey	*Secrete Key*
α	*Signature*
t	*Time*

Algorithm 1 Forward secrecy

procedure SIGNATURE
  Variables: User, Group Signature, Time, Admin.
Start:
  $U_i \leftarrow$ *User*, Log in to the cloud system.
  $A_i \leftarrow$*Admin*, Authentication permission to user U_i.
  At time T_i, user uploads a file F_i.
  $G_s \leftarrow$ Group Signature, generated by the cloud authority,
    where $G_s \in (U_i, T_i)$.
  At T_{i+1}, G_s is invalid.
  User is not able to access the data using other person's signature key.
End.

C. Algorithms

Our COIRS scheme proposed Algorithm 1 provide a better secrecy to user's files. In forward secrecy algorithm as name depicts it provides one step more secrecy for being accessed by the unauthorized users. We use asymmetric cryptographic technology with random variables. In forward secrecy technique at each stage the group signature is being produced, it means, if the secrete key holder compromised with others the secrecy of current file as well as past signatures being exposed by unauthorized users. To overcome this problem, asymmetric cryptography technique is used to generate different signature at every encryption and decryption process. Our cost calculation algorithm 2 calculates the overall cost required by the user to upload as well as download the file. As the size of the file increases the cost for that particular file is going to increase. Let F_i is the file size in bytes, A_i is the cost value per byte. Overall cost required to upload and download the files is given by,

$$C_{upload} = F_i * A_i \tag{1}$$

$$C_{download} = F_i * A_i \tag{2}$$

Algorithm 2 Cost Calculation

procedure COSTCOMPUTINGN
  Variables: Cost, File, Amount.
Start:
  $F_i \leftarrow$ File size *i*, bytes or kb.
  $A_i \leftarrow$ Amount or cost/byte or kb.
  For upload A_i= A_i/byte or kb.
  C_{upload} = $F_i * A_i$.
  For download A_i = A_i*/byte or kb*.
  $C_{download}$ = $F_i * A_i$.
End.

Algorithm 3 Average time calculation for the file size F_i

procedure AVERAGETIME
  Variables: System Time, Time periods, Amount.
Start:
  $T_i \leftarrow$ System Time in ms.
  Time periods T are divided into four time slots like 100, 200, 300, 400
    ms.
  $A_i \leftarrow$ Average time.
  $C_i \leftarrow$ Count of the group.
  $TotalTime \leftarrow$ = A_i / C_i.
  Result= TotalTime * F_i, where i is an integer value
    i.e F_1 = 1024 kb and F_2 = 2048 kb.
  For Upload or download a file of size F_1 = 1024 kb and F_2 = 2048 kb.
  Compute $T_i \leftarrow$ Result/Time period time slot, where *i* is integer
End.

Algorithm 3 computes the average time required for our COIRS model to upload the file where file sizes are 1024 kb and 2048 kb. As the size of the file increased the time required by the data owner to upload the files to the cloud becomes increases. The Time period is divided into 4 time slots, 100, 200, 300 and 400 ms. Total time is calculated separately for all time periods. It is calculated as,

$$TotalTime = A_i \div C_i \tag{3}$$

Here we are considering two constant file sizes are 1024 kb and 2048 kb. We calculate the average time to both these files is shown in fig. 6 and fig. 7.

V. EXPERIMENTAL ANALYSIS

In this section, we are analyzing our COIRS scheme on the bases of *Time* and *Cost* evaluation.

A. Time and Cost analysis

In our COIRS model, we evaluate the time and cost analysis using two entities *Data owner* and *Data center*. For both time and cost analysis, experiments were conducted by taking some constant files to generate the accurate analysis. Our analysis

for uploading time for each file when user uploads different file sizes. We are taking some constant file sizes i.e., 100 kb, 200 kb, 300 kb, 400 kb, 500 kb, 1000 kb, 1024 kb, 2000 kb and 2048 kb for uploading and downloading a file. In fig.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE II: Average time for the PKG to setup in COIRS system.

M (in kb)	Time (in ms)
1024	80
2048	1040

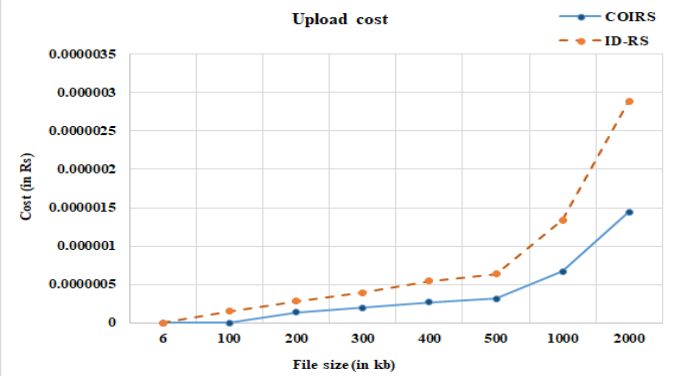


Fig. 4: The different file size upload cost using COIRS and ID-RS schemes.

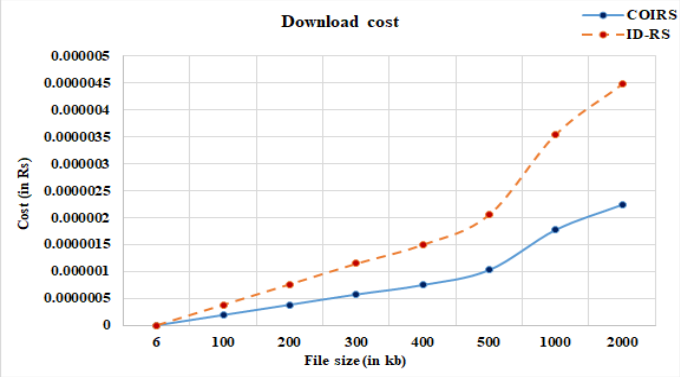


Fig. 5: The different file size download cost using COIRS and ID-RS schemes.

4, we depicts the cost required by the data owner to upload his different file sizes using COIRS and ID-RS model, as the size of the file increases, the cost of that file is going to increase and vice versa. At certain file size like 50 TB the cost value becomes threshold value. Above this threshold, the cost is depends upon slab values of the different file sizes. In fig. 5, illustrated the cost required for the data center using COIRS and ID-RS model to download the files for the user. The experiments were conducted on DELL i5 workstation inbuilt with 2.0 GHz, Intel Xeon dual-processor with 8 GB RAM and running on Windows 8 Professional 64-bit OS.

B. Implementation and Experimental Results

We calculated the analysis of our COIRS model with respect to 3 entities: *Data owner*, *Data center* and *Private key generator*. All analysis were conducted 20 times to gain an average results. The average upload time for the data owner using COIRS and ID-RS scheme, when F = 1024 kb is depicted in fig. 6 and time consumption is depicted in table III.

TABLE III: The average time for the data owner to upload file F = 1024 kb.

Group Name	Count	Time in COIRS (ms)	Time in ID-RS (ms)
Group 1	5	28	35
Group 2	10	30	40
Group 3	20	43	66

TABLE IV: The average time for the data owner to download file F = 1024 kb.

Group Name	Count	Time in COIRS (ms)	Time in ID-RS (ms)
Group 1	5	40	52
Group 2	10	45	61
Group 3	20	55	82

The average download time for the data center using COIRS and ID-RS scheme, when F = 1024 kb is as shown in fig. 7, for different groups with different file sizes and time consumption is depicted in table IV. The different users in Group 1, Group 2 and Group 3 contains 5, 10 and 20 users respectively. The count 'C' increases group by group. The average time required by the groups to upload and download their file increases as the number of users increased in the group.

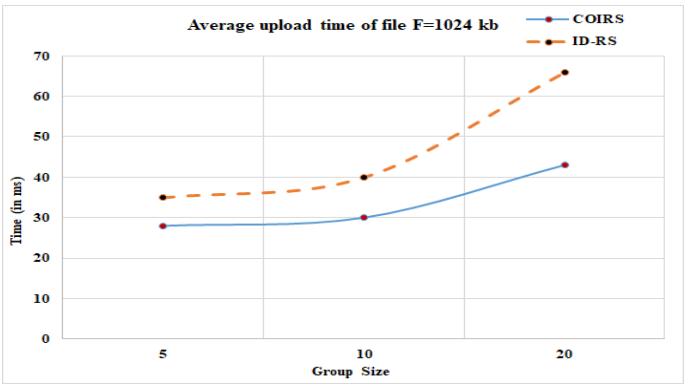


Fig. 6: The average upload time for the data owner using COIRS and ID-RS scheme, when F = 1024 kb.

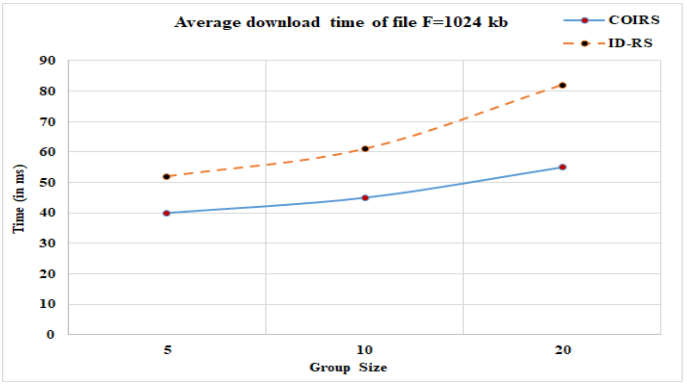


Fig. 7: The average download time for the data center using COIRS and ID-RS scheme, when F = 1024 kb.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

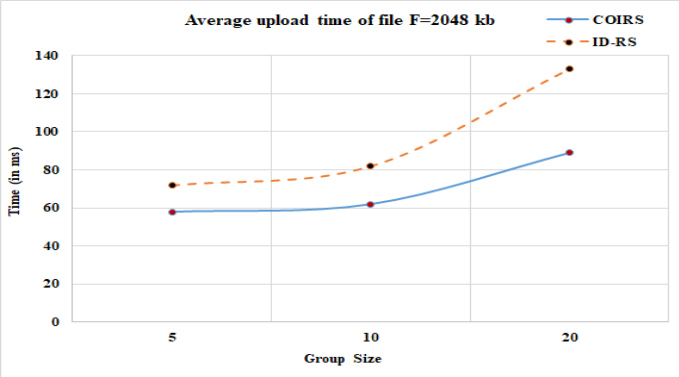


Fig. 8: The average upload time for the data owner using COIRS and ID-RS scheme, when F = 2048.

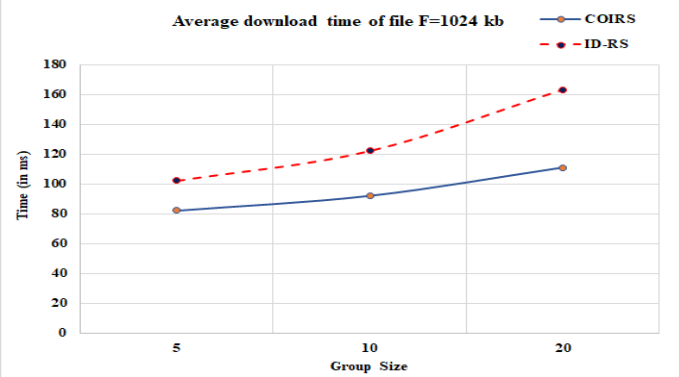


Fig. 9: The average download time for the data center using COIRS and ID-RS scheme, when F = 2048 kb.

Experiments were taken for the two constant file sizes F = 1024 kb and F = 2048 kb. In table II shows the average time for the private key generator to setup the system. PKG took 80 and 1040 ms to setup the whole system for F = 1024 kb and F = 2048 kb respectively. The average upload time for the data center to upload file with different choices of M and T, for F = 2048 kb is as shown in fig. 8. The average download time for the data center using COIRS and ID-RS scheme, when F = 2048 kb is illustrated in fig. 9. This requires authenticated users only upload or download files. The Time slices were increased by multiples of 100 up to 400. In group sharing decreases *cost* and *time*. The test bed for the user is a personal computer built in with 2 GHz Intel CPU with 3 GB RAM and running Windows 8 OS.

VI. CONCLUSIONS

In group sharing scheme, to create an authentic and anonymous data sharing, *Ring signature* is one of the promising technique. Ring signature scheme permits the manager or data owner to authenticate into the system in anonymous manner. In conventional sharing scheme certificate authentication becomes a bottleneck because of high cost. To avoid this problem COIRS scheme is constructed. This scheme describes, suppose, the secret key holder has been compromised, all generated past signatures still remain valid. Discussed about how to optimize the time and cost when sharing the files to the cloud. Provide a protection to this scheme from collision attack, it means that revoked users cannot get the original documents and to reach high efficiency, implies that previous users not necessary to update their secret keys for the condition while new user enters the group or exit from the group. In generally high secrecy can be provided for group sharing, by applying all these approaches. COIRS scheme reduces *cost* of file sharing, *time* of file upload or download and provides high *security* using Ring signature.

REFERENCES

[1] P. Muthi Reddy, S. H. Manjula, and K. R. Venugopal, "Secure data sharing in cloud computing: A comprehensive review," *International Journal of Computer (IJC)*, vol. 25, no. 1, pp. 80–115, 2017.

[2] C. Wang, S. S. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for secure cloud storage," *IEEE transactions on computers*, vol. 62, no. 2, pp. 362–375, 2013.

[3] X. Liu, Y. Zhang, B. Wang, and J. Yan, "Mona: Secure multi-owner data sharing for dynamic groups in the cloud," *ieee transactions on parallel and distributed systems*, vol. 24, no. 6, pp. 1182–1191, 2013.

[4] Z. Zhu and R. Jiang, "A secure anti-collusion data sharing scheme for dynamic groups in the cloud," *IEEE Transactions on parallel and distributed systems*, vol. 27, no. 1, pp. 40–50, 2016.

[5] S.-H. Seo, M. Nabeel, X. Ding, and E. Bertino, "An efficient certificateless encryption for secure data sharing in public clouds," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2107–2119, 2014.

[6] K. Liang, M. H. Au, J. K. Liu, W. Susilo, D. S. Wong, G. Yang, T. V. X. Phuong, and Q. Xie, "A dfa-based functional proxy re-encryption scheme for secure public cloud data sharing," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 10, pp. 1667–1680, 2014.

[7] P. Muthi Reddy, K. Dhanakshirur, R. Tanuja, S. H. Manjula, and K. R. Venugopal, "Security and privacy of data sharing in cloud computing," in *in Proceedings of Fifth International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST)*. CRS Press and Taylors, 2018.

[8] A. Shamir, "Identity-based cryptosystems and signature schemes," in *Workshop on the theory and application of cryptographic techniques*. Springer, 1984, pp. 47–53.

[9] P. Muthi Reddy, S. H. Manjula, and K. R. Venugopal, "Sdspg: Secured data sharing with privacy as a group in cloud computing," *International Journal of Current Advanced Research (IJCAR), ISSN: 2319-6505*, vol. 7, no. 2, pp. 10036–10043, 2018. DOI: http://dx.doi.org/10.24327/ijcar.2018.10043.1683.

[10] X. Huang, J. K. Liu, S. Tang, Y. Xiang, K. Liang, L. Xu, and J. Zhou, "Cost-effective authentic and anonymous data sharing with forward security," *IEEE Transactions on computers*, vol. 64, no. 4, pp. 971–983, 2015.

[11] E. Bresson, J. Stern, and M. Szydlo, "Threshold ring signatures and applications to ad-hoc groups," in *Annual International Cryptology Conference*. Springer, 2002, pp. 465–480.

[12] C. Liu, J. Chen, L. T. Yang, X. Zhang, C. Yang, R. Ranjan, and R. Kotagiri, "Authorized public auditing of dynamic big data storage on cloud with efficient verifiable fine-grained updates," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 9, pp. 2234–2244, 2014.

[13] K. Yang and X. Jia, "An efficient and secure dynamic auditing protocol for data storage in cloud computing," *IEEE transactions on parallel and distributed systems*, vol. 24, no. 9, pp. 1717–1726, 2013.

[14] M. Nabeel, N. Shang, and E. Bertino, "Privacy preserving policy-based content sharing in public clouds," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 11, pp. 2602–2614, 2013.

[15] P. Muthi Reddy, S. H. Manjula, and K. R. Venugopal, "Secured privacy data using multi key encryption in cloud storage," in *in Proceedings of Fifth International Conference on Emerging Applications of Information Technology (EAIT 2018)*. IEEE, 2018.

[16] X. Dai, J. M. Wang, and B. Bensaou, "Energy-efficient virtual machines scheduling in multi-tenant data centers," *IEEE Transactions on Cloud Computing*, vol. 4, no. 2, pp. 210–221, 2016.

[17] S. Bera, S. Misra, and J. J. Rodrigues, "Cloud computing applications for smart grid: A survey," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 5, pp. 1477–1494, 2015.

[18] S. Li, Y. Zhou, L. Jiao, X. Yan, X. Wang, and M. R.-T. Lyu, "Towards operational cost minimization in hybrid clouds for dynamic resource provisioning with delay-aware optimization," *IEEE Transactions on Services Computing*, vol. 8, no. 3, pp. 398–409, 2015.

[19] S. Yang, D. Kwon, H. Yi, Y. Cho, Y. Kwon, and Y. Paek, "Techniques to minimize state transfer costs for dynamic execution offloading in mobile cloud computing," *IEEE Transactions on Mobile Computing*, vol. 13, no. 11, pp. 2648–2660, 2014.

[20] J. Yao, H. Zhou, J. Luo, X. Liu, and H. Guan, "Comic: Cost optimization for internet content multihoming," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 7, pp. 1851–1860, 2015.

[21] A. Trombetta, W. Jiang, E. Bertino, and L. Bossi, "Privacy-preserving updates to anonymous and confidential databases," *IEEE Transactions on Dependable and Secure Computing*, vol. 8, no. 4, pp. 578–587, 2011.

[22] L. Zhou, V. Varadharajan, and M. Hitchens, "Achieving secure role-based access control on encrypted data in cloud storage," *IEEE transactions on information forensics and security*, vol. 8, no. 12, pp. 1947–1960, 2013.

[23] A. C. Zhou, B. He, and C. Liu, "Monetary cost optimizations for hosting workflow-as-a-service in iaas clouds," *IEEE transactions on cloud computing*, vol. 4, no. 1, pp. 34–48, 2016.

[24] J. Yu, F. Kong, H. Zhao, X. Cheng, R. Hao, and X.-F. Guo, "Non-interactive forward-secure threshold signature without random oracles," *Journal of Information Science and Engineering*, vol. 28, no. 3, pp. 571–586, 2012.

[25] M. Bellare and S. K. Miner, "A forward-secure digital signature scheme," in *Annual International Cryptology Conference*. Springer, 1999, pp. 431–448.

[26] P. Muthi Reddy, R. D. Rekha, R. Tanuja, S. H. Manjula, and K. R. Venugopal, "Forward secrecy in authentic and anonymous cloud with time optimization," in *in Proceedings of Fifteenth International Conference on Wireless and Optical Communications Networks (WOCN 2018)*. IEEE, 2018.

[27] A. Ramachandra, K. Pavithra, K. Yashasvini, K. Raja, K. Venugopal, and L. M. Patnaik, "Offline signature authentication using cross-validated graph matching," in *Proceedings of the 2nd Bangalore Annual Compute Conference*. ACM, 2009, p. 7.

Muthi Reddy P is a full time Research Scholar in the Department of Computer Science and Engineering from University Visvesvaraya College of Engineering, Bangalore University, Bengaluru, India. He was awarded Master of Technology in Computer Science and Engineering from Davangere University, Davangere. He obtained his Bachelor of Engineering degree in Computer Science and Engineering from BMS College of Engineering, Visvesvaraya Technological University. He was received the Diploma in Computer Science and Engineering from Govt. Polytechnic, Board of Technical Examinations, Karnataka. His research interests are in the field of Data Security, Data Sharing and IOT in the Cloud Computing.

Rekha Rangappa Dasar is obtained her ME degree from University Visvesvaraya College of Engineering, Bangalore University, Bengaluru, India. She was received the BE degree in Computer Science and Engineering from Basaveshwar Engineering College, Visvesvaraya Technological University. Her areas of interests include Networking, Data processing and Cloud Computing.

Tanuja R is currently the Assistant Professor, Department of Computer Science, University Visvesvaraya College of Engineering, Bangalore University, Bengaluru. She was obtained her Bachelor of Engineering from BMSCE, Bengaluru. She received her Masters degree in Computer Science and Engineering from UVCE, Bengaluru. Her research interests are in the field of Wireless Sensor Networks, Cryptography and Network security.

Manjula S H is currently Associate Professor, Department of Computer Science and Engineering, UVCE, Bangalore University, Bengaluru. She has obtained BE, M.Tech. Ph.D. in Computer Science and Engineering, Chennai. Her research interests are in the field of Wireless Sensor Networks and Data mining.

Venugopal K R is currently the Principal, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. He obtained his Bachelor of Engineering from University Visvesvaraya College of Engineering. He received his Masters degree in Computer Science and Automation from Indian Institute of Science Bangalore. He was awarded Ph.D in Economics from Bangalore University and Ph.D in Computer Science from Indian Institute of Technology, Madras. He has a distinguished academic career and has degrees in Electronics, Economics, Law, Business Finance, Public Relations, Communications, Industrial Relations, Computer Science and Journalism. He has authored and edited 64 books on Computer Science and Economics, which include Petrodollar and the World Economy, C Aptitude, Mastering C, Microprocessor Programming, Mastering C++ and Digital Circuits and Systems etc., He has filed 101 patents. During his three decades of service at UVCE he has over 640 research papers to his credit. His research interests include Computer Networks, Wireless Sensor Networks, Parallel and Distributed Systems, Digital Signal Processing and Data Mining. He is a Fellow of IEEE, ACM and ISTE.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Server Consolidation through Virtual Machine Task Migration to achieve Green Cloud

Geetha Megharaj
Associate Professor
Department of CSE
Sri Krishna Institute of Technology
Bengaluru, India
E-mail: geethagvit@yahoo.com

Mohan G. Kabadi
Professor and HOD Department of CSE
Presidency University
Bengaluru, India
E-mail: kabadimohan60@gmail.com

***Abstract*—Power Consumption in cloud centers is increasing rapidly due to the popularity of Cloud Computing. High power consumption not only leads to high operational cost, it also leads to high carbon emissions which is not environment friendly. Thousands of Physical Machines/Servers inside Cloud Centers are becoming a commonplace. In many instances, some of the Physical Machines might have very few active Virtual Machines, migration of these Virtual Machines, so that, less loaded Physical Machines can be shutdown, which in-turn aids in reduction of consumed power has been extensively studied in the literature. However, recent studies have demonstrated that, migration of Virtual Machines is usually associated with excessive cost and delay. Hence, recently, a new technique in which the load balancing in cloud centers by migrating the extra tasks of overloaded Virtual Machines was proposed. This task migration technique has not been properly studied for its effectiveness w.r.t. Server Consolidation in the literature. In this work, the Virtual Machine task migration technique is extended to address the Server Consolidation issue. Empirical results reveal excellent effectiveness of the proposed technique in reducing the power consumed in Cloud Centers.**

Keywords–Cloud Center; Server Consolidation; Virtual Machines; Task Migration;

I. Introduction

The Cloud Center (CC) is a computational resource repository [1], which provides on-demand computational services to the clients. The computational servers in CC are referred as Physical Machines (PMs). The required services are provided to the client through Virtual Machines (VMs), which abstracts these PMs, and each PM might host multiple VMs.

A. Overview on Server Consolidation

Cloud Computing is becoming widespread due to the reduction of cost and effort in maintaining servers in the client organizations. As more and more operations are migrated to the Cloud, the CCs expand in-terms of PMs, and this expansion leads to significant increase in total power consumption of CCs. In some situations, some of the PMs have limited active VMs, and recently it was demonstrated that [2], even a single active VM can contribute 50% power consumption in the corresponding PM. Veritably, shutting down such lightly loaded PMs by migrating their corresponding VMs can aid power consumption reduction in CCs. The process of running the CC by shutting down lightly loaded PMs is known as *Server Consolidation* (SC)

Currently, many efficient VM migration techniques for SC have been proposed in the literature. VM migration techniques are also used in load balancing inside CCs, wherein, the overloaded VMs are migrated to other PMs, so that, in these new PMs, sufficient resources can be provided for the efficient execution of tasks inside such migrated VMs. However, it was highlighted in [3] that, VM migration has significant drawbacks in achieving efficient load balancing or SC:

1. VM migration requires halting the current functionality of the VM, which is associated with significant memory consumption and task execution downtime.

2. There is chance that, customer activity information can be lost during the VM migration process, and which may increase the monetary expenditure.

3. Significant increase in dirty memory is associated with VM migration.

B. Motivation

In [3], [4], the new/extra tasks for overloaded VMs are migrated instead of migrating the actual VMs to achieve load balancing; however, this migration framework has not been applied to address SC problem. The merits of VM task migration technique obtained for load balancing, also need to be achieved for SC. The current framework of VM task migration presented in [3] requires extensive modifications to make it adaptable for addressing the SC problem.

C. Paper Contributions

The following contributions are made in this paper:

1. A new technique for VM task migration for SC is proposed. This new technique identifies the potential PMs which need to be shutdown. The extra tasks arriving for the VMs present in the potential PMs are migrated to other resourceful PMs, and this migration is achieved through a cost function which utilizes estimated parameters such as–probable task execution time and the cost of task migration. The VMs from which extra tasks are migrated, continue to be active until all the running tasks finish their execution, and then, the corresponding PMs can be shutdown.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

2. The proposed VM task migration technique is simulated using MATLAB. Empirical results demonstrate excellent power consumption reduction achieved by the proposed technique.

The paper is organized as follows: in Section 2, the related work in the area of the addressed problem is described. The proposed VM task migration technique for SC is presented in Section 3. The simulated results and corresponding discussions are presented in Section 4. Finally, the work is concluded with future directions in Section 5.

II. Related Work

Extensive contributions have been made to achieve SC through VM migration technique. Various techniques for SC in virtualized data center has been discussed in [5]. In [6], two VM migration techniques namely–*Hybrid* and *Dynamic Round Robin*(DRR) was presented. Two states were defined in the solution framework called–retiring and non-retiring. If a PM contains limited number of active VMs which are about to finish their task, then, the PM is in retiring state, else, it is in non-retiring state. The retiring PMs will not accept new tasks, and the active VMs are migrated to suitable PMs. Both, Hybrid and DRR exhibit excellent performance w.r.t. reducing power consumption in CCs.

Most of the VM migration techniques for SC are modeled through *Bin Packing Problem* (BPP), which is NP-complete. An approximation scheme based on *First Fit Decreasing* algorithm was proposed in [7] to effectively migrate VMs. Each bin is considered as a PM, and the highest priority PMs are subjected to VM migration.

The *Magnet* scheme proposed in [8], performs selection of suitable subsets of available PMs which can guarantee the expected performance levels. The PMs outside the selected subset are shutdown.

A CC management tool was presented in [9]. This tool not only provides continuous monitoring facility, it also provides facility to perform live migration of VMs.

In [2], it was emphasized that, VMs can be broadly classified as data intensive or CPU intensive based on their respective workloads. For this new framework, the BPP was modified, and suitable approximation schemes were presented.

The placement of migrated VMs for SC was performed through assigning priority levels to the candidate PMs in [10]. The PMs which consume low power were given higher priority.

Non-migratory technique for reduction of power consumption in CCs was presented in [11]. Energy efficiency model and corresponding heuristics were proposed to reduce power consumption in CCs. Similar techniques were presented in [12] which utilized green computing framework.

Resource scheduling techniques for SC were presented in [13]. Here, a new architectural model was presented to calculate energy expenditure for different resource scheduling strategies.

All the described VM migration techniques, even though they achieve noticeable performance in reducing power consumption, they all suffer from excessive down times in completing VM migration, and increase in dirty memory as explained before.

The initial work on VM task migration for load balancing in CCs was proposed in [3], [4], [14]. Different quality parameters such as–task execution time, task transfer cost and task power consumption were utilized in designing the scoring function for task migration. The optimal solution for performing VM task migration was searched through *Particle Swarm Optimization* (PSO) technique. Since, the VM task migration framework proposed in [3], [4], [14] was specifically designed to address load balancing issue, it requires suitable adaptations to address the SC problem.

III. VM Task Migration Technique for SC

The first step in SC is to identify suitable PMs which can be considered for shutting down. Let, PM_k indicate the k^{th} PM in the CC, $num(PM_k)$ indicate the number of active VMs in PM_k. Each PM is defined with a corresponding threshold indicated by $SD(PM_k)$, which indicates the required minimum number of VMs running in the PM to prevent it from shutting down. This case is represented in Equation 1. Here, $shutdown(PM_k) = 1$ indicates that, PM_k should be shutdown, and $shutdown(PM_k) = 0$ indicates that, PM_k should be kept active.

$$shutdown(PM_k) = \begin{cases} 1, & \text{if } num(PM_k) < SD(PM_k) \\ 0, & otherwise \end{cases} \tag{1}$$

A. Task Migration Framework

Let, $\widehat{SD}$ indicate the set of PMs which are eligible to be shutdown, and $\widehat{VM}$ indicate the set of active VMs hosted inside those PMs $\in \widehat{SD}$. The extra or new tasks which are submitted to $\widehat{VM}$ will be migrated to other suitable PMs. Once, the running tasks $\in \widehat{VM}$ finish their execution, all the PMs $\in \widehat{SD}$ can be shutdown.

Let, t_{iy} indicate the i^{th} extra task submitted to $VM_y \in \widehat{VM}$, and suppose it can be migrated to VM_z which is hosted in that PM $\notin \widehat{SD}$. The migration of t_{iy} also requires the migration of data associated with t_{iy}. The merit of this migration is analyzed through a scoring function represented in Equation 2. Here, $score(t_{iy}, VM_z)$ indicates the score of migration strategy which migrates t_{iy} from VM_y to VM_z, exe_{iz} indicates the estimated execution time of t_{iy} inside VM_z, $transfer(t_{iy}, VM_z)$ indicates the task transfer time from VM_y to VM_z, and both these metrics are represented in Equations 3 and 4 respectively. Here, c_z indicates the number of CPU nodes present in VM_z, m_z is the memory capacity of VM_z, d_{iy} indicates the size of data used by t_{iy}, and bw_{yz} indicates the bandwidth available between VM_y and VM_z. The metric formulation represented in Equation 3 is based on the intuition that, increase in data size of a task results in increased execution time, and presence of rich computational resources in VM influences the decrease in task execution

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

time. It is evident from Equation 2 that, higher values of $score(t_{iy}, VM_z)$ indicates unattractive options.

$$score(t_{iy}, VM_z) = exe_{iz} + transfer(t_{iy}, VM_z) \quad (2)$$

$$exe_{iz} = \frac{d_{iy}}{c_z \times m_z} \quad (3)$$

$$transfer(t_{iy}, VM_z) = \frac{d_{iy}}{bw_{yz}} \quad (4)$$

The extra task migration is performed batch-wise, rather than on a single task in-order to reduce computational overheads. All the extra tasks submitted to $\widehat{VM}$ in a specific time interval indicated by I_e are batched together for migration. Consider the scenario, where the batch of extra tasks $[t_{i_1y_1}, t_{i_2y_2}, ...t_{i_sy_s}]$ submitted to $\widehat{VM}$ need to be migrated. Suppose, $[VM_{z_1}, VM_{z_2}, ...VM_{z_s}]$ is a candidate solution for the required migration of tasks, wherein, $t_{i_jy_j}(1 \leq j \leq s)$ is considered to be migrated from VM_{y_j} to VM_{z_j}, and this candidate solution is denoted as S. Also, there is no restriction that, the VMs in the candidate solution should be distinct. The score of this migration scheme is represented in Equation 5.

$$\boxed{migration\_score(S) = \frac{\sum_{j=1}^{s} score(t_{i_jy_j}, VM_{z_j})}{s}} \quad (5)$$

The goal of the task migration scheme is represented in Equation 6, wherein, the most optimal candidate solution has to be discovered. It is evident that, the problem of finding the optimal migration scheme has combinatorial search complexity. To perform efficient search in polynomial search complexity, utilization of meta-heuristic techniques for finding near optimal approximate solutions becomes attractive.

$$\text{optimization condition} = \arg\min_{S} migration\_score(S) \quad (6)$$

B. Algorithm

PSO technique is a meta-heuristic technique which provides an approximate solution to the optimization problems, and it is inspired by the social behavior of birds. The search for optimal solution is carried out by group of particles, wherein, each particle has an exclusive zone in the candidate solution space, and union of all particle zones is equal to the candidate solution space. Each point in the candidate solution space represents a candidate solution vector. The particles are continuously moving in their corresponding candidate solution space to identify the optimal solution, and are involved in continuous communication for exchanging their locally discovered best solution, which in-turn decides the corresponding velocity of the particle for navigation. The particles continue their search until acceptable solution is obtained.

The PSO based solution technique for SC through VM task migration technique utilizes r particles. Here, the current position of the i^{th} particle at iteration t is indicated by $\overrightarrow{X}_i(t)$, and the position for the next iteration is indicated by $\overrightarrow{X}_i(t+1)$, which is calculated as represented in Equation 7. Here, $\overrightarrow{V}_i(t)$ indicates the velocity of i^{th} particle for $t+1$ iteration, and it is calculated as represented in Equation 8. Here, D_1 and D_2 indicate the degree of particle attraction towards individual and group success respectively, $\overrightarrow{x}_{gbest}$ and $\overrightarrow{x}_{pbest_i}$ indicate the global best solution obtained by all the particles until the current iteration and the local best solution obtained by the i^{th} particle until the current iteration respectively, W indicates a control variable, and $r_1, r_2 \in [0, 1]$ indicate the random factors.

$$\overrightarrow{X}_i(t+1) = \overrightarrow{X}_i(t) + \overrightarrow{V}_i(t+1) \quad (7)$$

$$\begin{aligned}\overrightarrow{V}_i(t+1) = W\overrightarrow{V}_i(t) + D_1r_1(\overrightarrow{x}_{pbest_i} - \overrightarrow{X}_i(t)) + \\ D_2r_2(\overrightarrow{x}_{gbest} - \overrightarrow{X}_i(t))\end{aligned} \quad (8)$$

The PSO based solution technique for SC through VM task migration technique is outlined in Algorithm 1. Here, $initialize\_PSO(P)$ divides the candidate solution space among the r search particles indicated by $P = p_1, p_2, ...p_r$, and assigns each particle to some arbitrary positions in their corresponding candidate solution space. Each particle calculates its candidate solution for the corresponding current position through $compute\_score(\overrightarrow{X}_i(t))$, which utilizes Equations 7 and 8. The values for $\overrightarrow{x}_{pbest_i}$ and $\overrightarrow{x}_{gbest}$ are calculated through $local\_best(score_i)$ and $global\_best(P, \overrightarrow{x}_{pbest_i})$ respectively. The particles continue to search until the acceptable solution is found, and which is calculated through $acceptable(\overrightarrow{x}_{gbest})$.

Algorithm 1 PSO Algorithm for SC

```
P = p_1, p_2, ....p_r
initialize_PSO(P)
flag = 0
t = 0
while flag == 0 do
    t = t + 1
    for i = 1 to r do
        score_i = compute_score(X_i(t))
        x_pbest_i = local_best(score_i)
        x_gbest = global_best(P, x_pbest_i)
        if acceptable(x_gbest) then
            flag = 1
        end if
    end for
    t = t + 1
end while
```

C. Simulation Setup

The proposed VM task migration technique for SC is implemented in MATLAB, and for the ease of reference it will be referred as *VMSC*. The corresponding simulation parameter settings are outlined in Table I. Here, the power

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Simulation Parameter	Set value
Number of PMs	Varied between 5×10^3 to 10^4
Number of VMs present in each PM indicated by $tvm(PM_k)$	Varied between 2 to 200 (randomized)
$nvm(PM_k)$	$0.5 \times tvm(PM_k)$
Number of extra tasks for a VM during I_e	Poisson distributed with $\lambda = 5$
Number of computing nodes/CPUs in each VM	Varied between 5 to 20
Main memory capacity for each VM	Varied between $4GB$ / $8GB$ /$16GB$
$\min SD(PM_k)$	Varied between $[5 - 25]$
Bandwidth between any 2 VMs	Varied between $100mbps$ to $500mbps$
Number of PSO search particles	Varied between 5 to 25
Computing nodes used for PSO technique execution	One computing node per particle
Size of task data	Varied between $1GB$ to $10GB$
Power consumed by each VM	Varied between 0 to 1(normalized)

TABLE I
SIMULATION PARAMETER SETTINGS

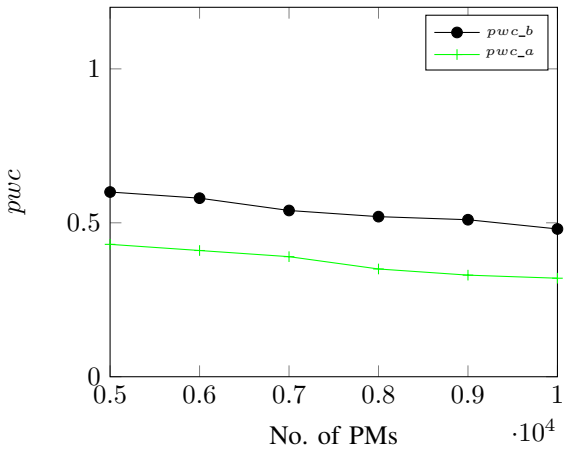


Fig. 1. No of PMs vs pwc

consumption of each VM is normalized to $[0 - 1]$ for the sake of convenience, wherein, 1 indicates the maximum power consumption, and 0 indicates the VM is inactive, also, the number of VMs present in each PM is decided randomly in-order to reflect realism. The effectiveness of *VMSC* is analyzed through two metrics, which are represented in Equations 9 and 10. Here, pwc_b indicates the average power consumption by all the PMs inside the addressed CC indicated by CC_r before *VMSC* is executed, $pwc(PM_k)$ indicates the average power consumed by PM_k, $|CC_r|$ indicates the number of PMs present in CC_r, $\widehat{CC}_r$ indicates CC_r after execution of *VMSC*, pwc_a indicates the average power consumption by all the PMs inside CC_r after the execution of *VMSC*, and $|\widehat{CC}_r| = |CC_r|$.

The metric $pwc(PM_k)$ is calculated as represented in Equation 11. Here, $pwc(VM_j)$ indicates the power consumed by the j^{th} VM, and $|PM_k|$ indicates the number of VMs present in PM_k. It is clear from Equations 9 and 10 that, $0 \leq pwc(a), pwc(b) \leq 1$.

$$\text{pwc\_b} = \frac{\sum_{PM_k \in CC_r} pwc(PM_k)}{|CC_r|} \tag{9}$$

$$\text{pwc\_a} = \frac{\sum_{PM_k \in \widehat{CC}_r} pwc(PM_k)}{|\widehat{CC}_r|} \tag{10}$$

$$pwc(PM_k) = \frac{\sum_{VM_j \in PM_k} pwc(VM_j)}{|PM_k|} \tag{11}$$

IV. EMPIRICAL RESULTS AND DISCUSSIONS

The first experiment evaluates the performance of *VMSC* when the number of PMs are varied. The analysis result w.r.t. pwc and execution time is illustrated in Figures 1 and 2 respectively. Due to the increase in PMs and the random number of VMs present in each PM, the number of PMs suitable for shutdown tends to increase, hence, pwc_b and pwc_a exhibit monotonically non-increasing behavior. The monotonically non-decreasing behavior w.r.t. execution time is majorly due to increase in computational load. It is clear that, *VMSC* provides significant benefits in optimizing power consumption in CCs, and exhibits its merit in identifying the approximate solution in appreciable execution efficiency.

The second experiment analyzes the execution time of *VMSC* when the number of PSO search particles are varied, and the number of PMs is fixed at 10^4, which corresponds to the highest load case utilized in empirical analysis. The analysis result is illustrated in Figure 3. As the number of PSO search particles increase, corresponding increase in parallelism results in better execution efficiency.

The third experiment analyzes the performance of *VMSC* when $\min SD(PM_k)$ is varied. The analysis result w.r.t. pwc_a and execution time is illustrated in Figures 4 and 5. The increase of $\min SD(PM_k)$ creates an opportunity to include more number of PMs for shutdown, which in-turn improves pwc_a, and for the same reason, which also increases the computational load, execution efficiency decreases.

The final experiment analyzes the execution time of *VMSC* when the number of PSO search particles are varied, and $\min SD(PM_k) = 25$. The analysis result is illustrated in Figure **??**. The performance reasoning of *VMSC* is similar to the second experiment.

V. CONCLUSION

In this work, the importance of SC in CCs was described. The drawbacks of VM migration techniques for SC were outlined. A new SC approach using VM task migration concept was presented, which utilized PSO based search technique. Empirical results demonstrated the effectiveness of the proposed technique in reducing power consumption in CCs, and appreciable execution efficiency. In future, design of probabilistic models for SC, which predict the load behavior of PMs can be investigated for implementing effective preemptive actions.

REFERENCES

[1] Zeng Wenying, Zhao Yuelong and Zeng Junwei, "Cloud Service and Service Selection Algorithm Research", *Proceedings of the first*

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

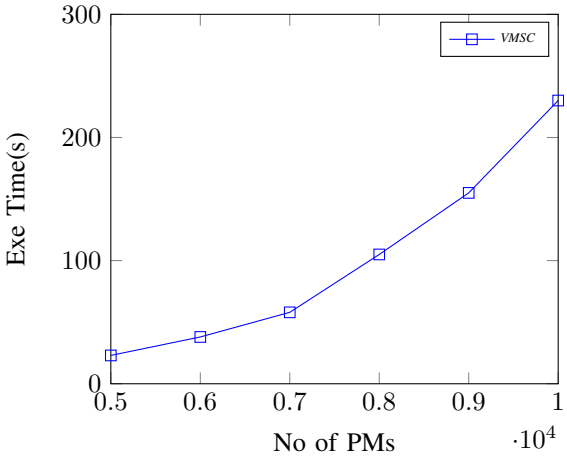


Fig. 2. No of PMs vs Exe Time

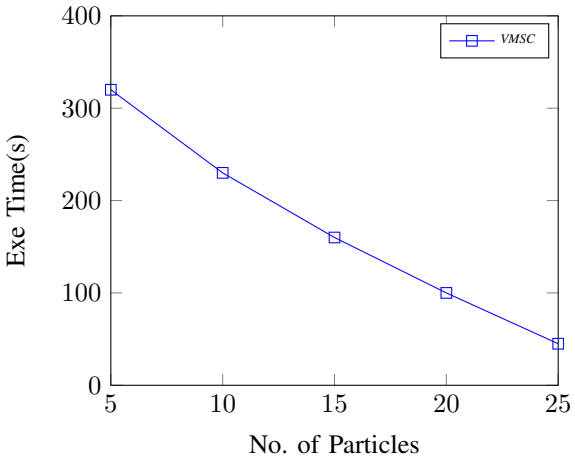


Fig. 3. No of Particles vs Exe Time

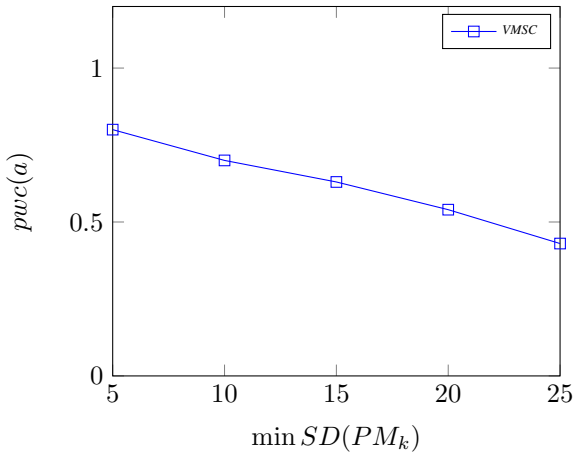


Fig. 4. $\min SD(PM_k)$ vs pwc_a

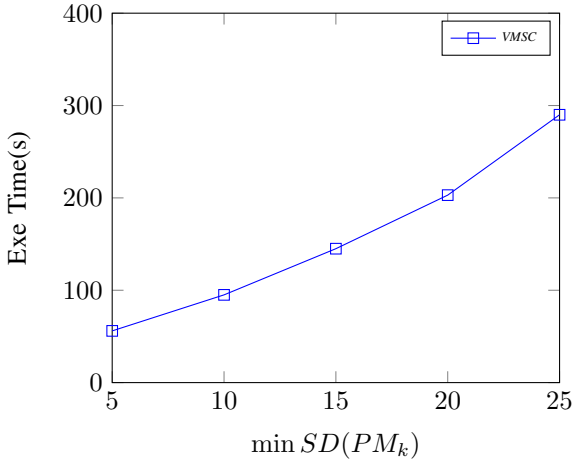


Fig. 5. $\min SD(PM_k)$ vs Exe Time

ACM/SIGEVO Summit on Generic and Evolutionary Computation, GEC 09, 978-1-60558-326-6, Shanghai, China, 1045-1048.

[2] Jyun-Shiung Yang, Pangfeng Liu and Jan-Jan Wu, "Workload Characteristics-aware virtual machine consolidation algorithms", *2012 IEEE 4th International Conference on Cloud Computing Technology and Science*, 978-1-4673-4510-1.

[3] F.Ramezani, J. Lu, and F.K.Hussain, "Task Based System Load Balancing in Cloud Computing Using particle Swarm Optimization", *International journal of parallel programming*, Vol.42, No.5, pp.739-754, 2014.

[4] Geetha Megharaj and Dr. Mohan G.Kabadi, "Metaheuristic Based Virtual Machine Task Migration Technique for Load Balancing in Cloud", *Second International conference on Integrated Intelligent Computing, Communication and Security(ICIIC-2018)*, on 24th,25th January 2018, SJBIT, Bengaluru.

[5] Amir Varasteh and Maziar Goudarzi, "Server Consolidation Techniques in Virtualized Data Centers: A Survey", *IEEE SYSTEMS JOURNAL*, VOL. 11, NO. 2, June 2017.

[6] C.C. Lin, P.Liu and J.J Wu, "Energy-efficient Virtual Machine provision Algorithms for cloud systems", *Utility and Cloud Computing (UCC)*, 2011 Fourth IEEE International Conference, 2011,81-88.

[7] Shingo Takeda and Toshinori Takemura, "A Rank based VM Consolidation Method for Power Saving in Datacenters", *IPSJ Online Transactions 3(2):88-96*, January 2010.

[8] Liting Hu, Hai Jin, Xianjie Xiong and Haikun Liu, "Magnet: A novel scheduling policy for power reduction in cluster with virtual machines", *2008 IEEE International Conference on Cluster Computing*, 13-22.

[9] Liu Liang, Wang Hao, Liu Xue, Jin Xing, He Wen Bo, Wang Qing Bo and Chen Ying, "Green cloud: A new architecture for Green Data Center", *Proceedings of the 6th International Conference Industry Session on Automatic Computing and Communication Industry Session 2009*, 978-1-60558-612-0, 29-38.

[10] Kamali Gupta and Vijay Katiyar, "Energy Aware Virtual Machine Migration Techniques for Cloud Environment", *International Journal of Computer Applications*, Volume 141, May 2016.

[11] Pragya and G.Manjeey, "A Review on Energy Efficient Techniques in Green Cloud Computing", *International Journals of Advanced Research in Computer Science and Software Engineering*, Volume 5, Issue (3), 2015,550-554.

[12] A.Banerjee, P.Agrawal and N.ch S.N. Iyengar, "Energy Efficient Model for Cloud Computing", *International Journal of Energy Information and Communications*, Volume 4,Issue (6), 2013,29-42.

[13] Kamali Gupta and Vijay katiyar, "Energy Aware Scheduling Framework for resource allocation in a virtualized cloud data centre", *International Journal of Engineering and Technology*, DOI: 10.21817/ijet/2017/v9i2/170902032.

[14] F. Ramezani, J.Lu, and F.Hussain, "Task Based System Load Balancing Approach in Cloud Environments", *Knowledge Engineering and Management*, pp.31-42, 2014.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Corelation Analysis Of Forensic Metadata For Digital Evidence

Zaenudin
Departement of Informatics
Universitas Islam Indonesia,
Yogyakarta, Indonesia
STMIK Mataram,
Indonesia
15917124@students.uii.ac.id

Bambang Sugiantoro
Departement of Information System
UIN Sunan Kalijaga
Yogyakarta, Indonesia
bambang.sugiantoro@uin-suka.ac.id

Yudi Prayudi
Departement of Informatics
Universitas Islam Indonesia
Yogyakarta, Indonesia
prayudi@uii.ac.id

***Abstract*— Metadata is the information that is embedded in a file whose contents are the explanation of the file. In the handling of the main evidence with a metadata-based approach is still a lot of manually in search for correlation related files to uncover various cases of computer crime. However, when correlated files are in separate locations (folders) and the number of files will certainly be a formidable challenge for forensic investigators in analyzing the evidence. In this study, we will build a prototype analysis using a metadata-based approach to analyze the correlation of the main proof file with the associated file or deemed relevant in the context of the investigation automatically based on the metadata parameters of Author, Size, File Type and Date. In this research, the related analysis read the characteristics of metadata file that is file type Jpg, Docx, Pdf, Mp3 and Mp4 and analysis of digital evidence correlation by using specified parameters, so it can multiply the findings of evidence and facilitate analysis of digital evidence. In this research, the result of correlation analysis of digital evidence found that using parameter of Author, Size, File Type and Date found less correlated file while using parameter without Size and File Type found more correlated file because of various extension and file size.**

Keywords: ***Metadata, Forensic, Correlation, Digital, Evidence***

I. Introduction

As the heterogeneity of digital evidence in investigation continues to evolve with technological advances, we are faced with newer digital devices, more artifacts and a variety of file formats, these developments bring benefits, while at the same time providing new opportunities for crime in information technology [1]. In many cases, there is a digital evidence that can assist the officer in uncovering a criminal case. One of them through information about the contents of a data or file called file metadata.

Metadata is information that is embedded in a file in the form of annotation of the file. Metadata contains information about the contents of a data used for the purpose of file or data management that later in a database [2]. Metadata is often called "information about information" or "data about data" [2].

So far, investigators of forensic analysis in the handling of major evidence with a metadata-based approach are still manually in search of correlation of related files. However, when correlated files are in separate locations (folders) and the number of files will certainly be a formidable challenge for forensic investigators in analyzing such digital evidence [1].

Metadata-based researches have been conducted, among others, by [3] linking data with other information, the user accessing it, the file directory where it was stored, the last time it was copied, and so on. Subsequent research Conducting analysis to verify metadata associated with images and track using GPS features [7].

To facilitate the process of correlation analysis, In his research build an AssocGEN analysis system using metadata to determine the association between user file artifacts, logs, and disposal of network packets and identify metadata to classify and determine correlations between artifacts and related artifact groups [5]. Forensic metadata has been done by previous research but by building different tools and parameters. Research with metadata-based forensics has been done by [4]. In his research, a forensic metadata system is used to read metadata characteristics in general and look for metadata correlation files with one parameter: file owner, file size, file date and file type. According to [5]. By using forensic metadata tools will greatly facilitate investigators in analyzing the correlation of digital evidence.

So in this study will build a prototype to understand and read the characteristics of metadata in general and detail the specific metadata and identify, analyze the metadata correlation to group related files or relationships that are considered relevant in the context of investigation automatically based on metadata parameters that is Author, Size , File Type and Date. By using some and all parameters that have been determined, so as to multiply the findings of evidence and facilitate analysis of digital evidence. With this research is expected to contribute to forensic analysts in analyzing the correlation of digital evidence with a metadata-based approach.

II. Literature Review

Several previously conducted studies related to forensic metadata serve as a reference in the writing of this research, among others;

In his research build an AssocGEN analysis system using metadata to determine the association between user file artifacts, logs, and disposal of network packets and identify metadata to group and specify correlations between artifacts and related artifact groups [5].

Other studies use various formats and metadata types to validate different types of documents and files that have a

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

number of formats and metadata types, which can be used to find properties of a file, document or activity of a network. In addition, metadata is widely used in any condition, where metadata can provide a variety of evidence between a group of people, as some do not know the type of information stored in their documents [6].

In his research aims to forensic examination of metadata that is linking data with other information, users who access it, file directory where the storage, last copied, and so forth. In a case, Metadata can produce indirect evidence to support evidence [3]. Next research Perform analysis to verify metadata associated with images and track using GPS features based on GPS Height, Latitude GPS, GPS Longitude and GPS position using Geo tagging feature) [7].

Subsequent research analyzed the BitCurator project to develop an extensible strategy for converting and combining digital forensic metadata into the archive metadata scheme and focusing on metadata generated by the open-source Digital Forensic (DFXML) tool [8]. Related research creates a metadata application for reading file metadata in general and can find files based on file correlation with one of the parameters of the file metadata [4].

From the above literature studies, in this study, will build a prototype for understanding and reading metadata characteristics in general and specific metadata detail and identifying, analyzing metadata correlations for grouping related files or relationships deemed relevant in the context of investigation automatically based on metadata parameters ie Author, Size, File Type and Date. By using some and all parameters that have been determined, so as to multiply the findings of evidence and facilitate analysis of digital evidence. With this research is expected to contribute to forensic analysts in analyzing the correlation of digital evidence with a metadata-based approach.

III. BASIC THEORY

A. Tools

The tools used to build forensic metadata are netbeans. Netbeans is a Java-based Integrated Development Environment (IDE) application from Sun Microsystems that runs on swing. Swing is a Java technology for desktop application development that can run on various platforms such as windows, linux, Mac OS X and Solaris. An IDE is a programming scope that is integrated into a software application that provides a Graphic User Interface (GUI), an editor or text code, a compiler and a debugger [9]

B. Classification of Digital Evidence

In the investigation of the evidence is very important for the sustainability of the case being investigated, because with the evidence that will be analyzed to reveal the motives and perpetrators of the crime. Investigators are expected to understand the types of evidence so that at the time of investigation they recognize the priority of priority evidence. There are several similar terms, namely electronic evidence, digital evidence and evidence findings.

Electronic evidence is physical and visually recognizable (computer, hand phone, camera, CD, hard drive, Tablet, CCTV etc.). While digital evidence is evidence that is extracted or recovered from electronic evidence (file, email, sms, image, video, logs, text). Digital Proof of Evidence is a proof taken from electronic evidence conducted analysis of the evidence, type of digital evidence, among others, Email / Email Address, Web History / Cookies, Image File, logical file, Deleted File, Lost File, Slack files, File Logs, Encrypted Files, Steganography files, Office files, Audio Files, Video Files, User ID and Password, Short Message Service (SMS), Multimedia Message Service (MMS), Call Logs.

Findings of evidence is a digital evidence more meaningful as the output analysis obtained by investigators who directly lead to the reconstruction of the case being faced. In this case, digital evidence is information directly related to the data required by the investigator in the investigation process [10].

C. Metadata Concepts

Metadata can be interpreted as "data (spatial) data", containing information about data characteristics and plays an important role in data exchange mechanisms. Through metadata information expected data users can interpret the data in the same way, when users see directly spatial data. The metadata document contains information that describes the characteristics of the data, especially the content, quality, condition, and manner of obtaining it. Metadata is used to perform pertinent spatial data documentation about who, what, when, where, and how spatial data is prepared.

There are several types of metadata files such as Descriptive Metadata is Data that can identify the source of information so that it can be used to facilitate the process of discovery and selection. Coverage included in this data is the author, title, year of publication, subject or keyword headers and other information that the process of filling is the same as the traditional catalog. Administrative Metadata is Data that can not only identify the source of information but also how it is managed. The scope of this data is the same as the descriptive data only with the data maker, the time of manufacture, the file type, other technical data. In addition, this data also contains information about access rights, intellectual property rights, storage and preservation of information resources. Structural Metadata is Data that can make between the related data can be related to each other. More explicitly, this metadata is used to determine the relationship between physical files and pages, pages and chapters and chapters with books as the final product [11].

D. Test Flow Metadata Forensic Systems

In forensic metadata research for the analysis of evidence, correlation includes several stages of testing is the testing phase to read the characteristics of metadata and testing to perform metadata correlation.

a) Metadata File Characteristic Reading Flow

Here is described in detail the steps of use of this application in viewing the characteristics of the metadata file in Figure 2 flowchart below:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

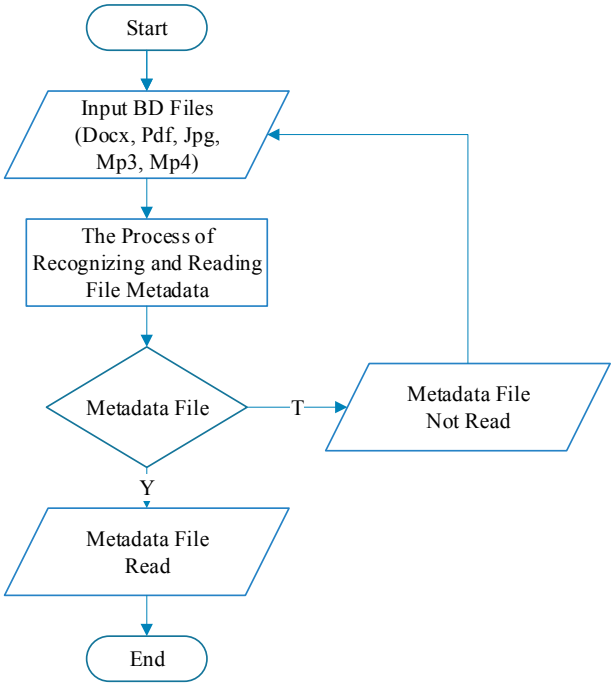


Figure 2. Flowchart Reading Characteristics of Metadata File

Explanation of the testing process to read the characteristics of metadata file using forensic metadata system that is built first to start or forensic metadata system is run, then input file digital evidence that will read metadata, the process of multiplying and reading the metadata file, there are conditions where the metadata file cannot read will return to the input file object evidence, then metadata that can be read metadata will be directly displayed metadata last program in closing or finished.

b) Metadata File Correlation Testing Flow

Here is described in detail the steps of the use of this application program to perform the correlation of the file in figure 3. flowchart below:

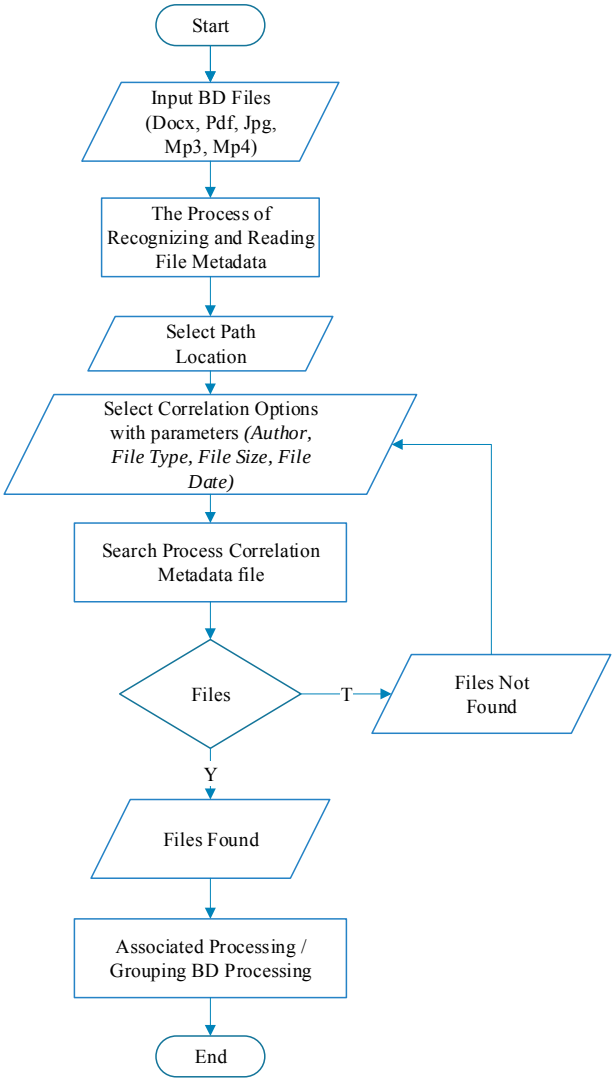


Figure 3. Flowchart Process Testing System / Tools Correlation Metadata file

First start the forensic metadata system, then do input the main evidence file to read metadata, then the process of understanding and reading the metadata file, then select the location of the correlation path and then select the correlation option with parameters, than the system will find the metadata correlation based on parameter selection, if the file is not found it will return to the correlation option but the correlation file found then will proceed to the analysis process and the last system is completed.

IV. RESEARCH METHODS

The method used in forensic metadata research for this correlation analysis of digital evidence can be seen in Figure 1 below:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

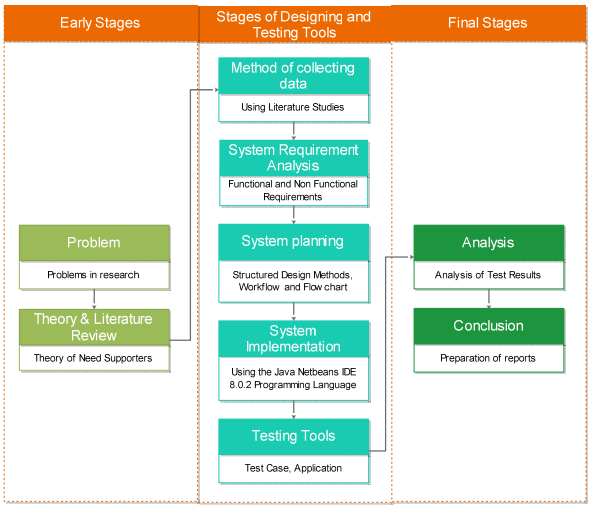


Figure 1. The Proposed Methodology

research methodology that will be built outline is divided into three stages, namely the first stage consists of problem identification and literature review, second stage or stage design and testing tools consist of data collection methods, system requirements analysis, system design, system implementation and testing tools, analysis of test results and the final stage of the completion stage of the conclusion contains the preparation of research reports.

V. ANALYSIS AND RESULT

In this study, the prototype has been built from the implementation until the results of analysis and discussion. Test prototypes built with some predefined files and to analyze the metadata correlation with specified parameters.

A. Results Read File Characteristics File Method

The main evidence file that will read metadata first in browse after the program will process until identified metadata then will appear metadata in general table, checksum and detail as in table 1 below:

Table 1. The result of reading metadata image file TTD.jpg

No	Kind of Metadata	Value
1	Location file	E:\Bahan-Bahan\TTD.jpg
2	Name File	TTD.jpg
3	Type File	Jpg
4	Author	Zen Alkarami
5	Computer	DESKTOP-HJQGNJT
6	Owner	46 DESKTOP-HJQGNJT\Zen

B. Results of File Metadata Correlation Analysis

The result of correlation analysis of metadata file based on parameter ie; Author, Size, File Type and Date. By testing files with extension Jpg, Docx, Pdf, Mp3, and Mp4. In one folder As follows:

a) Correlation Results with Author, Size, File Type and Date Parameters

The result of metadata analysis of correlated file is TTD.jpg file which metadata Author "Zen Alkarami", File Size "327946 byte", file type "Jpg" and with date in file TTD.jpg i.e. "January 24, 2018", conducted file- files are located in the materials folder with the option "equals", then found 2 files that its Author "Zen Alkarami", File size "327946 bytes", Extension file "Jpg" and the date is the same as "January 24, 2018" from metadata the date of the existing TTD.jpg file in that location. The following can be seen in the implementation view in Figure 4 and the results of the analysis from table 4 below:

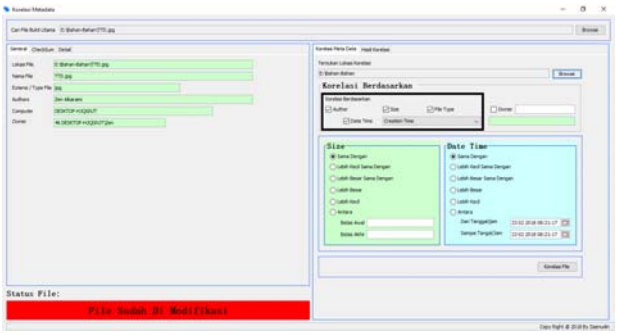


Figure 4. Display of Correlation Implementation with Author, Size, File Type and Date Parameters

Table 4. Correlation Results Based on Author, Size, File Type and Date Parameters

Nama File	Size	Date Creation	Date Modification	Path
gambar.jpg	327946	2018-01-24 04:13:54	2018-01-25 10:51:09	E:\Bahan-Bahan\gambar.jpg
TTD.jpg	327946	2018-01-24 04:13:52	2018-01-24 04:13:54	E:\Bahan-Bahan\TTD.jpg

b) Correlation Results Without Parameters Size and File Type

Results Correlation Analysis Without Parameters Size and File Type in question is to search for various types of files and sizes so obtained correlation results that vary or more with the evidence file TTD.jpg. Then got 6 file result of analysis which metadata Author its "Zen Alkarami", date "24-Januari-2018" with file type in the form of "Mp3, Pdf, Jpg and Docx" and file size different Here can be seen view implementation at Figure 5 and the results in table 5 below:

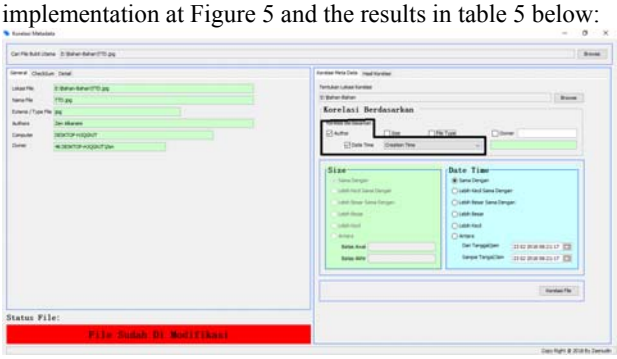


Figure 5. Show Correlation Implementation Without Parameter Size and File Type

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 5. Results Correlation Without Parameters Size and File Type

Nama File	Size	Date Creation	Date Modification	Path
audio.mp3	327946	2018-01-24 04:13:54	2018-01-25 07:03:23	E:\Bahan-Bahan\audio.mp3
Daftar TTD.pdf	6507	2018-01-24 04:17:18	2018-01-24 04:17:17	E:\Bahan-Bahan\Daftar TTD.pdf
format.pdf	327946	2018-01-24 04:13:54	2018-01-25 07:03:23	E:\Bahan-Bahan\format.pdf
Gambar.jpg	327946	2018-01-24 04:13:54	2018-01-25 10:51:09	E:\Bahan-Bahan\Gambar.jpg
Surat Pernyataan.docx	12490	2018-01-24 04:17:00	2018-01-24 04:16:59	E:\Bahan-Bahan\Surat Pernyataan.docx
TTD.jpg	327946	2018-01-24 04:13:52	2018-01-24 04:13:54	E:\Bahan-Bahan\TTD.jpg

VI. CONCLUSION

Based on the results obtained in the discussion, the forensic metadata research for the correlation analysis of digital evidence can be deduced as follows. Built-in forensic metadata can read all file types specifically on the computer both in general and in detail including the tested file as sample. Based on the test to read the characteristics of metadata can be understood in general that is divided into three main parts; General Metadata ie File location, File name, File type / Extension file, Outhors Owner and Computer. Metadata Checksum is MD5 and SHA-256 Value. Metadata detail is cration time, last access time, last modified time, directory, other, regular file symbolic link, size, Make, Model, Orientation, X Resolution, Y Resolution, Resolution Unit, Software, Date / Time, Positioning, Exposure Time, F-Number, Exposure Program and so on. The method used to find metadata and metadata correlation characteristics is by forensic metadata tools. Tools used are the work of the researchers themselves. Based on the test of metadata correlation analysis with parameter of Author, Size, File Type, and Date then found fewer file compare to without parameter size and file type hence found file with various extension and file size.

VII. FUTURE WORK

The suggestions that need to be developed for further research are as follows. In the next research need to be done correlation analysis not only with parameter of metadata. Further development and research needs to be added multi local or multi drive option to browse the main evidence file.

REFERENCES

[1] S. Raghavan and S. V. Raghavan, 2014. “AssocGEN: Engine for analyzing metadata based associations in digital evidence,” Int. Work. Syst. Approaches Digit. Forensics Eng., SADFE,

[2] J.Riley, 2017 Understanding Metadata: What Is Metadata, and What is it for?.

[3] A. Spore, 2016.“Report Information from ProQuest,” no. June,

[4] Subli, Sugiantoro & Prayudi, 2017. “ Forensic Metadata to support the investigation process of the "scientific journal DASI

[5] S. Raghavan and S. V Raghavan, 2013. “A study of forensic & analysis tools,” 2013 8th Int. Work. Syst. Approaches to Digit. Forensics Eng., pp. 1–5,

[6] F. Alanazi and A. Jones, “The Value of Metadata in Digital Forensics,” Proc. - 2015 Eur. Intell. Secur. Informatics Conf. EISIC 2015, vol. 8, no. 2011, p. 182,

[7] P. R. Kumar, C. Srikanth, and K. L. Sailaja, 2016. “Location Identification of the Individual based on Image Metadata,” Procedia Comput. Sci., vol. 85, no. Cms, pp. 451–454, 2016.

[8] L. Drive, M. Hall, C. Hill, K. Woods, A. Chassanoff, and C. a Lee, 2013. “Managing and Transforming Digital Forensics Metadata for Digital Collections,” 10th Int. Conf. Preserv. Digit. Objects, no. November, pp. 203–208,

[9] R. Sharma and S. Koshy, 2011. “Promoting Open Source Technology in Education : NetBeans : The Perfect Open Source IDE,” vol. 4333, pp. 571–575,

[10] Y. Prayudi, 2014 “Problema Dan Solusi Digital Chain Of Custody Dalam Proses Investigasi,”April,

[11] U. Salama, V. Varadharajan, M. Hitchens, and DUMMY, 2012. “Metadata Based Forensic Analysis of Digital Information in the Web,” Annu. Symp. Inf. Assur. Secur. Knowl. Manag., pp. 9–15,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Hybrid Algorithm for Enhancing and Increasing Image Compression Based on Image Processing Techniques

Khalil Ibrahim Mohammad Abuzanouneh
IT Department, College of Computer, Al Qassim University,
Al Qassim, Saudi Arabia,
ka.abuzanouneh@qu.edu.sa

***Abstract*—The data compression and decompression play a very important role and are necessary to minimize the storage media and increase the data transmission in the communication channel, the images quality based on the evaluating and analyzing different image compression techniques applying hybrid algorithm is the important new approach. The paper uses the hybrid technique applied to images sets for enhancing and increasing image compression, and also including different advantages such as minimizing the graphics file size with keeping the image quality in high level. In this concept, the hybrid image compression algorithm (HCIA) is used as one integrated compression system, HCIA has a new technique and proven itself on the different types of file images. The compression effectiveness is affected by the quality of image sensitive, and the image compression process involves the identification and removal of redundant pixels and unnecessary elements of the source image.**

The proposed algorithm is a new approach to compute and present the high image quality to get maximization compression [1].

In This research can be generated more space consumption and computation for compression rate without degrading the quality of the image, the results of the experiment show that the improvement and accuracy can be achieved by using hybrid compression algorithm. A hybrid algorithm has been implemented to compress and decompress the given images using hybrid techniques in java package software.

***Index Terms*—Lossless Based Image Compression, Redundancy, Compression Technique, Compression Ratio, Compression Time.**

Keywords

Data Compression, Hybrid Image Compression Algorithm, Image Processing **T**echniques.

I. INTRODUCTION

Data compression processes are one of the important topics in the nowadays, and the image processing has become a ground of it research today. There are different types of data must be stored in data warehouses, archives, and they must be transmitted through communication channels, and therefore several of data compression algorithms were designed for image processing[2].

The main compression techniques are a lossy and lossless compression. The lossless compression is applied when the file information has to be uncompressed as the same it was before compression. The files can be stored using the lossless compression technique, losing any data or character could be made the data misleading the in the worst case. So there are limits to the amount of space saving that can be gotten with lossless compression. In general, the ratios of Lossless compression the range from 20% to 60%, while the lossy compression process on the image file doesn't have to be stored completely [3]. Based on the lossless method a lot of bits can be thrown away from some images, such as audio data and video when the uncompressed process is done, and the total of data can be acceptable quality.

Compression image is an encoding rules process for decreasing the number in the original image to store or transmit it, the image compression can be identified removed unnecessary pixels of the source image, by reducing the memory size that needed for keeping the image high quality.

The lossy compression can be more meaningful when the compressed images have high quality and in the general can be satisfactory in the most cases [4].

The goal of the image compression process is to get the minimum number of bits for storing process and transmission. In the final experiments, the data encoding could possibly reach a 30-80% reduction in the size of data.

II. IMAGE PROCESSING AND COMPRESSION

Image processing is the image operation into a collection of pixels connected together, and it also the most significant task in the image compression to get better image analysis, the original image can be formed into different sizes and pieces, so the most important task in the image compression is to explore and apply the appropriate algorithms and parameters selection [5].

Any Image is characterized by a set of pixels, and in the images, there are a lot of common parts existing with the neighboring pixels and correlated together include a lot of redundant pixels. The two supporting components of compression are irrelevancy reduction and redundancy reduction.

1. Redundancy Reduction: The property in an image due to redundant bits is called redundancy. It means the data duplication in the images.

Data elimination is called redundant data reduction, it helps to get a minimal storages spaces and results in the image compression.

The image compression can apply a set of methods for reducing the total number of bits required for representing an image achieved by eliminating the difference of redundancy pixels existing in the pixels of an image[6]. There are three

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

basic types of redundancies in the digital image as the following:

a. Psycho Visual Redundancy is a type of redundancy that correlates with different sensitivities to whole image signals using human eyes. So the eliminating process for some pixels in the image processing is acceptable. Therefore, the deleting of some bits which less relative important colors of the human vision can be acceptable

b. Inter-pixel Redundancy is a redundancy corresponding to statistical dependencies among pixels, especially between neighboring pixels, the information related to individual pixels is comparatively small, the neighboring pixels of image are not independent, due to the correlation between the images neighboring pixels, the dependencies between values of pixels in the image is called inter-pixel redundancy or spatial redundancy [7].

c. Coding- Redundancy: The uncompressed image generally is coded each pixel in a fixed length, and is implemented using lookup tables to be reversible in the compression process. The uncompressed image is coded with each pixel, Hoffman code technique is the main algorithm to found coding redundancies.

2. Irrelevancy Reduction– In the image compression can be applied Irrelevancy technique with the actual information that is removed to reduce the required bits for compression image process, the elimination of irrelevant bits will be lost the information that cannot be recovered. In order to justify this, the removal of information which is least perceivable by the human visual system is performed. Irrelevance reduction bits are used in lossy compression.

The success compression of recognition original image is related to the edges marked quality.In this research will be investigated and evaluated edge detection techniques and active contour model to enhance and detect the image color at different levels. there are several algorithms are applied such as Prewitt, segmentation algorithm, and Canny edge detection based on comparison criteria with artificially generated images, the edge quality, and map quality are very important parameters in this stage [8]. The experiments results are explained that apply to these criteria could be utilized for further analysis and find the best edge detector in the compression image.

There are different types of image compression based on segmentation algorithm are:

1. The regions segmentation compression (RSC) can be used to cover the image coordinates.

2. The linear structures segmentations compression (LSSC) that including line segments and curve segments used active contour model.

3. The two-dimensional shapes segmentations compression (2DSSC), such as ellipses, circles, and strips (regions, long, symmetric), the cluster pixels inside salient image boundaries, the regions corresponding to objects surfaces, or objects natural parts.

The image segmentation compression (ISC) can be used in different felids such as the image recognition that is using for face recognition, the medical image such as diagnosis operations, locating diseases and other dangers pathologies.

For the video, image systems can be applied ISC in the traffic control system that is focusing to identify the shapes, sizes, and moving scene objects. The video image compression is divided into two approaches of segmentation: the region based compression, and boundary based compression, in the first segmentation, the purpose is to determine when a pixel belongs to an object or not [96], in the second segmentation the goal is to locate the boundary curves between the background and the objects.

The region segmentation algorithms can be applied as the following:

a) The thresholding method technique of region-based segmentation can be used to segment the original image for the objects separating from the background using a colors features values to compare with a threshold values in order to extract the color pixels class, a method starts from the first one pixel of a potential region and expands by adding adjacent pixels for an image that contains different types regions, the image should be segmented based on the image different areas which each area has a features values range, the thresholds are important to select a colors features values of the image regions to be very useful and effective in the images segmentation quality and compression process [10] , after this stage, the statistical test can be used to take a set of decisions related to which the pixels can be deleted into a region segmentation or not for increasing the image compression ratio.

b) Clustering-based colors segmentation technique

Any image can be divided into different classes or the same type of classes, the redundancy pixels of the image colors should be collected together in the similar classes for building the compression algorithm, and the different colors that contain a different type of pixels will be in different classes.

c) Edge-based colors segmentation technique is the main features technique of the colors image, which includes valuable pixels in the image analysis and diagnosis classification to explore the boundaries detection between the various region's colors by using the selected features of the pixels values as textures, and intensities of the image colors.

III. LOSSLESS METHOD OF IMAGE COMPRESSION

Lossless methods are usually having two stages of algorithms operations. The first stage is transforming the source image into another format for reducing the redundancy colors. In the second stage can be used an entropy encoder for removing the coding redundancy. The lossless decompressors are strong inverse processes of the lossless compressors [11].

In the medical images can be used lossless compressors methods to get more than 50 % of original image size.

While can be applied entropy methods in the lossless compression with several an application to compute MSE (mean square error) and PSNR (peak signal to noise ratio) between images and digitized radiographs, X-rays, and gamma rays to found a bit rate from 4 to 5 bpp (bytes/pixel).

In the Lossless compression can be applied various methods such as linear transformation, multiresolution methods, and investigated prediction in the decorrelation medical images before the coding stage, and getting best results.

The compression results were 3:1 for angiograms images, a less than 2:1 for MRI images.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The interpolation techniques and found-linear prediction interpolation techniques can be given good results in the compression ratios [12].

IV. IMAGE SEGMENTATION USING HICA ALGORITHM

The parameters selection and different types of algorithms will be applied to enhance the images segmentation and improve outputs file. The pixel scale and level of segmentation implement based on HICA algorithm to complete region labeling tasks of the image segmentation processes, the HICA algorithm should be used the image adaptive segmentation included the following steps [11]:

1. Compute the image statistics tables give us the probability for a given degree of a confidence level and identically distributed normally colors to select suitable threshold.
2. The segmentation process will generate initial values for image pixels.
3. Compute the segmentation of image pixels based on quality measures to satisfy conditions of the segmentation function.
4. The image segment should be used new parameters to calculate and keep the image segmentation quality.
5. The analyzing and modifying process based on the knowledge structures of the new image will be implemented by calculating MSE and PSNR for each image.

V. IMAGE COMPRESSION AND CHROMATIC FEATURES

In the research paper will apply HICA algorithm using chromatic features to determine and describe RGB colors distribution and the grey-level of an image, which are the most discriminative features of the image compression. the image pixels are represented a segmented object, The selection parameters are used to detect the edge of image boundaries that have the same colors of pixels from the current image that will be used extracted chromatics features.

The convergence process can be completed and achieved in under the number iterations required to detect the chromatic features and complete colors counted for the image compression.In the next step, the solutions have represented the intensity of colors pixels and chromatic features which can be detected and computed using the Hybrid algorithm of image compression. In this stage of research will improve its searching capacity for the image process environment. The image process is a stochastic process where pixels values are modeled as random variables, for calculating the probability density of grey level and color distribution as its image compression [13].

In this stage of image processing can get robust convergence as building simulations for image compression as possible with the reliable and high convergence of the original image.

the compression hybrid algorithm efficiently improves the performance compression in the image processing environment, and the best selection individual of color pixels based on features function to finding probability density of grey level, colors gradient, colors distribution, pixels color and boundaries shapes into the original image. The selection operators of HCA will be selected set of colors pixels to be the best solutions that have a better classification, based on a features function for reconstruction new image [14].

VI. PROPOSED SYSTEM

The proposed analysis system (HICA) of the image compression process will be explained the phases are shown in figure 1. The Transform process is applied for improving contrast variation and luminance in the original images. In the second phase, segmentation processes are applied and implemented to explore and isolate the interest pixels colors and remove noises before the image compression. The third phase goal is to extract the image characters to be used in the next phase of the compression process; the Features selection method is applied to decrease the redundant pixels and built a classification of the new Image. The selected features are selected for input to the classification method and take the decision about the class assignment by using the hybrid algorithm as shown in figure 1, the goal of the segmentation process in image compression is to separate colors from another different ingredient of the image[15].

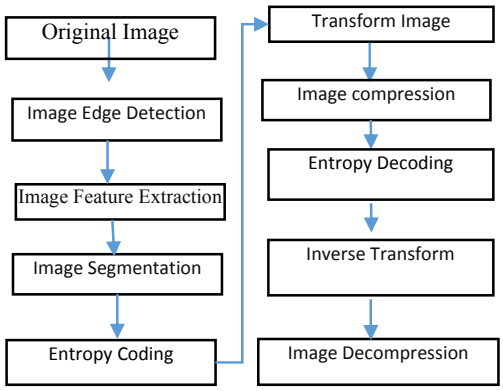


Fig 1: Illustrates the block diagram of the HICAs proposed a system based on Lossless image compression.

VII. ACTIVE CONTOUR AND IMAGE SEGMENTATION MODELS

In the last years, there are recent developments in the image compression fields with a new techniques research for improving features analysis of image compression images. the techniques have been developed to identify specific structures in image colors. The active contour one of the main methods can be adaptable for the required and desired features in the image compression, an active contour is using to delineate an object outline of a possibly noisy for 2D images, and can be applied in several fields such as shape recognition, object tracking, edge detection and segmentation image[15].

In this research, experiments have a set of forms and several types of images. The selection process of an appropriate method for image variable colors and segmenting for specific type image have been always and forever a challenge to select the algorithm of the image compression. In the active contour method there are a lot of enhancements and implemented in

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the image colors segmentation, in some images should be applied active contour method in the changeable curves and change forms to avoid distort colors of image boundaries in the segmentation process[16].

The active contour models can be moved based on internal or external forces extracted based on different characteristics of the image colors. The active contour adaptation occurs in response to both internal and external forces, the external forces model has described the grayscale level gradient, the active contour models can be divided into two types: the parametric models like the Snakes model, which defines a flexible contour that can dynamically adapt to required edges of the image colors, and the geometric models, such as the Level Set model it embeds the front to be zero level set in the higher dimensional function, to calculate the new function evolution, this evolution operation is dependent on the image characteristics extracted and geometric restrictions of the function, the segmentations of image colors are implemented on sub-images, the parametric active contour model is a curve x(s) defined in Equation.1 [17], to be moved in the image spatial domain to minimize the time and energy function Emin(t,s) is defined in Eq.1, and therefore the ccompression time will be decreased as possible is defined by Eq.1.

$$v(s) = [x(s) \quad , \quad y(s)] \quad , \qquad s \in \{1,0\} \tag{1}$$

$$E(s) = \int_0^1 \left[\frac{1}{2}(a|x'(s)|^2)\right| + B(|x''(s)|^2) + E_x(x(s))\,\Big|\,\Big] ds \tag{2}$$

Where x'(s), x″(s) use to found the first and second derivative, of x(s) respectively. α, β donate the weighting parameters of active contour model. Ext is the function of external energy which is derived from the image to take smaller values of boundaries features [18].

Energy Surface and Optimum Thresholding is the basics approach to image segmentation is an amplitude thresholding, a threshold T is chosen to separate the two regions modes, the image point for I(x,y) >T is considered as object points[19], otherwise, the point is called a background point. The threshold method is defined by Eq.3.

$$g(x,y) = \begin{cases} 1 \ , & I(x,y) \geq T \\ 0 \ , & I(x,y) < T \end{cases} \tag{3}$$

Where T is set on the entire image basis I(x, y), and the threshold is global. When T depends on spatial coordinates x and y, based on a dynamic threshold, when T depends on both I(x,y) and set property p(x,y) of local image[19], the average of gray level in a neighborhood centered on I(x,y), the threshold will be local and T is set according to a fitness function is defined by Eq.4. :

$$f(y,x) = T\,[p(x,y), I(x,y)] \tag{4}$$

The object locating will be described in the image I[x,y], using a template T[x,y], The best match Searching to minimize the mean squared errors (MSE) is written below:

$$\boldsymbol{E[p,q]} = \sum_{x=-\infty} \sum_{y=-\infty} \left[\boldsymbol{I}\,[x,y] - \boldsymbol{T}[(x-p),(y-q)]\right]^2 \tag{5}$$

VIII. COMPUTE MSE, PSNR, AND ENTROPY

A. Mean Square Error (MSE)

In this part of the research, the MSE will be computed MSE between the source image and the compressed image. The MSE lower values mean minimum error as seen in the inverse relation among the PSNR and MSE.To find the PSNR, in the first, it should calculate the mean-squared error is defined as.

$$MSE = \sum_{x,y} \frac{|I_1(x,y) - I_2(x,y)|^2}{x.y} \tag{6}$$

In this equation, x, y denote the rows number and columns in the original images, respectively.

B. Peak Signal-to-Noise Ratio (PSNR)

The PSNR block computes PSNR between two images. The ratio will be used as a quality measurement of the source image and a compressed image. PSNR will be used as a quality measure of the image reconstruction, the higher PSNR should be evaluated the higher quality reconstruction.

The PSNR is computed as in the following equation:

$$PSNR = 10log_{10}\left(\frac{R^2}{MSE}\right) \tag{7}$$

In the above equation, where R denotes the maximum range in the data type of original image. For example, the data type has an 8-bit integer of the image, so R = 255.

C. Image Entropy

Entropy image encoding is a lossless compression that can be implemented on an image color after the quantization process to represent image colors in a more efficient with minimum memory for saving stage or transmission. In the research paper will be applied the entropy encoding to increase the image compression ratio.

In this paper will use the entropy encoding to evaluate and describe the image quantity, the pixels amount that should be coded by a compression algorithm. The low entropy images which containing a black sky, and has little bit contrast with pixels large run with similar digital numbers. In the image is perfectly flat, the entropy will equal zero. So, the image can be compressed to a small size.

In the terrain images that have a lot of difference in a contrast from one pixel to another, it has very high entropy and cannot be compressed as much as the low entropy of image.

In the image, colors have corresponded to the gray levels and the individual pixels can be adopted. In the image pixels that have been perfectly histogramming equalized, the pixels should be equally occupied in all states of pixels, and the spread of pixels is a maximum for the image entropy. On the other hand, the image colors which have been a threshold, there are two states are occupied and the entropy will be very low.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The image entropy is equal zero when the pixels have the same values. in this research progression, we can note that the image entropy is decreased, in this case, the image colors should be moved from a full grayscale image to a threshold binary image (high entropy to zero entropy). So the compression ratio will be increased.

All changes in the image colors are meaningful, the Changes of image pixels that due to noise to be part of an image, and it represents the image as more than required. the heat and Noise can be played the same roles in the entropy increased of Images. The image entropy H is defined as.

$$Entropy = -\sum_{i=0}^{M-1} Pi\ log_2(Pi) \qquad (8)$$

Where M is equal the gray levels number, while pi is the correlated probability with gray level i.
Maximum entropy is achieved in the case of a uniform probability distribution, where M=2n and Pi is constant, given by

$$Pi = -\frac{1}{2^n} \qquad (9)$$

The maximum entropy is found from

$$Entropy_{max} = \sum_{i=0}^{M-1} \left(\frac{1}{M}\right) log_2\left(\frac{1}{M}\right) \qquad (10)$$

The image minimum entropy will be achieved if the image is not a variable, and all pixels have identical gray level i. For where the gray level pi = 1, and H = - log (1) = 0. The image entropy can restrict the lower bound on the bits average number per pixel for an image encoding without distortion and can apply to uncorrelated images is shown in figure 2.

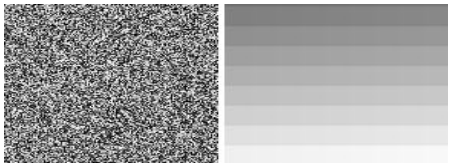


Fig 2. Illustrates a two image, the right has random noise; the right image has the same gray levels distribution.

The left image has random noise and its entropy contains 8 bits and is uncompressible. The right image has the same gray levels distribution with strongly spatially correlated.

IX. Experiments and Results

The proposed algorithm is applied to various types of Images such as medical images, classical images, commercial digital camera images etc. The experiments results show that the proposed algorithm has better efficiency and activity in the images compression from other algorithms; the results of the experiment were implemented in the Java application for image compression algorithms Package as illustrated above [1].
In the first experiment of the HICA proposed, the results quality and the compression ratio among the compressed image and the uncompressed image should be taken into consideration.
The final size of compressed images is compared with the original images between Huffman Algorithm and HICA. Also, the compression time increases as the original size increases, and in contrast, the compression ratio decreases as the original file size increases. The algorithm gives a good compression ratio that lies between 30% and 58%. The results experiments are explained in table 1.

TABLE1.
Shows results analysis between Huffman Algorithm and HICA for compression images and space saving.

Images	FileSize (bits)	Huffman Algorithm		HCIAs	
		Compression Size(bits)	SpaceSaving (Huffman)%	Compression Size(bits)	SpaceSaving (HCIAs)%
ImgA	37492	26591	0.29	24174	0.36
ImgB	27819	12716	0.54	11560	0.58
ImgC	32591	23645	0.27	21495	0.34
ImgD	24088	17409	0.28	15826	0.34
ImgE	16564	11055	0.33	10050	0.39
ImgF	13190	8885	0.33	8077	0.39
ImgG	31902	22909	0.28	20826	0.35
ImgH	22081	15040	0.32	13673	0.38
ImgI	42365	32124	0.24	29204	0.31
ImgJ	28177	18806	0.33	17096	0.39
ImgK	9070	6095	0.33	5541	0.39
ImgL	23524	18592	0.21	16902	0.28
ImgM	20183	13108	0.35	11916	0.41
ImgN	27138	19693	0.27	17903	0.34
ImgO	50885	26701	0.48	24274	0.52
ImgP	20033	14062	0.30	12784	0.36
ImgQ	22384	15728	0.30	14298	0.36

In the second experiment, we compared the results analysis of the space saving and compression images between LZW Algorithm and HICA. The results obtained after comparing and it was in an acceptable range, we can observe that the HICAs algorithm performs in an efficient way and gives better results is shown in table 2.

TABLE 2.
Shows results experiments between LZW Algorithm and HICA of space saving for compression images is shown in table 2.

No	FileSize	LZW Algorithm		HCIAs	
		Compression Size(bits)	SpaceSaving (LZW)%	Compression Size(bits)	SpaceSaving (HCIAs)%
ImgA	37492	29009	0.23	24174	0.36
ImgB	27819	13872	0.50	11560	0.58
ImgC	32591	25794	0.21	21495	0.34
ImgD	24088	18991	0.21	15826	0.34

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

ImgE	16564	12060	0.27	10050	0.39
ImgF	13190	9692	0.27	8077	0.39
ImgG	31902	24991	0.22	20826	0.35
ImgH	22081	16408	0.26	13673	0.38
ImgI	42365	35045	0.17	29204	0.31
ImgJ	28177	20515	0.27	17096	0.39
ImgK	9070	6649	0.27	5541	0.39
ImgL	23524	20282	0.14	16902	0.28
ImgM	20183	14299	0.29	11916	0.41
ImgN	27138	21484	0.21	17903	0.34
ImgO	50885	29129	0.43	24274	0.52
ImgP	20033	15341	0.23	12784	0.36
ImgQ	22384	17158	0.23	14298	0.36

In the third experiment, we calculated and estimated the space saving for HICA, LZW, and Huffman algorithm side by side the results have been increased estimated the space saving and getting better space saving after improving and minimizing the values of the image pixels in the transform image stage using HICA, the probability increment of an image pixels gives more flexibility and increases the code word and the space saving. The experiment results were satisfied and get good results is illustrated in Figure 3.

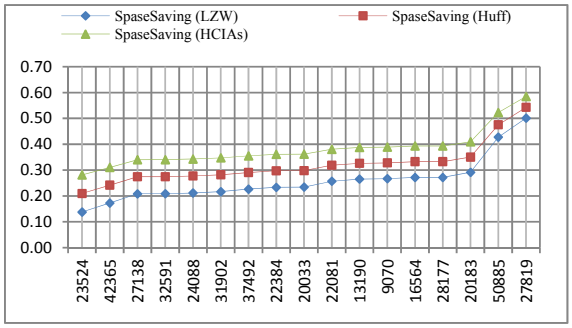


Fig 3. Illustrates a set of experiment analysis for comparing and estimating the space saving for HICA, LZW, and Huffman algorithm in the different types images and sizes.

In the fourth experiment, we calculated and compared the data compression ratio between the uncompressed and compressed of different files size. The compression ratio values were between 1.39 to 2.41. The values obtained after comparing are in an acceptable range. The experiment of the compression ratio was satisfied the results are illustrated in Table 3.

TABLE 3.
Shows results experiments of the data compression ratio between different files size for HICA algorithm is shown in table 3.

	Hybrid Image Compression Algorithms (HICAs)			
FileSize	CompSize	CompRatio	CompTime	DecompTime
23524	16902	1.39	2.8	3.20
42365	29204	1.45	2.9	3.34
27138	17903	1.52	3.0	3.49
32591	21495	1.52	3.0	3.49
24088	15826	1.52	3.0	3.50
31902	20826	1.53	3.1	3.52
37492	24174	1.55	3.1	3.57
22384	14298	1.57	3.1	3.60
20033	12784	1.57	3.1	3.60
22081	13673	1.61	3.2	3.71
13190	8077	1.63	3.3	3.76
9070	5541	1.64	3.3	3.76
16564	10050	1.65	3.3	3.79
28177	17096	1.65	3.3	3.79
20183	11916	1.69	3.4	3.90
50885	24274	2.10	4.2	4.82
27819	11560	2.41	4.8	5.53

In the fourth experiment, we compared between HICA, LZW, and Huffman algorithm after calculating the compression ratio in the different sizes of images files, there was effectiveness in this experiments of HICA algorithm, The values obtained after comparing are in an acceptable range, the results analysis is in an acceptable range.

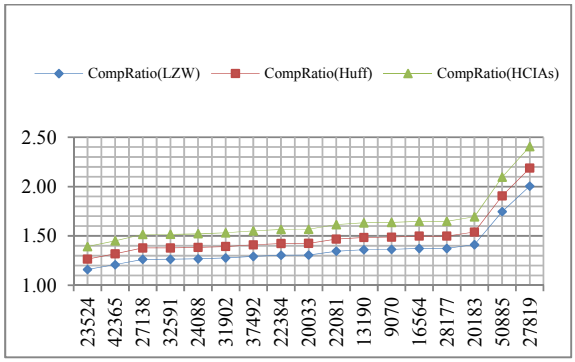


Fig 4. Illustrates a set of experiment analysis for comparing and estimating the space saving for HICAs, LZW, and Huffman algorithm in the different types images and sizes.

In the fifth experiment, we calculated the compression times for HICAs, we applied Feature Extraction method to enhance time ratio in the image compression, the results experiment of CPU time utilized in a compression and decompression was satisfied and in an acceptable range from 2.8 to 4.8 is illustrated in Figure.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

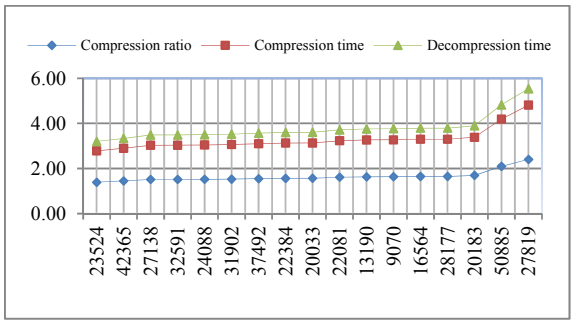


Figure 5. The comparison image based on CPU Time utilized using feature extraction method to enhance time ratio in the compression and decompression of the image colors.

In the sixth experiment, the compression ratio will be calculated to define the total of bits number to extract size of the original image, and compared to the bits number to represent the compressed image size, and display how much time an image can be compressed, and determines the distortion process in the image will be compared with the source image . The matrices quality measurement can be implemented used the following tools measures is illustrated in Table 4.

TABLE 4.
Shows the results analysis are calculated using the quality measurement matrices of the MSE, PSNR, and Entropy method.

No	MSE	PSNR	Entropy
Image1	19.594	35.210	7.915
Image2	39.188	32.199	7.437
Image3	58.782	30.438	7.765
Image4	78.376	29.189	5.727
Image5	97.970	28.220	7.680
Image6	117.564	27.428	6.621
Image7	137.158	26.759	7.449
Image8	156.752	26.179	6.928
Image9	176.346	25.667	7.610
Image10	195.940	25.210	7.751
Image11	215.533	24.796	7.549
Image12	235.127	24.418	6.026
Image13	254.721	24.070	6.884
Image14	274.315	23.748	7.246
Image15	293.909	23.449	6.915
Image16	313.503	23.168	7.611
Image17	333.097	22.905	4.404

In the sixth experiment, the PSNR will be used as a measure of the quality of the image reconstruction. In the higher PSNR

Should be indicated the reconstruction of a higher quality of image colors is illustrated in Figure 6.

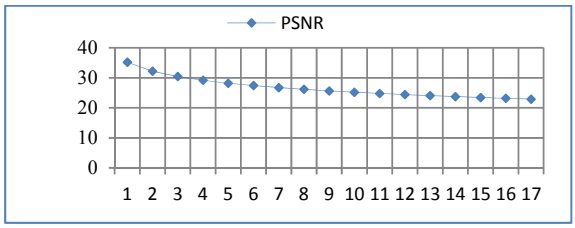


Figure. Illustrates a set of experiment analysis of the Peak Signal To Noise Ratio in the different types and sizes of images.

In the seventh experiment, we calculated the MSE of cumulative-squared error between the compressed image and the original source as shown in figure 7.

The MSE lower values mean minimum error as seen in the inverse relation among the PSNR and MSE as illustrated in figure 7.

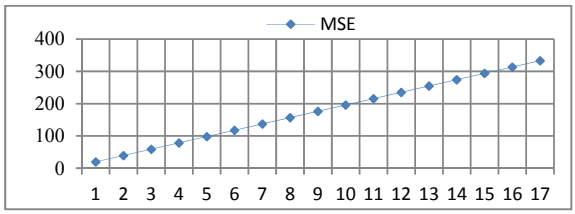


Figure 7. Illustrates a set of experiment analysis of Mean Square Error in the different types and sizes of images.

In the eighth experiment, the final results experiments of the Entropy was satisfied the probability of the image pixels after using active contour model to minimize energy function for image edge detection and boundaries, the entropy method in compression image, can be determined the total of pixels are actually present in the image colors, and the symbol entropy is the negative logarithm and its probability.
The entropy result is illustrated in Table 8.

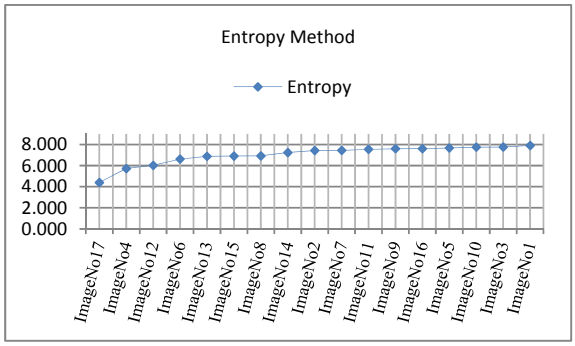


Figure 8. Illustrates a set of experiment analysis of Entropy method in the different types and sizes of images.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

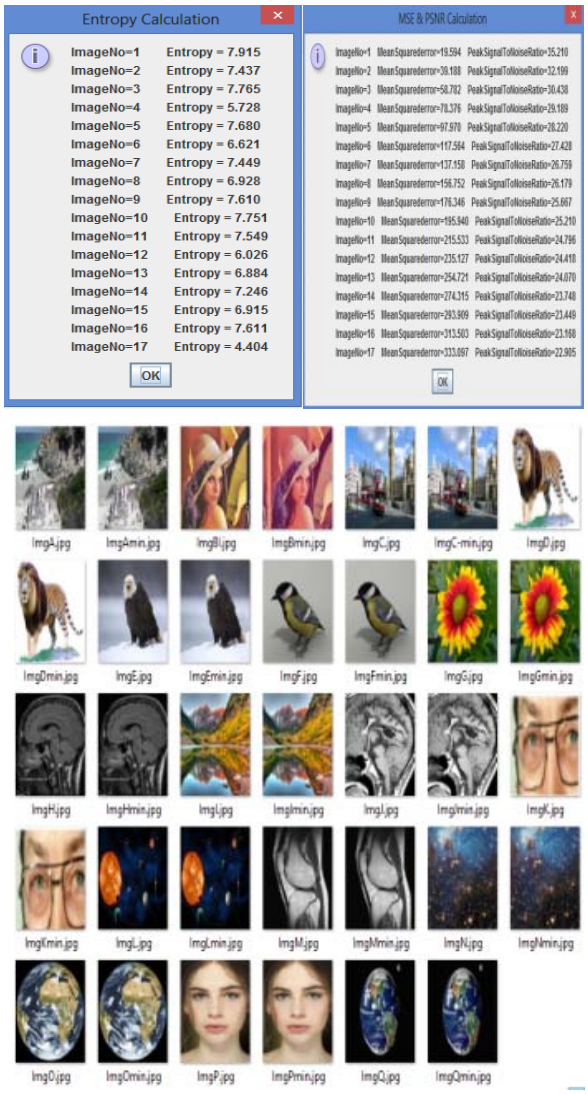
X. CONCLUSION AND FUTURE WORK

In this research, .we applied Hybrid image compression algorithms on different images. The results of an experiment for CPU time utilized in a compression and decompression was satisfied and an acceptable.

We developed Hybrid image compression algorithms to be capable of extending the image compression based on image processing techniques for better space saving.

we recommend using this approach of Hybrid image compression algorithms (HICA) for enhancing and increasing compression images.

In the Future work will apply the multi-objective of genetic algorithm based on image techniques with different types of optimization to improve the performance of Image compression in all entire domains.



REFERENCES

[1] Khalil Ibrahim Mohammed Abuzanounneh," New Image Processing Techniques Using Elitism Immigrants Multiple Objective of Genetic Algorithms for Disease Detection", International Journal of Computer Science and Information Security (IJCSIS), Vol. 15, No. 12, December 2017

[2] Khalil Ibrahim Mohammad Abuzanouneh.," Develop and Design Hybrid Genetic Algorithms with Multiple Objectives in Data Compression", IJCSNS International Journal of Computer Science and Network Security, VOL.17 No.10, October 2017 32

[3] Huffman D. A., "A Method for the Construction of Minimum Redundancy Codes",'Proceedings of the IRE, Vol.40, pp.1098-1101,1952.

[4] Jeffrey N.Ladino,"Data Compression Algorithms", http://www.faqs.org/faqs/compression-faq/part2/section1.html.

[5] Article-Compressing and Decompressing Data using Java by Qusay H.Mahmoud with contributions from KonstantinKladko. February 2002.

[6] Mark Nelson, Jean-loup Gailly,"The Data Compression Book", 2nd edition M&T Books, NY 1995, ISBN 1-55851- 434- 1. NewYork,

[7] Welch T . A., "A Technique for High -Performance Data Compression," Computer, pp. 8--18, 1984.

[8] Object enhancement and extraction. Picture Processing and Psychopictorics by Prewitt JMS (1970), pp. 75–149.

[9] Otsu N (1975) A threshold selection method from gray-level histograms. Automatic 11:285–296.

[10] Woods RE, Gonzalez RC, Eddins SL (2003) Digital image processing using MATLAB. Prentice-Hall, Upper Saddle River.

[11] SobelI (1978) Neighborhood coding of binary images for fast contour following and general array binary processing. Comput Graph Image Process 8:127–135.

[12] Mark. Nelson, Interactive Data Compression Tutor & The data compression book2nd Ed. by M&T books, HTTP: www.bham. ac.uk.

[13] Dr. Dobb's Mark Nelson," LZW Data Compression", by Journal October 1989.

[14] P. D. Johnson, D. Hankerson and G. A. Harris, "Introduction to Information Theory and Data Compression ".

[15] J. C. Vitter, P. G .Howard,"Arithmetic Coding for Data Compression," Proceedings of the IEEE, vol. 82, no 6, 1994 , pp 857 - 865.

[16] Terzopoulos D Kass M, Witkin A,"Snakes: active contour models". Comput Graph Image Process 1: 321.1988

[17] Xu C and Prince J "Gradient vector flow: a new external force for snakes". In 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Proceedings, pp. 66 – 71.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[18] Canny J."A computational approach to edge detection.", IEEE Trans. Pattern Anal Mach Intel, pp 679 – 698, 1986.

[19] Ahmed Kattan Riccardo Poli, " Evolutionary lossless compression with GP- ZIP *," in Proceedings of the 10th annual conference on Genetic and evolutionary computation, Atlanta, Georgia, USA, 2008, 2008, pp. 1211 - 1218.

[20] Data compression Structures. http: \\www.gitta.info - Version from: 18.6.2010.

Dr. Khalil Ibrahim Mohammed Abuzanouneh is an assistant professor in Computer College at the Department of Information Technology, Qassim University. In teaching, he has been focusing on applying genetic algorithms and solves the problem based on genetics approaches in Computer Education. In research, his current interests include Image processing, optimization problem, data compression and data security. Dr. Khalil received his Ph.D. degree in Computer Engineering, "Computer Aided Design".

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Vulnerabilities of Fingerprint Authentication Systems and Their Securities

Tanjarul Islam Mishu
MS Research Student, Dept. of Computer Science & Engineering
Jatiya Kabi Kazi Nazrul Islam University
Mymensingh, Bangladesh
Email: tanjarul26@gmail.com

Dr. Md. Mijanur Rahman
Associate Professor, Dept. of Computer Science & Engineering
Jatiya Kabi Kazi Nazrul Islam University
Mymensingh, Bangladesh
Email: mijanjkkniu@gmail.com

***Abstract*—The security of biometric fingerprint is a big challenge now-a-days, as it has world-wide acceptance. Compromised fingerprint templates may raise terrible threats to its owner. Because of the vulnerabilities of fingerprint authentication system, security issues about fingerprint have been a matter of great concern. This study summarizes the vulnerabilities of fingerprint authentication system and highlights the type of securities available against those challenges. It includes much classified knowledge about security of fingerprint template. This work is an endeavor to provide a compact knowledge to the research community about the security issues regarding fingerprint authentication system.**

Keywords: Attacks; Vulnerabilities; Cryptosystems; Fingerprint Templates; Template Security.

I. INTRODUCTION

Fingerprint authentication system is very popular all over the world because of its uniqueness, usability, reliability etc. It has wide application areas such as border control, airports, business, healthcare, logical access systems, criminal detection, security management, smart phones etc. So, the security of this area is a matter of great concern. Because, the system is vulnerable to several attacks. Ratha[1] presented a model for possible attacks on a biometric system. The model introduced varieties of vulnerable points of the system. This work will focus on the points mentioned in the model. The motive of the present study is to detect different kinds of attacks on each point of this sophisticated model and also to identify the existing security techniques to protect against such kind of the attacks. Although several studies have been done over the attacks and the security approaches, most of them focused on attacks and solutions separately. Very few of them are on both but they are not sufficient. They didn't expose some existing rare solutions. This study will depict the whole scenario of attacks on entire system and securities against the attacks existing now.

This paper is organized as follows. There are eight subsections in Section 2. Each subsection firstly introduces the attacks followed by the solutions against the attacks. As template database attacks contain rich data, Section 2.6 is divided into two parts. Finally, the conclusion is drawn in Section 3.

II. TYPES OF ATTACKS ON FINGERPRINT SYSTEM

Ratha et al. [1] and Anil et al. [2] showed eight points of attack in a biometric system (see Figure-1). Each point and its attacks and regarding solutions has been explained in the following subsections.

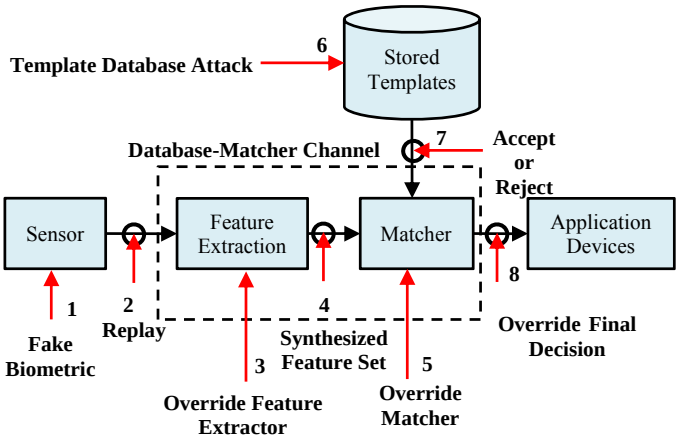


Figure 1. Points of attack in a biometric System

A. *Fake Biometric*

A fake or artificial fingerprint, called spoof, is given to the scanner to get access to the system. The scanner remains unable to distinguish between fake and genuine traits. So, the intruder easily gets access to the system. [2] Putte and Keuning [3] created dummy fingerprint with and without the co-operation of the owner and tested on several sensors. They showed a result that almost every sensor accepted the dummy fingerprint as real at first attempt. Matsumoto et al. [4] experimented gummy (fake) fingers on 11 types of different fingerprint system. In their experiment, about 68-100% gummy fingers were accepted by the system in their verification procedure. They also showed following ways how an attacker may deceive the system at scanner.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

(i) Fingerprints Known to System

The actual registered finger is presented at the scanner by evil way such as external force by the criminals, using the fingerprints when user sleeping etc.

(ii) Fingerprints Unknown to System

If the imposter can know about the category of actual fingerprint (whorls, arches, loops etc), he may use the similar fingerprints unknown to system. Though it is almost impossible, it may harm the systems which are developed on the basis of insufficient features of fingerprint. It may effects on False Acceptance Rate (FAR) of the system. So, the authentication should be based on sufficient features.

(iii) Severed Known Fingerprints

It is similar to the known fingerprint mentioned earlier. But, it is a horrible attack done by a criminal to severe the fingerprint from the real user's hand. To be protected, we should detect is the finger alive or not.

(iv) Genetic Clone of Known Fingerprints

Identical twins do not have same fingerprints. Because, the patterns of fingerprint are determined by the genetic mechanism and the nerve growth. So, they are not same but still very close. So, a genetic clone may be tried to deceive the system. To be protected from this kind of threats, we should keep tracking a genetic engineering on possibility of creating clones.

(v) Artificial Clone of Known Fingerprints

The attacker can make a 3D printed fingerprint or can make a mold of the known finger by which an artificial finger can be produced.

(vi) Printed Image of Known Fingerprints

This is very similar to the previous one. By the help of spraying some materials on the surface of the scanner to feel like actual finger, imposter can use printed image of fingerprint.

Liveliness detection can be solution to fake biometric traits. There are two separate methods, such as, passive (non-stimulating) and active (stimulating) automated liveliness detection methods [5]. Generally, passive detection techniques make use of biometric probes recorded through a biometric sensor such as pulse measurement, temperature measurement, active sweat pores detection, skin resistance detection, electrical conductivity etc.[16] Active detection techniques normally require additional interactions that should requested using challenge response procedures. The different challenge response approaches can be used such as request of different fingers in random order.

B. Replay Attack

After acquisition of raw biometric data, it sends the raw data (e.g. fingerprint raw image) to the feature extraction module. The imposter steals the biometric trait raw data by seizing the channel and stores the trait. The imposter can reply the previously stored biometric trait to the feature extraction module to bypass the sensor. Fingerprint images are sent over channel usually compressed using WSQ. Because of the open compression standard, transmitting a WSQ compressed image over the Internet is not particularly secure. If the image can be seized, it can be decompressed easily which can cause Replay Old Data [1].

Data hiding techniques such as steganography can be applied when the raw image is sent to feature extractor.

C. Override Feature Extractor

The hackers, by Trojan Horse, take control over the feature extractor to produces feature sets as they wishes [1].

When installing or updating programs in a device it should be verified and should be aware of using third party programs.

D. Synthesized Feature Set

If the imposter can intercept the channel between the feature extraction module and matcher, he can replace the original set with a different synthesized feature set (assuming the representation is known to imposter) [1]. Insecure communication channel may face the 'Hill Climbing Attack' [2].

Hill Climbing Attack

Uludag & Anil have developed an attacked for minutiae base fingerprint authentication system [6]. The location (c, r) and orientation Θ of minutiae points has been used by the attack. The system will works as the attackers knows the format of templates but not the information of templates. It uses the match score returned by the matcher and tries to generate minutiae set that results in successfully high matching score to be positive in identification. Figure-2 describes the Hill Climbing attack.

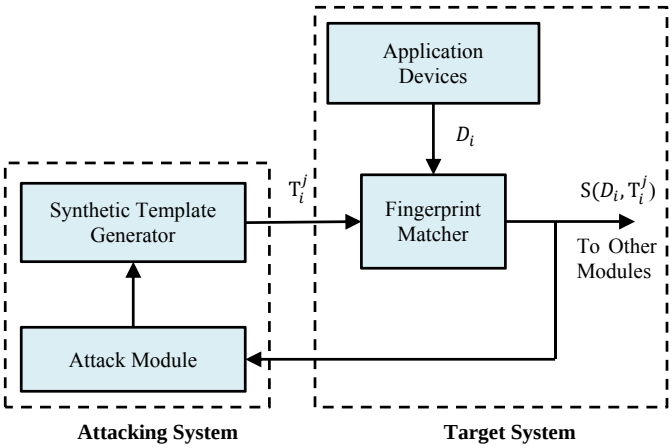


Figure 2. Block Diagram of Hill Climbing Attack

D_i refers to the database template corresponding to user i , i =1, 2,3,....N , where N is the total number of user. n_i is

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the total number of minutiae in D_i . T_i^j is the j^{th} synthetic template generated by the attacking system for user i . $S(D_i, T_i^j)$ is the matching score between D_i and T_i^j. $S_{threshold}$ refers to the decision threshold used by the matcher. Note that the attacking system does not know this value.

At the beginning of the attack, it generates several synthetic templates. Then begins attack with these templates and accumulate the matching scores returned by the matcher. It chooses the template having highest matching score. Then tries modification (perturbing, adding, replacing or deleting of minutiae) to get larger match score and chooses the larger one as the best template T_i^{best}. This modification continues until the matcher accept the current best score where $S^{best}(D_i) > S_{threshold}$.

To be safe from hill climbing attack, we can add some extra features in the matcher of authentication system. These may include-

i) tracking the number of failures within specific time.
ii) limiting the number of tries within specific time.

E. Override Matcher

The hackers replace the matcher by a Trojan horse program that generates very high or low matching scores as the hackers want, regardless of original scores [1].

The Matcher is also a program like feature extractor. Attacks to this point can be solved in the similar way as feature extractor described in section 2.3.

F. Template Database Attack

(i) Type of Attacks

The template databases can lead to three kinds of threats [3] as describe below.

a. Template Replaced by The Imposter's Template

The imposter can replace the original template with new one to gain the unauthorized access to the system whenever he wants like an authorized user.

b. Masquerade/Physical Spoof Created from Templates

Minutiae information is unique to each individual. The view of non-reconstruction was dominant in the biometrics communities, until some recent researches. Over last few years, some works were done that showed that a fingerprint image can be reconstructed from a minutiae template. The fingerprint image reconstructed from the minutiae template, known as a "masquerade" image since it is not an exact copy of the original image, will likely fool the system when it is submitted [7]. In 2007, Cappelli at al [8] did some amazing experiments. The authors analyzed the ISO/IEC 19794-2 minutiae standard template. They took different ways of test. In one experiment, they used basic minutiae

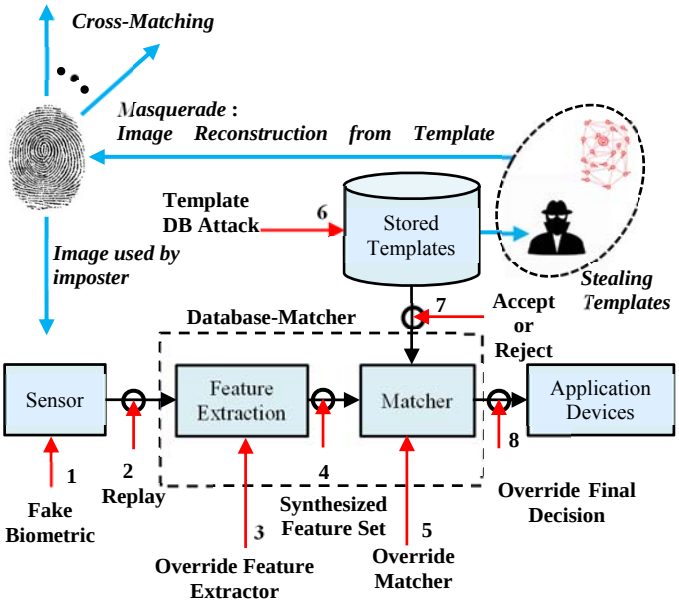


Figure 3. Image Reconstruction (Masquerade) from stored templates

information only (i.e. positions x, positions y, and directions). In another test, they also used optional information: minutiae types, Core and Delta data, and proprietary data (the ridge orientation field in this case. In their experiments, nine different systems were tested and the average percentage of successful attacks was 81% at a high security level and 90% at a medium security level. Image Reconstruction with points of attack in fingerprint is shown in figure-3. Masquerade can be very threatening fact to the owner. Because, hackers may track the owner where he/she is using the fingerprint. They may hack bank accounts and other secured accesses. They may use the masquerade to databases at other organizations to get unauthorized access, though they use different templates and algorithms, called Cross-Matching.

c. Stolen Templates

Imposter can steal the template and replay that on matcher. The stolen template can be used as synthesized feature set.

(ii) Template Protection Techniques

All the template protection techniques can be categorized in two major categories, such as, (a) feature transformation and (b) biometric cryptosystem. Figure-4 shows a graphical representation of biometric template protection techniques. Other types of template protection techniques are water marking [14], steganography [15], system on card/match on card [2] etc.

a. Feature Transformation

For the protection, the features generated from the input image are transformed to a new form. It is not kept in real

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

form rather stored in transformed form. The generated transformed form can be invertible or non-invertible.

1. Invertible Transformation (Bio Hashing)

In invertible feature transformation, the template is transformed with some parameter of user. At the site of authentication, the template is inverted again with the secret parameters. The scheme can't provide high security without

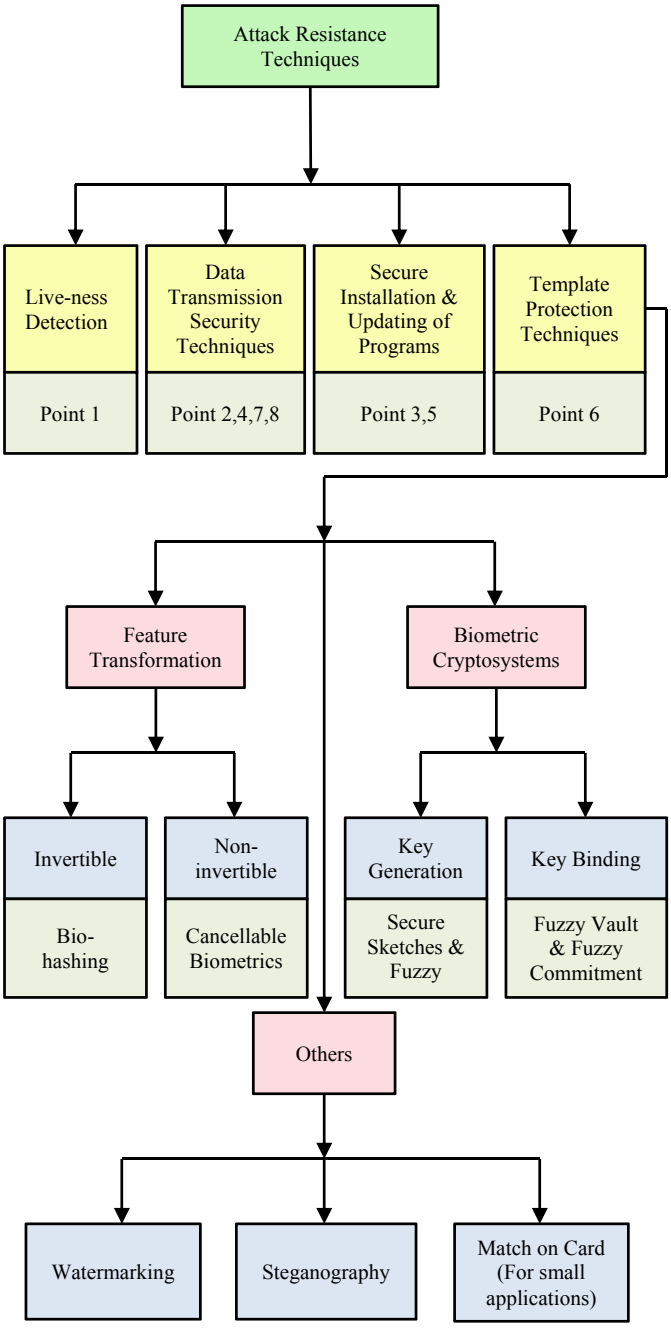


Figure 4. Attacks and Solutions on Fingerprint Authentication System.

the secret transformation. Because if the secret key(transformation parameters) is compromised with imposters, they can revert the template. So, the key should be secured enough.[17]

2. Non-invertible Transformation (Cancellable Biometrics)

Cancellable biometrics scheme is an intentional and systematic repeatable distortion of biometric template data with the purpose of protecting it under transformational-based biometric template security. In the verification site, the query image is transformed in same the manner, then compared. In the concept of cancellable transformation, a transformed template can be cancelled and re-issued by changing transformation parameters if problem issued [9].

b. Biometric Cryptosystems

Cryptosystem technique on biometric data is called biometric cryptosystem where a key (or keys) is used to encrypt the biometric data. The key can be generated from biometric data itself or from an external data. At the matcher, the key is used to decrypt the biometric data. Observing the literature, we divide Biometric Cryptosystems into two major parts: Key Generation or Key Binding.

1. Key Generation

At the time of enrolment, a unique key is chosen from the features extracted from the fingerprint. This key is not stored in the database [10].

A *Secure Sketch* reliably reproduces the biometric secret without leaking any information. It works in two phases: Generation & Reconstruction. It takes biometric data as input and creates a sketch of that data. Later, at reconstruction, the generated sketch and the data sufficiently similar (query image) to original the input data are given. Then, it reproduces the original input data. Thus, it can be used to reliably reproduce error-prone biometric inputs without incurring the security risk inherent in storing them [11].

Fuzzy Extractor reliably extracts almost uniform randomness R from its input. It is error-tolerant because if we change deliver different template from same finger, R will not change. The resultant R is almost similar to the original R. This R is used as a key in cryptographic application [9].

2. Key Binding

In key binding, cryptographic key is tightly bound with the biometric template so that it cannot be released without a successful biometric authentication and without accessing template directly [12]. The key Binding can be categorized as Fuzzy Vault and Fuzzy Commitment.

Fuzzy Vault is first introduced by Juels and Sudan [13] as a cryptographic construct. There are used two set of points : fuzzy unsorted points and chaff points. The unsorted

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

data set is taken from biometric data. Meenakshi[11] explained the fuzzy vault with biometric system. In fuzzy vault framework, the secret key S is locked by G, where G is an unordered set from the biometric sample. A polynomial P is constructed by encoding the secret S. This polynomial is evaluated by all the elements of the unordered set G. A vault V is constructed by the union of unordered set G and chaff point set C which is not in G. The vault, V = G U C. The union of the chaff point set hides the genuine point set from the attacker. Hiding the genuine point set secures the secret data S and user biometric template T. The vault is unlocked with the query template T'. T' is represented by another unordered set U'. The user has to separate sufficient number of points from the vault V by comparing U' with V. By using error correction method, the polynomial P can be successfully reconstructed if U' overlaps with U and secret S gets decoded. If there is not substantial overlapping between U and U' secret key S is not decoded. This construct is called fuzzy because the vault will get decoded even for very close values of U and U' and the secret key S can be retrieved. Therefore fuzzy vault construct becomes more appropriate for biometric data which possesses inherent fuzziness.

A *Fuzzy Commitment* scheme is one where a uniformly random key of length 1 bits (Binary vector) is generated and used to exclusively index an *nbit* codeword of suitable error correcting code where the sketch extracted from the biometric template is stored in a database [9].

G. Database-Matcher Channel Attack

On this type of attack, the stored templates coming from database is being modified before reaching to matcher. So, the matcher gets modified templates.

Maintaining secure data transmission can solve the problem. Different error detection techniques such as parity check, checksum, cyclic redundancy checks can be used to identify the transmitted template is modified or not.[18]

H. Override Final Decision

Final result coming from the matcher is modified by the imposters. It changes the original decision (accept/reject) by changing the match scores.

Sending the result through a trusted channel and using a secure delivery can be used to get the correct result.

III. Conclusion

This study conveys a prominent analysis on the vulnerabilities of Fingerprint Authentication System of each point of the model and shows the effective security system existing now. This work brings vulnerabilities and securities, compacted together, of fingerprint authentication system. Different types of attack such as fake biometric, replay data, synthesized feature set and template database have been explained about how they occur. The paper also contains the prevention techniques against the corresponding attacks. As the template database is very sensitive part of the system, its protection techniques are have been analyzed with high significance. This paper even shows very small attempts taken such as match on card for the security of fingerprint template. In the analysis, it has been learnt that attack on template is very severe. If the templates are compromised, the security of their owner will be violated. So, template security requires more attention of research authority. Though several types of work have been done on the template security, they are not able to satisfy all the requirements such as recoverability, security, privacy, high matching accuracy etc. So, our next work is to generate an efficient template security scheme.

References

[1] Ratha, Nalini K., Jonathan H. Connell, and Ruud M. Bolle. "An analysis of minutiae matching strength." International Conference on Audio-and Video-Based Biometric Person Authentication. Springer Berlin Heidelberg, 2001.

[2] Jain, Anil K., Karthik Nandakumar, and Abhishek Nagar. "Biometric template security." EURASIP Journal on Advances in Signal Processing 2008 (2008): 113.

[3] T. Putte and J. Keuning, "Biometrical fingerprint recognition: don't get your fingers burned", Proc. IFIP TC8/WG8.8, Fourth Working Conf. Smart Card Research and Adv. App., pp. 289-303, 2000.

[4] Tsutomu Matsumoto, Hiroyuki Matsumoto, Koji Yamada, Satoshi Hoshino. "Impact of artificial 'gummy' fingers on fingerprint systems." Proc. SPIE 4677, Optical Security and Counterfeit Deterrence Techniques IV, April 2002.

[5] Rogmann, Nils, and Maximilian Krieg. "Liveness Detection in Biometrics." Biometrics Special Interest Group (BIOSIG), International Conference of the. IEEE, 2015.

[6] Uludag, Umut, and Anil K. Jain. "Attacks on biometric systems: a case study in fingerprints." Proceedings of SPIE. Vol. 5306. 2004.

[7] "Fingerprint Biometrics: Address Privacy Before Deployment." https://www.ipc.on.ca/wp-content/uploads/2008/11/fingerprint-biosys-priv.pdf, 2008.

[8] Cappelli, Raffaele, Dario Maio, Alessandra Lumini, and Davide Maltoni. "Fingerprint image reconstruction from standard templates." IEEE transactions on pattern analysis and machine intelligence 29, No.9. 2007.

[9] Mwema, Joseph, S. Kimani, and M. Kimwele. "A Simple Review of Biometric Template Protection Schemes Used in Preventing Adversary Attacks on Biometric Fingerprint Templates." International Journal of Computer Trends and Technology 20.1 (2015): 12-18.

[10] Khandelwal, Sarika, P. C. Gupta, and Khushboo Mantri. "Survey of Threats to the Biometric Authentication Systems and Solutions." International Journal of Computer Applications 61.17. 2013.

[11] Meenakshi VS, "Secure And Revocable Biometric Template Using Fuzzy Vault For Fingerprint, Iris And Retina." A Thesis Submitted To Avinashilingam Deemed University For Women Coimbatore – 641043, 2010.

[12] Huixian, Li, et al. "Key binding based on biometric shielding functions." Information Assurance and Security. IAS'09. Fifth International Conference on. Vol.1. IEEE, 2009.

[13] Juels, Ari, and Madhu Sudan. "A fuzzy vault scheme." Designs, Codes and Cryptography 38.2 (2006): 237-257.

[14] Patel, Monika, and Priti Srinivas Sajja. "The Significant Impact of Biometric Watermark for Providing Image Security using DWT based Alpha Blending Watermarking Technique."

[15] Rubal Jain and Chander Kant. "Attacks on Biometric Systems: An Overview." International Journal of Advances in Scientific Research 2015; 1(07): 283-288.

[16] M. Villa and A. Verma. "Fingerprint Recognition," in Biometrics in a Data Driven World: Trends, Technologies, and Challenges, M. Gofman and S. Mitra Eds., CRC Press, USA, pp. 265-281, 2017.

[17] Ramu, T., and T. Arivoli. "Biometric Template Security: An Overview." Proceedings of International Conference on Electronics. Vol. 65. 2012.

[18] Siwach, Ajay, Sunil Malhotra, and Ravi Shankar. "Analysis of Different Error Detection Schemes over OFDM Signal." International Journal of Engineering Trends and Technology- Volume4 Issue4-2013

AUTHORS PROFILE

Tanjarul Islam Mishu received his B.Sc. (Engg.) in Computer Science and Engineering from Jatiya Kabi Kazi Nazrul Islam University, Mymensingh, Bangladesh in 2016. Currently, he is student in M.Sc. of the institute. His research interest is focused on Biometrics Systems, Pattern Recognition, Image Processing, and Data Mining.

Dr. Md. Mijanur Rahman is a faculty member of the Dept. of Computer Science and Engineering, Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymensingh, Bangladesh, since April 2008 (very beginning of the university). Now, Dr. Rahman is working as an Associate Professor of the CSE department. He also served as Lecturer and Assistant Professor in the same department. He served before as an Instructor (Tech) Computer in Govt. Polytechnic Institute from December 2005 to April 2008. Dr. Rahman obtained his B. Sc. (Hons) and M. Sc. degree both with first class first in CSE from Islamic University, Kushtia, Bangladesh. He also obtained his PhD degree in Computer Science and Engineering from Jahangirnagar University, Savar, Dhaka, Bangladesh in August 2014. His teaching and research interest lies in the areas such as Digital Signal Processing, Digital Speech Processing, Biometric Technology, Pattern Recognition, etc. Many of his research papers have been published in both national and international journals.

Mobile First Companies: A Case of App Usability in Kuwait

Bareeq A. AlGhannam, Ph.D.
Computer Science and Information Systems Department, College of Business Studies
The Public Authority for Applied Education and Training
Kuwait, Kuwait
ba.alghannam@paaet.edu.kw.

Eiman AlEssa, MSc.
Computer Science and Information Systems Department, College of Business Studies
The Public Authority for Applied Education and Training
Kuwait, Kuwait
eao.alessa@paaet.edu.kw

Manal Almukhaizim, MSc.
Computer Science and Information Systems Department, College of Business Studies
The Public Authority for Applied Education and Training
Kuwait, Kuwait
ma.almukhaizeem@paaet.edu.kw

***Abstract*—This study investigates the importance of the usability of a Mobile First Company (MFC) app. The number of MFCs is growing rapidly worldwide, and the existence of such companies primarily rely on their apps being used. There is a broad range of usability literature, however scarce data exists, that describes how app usability contributes to the success of MFCs. This research uses a case study to empirically extract an initial link between MFC success and the**

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

perceived usability of its app. Arabic-System Usability Scale (A-SUS) is employed to evaluate the usability of an MFC app in Kuwait. The results are used to start collecting data in order to initiate a correlation between MFC success with its app perceived usability.

Keywords-App Usability; System Usability Scale; SUS; Mobile First Companies; Standard Usability Questionnaire; Arabic-System Usability Scale

I. Introduction

Mobile First Companies (MFC)s are companies that provide content or services or products via mobile application. Mobile application is shortened by the term "*app*"; where "*apps refers to software applications designed to run on smartphones, tablet computers and other mobile devices* [1]". Reference [1] states that "*The proliferation of mobile-app enterprises, along with that of smartphone usage, has dramatically changed traditional business models*". This is evident by the number of smartphone users that is increasing tremendously [2]. In 2014 it reached 2.1 billion users and is estimated to reach 6.1 billion users by the year 2020 [3]. Statistics dated to the published literature shows that App Store has a lead of 2.2 million apps, followed by iTunes with 2.0 million apps, however Android is not as mature [4].

MFCs targets mobile users through mobile applications. In time these mobile applications could expand into websites for desktop/laptop. These apps are prominently sweeping the market, especial in this era of ever evolving technology[1] [5]. For these apps to succeed in the industry they need to be usable.

App usability anchors the success of these companies because their existence relies on them being used. Usability measures integrated within the design and the development of such software improves the app quality. Quality of an app can be incorporated by following a proper process for app development [5]. Quality assurance of an app ensures that the app doesn't fall under the category of being a "bad app"; which in turn increases the potential of the apps' success. If an app is of a good quality, users will keep using it [5], and therefore will increase the success of a MFC.

This research investigates the concept of MFCs in Kuwait. It evaluates the perceived usability of "PayLe Collection Company" app, which is called "PayLe". The research then goes further into details, investigating whether it's perceived usability results of its app are satisfactory. It goes further to initiate an initial relationship between the apps' perceived usability and the company success.

The paper begins with an outline of the concept MFC. This outline is followed by a literature review of critical app issues in the industry in conjunction with a review of app usability. The literature review is critically analyzed

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

and leads to the best fit tool to measure an apps perceived usability. This tool is applied on the chosen PayLe app in hope of building a correlation between perceived usability and mobile app success. The paper elaborates with a conclusion and expands on recommendations for further research.

II. Litrerature Review

A. Mobile First Companies (MFC)

Mobile First Companies (MFC)s are companies or enterprises that provide content, services or products through apps or mobile sites. Most MFCs are start-up companies with small to mid-range capital. MFCs might also be a new services offered by established companies. Diverse enterprises are making use of this philosophy, because they noticed the rapid growth of smartphone usage. The mobile industry is expanding and the number of app downloads increases tremendously every day [6]. Numbers show that app downloads increased from 149 billion to 197 billion mobile from the year 2016 to 2017, and it is predicted to overcome 350 billion app downloads by the year 2021 [7].

This recognized growth in mobile application usage is due to many reasons: apps are developed faster through agile software development process [8]; where agile development has a lower life cycle than conventional software development, needs a smaller team to develop, and can be fully launched within a couple of weeks. Other factors that contribute to the popularity of these apps is in relation to the innovative use of the features in smart phones [5] [9], as well as the convenience of using the app even when there is no internet connection [5]. Furthermore, app usage is faster than mobile website usage [1].

MFCs exist as long as users use them continuously. This is achieved once users are able to obtain the content/service/product they are seeking in an easy to use and efficient manner. In software engineering ease of use and efficiency are factors of usability. Therefore MFC app usability is vital. And the employment of good development practices is essential to ensures high usability [5].

B. Mobile Application Issues

The mobile industry is challenged to find innovative ways to take advantage of the distinguished features of smartphones and mobile devices [1] [5] [9].

Although there is a noticeable advantage in employing smart phones, the level of security is a major concern in the industry [10]. Design needs to emphasise more on mobile security issues to raise it to the standards of personal computers.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Other challenges that face mobile app designers is the different versions of operating system platforms of smart phones and mobile devices. Android and iOS are the two main streams, each with its advantages and drawbacks. Mobile apps makes ultimate use of these two main platforms when the programming is done at the device level. However, this is not feasibly economical as it is difficult to keep track of what mobile users are currently using, especially with the emergence of other mobile operating systems such as blackberry and windows.

Usually mobile developers use HTML5 web-based [11] or Native apps. For HTML5 one tool is used for app development, however it has a lesser user experience. On the other hand native apps use different fragmented tools for each mobile platform which provide better user experience. The optimal solution is by the use of Hybrid architecture; app developers can make one app to fit all platforms for all type of devices. Hybrid is a better option because it provides a compromise between HTML5 web-based and native apps.

Moreover, the different sizes of displays and screens of mobile devices make it difficult to target a specific dimension. This could be resolved by the use of responsive design; where it ensures that the design of the website fits any screen [11]. Since time is an issue in the case of mobile applications, responsive design is not recommended to be used.

It is also crucial to point that the attention span of mobile users is tremendously low. Therefore speed needs to be considered and attenuated in addition to good mobile app design. Certain development process and measures need to be evaluated to ensure the good design of the mobile app and thus its success [5].

C. The Need for App Usability

In the mobile application realm, usability has been and is still a major interest of practitioners [12] [13] [14] [15]. Mobile app usability is considered crucial and very important, however literature shows that it is untapped territory and there is scare literature conducted specifically on app usability [16]. Reference [16] called for and developed a processes to be followed by software researchers and practitioners to specifically measure the mobile usability taking into consideration the context. A rigorous process when followed will ensure both quality and efficiency of the app.

The available software usability measurement tools in both the literature and industry are originally developed for desktops and laptops. Literature shows that no specific usability measure has been developed for mobile devices. Smart phones and mobile devices have features that evolve constantly [9]. Examples of current existing features include making use of the digital camera to share pictures, sensing capabilities, utilising the global positioning system (GPS) such as sharing and finding locations in maps. Mobile apps need to integrate these features seamlessly, they also need to grow adaptively with the vast growing changes in the mobile technology industry. And usability tools need to recognise these changes and be able to include them in its measures.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

As was stated before our outmost goal is to reflect on how mobile users perceive usability of apps. Some practitioners and researchers call it "mobile experience", this is because different mobile users experience apps differently due to the differences in their devices capabilities [5] [9]. Usability is essential, and there is a need to make sure of the continuance of the app usage; this is vital but unfortunately the literature is scarce in that matter [5].

Usability is a major factor of app success which in turn leads to MFC success. Earning profit from the app is manifested through users being loyal to the app and using it continuously for content/services/products. This keeps the MFC business viable by stimulating profit. Profiting from the app is called "Monetisation" [1].

User perceptions of apps is crucial, especially when the existence of the company relies on the app being used. In general it is vital to realise the importance of software usability [14] [17], and it is more vital to realize the importance of usability for app acceptance [18].

Usability practitioners and researchers face many challenges when it comes to mobile application [19]. Just as any software system, it is difficult to identify which usability tool, approach or method is better than the other [17], that is why the goal of the study needs to dictate which tool is most suitable to be used as a usability measurement [20]. Reference [21] presents an analysis of the different software usability measurements and gives reasoning of what to use for what. And the usability tool chosen depends on the goal of the research or study [22]. Other researchers have also presented the various way to evaluate such a measurement of usability [17] [23] [24] [25] [26].

Usability can be performed at different points of the software development life cycle [23] [25], and results of usability can be used to enhance the software[27] [28]. Mobile applications benefit from the use of standardised usability questionnaires for their evaluation [20]. There is a need for specific process to document the usability difficulties. Following a process ensures improved app usability and this will enable developers to cope with the evolving next generation of mobile technology.

In our research, the goal is evaluating perceived usability. Literature shows that standard questionnaires are widely used in practice to evaluate such usability [29]. Evidence shows that they are adequate tools to be used in order to satisfy the goal of the research of perceived usability. One such standard usability questionnaire is called "The System Usability Scale" (SUS) [30]. We believe that SUS best fits our research because of the following reasons: (1)Short and fast; which appeals to the respondents, (2)Easy to administer, (3)Easy to complete, (4)Has guided analysis process, and (5)Psychometrically evaluated

SUS is a standard questionnaire for system usability that is psychometrically proven [30]. In 1986 SUS was developed by Digital Equipment Corporation (DEC) in the UK by John Brooke. It consists of a ten statements questionnaire. The questionnaire starts with a negative statement followed by a positive statement alternatively. Respondents choose from a five level Likert scale ranges from (1) being least agree (5) being most agree. The SUS

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

questionnaire is then analysed using specific guidelines to obtain a single numeric value that represents the subjective measure of perceived usability. The SUS usability single value is interpreted differently depending on the users and what genres [27] [31].

To calculate the SUS score, for all odd statements a one is to be subtracted from the choice of the evaluator, however, the even statements the evaluators' choice is subtracted from five. The results will be in the range of (0 to 4); these values are considered transformed values with four representing the most positive response. Then, the responses for each evaluator are accumulated and multiplied by 2.5, this last process visualises a result between a zero and one hundred. This value is not a percentile, and care should be taken not to perceive it as a percentage value. If a percentile ranking is needed, then a process of normalisation needs to be undertaken. Since these results are not percentiles, they are interpreted differently; reference [33] provides us with detailed interpretations of SUS scores. It should be noted that the genre and environment in which the evaluation is conducted is crucial. This importance of environment reflection on usability is in sync with [17] who stressed on the effect of the environment on usability measures. As a conclusion identical scores for two different users in different genres might give different indications of usability. To make the evaluation more comprehendible, researchers have transformed the numerical value to an adjective representation [34]. Usability measured as perceived user satisfaction is crucial for developed apps [13], and thus crucial for MFCs. MFCs thrive on what makes them continuously usable, where a usable app is a factor that leads to MFC success.

An approach is needed to find the correlation between mobile perceived usability and MFCs success. Significance of such correlation is essential for start-up small companies because they rely on the success of mobile-first strategies. Limited literature is present relating the success of mobile applications to perceived usability measures in any phase of the app development cycle. More specifically finding data supported by scientific research and studies is even harder to find in developing countries such as Kuwait. Small start-up companies in Kuwait make the majority of MFCs; which is the target of this research. And at this era in Kuwait, MFCs are at most importance because they open new markets in Kuwait that promotes the local economy. MFCs are considered in Kuwait as micro to small businesses depending on their capital. And the government is supporting such businesses by a fund called “Small Projects Fund”. This research gives a better understanding to the concept of MFC, how to measure their apps perceived usability, and initiate an initial correlation between the success of the mobile app and the result of the perceived usability measures.

A case study is chosen as a first step to better understand MFCs perceived usability grounded by findings from a usability tool applied on an app in Kuwait market. Details are provided in the methods section below.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

III. Methods

A. Process and Tools

The following steps make up the process used and tools employed:

1) *A successful MFC was chosen from the Kuwaiti market.*
2) *Success factors, related to its app usability, were encapsulated via in-depth interview.*
3) *Adequate Usability Evaluation tool is chosen depending on the research goal and used to measure perceived app usability.*
4) *A tentative correlation is established between MFC success and its app usability*.

B. Participants and Setting

PayLe Collection company is chosen as a case study to measure the mobile app usability of a MFC in Kuwait, as it represents a mobile application of a small successful company. It provides an easy payment process for the value of sales without any paper work as a solution for small home businesses or individuals to collect sales value through debit or credit cards in an innovative way. It offers "PayLe" payment service through the mobile phone application which is at the present time a revolution in the world of electronic payment in Kuwait. The app is characterized by the ease of payment between a merchant and a client through the provision of the service of debit/credit cards on smart phones without the need to commit to points of sales or the need of the seller to meet with the buyer.

From an in-depth interview with the founder of PayLe app Mr. Ameer Almansoor , he stated that the main goal of PayLe was to fill a gap in the Kuwaiti market by providing a payment gateway via an Arabic interfaced app. There was emphasis on the following: the constructed app is easy to use, secure, and has an-Arabic interface as the majority of the users are native Arabic speakers. Users use the Arabic PayLe app in confidence and without hesitance as opposed to other similar apps with English Interface. There was also emphasis that the payment gateway is through an app that is linked to a mobile phone text message, where according to the founders' experience mobile phone payment is preferred over website payment in Kuwait environment. Most importantly the app developed is highly secured and that adds to the users confidence while using it.

The app is legalized with a license and aligned with the regulations of the state of Kuwait for money transfer with regular checks from National Security of Kuwait. Admin side of the app supports transparency of transferred funds with detailed information of sender and recipient. PayLe app use is expanding and is now covering Saudi Arabia with furthermore intentions to expand to Gulf Council Countries (GCC).

C. Procedure and Analysis

SUS is chosen to be used to measure the perceived usability in this study. SUS is very popular and widely used by practitioners and researchers. The vast publications indicate various attempts to deepen our knowledge in SUS.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Studies include factor analysis [35], psychometric Evaluation of the SUS is conducted by measuring the validity and reliability of the tool.

Literature shows the importance of usability tools conducted in the native language of the user [20] where many scholars have used such language adaptations to various standard usability tools [36] [37] [38] [39] [40] [41]. And for that reason in this research we employ the Arabic adaptation of the SUS standard evaluation tool called Arabic-System Usability Scale (A-SUS) [20].

Psychometric Evaluation considers the validity, reliability, and sensitivity of a questionnaire [30]. It is essential to examine the psychometric evaluation of the usability tool [42], specifically if it is adapted in a different language. The literature shows many standard tools used in other languages have gone through psychometric evaluation [36] [43] The psychometric process conducted previously for A-SUS in conjunction with the communication disorder app [20] is to be followed in this study. Reliability, validity, and sensitivity is of concern and once established, A-SUS results will indicate usability as perceived satisfaction [20].

The A-SUS score is calculated using the same procedure used to calculate SUS presented in the literature review. Psychometric evaluation of A-SUS ensures that the essence of SUS is reflected upon it; where psychometric evaluation results of A-SUS indicates similar results to previously conducted research using SUS [20].

IV. Results

A-SUS questionnaire was administered through google forms. The questionnaire was sent to PayLe users and was also pinned as a link in the app. A total of 296 responses was collected a period of one week. From the demographic data only 1.4% replied that they rather have the service on a PC, and 8.1% replied that they do not mind if the service was on a PC or mobile but prefer if it was on a PC. This very small percentage confirms with the literature that indicates users prefer using apps in mobile devices.

Reliability result of 0.80 alpha Cronbach is measured. This score is both valid and reliable where alpha coefficient greater than 0.70 indicates reliable results.

Construct Validity results of Pearson correlation ranges between the values 0.534 and 0.692 and this range is within the accepted range of valid results. This indicates that our tool is valid for measuring the SUS score.

A-SUS score of 77.516 and it was calculated as it was presented in the methods section. This score of 77.516 is considered an acceptable result of usability when compared to SUS benchmark scores, where scores of 68 and over are average [43]. Our scores of PayLes' perceived usability represents a system with above average usability; specifically if compared with software products where an average of 72 is documented [43]. This confirms what we

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

were hoping for; which is a tentative link between the success of an MFC and having an app with accepted perceived usability.

V. Discussion

Usability measures obtained from conducting A-SUS were performed on PayLe; the chosen mobile application. The collected data is analysed using precise instructions of SUS and similarly A-SUS. Results show good usability results and an initial correlation is initiated between the success of PayLe and its app high usability results. This is the first step to accumulating data related to MFC app usability using A-SUS.

VI. Conclusion and Further Studies

This research presents a single incident that correlates the success of a MFC with its measured app perceived usability. It is a first step into building a link between these two factors. However, this correlation has not reached the point of generalisation and further research of other MFCs need to be conducted to establish a pattern of cause and effect. A repetition of the study with similar correlation gives further evidence for generalisation. Once a generalisation is established, and an app usability is said to infer MFC success, then it is possible to predict MFC success based on its app usability in the very early stages of a soft launch or earlier stages in the development lifecycle.

This research stresses on the importance of a MFC app, and recommends a usability evaluation strategy to be administered in a systemic manner. Such strategies would increase the profit margin of a MFC and thus lead to its success.

Softwares in general are in need of standard usability questionnaires. Standard Usability tools used to measure the success of MFC apps can cautiously be used to indicate the success of such MFC. MFCs primarily rely on their apps for existence as other physical presence of the company is evident. If the app is not usable then there is no pathway for the company to gain profit.

In this study an adaptation of SUS; A-SUS questionnaire has been used as a feasible tool that gives a quick, valid and reliable indication of an app's usability. The results are used in two folds: to collect and analyse data related to A-SUS, and to elucidate further how usability can be used to reflect upon the success of MFC.

A-SUS was administered to evaluate the usability of PayLe app, which is a successful MFC in the Kuwaiti economical market. A-SUS results applied to PayLe indicate high usability when compared to benchmark studies. A blend of interviews in addition to the standard usability questionnaires are employed to gain better understanding of PayLes' success and its relation to the perceived usability of its app. Results indicate that PayLe developers basically relied on two aspects: First, simplicity; Second, the native language of the targeted users is used in the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

interface of its app. These two aspects ensured to engage users and gain their trust and loyalty. Both aspects promote the usability of the app and lead to the acceptable usability scored measuring of A-SUS.

This study reports results from a single case and findings do not infer generalisation. Therefore the same tool needs to applied on other apps of MFCs where usability results can be linked to an apps success. A collection of data in the future over time and on the diverse types of MFC apps will give sufficient databases where usability patterns can be depicted and further inference can imply generalisation.

Future research would also gain tremendous benefit from further usability evaluation, with emphasis on factor analysis. There is a need to depict specific aspects that affect usability, and they could be linked to further demographic attributes such as gender, education level...etc.

Also employing other usability tools to the same MFC app, and conducting a comparison of the results might give a better indication of what usability tools to use in the future to better estimate MFC success.

Acknowledgment

We would like to thank Mr. Ameer Almansoor founder of PayLe for collaborating with us in this research and posting the A-SUS questionnaire as a link in the PayLe app. Also we would like to thank all the participants who are users of PayLe both merchants and clients for volunteering to complete the questionnaire and who allowed us to collect their feedback from the app evaluation.

References

[1] A. Tang, "Mobile app monetization: app business models in the digital era", *International Journal of Innovation, Management and Technology*, vol. 7, no. 5, October 2016, Available:http://www.ijimt.org/vol7/677-MB00017.pdf, [Accessed January, 4, 2018].

[2] Ericsson Website, *Ericsson Mobility Report: 70 Percent of World's Population Using Smartphones by 2020*, 2015. Available:www.ericsson.com, [Accessed 16 December, 2017].

[3] Available: http://www.digital.com, [Accessed December, 30, 2015].

[4] Available: http://www.businessofapps.com, [Accessed December, 30, 2015].

[5] A. Inukollu1, D. Keshamoni, T. Kang and M. Inukollu., "Factors influencing quality of mobile apps: role of mobile app development life cycle", I*nternational Journal of Software Engineering & Applications (IJSEA)*, vol. 5, no. 5, September 2014, Available: http://airccse.org/journal/ijsea/papers/5514ijsea02.pdf, [Accessed January, 1, 2018].

[6] A. Logan. "Regulating the mobile app market", *Regulation*, Washington, vol. 37, no. 3, 2014, pp. 13-14.

[7] P. LaBerge, *The 2017 Mobile Growth Handbook*, 2017. Available: https://branch.io/attachment/2017-Mobile-Growth-Handbook.pdf, [Accessed January, 1, 2018].

[8] H. Flora, S. Chande and X. Wang, "Adopting an agile approach for the development of mobile applications", *International Journal of Computer Applications (0975 – 8887)*, vol. 94, no. 17, May 2014, doi: 10.5120/16454-6199, Available: https://pdfs.semanticscholar.org/aa3f/4916c79db8118729e1b0bc0039d248e5eb72.pdf, [Accessed January, 1, 2018].

[9] L. Wroblewski, *Mobile First*, Publisher: Jeffry Zeldman Designer: Jason Santa Maria Editor: Mandy Brown Copyeditor: Krista Stevens Compositor: Rob Weychert. ISBN 978-1-937557-02-7, 2011. Available: http://www.ferrispark.com/audio/DOCUMENTS/mobile-first.pdf, [Accessed December, 17, 2017].

[10] E. Alotaibi, A. AlBar and R. Hoque, "Mobile computing security: issues and requirements", *Journal of Advances in Information Technology,* vol. 7, no. 1, February 2016. Available: https://www.researchgate.net/profile/Md_Hoque9/publication/304743361_Mobile_Computing_Security_Issues_and_Requirements/links/5778f36208ae1b18a7e61cda/Mobile-Computing-Security-Issues-and-Requirements.pdf, [Accessed January, 5, 2018].

[11] Available: http//www.w3schools.com

[12] A. Kaikkonen, T. Kallio, A. Keklinen, A. Kankainen and M. Cankar, "Usability testing of mobile applications: A comparison between laboratory and field testing", *Journal of Usability Studies*, vol. 1, no. 1, 2005, pp. 4-16.

[13] B. Campbell, C. Tossell, M.D. Byrne and P. Kortum, "Voting on a smartphone: evaluating the usability of an optimized voting system for handheld mobile devices", *Proceedings of the Human Factors and Ergonomics Society,* 2010*,* pp. 1100-1104, Santa Monica, CA, Available: http://chil.rice.edu/research/pdf/CampbellTossellByrneKortum_HFES_(2011).pdf, [Accessed February, 1, 2017].

[14] P. Kortum and M. Sorber, "Measuring the usability of mobile applications for phones and tablets*", International Journal of Human–Computer Interaction*, vol. 31, no. 8, 2015, pp. 518-529.

[15] K. Blagec, K. M. Romagnoli, R. D. Boyce and M. Samwald, "Examining perceptions of the usefulness and usability of a mobile-based system for pharmacogenomics clinical decision support: A mixed methods study", *PeerJ* 4:e1671, 2016, Available: https://doi.org/10.7717/peerj.1671, [Accessed February, 1, 2017].

[16] R. Harrison, D. Flood and D. Duce, "Usability of mobile applications: literature review and rationale for a new usability model", *Journal of Interaction Science,* 2013, Available: http://www.journalofinteractionscience.com/content/1/1/1, [Accessed February, 1, 2016].

[17] N. Bevan, "International standards for usability should be more widely used", *Journal of Usability Studies.* 4, no. 3, 2009, pp. 106-113.

[18] F. Lettner and C. Holzmann, "Usability evaluation framework", In: Moreno-Díaz R., Pichler F., Quesada-Arencibia A. (eds) Computer Aided Systems Theory – EUROCAST 2011, EUROCAST 2011. Lecture Notes in Computer Science, vol. 6928, Springer, Berlin, Heidelberg, Available: doi https://doi.org/10.1007/978-3-642-27579-1_72

[19] F. Nayebi, J. Desharnais and A. Abran, "The state of the art of mobile application usability evaluation", *25th IEEE Canadian Conference on Electrical and Computer Engineering* (IEEE CD: 978-1-4673-6/12), Montreal, April 29-May 2, 2012.

[20] B. AlGhannam, S. Albustan, A. Al-Hassan, & L. Albustan, "Towards a standard Arabic System Usability Scale (A-SUS): psychometric evaluation using communication disorder app", *International Journal of Human-Computer Interaction, 2017,* Available: https://doi.org/10.1080/10447318.2017.1388099, [Accessed December, 30, 2017].

[21] J. Nielsen, *Usability 101: Introduction to Usability,* 2012, Available: https://www.nngroup.com/articles/author/jakob-nielsen, [Accessed January, 19, 2018].

[22] C. Rohrer, *When to Use Which User-Experience Research Methods,* Nielson Norman Group, 2014, Available : https://www.nngroup.com/articles/which-ux-research-methods/, [Accessed March, 30, 2017].

[23] J. Scholtz, " Usability Evaluation", *National Institutes of Standards and Technology*, 2010, Available : http://cse.yeditepe.edu.tr/~tserif/spring2010/cse489/papers/UsabilityEvaluation_rev1.pdf, [Accessed February, 1, 2016].

[24] K. Sagar and A. Saha, "A systemic review of software usability Studies", *International Journal of Information Technology*, 2017, pp. 1-24, doi.org/10.1007/s41870-.

017-0048-1, Available: https://link.springer.com/article/10.1007/s41870-017-0048-1#citeas, [Accessed February, 1, 2018].

[25] H. Hayat, R. Lock and I. Murray, (2015), "Measuring software usability", *In Proceedings Software Quality Management Conference, BCS*, Available: https://dspace.lboro.ac.uk/2134/18275, [Accessed September, 1, 2016].

[26] J. Sauro and J. Lewis, *Quantifying the User Experience: Practical statistics for user research*, Waltham, MA: Morgan Kaufmann. 2012.

[27] J. Brooke, "SUS: A retrospective", *Journal of Usability Studies*. vol. 8, no. 2, 2013, pp. 29-40.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[28] R. Tsopra, J.P. Jais, A. Venot and C. Duclos, "Comparison of two kinds of interface, based on guided navigation or usability principles, for improving the adoption of computerized decision support systems: application to the prescription of antibiotics", *Journal of the American medical informatics Association,* vol. 21, no. 1, 2014, pp. 107–116.

[29] J. Lewis, B. Utesch and D. Maher, "Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability", *International Journal of Human–Computer Interaction,* vol. 31, no. 8, 2015, pp. 484-495.

[30] J. Lewis, " IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for Use", *International Journal of Human-Computer Interaction*, vol. 7, no. 1, 1995, pp. 57-78.

[31] J. Brooke, "SUS: A 'quick and dirty' usability scale", In P.W. Jordan, B. Thomas, B.A. Weerdmeester, & I.L. McClelland (Eds.), *Usability Evaluation in Industry*, pp. 189-194, London: Taylor & Francis, 1996.

[32] J. Lewis and J. Sauro, *Quantifying the User Experience: Practical Statistics for User Research*, Morgan Kaufmann Publishers Inc., San Francisco, CA, 2012.

[33] A. Bangor, P.T. Kortum and J.T. Miller, " Determining what individual SUS scores mean: Adding an adjective rating scale", *Journal of Usability Study*. Vol. 4, no. 3, 2009, pp. 114-123. Available: http://uxpajournal.org/issue/volume-4-issue-3/, [Accessed 10 January, 2017].

[34] J. Lewis and J. Sauro, "The factor structure of the System Usability Scale", *In Proceedings of Proceedings of the 1st International Conference on Human Centered Design: Held as Part of HCI International 2009*. 2009, pp. 94-103. ACM doi: 10.1007/978-3-642-02806-9_12.

[35] O. Erdinç and J. Lewis, " Psychometric evaluation of the T-CSUQ: The Turkish version of the computer system usability questionnaire", *International Journal of Human-Computer Interaction*, vol. 29, no. 5, 2013, pp. 319–326. doi:10.1080/10447318.2012.711702

[36] K. Lohmann, "System Usability Scale (SUS)- An improved German translation of the questionnaire", 2013, Available: https://minds.coremedia.com/2013/09/18/sus-scale-an-improved-german-translation-questionnaire/, [Accessed February 02, 2017].

[37] I. Dianat, Z. Ghanbari and M. Asghari-Jafarabadi, "Psychometric properties of the Persian language version of the System Usability Scale" *Health Promote Prospect,*, vol. 4, no. 1, 2014, pp. 82-89.

[38] B. Blažica and J. R. Lewis, "A Slovene translation of the System Usability Scale: The SUS-SI", *International Journal of Human-Computer Interaction*, vol. 3, no. 2, 2015, pp. 112-117.

[39] A. I. Martins, A. F. Rosa, A. Queirós, A. Silva, and N. P. Rocha, " European Portuguese validation of the System Usability Scale (SUS)", *Procedia Computer Science,* vol. 67, 2015, pp. 293-300. Available: www.sciencedirect.com

[40] A. Borkowska, and K. Jach, " Pre-testing of Polish translation of System Usability Scale (SUS)", In: L. Borzemski, A. Grzech, J. Świątek, Z. Wilimowska, Ed. Information Systems Architecture and Technology: In Proceedings of 37th International Conference on Information Systems Architecture and Technology – ISAT 2016 – Part I, *Advances in Intelligent Systems and Computing*, vol. 521, pp. 143-153, 2016, Springer, Cham, ch. 12.

[41] J. Lewis, "Psychometric evaluation of PSSUQ using data from five years of usability studies", *International Journal of Human-Computer Interaction*, vol. 14, no. (3&4), 2002, pp. 463-488.

[42] M. Deniz and A. Alsaffar, "Assessing the validity and reliability of a questionnaire on dietary fibre-related knowledge in a 450 Turkish Student Population", *Journal of Health, Population, and Nutrition*, vol. 31, no. 4, 2013, pp. 497–503.

[43] J. Sauro, "The challenges and opportunities of measuring the user experience", Journal of Usability Studies, vol. 12, no. 1, 2016, pp. 1-7.

AUTHORS PROFILE

Bareeq AlGhannam is an Assistant Professor in the Computer Science and Information Systems Department, College of Business Studies, the Public Authority for Applied Education and Training in Kuwait. Dr. AlGhannam has a Bachelor of Science in Computer Engineering with a Ph.D. in Software Engineering focused on the realm of Stakeholder Collaboration within software requirements collection. Currently Dr. AlGhannam is conducting various research in Systems Usability Evaluation with emphasis on stakeholders perspectives.

Eiman Alessa is Instructor in the Computer Science and Information Systems Department, College of Business Studies, The Public Authority for Applied Education and Training in Kuwait. Alessa has a Bachelor degree of Science in Computer and Mathematics with minor in Accounting. A master degree in Database design . Currently Alessa is conducting research in Systems Usability Evaluation and payment gateways .

Manal AlMukhaizim is an Instructor in the Computer Science and Information Systems Department, College of Business Studies, The Public Authority for Applied Education and Training in Kuwait. Instructor AlMukhaizim has a Bachelor degree of Science in Computer and Mathematics with minor in Accounting. A master degree in computers developing a 4th generation language applied on Database. Currently AlMukhaizim is conducting research in Systems Usability Evaluation.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Use of Illumination Invariant Feature Descriptor For Face Recognition

Noorjahan Khatoon*
Dept. of Computer Engineering and General Science
Advanced Technical Training Centre (ATTC)
Bardang, East Sikkim, India
njk_gtk@yahoo.com

Mrinal Kanti Ghose
Dept. of Computer Applications
Sikkim University
Samdur, Tadong, Gangtok, East Sikkim, India
mkghose2000@yahoo.com

Abstract: **In this paper, an attempt has been made to extract texture features from facial images using an improved method of Illumination Invariant Feature Descriptor. The proposed local ternary Pattern based feature extractor viz., Steady Illumination Local Ternary Pattern (SIcLTP) has been used to extract texture features from Indian face database. The similarity matching between two extracted feature sets has been obtained using Zero Mean Sum of Squared Differences (ZSSD). The RGB facial images are first converted into the YIQ colour space to reduce the redundancy of the RGB images. The result obtained has been analysed using Receiver Operating Characteristic curve, and is found to be promising. Finally the results are validated with standard local binary pattern (LBP) extractor.**

Keywords—LBP; LTP; SIcLTP; ZSSD; face recognition; texture feature.

I. INTRODUCTION

Humans often use faces to recognize individuals and advancements in computing capability over the past few decades now enable similar recognitions automatically [9]. Early face recognition algorithms used simple geometric models, but the recognition process has now matured into a science of sophisticated mathematical representations and matching processes [17]. The characteristic that makes it a desirable biometric modality is its uniqueness, universality, acceptability and easy collectability. Face recognition can be used for both verification and identification. Its potentiality and applicability in the areas of security and surveillance makes it more lucrative to be studied as biometric modality. Also, its ease of acquisition from a distance via non-contact offers an added advantage over other biometric modalities.

Its use in biometric could provide access control to various internet of things which are used to protect house, property, child and non-adult population from dangerous predators and illegal hazards.

An excellent survey of existing Face recognition technologies and challenges is given by Li. et al [10]. The problems associated with illumination, gesture, facial makeup, occlusion, and pose variations adversely affect the recognition performance. While Face recognition is non-intrusive, has high user acceptance, and provides acceptable levels of recognition performance in controlled environments, the robust face recognition in non-ideal situations continues to pose challenges [17]. This of course is minimized a little by 3D technologies [11]. Sharma et al [11] have given a survey of different concepts and interpretations of biometric quality. To deal with low-resolution face problem, Choi et al [20] demonstrated that face colour can significantly improve the performance compared to intensity-based features. Experimental results show that face colour feature improved the degraded recognition rate due to low-resolution faces by at least an order of magnitude over intensity-based features.

Ahonen et al [21] experimented with chromatic information integrating them with an adaboost learner to address non linearity in face patterns and illumination variations in training databases for facial recognition. Uçar et al [3], presented colour face recognition algorithm by means of fusing colour and local information. Kalaiselvi et al [8] have made face recognition more reliable under uncontrolled lighting conditions by combining the strengths of robust illumination normalization, local texture based face representations, distance transform based matching and kernel based feature extraction and multiple feature fusion.

II. CHALLENGES IN FACE AS A BIOMETRIC MODALITY

One of the biggest challenges faced by human beings is that if the number of unknown faces is very large, it becomes very difficult for anyone to correctly identify the faces [19].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

In this area computers are quite efficient in terms of the high memory space, computational speed, accuracy and diligence. Some of the important challenges in face recognition are described hereunder [16].

(i)Pose variation happening due to subject's movements or camera's angle may result in deteriorating illumination condition and thus affecting the accuracy of face recognition.

(ii)Face occlusion due to presence of beards, glasses or hats causes high variability and poses hindrance in the feature extraction of the important parts of face such as eyes, nose, forehead and mouth. Face features can also be partially covered by objects or other faces present in the scenes.

(iii) Facial expression may influence the quality of an image, affecting the appearance of a face. Such situation also hampers the illumination condition of the images in consideration.

(iv)Illumination variation due to non-uniform lighting conditions may also pose a great challenge in facial recognition system. The stark dazzle and glare makes the process of feature extraction difficult leading to poor pattern identification.

III. Important Feature Extraction Techniques

Literature survey reveals that many recognition techniques involving various methods of feature extraction for biometric authentication have been devised over the years, but none of the techniques proposed are 100% safe and accurate.
The major feature extraction techniques are

a) PCA based approach [4]
b) SIFT based approach [5] and
c) SURF based approach.[6]

Each of them is having their advantages and disadvantages. Therefore further investigation into this field is a continued effort.

IV. Objective of the Present Study

In this paper, an improved illumination invariant feature descriptor has been investigated to extract the colour texture features from facial imageries. Analysis of textures has been an important factor in image processing having many applications such as object recognition, remote sensing and content based image retrieval tasks [2]. It is an integral part of machine vision and texture classification and is the direct implication of object recognition. The present study mainly exploits this issue.

A simple strategy which could be used to analyse image texture is to find changes in texture on a sliding window. Texture features are summed up as scalar values and features are assigned to each of the image pixel pertaining to window centres. For each pixel, a description of the 'texture" depends on the neighbouring pixels. Stochastic textures are usually natural and consist of randomly distributed texture elements, represented by lines or curves [22].

Most of the works that has been carried out so far pertains to the spatial statistics of the image gray level which is closer to the definition of texture.

The performance of different classifiers depends much on the feature data that have been used. The Local Binary Pattern (LBP) is considered to be simple yet efficient and less complex in implementation [15] but has the weakness such as sensitivity to noise. Very often LBP code defined over an image is used to describe the texture as a histogram of that image [23].

$$LBP_{P,R} = \sum_{P=0}^{P-1} s(x)2^{p} \quad where, x = g_p - g_c$$
$$and, S(x) = \begin{cases} 1, x \geq 0, \\ 0, x < 0, \end{cases} \quad (1)$$

where, g_c and g_p ($p = 0, P - 1$) denote the gray value of the centre pixel and gray value of the neighbour pixel on a circle of radius R, respectively, and $P$is the number of the neighbours.

The Local ternary pattern (LTP) is a variant of Local Binary Pattern and is found to be a very powerful feature descriptor [15].

V. Proposed Method

The LBP operator has two major points of weaknesses; firstly, if the images are deformed and the pattern is not uniform, it misses the local structure as it fails to consider the effect of centre pixel. Secondly, the flat image areas are having all pixels approximately the same gray values, the LBP operator will give some bits the value 0 and others the value 1, implying noise being added to these areas. This makes the operator unstable. Thus LBP operator becomes unsuitable for analyzing these areas [14].

Tan and Triggs [15] presented a new texture operator viz., Local Ternary pattern, which is more robust to noise. The problem of noise in the LBP has been resolved by introducing a user defined threshold say, t, to the central pixel and reassigning pixel values in the interval (-1, +1).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

One of the major challenges faced in object recognition is illumination variation. Liao et al. [18] proposed an efficient background subtraction framework that dealt with illumination variation, in which a pixel wise background subtraction algorithm with local patterns on monocular grey scale video sequences is used.

This present work is motivated by the same concept with improvement used in case of object recognition. The illumination invariant descriptor viz., SIcLTP has an improvement over LBP/LTP wherein, the constant value of LTP for obtaining thresholded ternary output is replaced with a value proportional to the intensity of the central pixel which is a predefined scale factor, indicating how much of the central pixel's intensity can be tolerated. Also the radius parameter of SIcLTP is determined by the value of central pixel making it illumination invariant. SIcLTP has an edge over LTP in terms of advantages offered which are:

1. The operator is computationally simple and efficient.
2. The presence of a tolerant value makes it robust in case of noisy images.
3. The scale invariance property makes it more robust to illumination changes.

Mathematically, given any pixel location, (x_c, y_c), SIcLTP encodes it as

$$SIcLTP^{t}_{N,R}(x_c, y_c) = \sum_{b=0}^{N-1} s_t(p_c, p_b) \qquad (2)$$

where,

P_c is the intensity value of the centre pixel,
P_b is that of its N neighborhood pixels
N neighborhood
R radius
$\sum$ denotes concatenation operator of binary strings
t is a scale factor indicating the comparing range

Since each comparison can result in one of three values, SIcLTP encodes it with two bits, and S_t is a piecewise function defined as

$$S_t = \begin{cases} 01, if\ P_b > (1+t)\,P_c \\ 10, if\ P_b < (1-t)\,P_c \\ 00, otherwise. \end{cases} \qquad (3)$$

VI. Facial Feature Extraction

In this paper, the experiment has been conducted using the Indian face database by Jain and Amitabha [12]. Ten instances of thirty male and female facial images making it a total of three hundred facial images have been considered for extracting texture features using a Local Ternary Pattern based texture feature descriptor named Steady Illumination colour Local Ternary Pattern (SIcLTP)as described in Equation (2) and (3) above. It is worth mentioning that the application of the said technique had yielded promising results for Iris images [1]. .

After extracting the features from Faces using SIcLTP, the similarity and dissimilarity between the equal sized images have been tested using the concept of Zero Mean Sums of Squared Differences (ZSSD) proposed by Patil et. al [7].

VII. Experimental Results and Discussion

The experimental results obtained for the above mentioned facial database is plotted in the form of Receiver Operating Characteristic (ROC) curve as the measure of the discriminating power of the classifier or object recognizer, which in turn describes the accuracy of a test to discriminate match and mismatch cases [13]. Some of the sample input images from the database are depicted in Fig 1 below.

Fig. 1: Sample face images from the database

The samples of extracted features using the proposed descriptor viz., SIcLTP, from the facial images are shown in Fig 2 below.

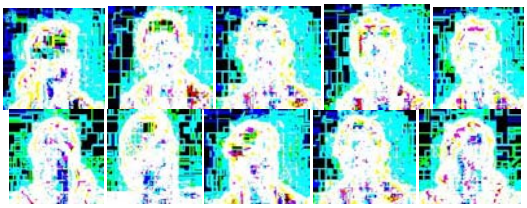


Fig 2 Face features extracted with proposed SIcLTP operator

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Validation of the result obtained has been carried out for the same database using LBP as a feature descriptor as mentioned in Equation (1) above. The process of finding the similarity match is also kept same using ZSSD in this case also.

The samples of the LBP extracted features are shown below in Fig 3.

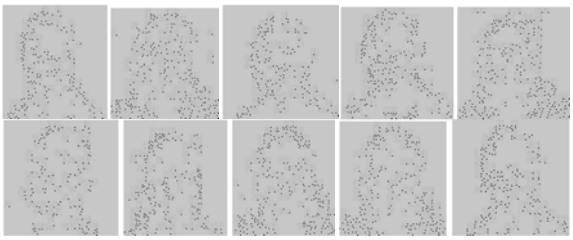


Fig 3 Face features extracted using LBP

The image similarity matching was done at random, picking any image from the database and matching that image with other images in the database at random. The sum of squared differences results in a scalar value which denotes how closely the images compared are similar. The scalar value 0 indicates the exact and symmetrical match and the lowest values indicate the closest and correct matches.

The ROC curves used to plot the results obtained is shown below in Fig 4 for SIcLTP and Fig 5 for LBP.

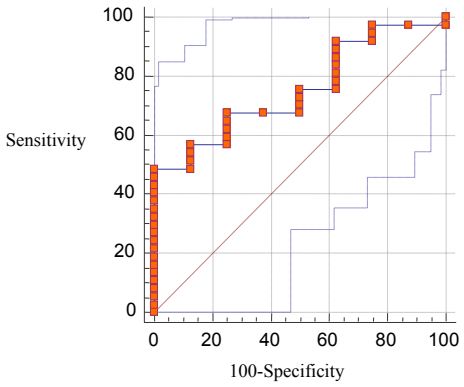


Fig 4 ROC curve using SIcLTP

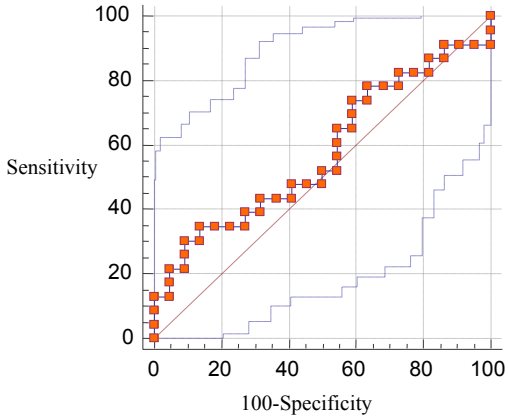


Fig 5 ROC curve using LBP

The tabulation for comparison of the accuracy with respect the SIcLTP and LBP method used has been shown inTable1 below.

TABLE I. COMPARATIVE RECOGNITION ACCURACY

Method	AUC	Correct matches in %
SIcLTP	0.753	82.2%
LBP	0.575	51.1%

The comparison of recognition accuracy makes it evident that the SIcLTP performs better than LBP, as the recognition accuracy using SIcLTP is much higher than using LBP.

VIII. CONCLUSION

In this paper, the experiments have been conducted for Indian Face Database by converting the RGB colour space of the data to YIQ colour space. The proposed SIcLTP operator has been applied. The recognition accuracy has been measured by using ZSSD and the efficiency of the proposed descriptor has been evaluated by using ROC curve. The results obtained are depicted in Table and Figures above. It is worth mentioning that the accuracy of the proposed descriptor is 82% in comparison to LBP being 51% only. The experiment conducted thus demonstrated the effectiveness of the operator SIcLTP, as a feature extractor for Face modality. Further, the Face modality could be used in context to the fusion of modality with other biometric traits to further enhance the accuracy in a multimodal scenario.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

REFERENCES

[1] Noorjahan Khatoon, Mrinal K Ghose, *Steady Illumination Color Local Ternary Pattern as a Feature Extractor in Iris Authentication*, Proceedings of SAI Intelligent Systems Conference (Intellisys) 2016, Lecture notes in Networks and Systems, Computational Intelligence, Springer International Publishing, Vol. 2, pp. 966-973, 2016.

[2] P. Perner, *Cognitive Aspects of Object Recognition - Recognition of Objects by Texture*, 19th International Conference on Knowledge based and Intelligent Information and Engineering Systems, Procedia Computer Science, Elsevier Publication, Vol. 60, pp. 391-402, 2015

[3] A. Uçar, *Color Face Recognition Based on Steerable Pyramid Transform and Extreme Learning Machines*, Research Article, The Scientific World Journal Volume, Hindawi Publishing Corporation, pp.1-15 , 2014

[4] Sudeep Thepade, Rik Das and Saurav Ghose, *Feature Extraction with Ordered Mean Values for Content Based Image Classification*, Advances in Computer Engineering, Hindawi Publishing Corporation, Vol. 15, pp.1-6, 2014

[5] Minal Mun and Prof. Anil Deorankar, *Implementation of Plastic Surgery Face Recognition using Multimodal Biometrics Features*, (IJCSIT), Vol. 5, Issue 3, pp. 3711-3715, 2014

[6] Priyanka and Y. Singh, *A study on facial feature extraction and facial Recognition Approaches*, International Journal of Computer Science and Mobile Computing, Vol. 4, Issue 5, pp. 166-174, 2014

[7] S. Patil, J. S. Nadar, J. Gada , S. Motghare and S. S Nair, *Comparison of Various Stereo Vision Cost Aggregation Methods, International Journal of Engineering and Innovative Technology*, (IJEIT) Volume 2, Issue 8, pp. 222-226, 2013

[8] P. Kalaiselvi and S. Nithya, *Face Recognition System under Varying Lighting Conditions*, IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 14, Issue 3,pp. 79-88, , 2013

[9] Tian J. Hu, Q. Ma, M. X. Ha, *An Improved KPCA/GA-SVM Classification Model for Plant Leaf Disease Recognition,* Journal of Computational Information Systems, Binary Information Press, Vo. 8, Issue 18, pp. 7737-7745, 2012

[10] Stan Z. Li, Face *Recognition: Methods and Practice*, CBSR and NLPR, ICB Tutorial Delhi, India, pp.1-186, 2012

[11] P. B. Sharma and M. M. Goyani , *3D Face Recognition Techniques - A Review*, International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 1, pp.787-793, 2012

[12] Jain Vidit and Mukherjee Amitabha, *The Indian Face Database*, IIT Kanpur, 2002, Accessed/Downloaded January 2012

[13] Z. Zhang and Y. Huang, *A Linear Regression Framework for the Receiver Operating Characteristic (ROC) Curve Analysis*, Open Access Research Article in Biometrics & Biostatistics, Vol. 3, issue 2, pp. 2-7, 2012

[14] C. Zhu and R. Wang, *Local Multiple Patterns based Multiresolution Gray-Scale and Rotation Invariant Texture Classification*, Information Sciences, Vol. 187, pp. 93-108, 2011.

[15] X. Tan and B. Triggs, *Enhanced Local Texture Feature sets for Face Recognition under Difficult Lighting Conditions*, IEEE Transactions on Image Processing, Vol. 19, Issue 6, pp. 1635-1650, 2010

[16] Biometrics Market Intelligence http://www.Biometricsmi.com. Accessed January 2010

[17] P. Buyssens and M. Revenu , Fusion *Levels of Visible and Infrared Modalities for Face Identification*, BTAS, Washington, US, Vol. 6, pp.1-6, 2010

[18] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikainen, and S. Z. Li, *Modelling Pixel Process with Scale Invariant Local Patterns for Background Subtraction in Complex Scenes*, Center for Biometrics and Security Research & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Machine Vision Group, University of Oulu, Finland, pp.1-6, 2008.

[19] Annu Rev Neurosci., Mechanism *of face perception*, doi: 10.1146/annurev.neuro.30.051606.094238, pp. 411–437, 2008.

[20] J. Young Choi, S. Yang, Y. M Ro and K. N Plataniotis, *Color Effect on the Face Recognition with Spatial Resolution Constraints*, Tenth IEEE International Symposium on Multimedia, IEEE Computer Society, pp. 294-301, 2008

[21] T. Ahonen, A. Hadid, and A. Pietikainen, *Face Description with Local Binary Patterns: Application to Face Recognition*, IEEE TPAMI, Vol 28, Issue 12, pp.1-15, 2006

[22] A. Materka and M. Strzelecki, *Texture Analysis Methods - A Review*, Technical University of Lodz, Institutes of Electronics, Cost B11 Report, Brussels, pp. 1-33, 1998

[23] T. Ojala, M. Pietikäinen and D. Harwood, *A Comparative Study of Texture Measures with Classification based on Feature Distributions*, Pattern Recognition. Vol. 29(1), pp. 51–59, 1996

Control algorithm to improve the efficiency of Solar Tracking System in Sultanate Oman

Rustom Mamlook[1], Omar Badran[2], Salem Nijmeh[3], Salah M. Abdallah[4], Abdullah Aljumah[5], Ismail Mohamed Ali Arafat [6]

[1]Electrical and Computer Department, Dhofar University, Salalah, Sultanate Oman.

rstmamlk@hotmail.com

[2]Department of Mechanical Engineering, Al-Balqa` Applied University, Amman Jordan.
[3]Department of Mechanical Engineering, The Hashemite University, Zarqa 13115, Jordan.

[4]Department of Mechanical and Industrial Engineering, Applied Science University, Amman 11931 Jordan.

[5]Dean of College of Computer Engineering & Sciences, Prince Sattam Bin Abdul-Aziz University, Saudi Arabia.

[6]Department of Electrical Technology, Alkharj Technical College, Al-Kharj, Saudi Arabia.

Abstract-- The aim of this research is to utilize the new control algorithm of the sun tracker and the developed computer capabilities to improve the efficiency of tracking. The new tracking method installed on new innovative approach of water distillation taking advantage of high possible concentration of parabolic trough collector to reach a new level of daily harvest per square meter. Water distillation yield is predicted to score high percentage output of distillate due to the high temperature average about 40 degrees as maximum and 30 degrees as minimum. Also the high sunny hours about 9-12 hours per day. Mechanical system will be designed and tested for high ability to withstand the extra loading also some imperfections are forecasted. The present study may found more reliable and trusting techniques in tracking and water distillation. Saline water distillation as predicted will score a noticeable level because of the use of parabolic collector and promoted the efficiency. Keeping good temperature difference between vapor and condensation surface will increase the output and reduce the capacity of temperature. The mechanical design must be convenient to Sultanate Oman climate conditions and have to running, smoothly and safe.

Keywords: New control algorithm; Sun tracker; Improve the efficiency of tracking; New innovative approach of water distillation; Parabolic trough collector.

Introduction

Research Problem

Oman is one of the countries that face shortage in fresh water sources. Recent years the rapid increase on fresh water demands affected the consumers and the government deeply, so that the search of new sources becomes fact that must be taken seriously. Distillation of waste-water or sea water is one of the steps to get the fresh water. Renewable energy that scores the two advantage of

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

reducing the oil usage and distillation of the waste water is one of the main purposes of this research; this will present the chances and measures for the appropriate usage of renewable sources and highlighting the importance of achieving distillate water to be used for industrial process and medication. In this research we will focus on utilization of the solar energy in distillation.

Solar distillation has the advantage of cost saving over other types of distillation such as reverse osmosis, because solar energy is limitless and easily available and likewise seawater is readily available, there is an abundance of these sources. Solar distillation has proved to be highly effective in cleaning up water supplies to provide safe drinking water (Ghosh.1991). As energy requirement to produce 1 liter (i.e. 1kg since the density of water is 1kg/liter) of pure water by distilling brackish water requires a heat input of 2260kJ. Distillation is therefore normally considered only where there is no local source of fresh water that can be easily offered (Malik, 1982).

Research Objectives

Extensive fossil fuel consumption in almost all human activities led to some undesirable phenomena such as atmospheric and environmental pollutions, which have not been experienced before in known human history. Consequently, global warming, greenhouse effect, climate change, ozone layer depletion and acid rain terminologies started to appear in the literature frequently (Guthrie, 2003). Solar radiation is an integral part of different renewable energy resources. It is the main and continuous input variable from practically inexhaustible sun. Solar energy is expected to play a very significant role in the future especially in developing countries such middle east.

Solar distillation application for the communities living in arid areas of almost places of Sultanate Oman is recommended due to the shortage of potable water and due to its simple technology and low cost, which can be easily adopted by local rural people. Solar distillation can be used to convert the available saline or brackish water into potable water economically.

Sultanate Oman has high solar radiation as far as utilization of solar energy is concerned. Also Saudi has an excellent mean solar radiation on horizontal surfaces of 2200 thermal kilowatt hours (kWh) per square meter while other near country such as Jordan has 5.5 – 6 kWh/m2/day and that of Europe and most of North America, which amounts to 3.5 kWh/m2/day — i.e., Also the solar isolation in Jordan occurs for about 2600-4200 sunshine hours in a year [39] (EL-Mulki,1986).

Recently different designs of solar still have emerged. The single effect solar still is a relatively simple device to construct and operate. However, the low productivity of such solar still leads one to look for ways to improve its productivity, and efficiency.

Many design variations exist, and a wide variety of construction materials are used. The amount of distilled water that can be produced varies quite dramatically with the geographical position, the sun's position, prevailing meteorological conditions, solar still design, and operational techniques. (Al-Hayek and Badran, 2004), (Badran and Al-Tahaineh, 2005), (Badran and Fayed, 2004), and (Badran and Abu Khader, 2007) found that other parameters such as water depth, salinity, black dye, solar insulation, wind speed and direction have an effect on the output of the solar stills.

(Abu Khader et al, 2008) found that the sun tracking methods can increase the solar still capability to capture more solar energy to be used later for higher production. These studies were behind the idea of conducting this work, and the need for a research student has a Mechatronics background to implement this work is necessary.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

It was found that for solar concentrator systems as well as for radiometric measurement of the solar radiation the tracking of the sun is necessary. The trackers will periodically update the orientation of the device to the actual position of the sun.

The present research is planned to concentrate on a single and two-axis (North-south axes, East-West axes). Electro-mechanical sun tracking system will be designed and constructed for solar stills applications. The measured variables will be compared with that at fixed axis still system. (Abu-Khader et al, 2008) found that the multi-axes sun tracking (MAST) system can be applied to all types of solar systems to increase their efficiency. While multi-axes sun tracking of the parabolic trough solar stills has not seen the intensive research and development activity; however, some researchers investigated the effect of using MAST systems controlled by a modern computerized control system such as PLC for PV and electric generation systems. With the tracking system, different types of passive solar stills may be used in parallel with the parabolic trough still as follows. The conventional single basin solar still (also known as roof type) is the simplest and most practical design for an installation to provide distilled drinking water for daily needs. It is suggested that the following conventional solar stills (CSS) can be used;

a. Symmetrical double-sloped

b. Nonsymmetrical double-sloped

c. Single-sloped

The choice among the three configurations depends on location, local expertise and the materials available for construction of the system.

Effects of the system design and climatic parameters, on the performance of the system will be investigated. It has been established that the overall system efficiency in terms of daily distillate output will increase by decreasing the water depth and the use of latent heat of condensation for further distillation. Further, increasing the temperature difference between the evaporating and the condensing surface can increase the daily distillate output of passive solar through the trough pipe. The condition can be achieved either by increasing the evaporating surface temperature or decreasing the condensing surface temperature or combination of both. Feeding the thermal energy into the basin from external source can increase the evaporating surface temperature. The water can be heated during sunshine hours and most of the thermal energy is stored in water mass (Sukhatme, 1991) (Duffie and Beckman, 1991).

The objectives of the entire study which try to achieved can be summarized in the following points:

- Design and implement a sun tracker for different distillation systems.
- Design and implement new innovative parabolic distillation system.

Research Significance

Solar energy exists everywhere; the efficiency of any solar system is directly proportional to the solar radiation fall on it. Maximizing the solar system performance is the main target of using solar tracker; there comes the importance of solar tracker. The main reason to use a solar tracker is to reduce the cost of the energy you want to capture. A tracker solar system produces more power over a longer time than a stationary system with the same area faced the sun. This additional output or “gain” can be quantified as a percentage of the output of the stationary system. Gain varies

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

significantly with latitude, climate, and the type of tracker you choose as well as the orientation of a stationary installation in the same location.

Climate is the most important factor. The more sun and less clouds, moisture, haze, dust, and smog, the greater the gain provided by trackers. At higher latitudes gain will be increased due to the long arc of the summer sun. In the cloudiest, haziest locations the gain in annual output from trackers can be in the low 20 percent range.

In many regions of the world, especially Middle-East, desalination has become a most reliable source of fresh water. The different methods used in desalination are based on thermal or membrane principles (Sayigh.1986). Among the thermal methods used is solar distillation, interest in solar distillation stems from the fact that areas of fresh water shortages have plenty of solar energy (i.e. Sultanate Oman). Moreover, it's low operating and maintenance costs made it an attractive method in areas away from the electricity grid lines. But most of them suffer from low productivity which put forward an initiative to look for ways to enhance its productivity and efficiency.

In the present research different designs of solar stills (i.e. cylindrical parabolic and simple solar stills) will be coupled with an innovative electro-mechanical sun tracking system to enhance the productivity. The new design of the sun tracking system will produce a significant enhancement in the still productivity, due to its capability to capture more solar radiations.

Based on the previous researches (Abdallah and Badran,2008) (Nayfey et al,2006) (Samee et al,2007) (Tiwari et al, 2003) on solar stills, it may be concluded that there are a limited number of previous studies published on the performance of sun tracking parabolic solar stills. Furthermore, the results published are very brief and of limited scope. In this study the solar still productivity will be modeled and a developed energy balances technique will be investigated for the new designs, also the thermal capacity of still elements will be accounted for the calculations. Moreover, performance analysis will be conducted under a wide range of parameters. The numerical simulations using mathematical analyses will be compared with the experimental results under different weather conditions for Amman city, in addition to different geometric and flow conditions.

Literature Review of the Research

The concept of sun tracking relays on identifying the location of the sun relative to earth at all times during the day. The rotation of the earth around itself causes the sequence of day and night where its rotation around the sun causes the variation of day and night lengths.

Early researches by (Neville, 1978) and (Hession and Bonwick, 1984) discuss the sun tracking mathematically and the multi usage of sun tracker coupled with collectors.

Many researchers devoted their study to use sun tracking systems as improvement factor yields increase in power. (Roth et al, 2004) designed and built an electromechanical system to follow the position of the sun based on four –quadrant photo detector sensor forming closed loop servo systems.

(Abdallah, 2004) study different types of trackers to investigate the effect in the voltage-current characteristics in the output power of PV panels, four types of trackers (two-axis, east–west, vertical and north–south) gains increase in the output power by (43.87%, 37.53%, 34.43% and 15.69%) respectively.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Also (Abdallah and Nijmeh, 2004) designed two-axis sun tracker based on open loop controller to investigate experimentally the effect of using sun tracking system, the result was a 41.43% increase in the collected power as compared with fixed surface collector with tilt angle 32°.

The new algorithms in Artificial Intelligence (AI) i.e. fuzzy and neuro-fuzzy also used in solar energy environment. (Alata et al, 2005) demonstrates the design and simulation of controller using first order sygeno fuzzy inference system, with full simulation in MATLAB – virtual reality toolbox.

(Al-Mohamed, 2004) achieved 20% increase in the output power of PV panel due to the use of automatic closed loop sun tracker using photo resistance as sensors, the controller was PLC with computerized monitoring capabilities through Recommended Standard 232 (RS232).

Another study by (Bakos, 2006) based on design and construction of a sun tracking system for parabolic trough, the study aims to investigate the continuous operation of two-axis tacking effect in the collected power, and the result showed that sun tracking increased the output by 46.46% using closed loop system.

(Abu-khader et al, 2008) investigated experimentally the effect of using multi-axis sun tracking on Flat Photovoltaic System (FPVS) to evaluate its performance under Jordan climate, the tracker based on time varying system – open loop system – on other words it doesn't use sensors, their result showed that an overall increase of about 30-45% in the output power was achieved.

(Lakeou et al, 2006) designed low cost 0.9 kW photovoltaic system with solar tracking system interfaced with 1 kW wind turbine. The control circuit made of low cost logic circuit to track the maximum sun radiation, but it is not easily adjustable for different climates.

(Abdallah and Badran, 2008) deployed sun tracking system for enhancing solar still productivity, the computerized tracker is an open loop controller based on time as the main variable to control the orientation of solar still, and they found a noticeable increase in the productivity of around 22% with an increase in the overall efficiency of 2%.

(Tomson, 2008) tested the high latitude angle – i.e. 60º- in the North European regions with low solar radiation levels, with comparison of continuous tracking and discrete two-positional tracker, the result shows the effect of using discrete systems in energy saving with increase in seasonal energy by 10-20%.

(Rubio et al, 2007) presented a control application that able to track the sun with high accuracy without the necessity of precise procedure or recalibration, the tracker is hybrid system with combination of open loop and dynamic closed loop, taking energy saving factors in considerations.

In astronomical studies researchers depend on the accurate evaluation of the sun angels. (Grena, 2008) proposed a new algorithm for accurate sun angles determination, his result indicates high precision tolerance around 0.0027 º over the period 2003-2023.

(Ming and Frank, 2004) applied image segmentation to detect sun flare properties and use it in sun tracking purposes, center of flare and boundaries and filtering are some of feature analyzed by image segmentation.

Solar Concentrators – parabolic trough

Parabolic trough technology is currently the most proven solar thermal electric technology in the world (Naeeni and Yaghoubi, 2007a). This is due to the nine large commercial scale solar power plants installed in USA (Price, 1999) (Yaghoubi et al, 2003)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Many researches were developed their study model based on the available solar parabolic plants (Yaghoubi et al, 2003) (Naeeni and Yaghoubi, 2007a) (Naeeni and Yaghoubi, 2007b). These researches aim to proof the cost efficient, improvement and investigating the factors that affecting the production rate and amount. (Price, 1999) found the cost for U.S market about 5.5¢/kWh based on advanced combined-cycle technology, some of authors based on the effect of the wind and thermal that affecting parabolic trough performance (Naeeni and Yaghoubi, 2007a).

(Geyer et al, 2002) investigated the effect of different direction of force in the parabolic collector, these forces applied to different types of parabolic structures also customize the cost of the system based on Euro trough types with new closed loop sun tracking system.

(Price, 2003) developed a model of parabolic trough solar power plant to help plants designers for the best optimization. Numbers of parabolic plants with different configuration are considered in his research to achieve the integration of system capital and operational and maintenance cost.

Research Methodology

The research is based on design and manufactured new system of tracking and water distillation. The system is being tested mechanically and electrically, also the results gathered are proofed. Many factors are taken in consideration such as water temperature, average temperature, average radiation, water depth… etc. Saline water distillation as predicted scored a noticeable level because of the use of parabolic collector and promoted the efficiency.

In our work, two pyrometers (Kipp & Zonen) are mounted on the two-axis advanced tracking and fixed Photovoltaic (PV) modules. The modules are connected to a variable resistor. Measurements of current, voltage output and radiation are recorded and stored into a computer. The data presented in this work is for a typical day in June in Sultanate Oman. Figure 1 shows the I-V characteristics curves for the two cases. The measured solar radiation values in W/m^2 are shown in Figure 2. It can be seen from the figure that the pattern of hourly variations is typical of a cloudless day, and that largest gains occur early and late in the day. The maximum power output of the PV panels is shown in Figure 3. The results (Figures 2 and 3 are similar to the results in (Mamlook et al, 2016) for Saudi Arabia.

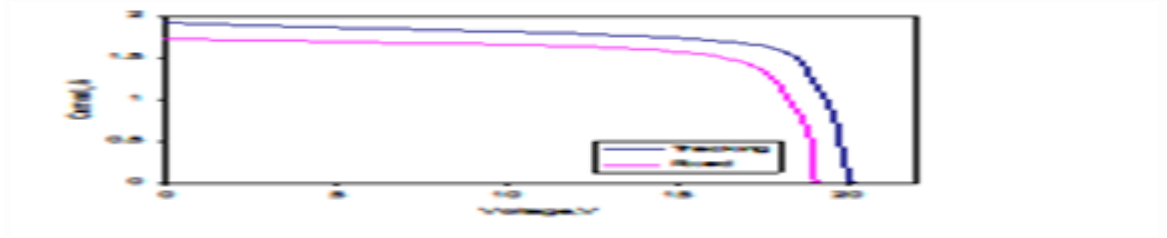


Fig. 1. I-V characteristics for the two cases of the PV modules.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

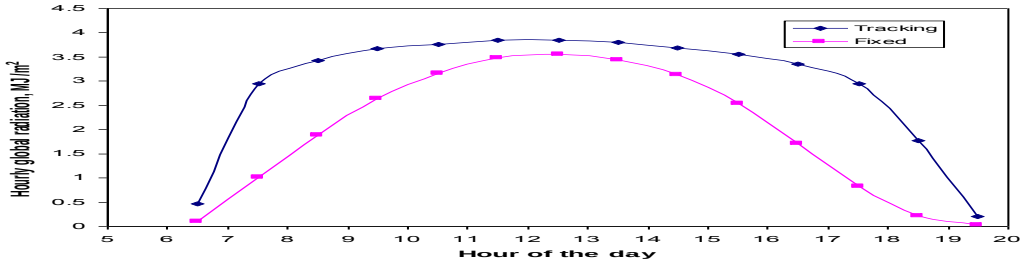


Fig. 2. Hourly global solar radiation comparison.

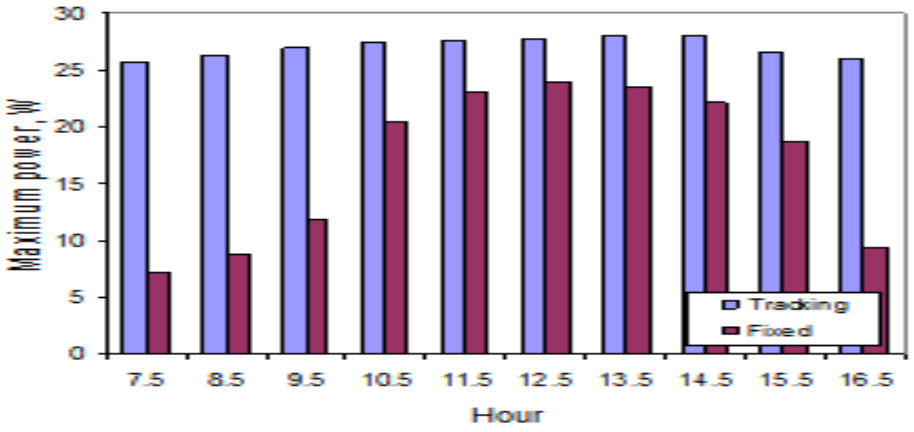


Figure 3. Power (Maximum) of PV panels mounted on moving and tracking surfaces

Comparison is made between advanced tracking and fixed surfaces based on percentage gain in daily radiation and power output as shown in Table 1. It can be noted that the gains are considerable and reach 45.9% and 48.7%, respectively, which can be used in solar distillation system [6 - 7], and we are working on new tracking method installed on new innovative approach of water distillation taking advantage of high possible concentration of parabolic trough collector [12] to reach a new level of daily harvest per square meter. Water distillation yield is predicted to score high percentage output of distillate due to the high temperature average about 40 degrees as maximum and 30 degrees as minimum. Also the high sunny hours about 9-12 hours per day. By keeping good temperature difference between vapor and condensation surface the output is increased and the capacity of temperature is reduced. The mechanical design is convenient to Sultanate Oman and Oman climate conditions and is running, smoothly and safe.

TABLE 1. MEASURED DAILY TOTAL SOLAR RADIATION IN MJ/M^2.

Date	2-axis tracking	Fixed @ 32 latitude	% Gain
12/06/2017	41.3	27.7	48.7%
11/06/2017	27.5	20.7	32.6%
10/06/2017	34.5	23.6	45.9%
8/06/2017	36.1	26.1	38.1%
Average	34.8	24.6	41.3%

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Conclusions

An electro-mechanical two-axis PV advanced tracker (azimuth and altitude) is designed and built. Two advanced tracking motors were used. One for the joint rotated around the horizontal axis to control β, and the other for the joint rotated around the vertical axis to control γ. β and γ are controlled using advanced fuzzy if-then rules model: a knowledge representation scheme for describing a functional mapping or a logic formula that generalized an implication in two-valued logic.

The system uses two electrically powered motorized actuators to move the PV modules. The actuators are controlled by an advanced programmable fuzzy logic controller (APFLC) device to control the motion of the sun-tracking surface. A program is developed and entered to achieve the required positioning based on solar geometry. A not advanced work has been done in (Mamlook and Nijmeh, 2005). And similar advanced work has been done for Saudi Arabia (Mamlook et al, 2016).

An experimental test is conducted to monitor the performance of the system, and measure the values of solar radiation and maximum power of the moving PV modules in Amman, Jordan. Measurements are compared with those on a fixed surface tilted at 32° oriented towards the south. Preliminary results indicate that the use of two-axis tracking will increase the daily power produced by more than 60 % in summer.

The system is characterized by an advanced not complicated in set-up and controls. It operates smoothly with precise positioning even in adverse weather conditions.

References

1. Abdallah. S and Badran O.O. 2008. Sun tracking system for productivity enhancement of solar still. Desalination 220(2008) 669-676.
2. Abdallah. S and Nijmeh. S. 2004. Two-axis sun tracking system with PLC control. Energy Conversion and Management 45 (2004) 1931–1939.
3. Abdallah. S. 2004. The effect of using sun tracking systems on the voltage–current characteristics and power generation of flat plate photovoltaics. Energy Conversion and Management 45 (2004) 1671–1679.
4. Abu-khader. M and Badran. O and Abdallah S. 2008. Evaluating multi-axes sun-tracking system at different modes of operation in Jordan. Renewable and Sustainable Energy Reviews 12 (2008), 864-873.
5. Alata. M and M.A. Al-Nimr and Yousef Qaroush. 2005.Developing a multipurpose sun tracking system using fuzzy control. Energy Conversion and Management. 46 (2005) 1229-1245.
6. Al-Hayek I, Badran O. O.2004. The effect of using different designs of solar stills on water distillation. Desalination 2004;169(2):121–7.
7. Al-Mohamed. A.2004. Efficiency improvements of photo-voltaic panels using a Sun-tracking system. Applied Energy 79 (2004) 345–354.
8. Badran, O. and Fayed, M.F. 2004. Water distillation by solar energy. Proceedings of International Mechanical Engineering Conference (IMEC2004), December 5-8, Kuwait. Pages 58-65.
9. Badran, O.O. Abu-Khader, M.M. 2007. Evaluating thermal performance of a single slope solar still. J. of Heat Mass Transfer, springer-verlag. Vol 43:985-995

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

10. Badran, O.O. and Al-Tahaineh, H. 2005. The effect of coupling a flat-plate collector on the solar still productivity. Desalination, Vol. 183:137-142.
11. Badran, O. and Fayed, M.F. 2004. Water distillation by solar energy. Proceedings of International Mechanical Engineering Conference (IMEC2004), December 5-8, Kuwait. Pages 58-65.
12. Bakos. G. 2006. Design and construction of a two-axis sun tracking system for parabolic trough collector (PTC) efficiency improvement. Renewable Energy 31 (2006) 2411-2421.
13. Duffie J. and Beckman W. A.1991. Solar Engineering of Thermal Processes. 2nd ed. Wiley Inc.
14. El-Mulki H. 1986. Solar and Wind Energy Potentials in Jordan. Proceedings of the International Seminar on Appropriate Technology in the Fields of Solar and Wind Energy Applications. Royal Scientific Society. Jordan.
15. Geyer. M, Lüpfert. E, Osuna. R, Esteban. A, Schiel. W, Schweitzer. A, Zarza.E, Nava.P, Langenkamp. J and Mandelberg.E. 2002. 11th International Symposium on Concentrating Solar Power and Chemical Energy Technologies, Switzerland.
16. Ghosh. G .1991. Solar Energy: The infinite Source, S.B. Nangia for Ashish Publishing House, New Delhi.
17. Grena. R. 2008. An algorithm for the computation of the solar position. Solar Energy 82 (2008) 462–470.
18. Guthrie. N. 2003. The solar revolution 'the road to tomorrow's power today'. United Arab Emirates. Zayed Intl. Center for Coordination & Follow-Up.
19. Hession PJ, Bonwick WJ. 1984. Experience with a sun tracker system. Solar Energy 32:3-11.
20. Lakeou. S, E. Ososanya, Ben O. Latigo, W. Mahmoud G. Karanga and W. Oshumare. 2006. Design of a Low-cost Digital Controller for a Solar Tracking Photo Voltaic (PV) Module and Wind Turbine Combination System. 21st European Photovoltaic Solar Energy Conference, Dresden, Germany.
21. Malik. A .1982. Solar Distillation. A comprehensive technical text. Pergamon Press
22. Ming.Q and Frank S. 2004. Automatic Solar Flare Tracking Using Image Processing Techniques.
23. Mamlook, R.; Nijmeh, S.; Sun Tracking Programmable Fuzzy Logic Control System, SCSC: Summer Computer Simulation Conference, Philadelphia, USA, July 24-28, 2005.
24. Mamlook, R et. al, Advanced Fuzzy Control innovative solar tracking system for solar distillation system, International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 11, pages: 144-160, November 2016.
25. Naeeni. N and M. Yaghoubi. 2007a.Analysis of wind flow around a parabolic collector (1) fluid flow. Renewable Energy 32(2007) 1898-1916.
26. Naeeni. N and Yaghoubi.M.2007b.Analysis of wind flow around a parabolic collector (2) heat transfer from receiver tube. Renewable Energy 32(2007) 1259-1272.
27. Nafey. S, Mohamad M. A, Sharaf M.A. 2006. Theoretical and experimental study of a small unit for solar desalination using flashing process. Proceeding of the 10th international water technology con Alexandria. Egypt (23-26) March.
28. Neville RC.1978. Solar energy collector orientation and tracking mode. Solar energy 202:7-11.
29. Price, H., and Kearney, D. 2003. Reducing the Cost of Energy from Parabolic Trough Solar Power Plants. ISES International Solar Energy Conference, Hawaii Island, Hawaii National Renewable Energy Laboratory NO. NREL/CP-550-33208.
30. Price. H. 1999. Parabolic trough solar power for competitive U.S. markets. Proceedings of the ASME Renewable and Advanced Energy Systems for the 21st Century Conference. Hawaii.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

31. Roth. P and Georgieve.A and Boudinov.H. 2005. Cheap two-axis sun following device. Energy conversion Manage 2005; 46(7-8):1179-1192.
32. Rubio. F and Ortega.M and Gordillo.F and Martinez.L. 2007. Application of new control strategy for sun tracking. Energy Conversion and Management 48 (2007) 2174-2184.
33. Samee. M, Mirza. U, Majeed. T, Ahmad.N .2007. Design and performance of a simple single basin solar still. Renewable and Sustainable Energy Reviews, Volume 11, Issue 3, Pages 543-549
34. Sayigh.A.1986. Solar' Energy Utilization in the Arab World. Proceedings of The International Seminar on Appropriate Technology in the Fields of Solar and Wind Energy Applications. Royal Scientific Society. Jordan.
35. Sukhatme, S. 1991.solar energy 'Principles OF Thermal Collection and Storage'. McGraw-Hill. 7th Edition.
36. Tiwari, G. N, H. N. Singh, Rajesh Tripathi.2003. Present status of solar distillation. Solar Energy, Volume 75, Issue 5, Pages 367-373.
37. Tomson. T. 2008. Discrete two-positional tracking of solar collectors. Renewable Energy 33 (2008) 400-405.
38. Yaghoubi. M, Azizian.K and Rahi.M. 2003. Use of Shiraz power plant for various alternative applications. Proceeding of the first international energy and environment symposium. Izmir. P (805-11).
39. Internet website: https://weather-and-climate.com/average-monthly-Rainfall-Temperature-Sunshine-region-muscat-om,Oman, 2018

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Prototype Models of IoTs Interoperability

Rachida Ait abdelouahid[1*]
Hassan II University of Casablanca- Faculty of science Ben M'sik, Laboratory of technology of information and modeling, Cdt Driss El Harti, BP 7955 Sidi Othman Casablanca.
Casablanca, Morocco
rachida.aitbks@gmail.com

Abdelaziz Marzak[2]
Hassan II University of Casablanca - Faculty of science Ben M'sik, Laboratory of technology of information and modeling, Cdt Driss El Harti, BP 7955 Sidi Othman Casablanca.
Casablanca, Morocco
marzak@hotmail.com

Nawal Sael[3]
Hassan II University of Casablanca - Faculty of science Ben M'sik, Laboratory of technology of information and modeling, Cdt Driss El Harti, BP 7955 Sidi Othman Casablanca.
Casablanca, Morocco
saelnawal@gmail.com

***Abstract*— An IoT platform is a fusion of physical resources such as connectors, wireless networks, smart phones and computer technologies viz; protocols, web service technologies, etc. the heterogeneity of used technologies generates a high cost at interoperability level. This paper presents a generic meta-model of IoT interoperability based on different organizational concepts such as service, compilation, activity and architectures. This model called M2IOTI, defines a very simple description of the IoT interoperability. M2IOTI is a meta-model of IoT interoperability by which one can build a model of IoT interoperability with different forms of organizations. We show that this meta-model allows for connected objects heterogeneity in semantic technologies, activities, services and architectures, in order to offer a high level at IoT interoperability. We also introduce the concept PSM which uses the same conceptual model to describe each interoperability model already existed. Such as conceptual, behavioral, semantic and dynamic models. We have also proposed a PIM model that regroups all the common concepts between the PSMs interoperability models.**

Keywords*— *Internet of Things (IoT), Interoperability Model, PSM Model, PIM Model, Meta-model.

I. INTRODUCTION

An IoT platform is a fusion of physical resources such as connectors, wireless networks, smart phones and computer technologies, protocols, web service technologies, etc. the heterogeneity of used technologies generates a high cost at interoperability level. This paper aims to present at the first section a PSM model for each existed interoperability models, a PIM model that regroups all the common concepts between the PSMs interoperability models and a Meta-model of IoT interoperability Models based on different organizational concepts such as conceptual, behavioral, semantic and dynamic models. The second section is dedicated to the creation of the models specific to every proposal model of interoperability Such as conceptual, behavioral, semantic and dynamic models. PIM. The purpose of this part is the creation of a high-level model of abstraction highlighting the concepts used in the part procedure PSM to define the models of interoperability of the IoTs. The result of these two phases is a generic meta-model of IoT interoperability based on different organizational concepts such as service, activity, compilation and architectures. This model called M2IOTI(Meta-Model of IOT Interoperability), defines a very simple description of the IoT interoperability shaping in metametamodel MOF and to allow to define the structure of these models. M2IOTI is a meta-model of IoT interoperability by which one can build a model of IoT interoperability with different forms of organizations. This meta-model allows for connected objects heterogeneity in semantic technologies, connectivity and architectures, in order to offer a high level at IoT interoperability. This paper is organized as follows. The second section consists of a balance sheet regarding the already existing interoperability models. The third section is dedicated to present a synthetic study of the interoperability models. The fourth section consists of a dependency graph associated to our contribution section. The fived section is dedicated to present our proposal PIM model of IoT interoperability. The sixth section is dedicated to present the M2IOTI meta-model of IoT interoperability. The seventh section presents the results of this paper. Finally, the last section presents a conclusion of the recapitulative of the study realized and future perspectives.

II. RELATIVE WORKS

Classically, the interoperability is the connection of the people, the data and the diversified systems [1]. Interoperability is of great importance and relevance in large systems and should be seen as a requirement. To be interoperable means to be able to exchange streams of various kinds and to share the elements realized by these flows with confidence in order to carry out an action that is independent of the environment with which these flows exchange [1]. In literature, many works were conducted in this area; most of them propose a model layer in order to define and clarified the term of interoperability. (Tolk et al., 2004) [11] proposed an interoperability model called LCIM composed of six main concepts namely; the conceptual level which presents a common vision of the established world, that is to say an epistemology of which several standards are applied by way of example; DoD architecture framework, UML, MDA and DEVS. The semantic level; it guarantees not only the exchange of data but also their contexts. The unambiguous meaning of the data is defined by common reference models as an example; C2IEDM, PDU, RPR FOM

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

and XML. The technical level; whose physical connectivity is established has allowed the exchange of bits and bytes as an example; TCP / IP, HTTP, SMTP and IIOP. Syntax data can be exchanged in standardized formats, that is, the same protocols and formats are supported as examples; HLA OMT, PDU, XML, SOAP and WSDL. The Dynamic level; which allows not only the exchange of information but also its use and its applicability, ie the knowledge, can be exchanged; the applicability of the information is here defined unambiguously. This level includes not only the knowledge implemented, but also the interrelationship between these elements as an example; UML, WEWS, MDA, DEVS. (Panstar et al., 2012) [10] proposed an interoperability model based on six concepts namely; the Communication level: which focuses on the syntactic part of the data information as a context integration object as an example; data formats, SQL, SOAP and XML markup. The Conceptual level, which focuses on abstraction and modeling of adaptation, generalization and transformation as means of integration as an example; reference styles and models. The Dynamic level; which focuses on the contextual changes of events as integration objects based on a set of standards as an example; UML, OWL and MDA. The Behavioral Level; which emphasizes the ability to match actions with each other and the process as an integration object as an example, the architectural structures specific to the field. The Semantic level that focuses on understanding data information as an integration object without its use that is based on a set of standards as an example; XML, RDF, Schemas, ontologism, semantics and Web technologies. The Connection level; it focus on the network connectivity channel as an integration object as an example; cable, Bluetooth and WI-Fi. (Lappeteläinen et al., 2008) [12] have proposed a model layer of interoperability based on three mainly concepts viz; device, service and information. (Jussi Kiljander et al., 2012) [9] Proposed an interoperability model based primarily on two concepts; connectivity interoperability that mainly covers the layers proposed in the traditional Open System Interconnection (OSI) model from the physical layer to the transport layer. This ensures the transformation of data between devices. However, they are not able to understand the meaning of the data. Semantic interoperability defines the technologies needed to enable communicating parties to share the meaning of information. As signaled in [10], (J. Honkola et al., 2010)[13] have proposes a smart-M3 Interoperability platform based on a blackboard architecture model, the M3(multi –device, multi device, multi domain) is baseline architecture for smart architecture, the M3 concept distinguishes three interoperability levels; device interoperability, service interoperability and information interoperability. The interoperability levels were further elaborated in order to match them better to the development of smart averment and their application. The three levels, device, service and information are quite similar to the levels of the C4IF [14] and [10] three first levels from bottom to top; connection interoperability, communication interoperability and semantic interoperability. (V. Peristeras et al., 2006) [14] have proposed a model interoperability layer based on four mainly levels; connection, communication, consolidation and collaboration. The framework (C4IF) explode the concepts of language, theories, such as the language form, syntax, meaning and use of symbols and interpretation, C4IF maps the linguistic concepts to the interoperability as follows; connection level that explain the ability to exchange signals and channel used as an object of integration without knowing anything about content. Communication level, it defines the ability to exchange data and use information as an object of integration, i.e; format as syntax of data but without knowing the context in which data is used , this level is provided between software entities, e.g., components and services by means of semantics. The collaboration level, offer the ability to act together and uses processes/tasks as an object of integration. This level is achieved between tasks, process and others.

III. PSMs Models Associated to IoT Plateforms

A. *Tolk Interoperabilty Model*

Figure 1 illustrates the PSM model specific to IoT interoperability platforms. The model corresponds to a class diagram. In which each fundamental concept is represented by means of a class and each existing relationship between concepts. It contains fifth main classes, conceptual, semantic, technical, syntactical and dynamic.

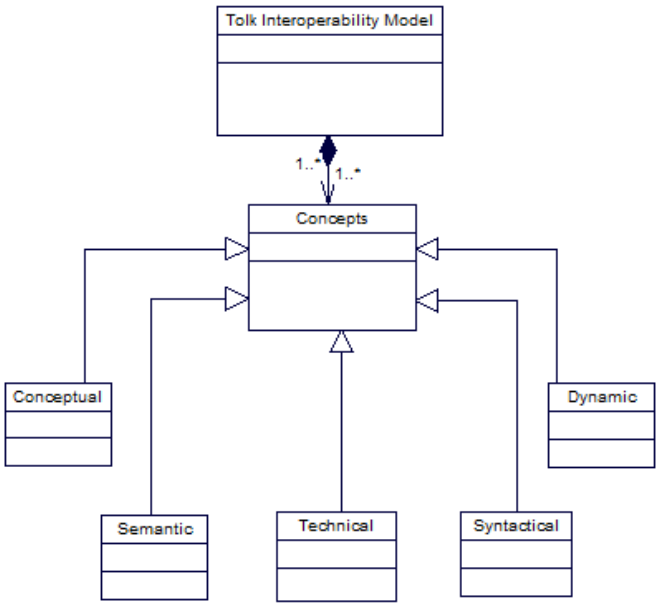


Fig. 1. PSM Interoperability Model of the interoperability defined by (Tolk et al., 2004) [11]

B. *Pantsar Interoperabiliy Model*

Figure 2 illustrates the PSM model specific to IoT interoperability platforms. It contains six main classes, communication, conceptual, dynamic, behavioral, and semantic and connection.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

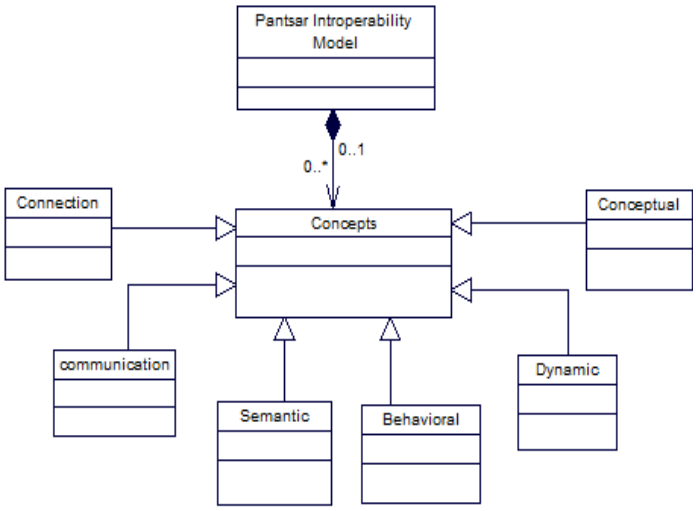


Fig. 2. PSM Interoperability Model of the interoperability defined by (S Pantsar-syvaniemi et al., 2012) [10]

C. *Lappeteläinen Interoperability Model*

Figure 3 illustrates the PSM model. It contains three main classes, device, service, information.

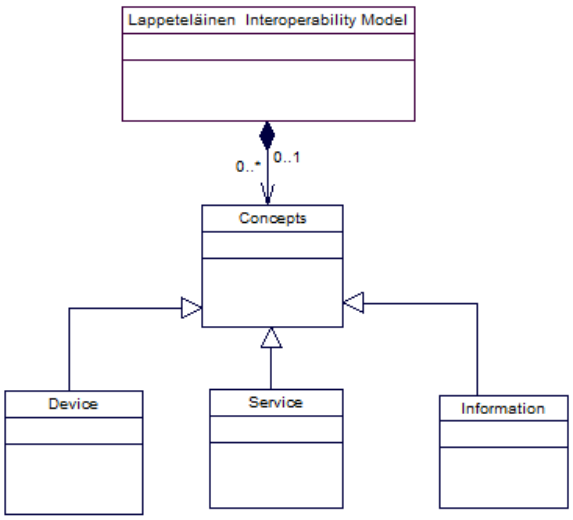


Fig. 3. PSM Interoperability Model of the interoperability defined by (Lappeteläinen et al., 2008) [12]

D. *jussi Interoperability Model*

Figure 4 illustrates the PSM model specific to IoT interoperability platforms. It contains two main classes, connectivity and semantic.

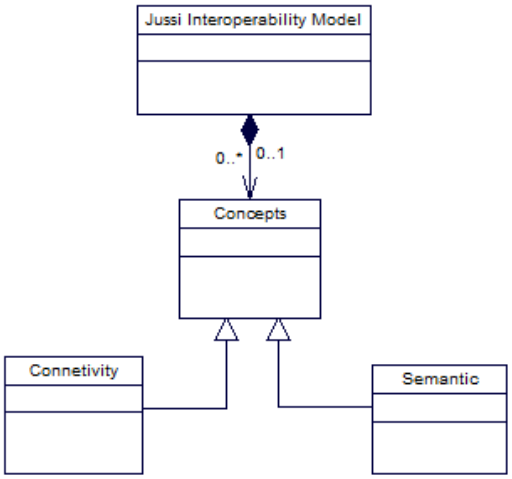


Fig. 4. PSM Interoperability Model of the interoperability defined by (Jussi Kiljander et al., 2012) [9]

E. *J. Honkola Interoperability Model*

Figure 5 illustrates the PSM model specific to IoT interoperability platforms. It contains three main classes, device, service and information.

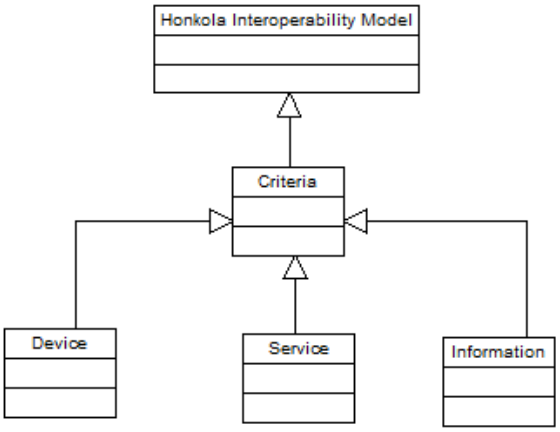


Fig. 5. PSM Interoperability Model of the interoperability defined by (J. Honkola et al., 2010) [13]

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

F. *V.Panstar Interoperability Model*

Figure 6 illustrates the PSM model specific to IoT interoperability platforms. It contains four main classes, connection, communication, consolidation and collaboration.

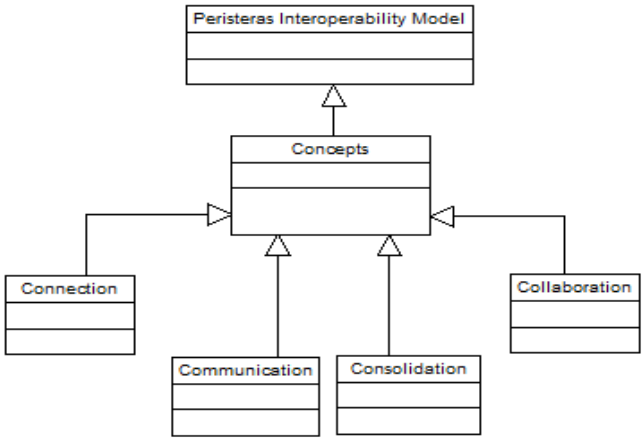


Fig. 6. PSM Interoperability Model of the interoperability defined by (V. Peristeras et al., 2006) [14]

IV. SYNTHETIC STUDY OF PSMS INTEROPERABILITY MODELS

In this section, we present a synthetic study of the IoT interoperability models studied in the previous section in tabular format. To compare these different models, we have based on one mainly characteristic viz; levels. And as we saw in the previous section, the second step is to compare the PSM models representing different interoperability models to move to a higher level of abstraction. Hence this second step, PSM models representing six proposed and revised interoperability models to identify common features and concepts. During the comparison of these models expressed in diagrams of classes UML there are several considered points to know; the concepts, their definitions as well as the relations between them. Once the models specific to the models of interoperability describe and to compare, the third stage is begun. The objective of the third stage is the production of a general model grouping (including) all which is in common between all the PSM models. The table 1 below presents a comparison of the structures of these models.

TABLE 1: SYNTHETIC STUDY OF INTEROPERABILITY MODELS

References / levels	(Tolk et al., 2004) [11]	(Pantsar-Syväniemi et al,2012) [10]	(Lappeteläinen et al. ,2008) [12]	(Jussi Kiljander et al.,2012) [9]	(V.Panstar-Syvanniemi et al. ,2006) [14]	(J. Honkola et al., 2010) [13]
connection		x			x	
technical	x					
syntactical	x					
semantic	x	x		x		
pragmatic/dynamic	x	x				
conceptual	x	x				
behavioral		x				
communication		x			x	
device			x			x
service			x			x
information			x			x
Connectivity				x		
Consolidation					x	
Collaboration					x	

As shown in the table above, we compared the PSMs representing the different interoperability models of IoT platforms, which aim to move to a higher level of abstraction. In this study, the PSM models representing the six interoperability models proposed by ((Tolk et al., 2004) [11], (Pantsar-Syväniemi et al, 2012) [10], (Lappeteläinen et al., 2008) [12], (Jussi Kiljander et al., 2012) [9]), (J. Honkola et al., 2010) [13] and (V. Peristeras et al., 2006) [14] will be reviewed and revised to identify common features and functionality. When comparing these models expressed in UML class diagrams, there are several points to consider knowing: the classes of each model, including relationships between them and the characteristics of each class. Once the interoperability models specific to IoT platforms have been described and compared, we will propose a generic hybrid model grouping all that is common to all PSM models. The table above shows a comparison of the structures of these models, which we draw, the following observations:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- Most of these PSM models include one common class; semantic class, that offers the ability to understand the meaning of the information exchange.
- Most of these PSM models include one common class; communication class, that offers the ability to guarantee the communication between two objects.
- The PSM model proposed by (J. Honkola et al., 2010)[13] based at the same classes used at the PSM model proposed by (Lappeteläinen et al. ,2008)[12].
- The three levels, device, service and information are quite similar to the levels of the [14] and [10] three first levels from bottom to top; connection, communication and semantic interoperability.
- The PSM model proposed by (Tolket al., 2004) [11] and the PSM model proposed by (Pantsar-Syväniemi et al, 2012)[10] are including the three common classes as follows; dynamic interoperability, semantic interoperability and conceptual interoperability.

V. INTERDEPENDENCE OF INTEROPERABILITY CONCEPTS

This section shows the interdependence of interoperability concepts mentioned in the PSM models presented in the section above and the definitions of each interoperability levels.

TABLE 1 : INTEROPERABILITY LEVELS DEFINITIONS

concept	definition
connection	Focus on network connectivity channel as an object on integration [10], The ability the exchange signals and the channel used as an object of integration without knowing anything about content, this level is a prerequisite for any interaction between physical entities.[13]
technical	Physical connectivity is established allowing bits and bytes to be exchange. [11]
syntactical	Data can be exchanged in standardized formats, the same protocols and formats are supported. [11]
semantic	Not only data but also its contexts, information, can be exchanged[11], Focus on understanding data information as an object of integration without its usage[10]
Pragmatic/ dynamic	Information and its use and applicability, i. e. knowledge, can be exchanged[11], Focus on changes of context Events as objects of integration[10], it means that the receiver of the information not only understand its meaning(semantic level), but also what to do with it.[11]
conceptual	Comprises not only the implemented knowledge, but also the interrelations between these elements. [11], Focus on abstraction and modeling Scoping, generalization and transformation as means of integration.[10]
behavioral	Focus on an ability to match actions together process as an object of integration.[10]
Communication	Focus on data information as an object of integration without context[10], The ability to exchange data and use information as an object of integration, i.e. format and syntax of dat.[13]
device	All Interoperable objects.
service	Services exchange between two software.
information	Data exchange between systems.
connectivity	This guarantees the transformation of data between devices.[9]
consolidation	The ability to understand data an its meaning and use information as an object of integration but without knowing how it is used.[13]
collaboration	Ability to act together and uses process/task as an object of integration, its achieved between tasks, processes, etc.[14]

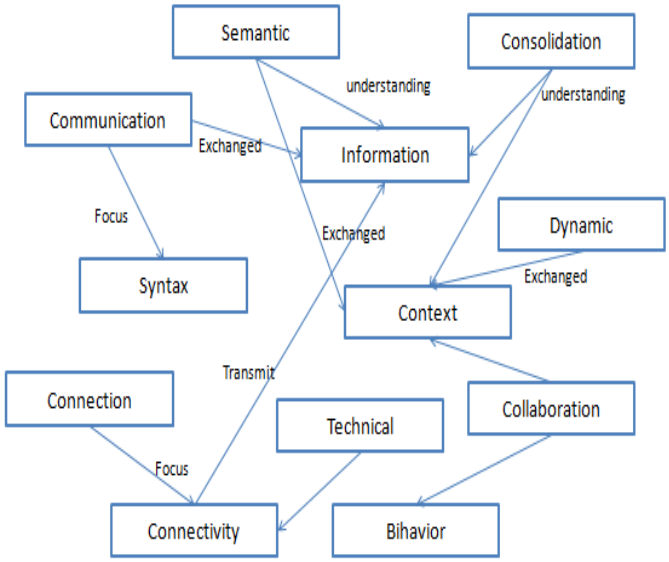


Fig. 7. Interdependence between interoperability concepts

In the figure 7 we have based at the table 2 above that offers the definitions of each interoperability level presented in the proposed PSM models, in order to present the interdependence between all the interoperability levels, as follows;

- When the devices are interoperable at the connectivity level then they are able to transmit data (information) with each other;
- The semantics level interoperability cover the technologies needed for enabling the meaning of information to be shared by communicating parties;
- The semantic level is able to understand and exchange data and their context;
- The connection level is focusing on the connectivity level to be able to exchange signals;
- The communication level is focused on data syntax in order to exchange the data.
- Consolidation level is dedicated to understand data and its context;
- Dynamic level can exchange not only the data but also its applicability and its knowledge;
- Conceptual level comprises not only the implemented knowledge, but also their interrelations;
- Technical level is the physical connectivity established between object to be able to exchanged bits and bytes.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- Consolidation level is dedicated the data and its context.
- Consolidation level has the same definition and utility as the semantic level.
- Collaboration level focuses at the behavior between actors.

VI. PIM INTEROPERABILITY MODEL

The objective of this section is the proposition of a general hybrid model combining all that is common between all PSMs models from the comparison made previously between the PSMs models. To define this model we have to integrated all the features and common features of all PSM models into a common model. Figure 8 illustrates the PIM model .That corresponds to a class diagram. In which each fundamental concept is represented by means of a class and each existing relationship between concepts the help of an association. It contains twelve main classes viz; connectivity, compilation, semantic, architecture, syntactical, communication, information, service, behavior, technical, conceptual and device.

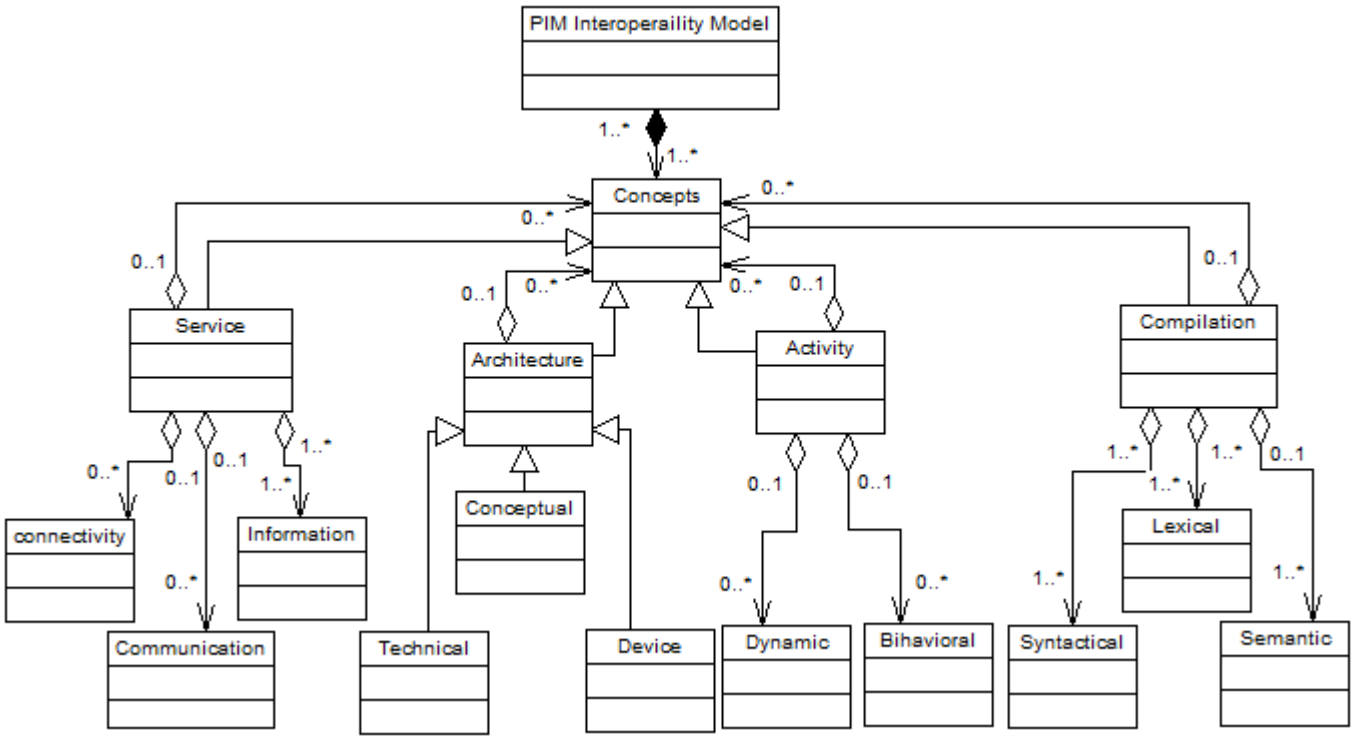


Fig. 8. PIM Interoperability Model

Model general PIM represents In figure above represents a general model. The general structure of model obtained shows clearly that all the model PSMs shares a common structure of the way in him defined the interoperability. The structure of this general model obtained (Figure 8) clearly shows that all PSM models share a structure that takes into account three main classes vis;

- The compilation class: this class groups three levels; lexical level which used to check the syntactic conformity of strings using a grammar, syntactical level that makes it possible to make a syntactic analysis on the information exchanged before doing their semantic analysis and the semantic level which is responsible for giving the interpretation of information and data exchanged between connected objects.
- The activity class: it based on two main concepts viz; dynamic interoperability; means that the receiver of the information not only understand its meaning (semantic level), but also what to do with it and behavioral interoperability which offers the possibility of exchanges the behavior of the connected objects.
- The service class: it combines three concepts; viz connectivity interoperability, information interoperability and communication interoperability. This class is responsible for ensuring communication between IoT platforms, as well as the transfer and retrieval of information exchanged; that is to say, it allows the support of

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the realization of a set of services related to data processing.

- The architecture class: This class is responsible for defining all the technical, conceptual and material needs as well as the structure of all hardware and software resources.

This model is not only used to illustrate the structure of the models PSMs representing the various models of interoperability, but it is also used to show clearly the necessary and sufficient concepts in the realization of such a model.

VII. META MODEL OF IOTS INTEROPERABILITY

We have proposed in this section a new meta-model of high level of abstraction which consists of a set of concept. This meta-model is based on the MOF model that aims to define the key concepts used for modeling interoperability models of IoTs. It breaks down the notion of interoperability of IoTs into three categories; the compilation class that groups together all the concepts related to the semantic as well as the interpretation of the exchanged data and to allow the connected objects to interact as simple users, the service class which gathers all the concepts related to the communication, the information and connectivity, the architecture class which gathers all the technical, material and conceptual concepts and the activity class which gathers all the concepts related to the behavioral and dynamic interoperability. Figure 9 below illustrates the proposed Meta model for interoperability of IoTs. The model corresponds to a class diagram in which each fundamental concept is represented by means of a class and each existing relationship between concepts the help of an association. It contains three main classes, connectivity, semantic, architecture.

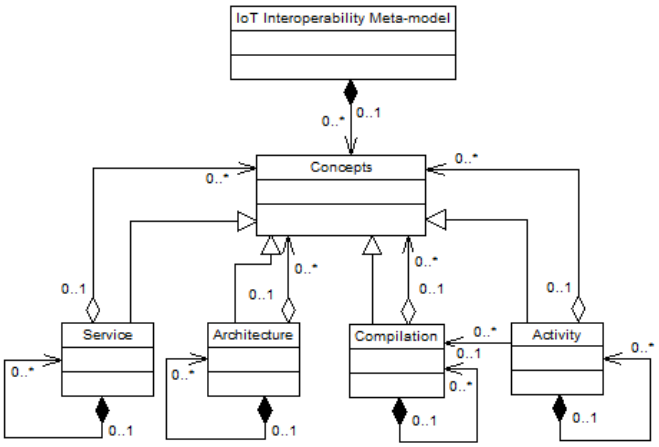


Fig. 9. Meta model proposed for the interoperability of IoTs

VIII. RESULT

The main purpose behind the creation of one Meta model is to allow to model systems belonging to certain domain. What to allow in our case to define all the concepts as well as the relations between them to define the interoperability, this Meta models allows of:

- Cover the concepts: this meta-model highlighted the main elements of the interoperability of the IoTs, as well as their high-level interaction;
- Highlight: the relations associating the diverse elements of the interoperability to specify in the domain of the internet of objects;
- Avoid the redundancy: make group all the elements which are the same senses in a single concept.

So this Meta-model can be used as continuation:

- It considers the absence as meta model for the interoperability dedicated specifically to the domain of the internet of objects (IoT), by offering him a language of modeling using its own terms;
- The definition of the models of the interoperability of the IoTs in a simple way thanks to the predefined structuring imposed by our meta models;
- It also offers a frame of understanding of the models of interoperability;
- A frame of understanding of the models of interoperability.

IX. CONCLUSION AND PERSPECTIVES

In this paper, we have proposed and described in detail the PSM model of different existing interoperability models. Basing on these PSMs models we have proposed a general hybrid model of IoT Interoperability called PIM model, combining all the common concepts between the PSMs models. As a result we have proposed a meta-model of high level of abstraction which consists of a set of concepts namely; service interoperability, compilation interoperability, activity interoperability and architecture interoperability, which corresponds to a class diagram. In which each fundamental concept is represented by means of a class and each existing relationship between concepts. We are planning in our future work to propose a new quality model for evaluating the interoperability of IoT platforms.

REFERENCES

[1] Chapurlat, V., Daclin, N.: System interoperability: definition and proposition of interface model in MBSE Context. In: the Proceedings of the 14th IFAC Symposium on Information Control Problems in Manufacturing Bucharest, Romania, May 23-25, IFAC(2012)

[2] Desai, P., Sheth, A., Anantharam, P.: "Semantic Gateway as Service Architecture for IoT Interoperability". In: the IEEE International Conference on Mobile Services, New York, NY, pp.313-319.(2015)

[3] Blackstock M. and Lea R.: IoT interoperability: A hub-based approach. In: the International Conference on the Internet of Things (IOT), Cambridge, MA, pp. 79-84.(2014).

[4] Ghofrane F.: A distributed and flexible architecture for Internet of Things. In: The International Conference on Advanced Wireless, Information, and Communication Technologies AWICT, pp 130 – 137 (2015)

[5] Aloi, G., Caliciuri, G., Fortino G., Gravina, R., Pace, P., Russo, W., Savaglio, C.: A Mobile Multi-Technology Gateway to Enable IoT Interoperability. In: the First International Conference on Internet-of-

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Things Design and Implementation (IoTDI), Berlin, pp. 259-264. (2016)

[6] Antonio L., Albertia M., Daniel M., Bontempo M., Lucio H. O., Rodrigo R. R., Arismar C. S. Jr.: Cognitive radio in the context of internet of things using a novel future internet architecture called Nova Genesis. In: Computers and Electrical Engineering 57 , pp 147–161 (2017)

[7] George C., Jean D., Tim K., Gordon B.: Distributed Systems: Concepts and Design (5th Edition). In: Boston: Addison-Wesley. ISBN 0-132-14301-1 (2011)

[8] Kashif D., Amir T. k., Harun B., Frank E., Kurt G.: A resource oriented integration architecture for the Internet of Things: A business process perspective. In: Pervasive and Mobile Computing (2015)

[9] Jussi K, Alfredo D, Francesco M, Pasi H, Janne T, Arto Y, Juha-pekka S, (member, ieee), and Tullio S.C.: Semantic Interoperability Architecture for Pervasive Computing and Internet of Things. In: IEEE Access: the journal for rapid open access publishing, Digital Object Identifier .(2014)

[10] Susanna P. S., Anu P., Eila O., Jarkko K., Antti E. : Situation-based and self-adaptive applications for the smart environment .In : Journal of Ambient Intelligence and Smart Environments 4 (2012) 491–516

[11] Dr. Andreas Tolk : Composable Mission Spaces and M&S Repositories - Applicability of Open Standards.In: Spring Simulation Interoperability Workshop (2004)

[12] A. Lappeteläinen, J.-M. Tuupola, A. Palin, and T. Eriksson, ``Networked systems, services and information the ultimate digital convergence," in Proc. 1st Int. NoTA Conf., Helsinki, Finland, 2008, pp. 1_7.

[13] V. Honkola, H. Laine, R. Brown and O. Tyrkko : Smart-M3 interoperability plarform," in Proc. Of the 1st Int. Workshop Semantic Interoperability for Smart Spaces(SISS 2010) in IEEE Sympsium onn Computers and Communications(ISCC 2010), IEEE Computer Society, 2010, pp. 1041-1046.

[14] V. Peristeras and K. Tarabanis, the connection, communication, consolidation, collaboration interoperability framwork(C4IF) for information systems interoperability, IBIS – Interoperbility in Business Information Systems 1(1) (March 2006), 61-72.

Removal Based Improved Replication Control and Fault Tolerance Method for Road Traffic Data Management

Dr Annaluri Sreenivasa Rao[1a], Dr Attili Venkata Ramana[2b], Dr. G. Suresh Reddy[3c], Dr C.Sunil Kumar[4d]
[a)]Assistant Professor in IT,VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India,
[b)]Assoc. Prof in CS, Dept. of ECM, Sreenidhi Institute of Science and Technology, Hyderabad, India,
[c)] Professor , Dept of IT,VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India,
[d)] Professor in ECM, Sreenidhi Institute of Science and Technology, Hyderabad, India
[a)]dr.attilivenkataramana@gmail.com,[b)] annaluri.rao@gmail.com, [c)] ithod@vnrvjiet.in, [d)] ccharupalli@gmail.com

***Abstract*— The massive growth in road traffic and subsequence generation of traffic related data insisting the researcher to proceed for the analytical research on the traffic prediction. However the gigantic size of the data and chances of storage failure may cause the purpose inefficient.The advancement in technologies and high demand for fault tolerant storage solutions most of the cloud based commercial storage service providers are now equipped with Erasure based Reed – Solomon fault tolerance mechanism. However the additional cost for replication is still an overhead for service providers and customers. In this work, we propose a novel erasure based code and further optimization as shortening the proposedcode also for the digital storage formats. The work also results into a comparative study of cost analysis for commercial cloud based storage service providers. Finally the work demonstrates the improvement in code shortening and making the performance higher.**

***Keywords*— Erasure, Reed – Solomon, Code Shortening, Performance Comparison, Evolution Application, Response Time Comparison, Dropbox, Google Drive, Hightail, OneDrive, SugarSync**

I. Introduction

In the past years, the high upcoming demand for storage with high performance and reliability were been understood.The industry was approaching towards a phase where the lack of standardization of digital storagewas limiting the applications to make storage more reliable for commercial storage providers. The major bottleneck for the standardization was the non-standard storage solutions used by different service providers. In the early 80's, the industry adopted cloud computing for distributed storage solutions. The effort was well recognized and multiple companies came together to form a consortium in order to frame the standardization for digital storage.

As far as data storage is concerned, there are multiple schemes are available to improve file and data compression. The other most influencing parameters For instance, a data file that is uploaded and accessed on the server may seriously be effected by the network bandwidth as well as the server workload. This will degrade the efficiency [1]. Moreover the cloud storage services deals with a great scope and domain of the data being storage and retrieved along with the frequency of access varying depending on the mode of the operation performed on the data [2]. Offering unlimited storage container space might cause a high economic drawback on the cloud storage provider and as well as the users due to inefficient storage [3]. Hence, a technique or automation is needed to find the best suitable storage structure based on cost and other influencing factors. There are many free offerings of the cloud storage services; however they may not suite the application requirement to the best always [4].

Two major companies, Philips and Sony took the major initiative to define the standard storage formats in digital media. The standard is well accepted today and been referred as compact storage format. This standard format is majorly used for achieving any data, which also reduces the storage cost compared to the early storage formats. However the compact storage format has limitations in order to achieve high availability. It is difficult to predict how a storage media gets corrupted. In the earlier studies we have understood the reasons for storage device failure. Henceforth we realise the following errors for storage failures as

(1) The additional noise affecting the storage during transmission or during retrieval

And

(2) Mishandling of the removable devices

The most important improvement in the recent time for fault tolerance in digital media storages is the Reed – Solomon code. The basic benefit of the Reed – Solomon codes is to rearrange so that the timely restoration can be achieved for storage devices. Thus in this work we concentrate on further enhancement of the Erasure based fault tolerance mechanism.

The rest of the work is framed such as in Section II we understand the cost effectiveness of the commercial cloud storage solutions, in Section III we realise the basic Reed Solomon Fault Tolerance scheme, in Section IV we propose the novel Reed – Solomon based code, in Section V we propose the further optimization of the proposed code, in Section VI we discuss the implementation and results and in Section VII we conclude the work.

II. Commercial Cloud Storage Services

As the choice of storage services from cloud is not limited and most of those are configured to give best advantages for specific type of data and operation, we compare most of the services here [5 – 7].

A. Dropbox

The Dropbox is a storage service which is available for client side access for Windows systems, Linux Systems, Macintosh systems, Blackberry mobile operating systems, Android mobile operation systems and finally the IPhone operating systems. The free Basic account comes with a paltry

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

2GB of storage. For document based applications this is huge. The Storage service is good choice for applications using the container for read only data.

Table 1. Cost Comparison for Dropbox.

Data Load	Cost
Load in GigaBytes	Price in US Dollars
100	99 USD
200	99 USD
300	99 USD
400	499 USD
500	499 USD
1000	Not Available
> 1000	Not Available

Here we provide a graphical representation of the cost price comparison:

Fig.1. *Cost Comparison for Dropbox*

Table 2. Support for Mobile Based Cloud Applications in Dropbox

Client OS Type	Support
Apple IPhone Operating Systems	Available
Android Mobile Operating Systems	Available
Blackberry Operating Systems	Available
Microsoft Mobile Operating System	Available

B. Google Drive

The most popular cloud storage service is Drive storage from Google. The basic account comes with 15 Giga bytes of storage for a new customer account or an existing account created with Google Email. The highest rated benefit of the Google Drive is the service can be also be integrated with other existing google services for storing various types of data from other services.

Table 3.CostComparisonfor Google Drive

Data Load	Cost
Load in Giga Bytes	Price in US Dollars
100	60 USD
200	120 USD
300	120 USD
400	240 USD
500	240 USD
1000	600 USD
> 1000	1200 to 9600 USD

Here we provide a graphical representation of the cost price comparison:

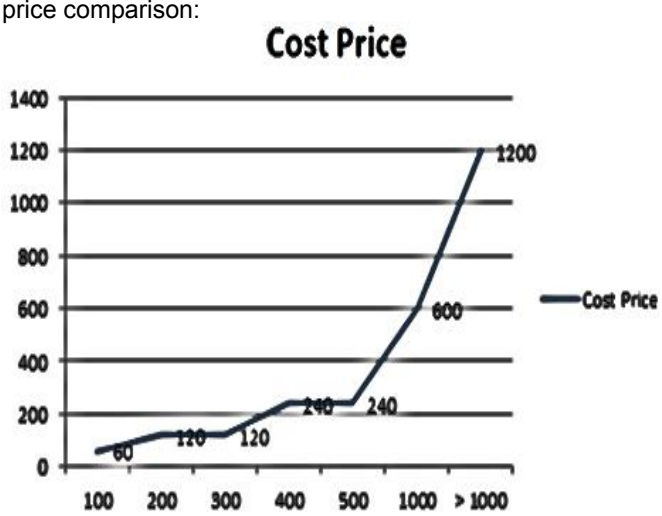


Fig.2. Cost Comparison for Google Drive

Table 4.Support for Mobile BasedCloudApplications in Google Drive

Client OS Type	Support
Apple IPhone Operating Systems	Available
Android Mobile Operating Systems	Available
Blackberry Operating Systems	Not Available
Microsoft Mobile Operating System	Not Available

C. Hightail

The previous version of business cloud storage of Hightail was popular by name of YouSendIt. The basic reason for creating the name was the core of the features that Hightail provides. Hightail is majorly known for sharing files, which can be digitally signed for verifications. The core technology behind this provider is link sharing, where the sender can upload a file and the link to that same file can be shared with the recipient. The recipient can click on the link to download the same. This service is popular for business users as it provides the private cloud storage and the desktop version of

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

the client, which can be used for syncing local files to the cloud storage.

Table 5.CostComparisonforHightail

Data Load	Cost
Load in Giga Bytes	Price in US Dollars
100	Free
200	Free
300	Free
400	Free
500	Free
1000	Free
> 1000	195 USD

Table 6.Support for Mobile BasedCloudApplications in Hightail

Client OS Type	Support
Apple IPhone Operating Systems	Available
Android Mobile Operating Systems	Not Available
Blackberry Operating Systems	Not Available
Microsoft Mobile Operating System	Not Available

D. OneDrive

The OneDrive was previously popular as SkyDrive. The functionalities are mostly same as Dropbox. The most important factor for this storage service is that the client version is available for Windows systems, Linux Systems, Macintosh systems, Blackberry mobile operating systems, Android mobile operation systems and finally the IPhone operating systems. Moreover the supports for social media plug-ins are also available here. This feature makes the application more compatible with other applications to access data directly.

Table 7.CostComparisonforOneDrive

Data Load	Cost
Load in Giga Bytes	Price in US Dollars
100	50 USD
200	100 USD
300	Not Available
400	Not Available
500	Not Available
1000	Not Available
> 1000	Not Available

Here we provide a graphical representation of the cost price comparison:

Fig.3. Cost Comparison One Drive

Table 8.Support for Mobile BasedCloudApplications in OneDrive

Client OS Type	Support
Apple IPhone Operating Systems	Available
Android Mobile Operating Systems	Available
Blackberry Operating Systems	Available
Microsoft Mobile Operating System	Available

E. SugarSync

The SugarSync is majorly popular among business users for its effective and fast online backup solutions. The service can also be used for complete folder and individual file syncing with multiple applications and multiple users. Moreover the service provides a unique function to share the stored content over multiple devices at same point of time but with different permission levels. The most important factor for this storage service is that the client version is available for Android mobile operation systems and also the IPhone operating systems.

Table 9.CostComparisonforSugerSync

Data Load	Cost
Load in Giga Bytes	Price in US Dollars
100	99 USD
200	250 USD
300	250 USD
400	250 USD
500	250 USD
1000	550 USD
> 1000	Pay Per Use

Here we provide a graphical representation of the cost price comparison:

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Fig.4. Cost Comparison for Sugar Sync

Table 10.Support for Mobile BasedCloudApplications in SugerSync

Client OS Type	Support
Apple IPhone Operating Systems	Available
Android Mobile Operating Systems	Available
Blackberry Operating Systems	Available
Microsoft Mobile Operating System	Available

III. Reed – Solomon Code For Fault Tolerance

The most important factor that makes Reed-Solomon framework to implement is the simplicity. Here in this work we consider the scenario to compare the performance of Reed – Solomon and Proposed Encoding technique [8].

We consider there will be K storage devices each hold n bytes of data such that,

$$D = \sum D_1 , D_2 . D_3 .....D_k \quad \text{...Eq1}$$

Where D is the collection of storage devices

Also there will be L storage devices each hold n bytes of check sum data such that,

$$C = \sum C_1 , C_2 , C_3 ....C_L \quad \text{...Eq2}$$

Where C is the collection of Checksum devices

The checksum devices will hold the calculated values from each respective data storage devices.

The goal is to restore the values if any device from the C collection fails using the non – failed devices.

The Reed – Solomon deploys a function G in order to calculate the checksum content for every device in C. Here for this study we understand the example of the calculation with the values as K = 8 and L = 2 for the devices $C_1$and C_2 with G_1 and G_2 respectively [9].

The core functionalities of Reed – Solomon is to break the collection of storage devices in number of words [10] [11]. Here in this example we understand the each number of words is of u bits randomly. Hence the words in each device can be assumed as v, where v is defined as

$$v = (nbytes). \frac{8bits}{byte} . \frac{1word}{u\ Bits} \div \quad \text{... Eq3}$$

Furthermore, v is defined as

$$V = \frac{8n}{u} \quad \text{...Eq4}$$

Henceforth, we understand the formulation for checksum for each storage device as

$C_i = W_i .(D_1 , D_2 , D_3 ...D_k)$ …Eq5 Where the coding function W is defined to operate on each word

After the detail understanding of the Erasure fault tolerance scheme, we have identified the limitations of the applicability to the cloud storage services and propose the novel scheme for fault tolerance in this work in the next section.

IV. Proposed Novel Fault Tolerance Scheme

With the understanding of the limitations of existing erasure codes to be applied on the cloud based storage systems as the complex calculations with erasure codes will reduce the performance of availability measures significantly. Thus we make an attempt to reduce the calculation complexities with simple mathematical operations in the standard erasure scheme.

The checksum for storage devices are considered as C_i from the Eq5. We propose the enhancement as the following formulation for checksum calculation:

$$C_i = W_i .(D_1 , D_2 , D_3 ... D_k) = W_i (D_1 \oplus D_2 \oplus D_3 ... \oplus D_k) \quad \text{...Eq6}$$

Here the XOR operation being the standard mathematical operation most suitable for logical circuits used in all standard hardware makes it faster to be calculated.

Also we redefine the function to be applied on each word for the storage devices D as following:

$$W = \begin{matrix} w_{1,1} & . . . & w_{1,L} \\ . & . . . & . \\ . & . . . & . \\ . & . . . & . \\ w_{K,1} & . . . & w_{K,L} \end{matrix}_{K X L} \quad \text{...Eq7}$$

The proposed matrix will be stored on one of the devices and will be recalculated only once. As the modified checksum formulation is an XOR operation, thus which will automatically notify in case of any change.

Furthermore, we optimize the proposed code framework in the next section.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

V. Optimizing Proposed Novel Fault Tolerance Scheme

The Reed Solomon code is expressed by the power of coefficient denoted by n for the data blocks, where n is expressed as

$$n = 2^m - 1 \quad \text{...Eq 8}$$

and the code blocks are represented as

$$k = 2^m - 1 - 2t \quad \text{...Eq 9}$$

Where m represents the number of bits per data and t represents the capability of correcting errors. In general theReed – Solomon code considers an 8 bit data and 2 bit code, the error correcting code can be represented as (255,251) code.

Here in this part of the work, we try to optimize the code length further to reduce the replication cost. The steps of the optimization algorithm are explained here:

Step-1. First we consider the effective code in (255,251) block, where the code is consisting of zero and non-zero codes.

Step-2. Then we find the number of zero codes in the segment. For instance the numbers of zero codes are227 in the code block. These codes will not have any effect in the error correction and fault tolerance mechanism.

Step-3. Then we find the effective block of the code as (28,24) for a 2 bit error correction code.

Step-4. Hence as a final outcome of the optimization technique, we got the optimized code block.

VI. Implementation and Results

To simulate and understand the improvement in the outcomes we implement the Reed – Solomon code with the enhancement and optimization proposed in this work.

We accept any random data as the initial data block for the testing [Table -11].

Table 11.Initial Data Block

0 0 0 0
1 0 0 0
0 1 0 0
0 0 1 0
0 0 0 1
1 1 0 0
0 1 1 0
0 0 1 0
1 1 0 1
1 0 1 0
0 1 0 1
1 1 1 0
0 11 1
1 1 1 1
1 0 1 1
1 0 0 1

Based on the modified fault tolerance scheme, we realise the addition and multiplication table [Table -12 & 13].

Table 12.Addition Table

0 a^0 a^1 a^2 a^3 a^4 a^5 a^6 a^7 a^8 a^9 a^10 a^11 a^12 a^13 a^14
0 |0 a^0 a^1 a^2 a^3 a^4 a^5 a^6 a^7 a^8 a^9 a^10 a^11 a^12 a^13 a^14
a^0 |a^0 0 a^4 a^8 a^14 a^1 a^10 a^13 a^9 a^2 a^7 a^5 a^12 a^11 a^6 a^3
a^1 |a^1 a^4 0 a^5 a^9 a^0 a^2 a^11 a^14 a^10 a^3 a^8 a^6 a^13 a^12 a^7
a^2 |a^2 a^8 a^5 0 a^6 a^10 a^1 a^3 a^12 a^0 a^11 a^4 a^9 a^7 a^14 a^13
a^3 |a^3 a^14 a^9 a^6 0 a^7 a^11 a^2 a^4 a^13 a^1 a^12 a^5 a^10 a^8 a^0
a^4 |a^4 a^1 a^0 a^10 a^7 0 a^8 a^12 a^3 a^5 a^14 a^2 a^13 a^6 a^11 a^9
a^5 |a^5 a^10 a^2 a^1 a^11 a^8 0 a^9 a^13 a^4 a^6 a^0 a^3 a^14 a^7 a^12
a^6 |a^6 a^13 a^11 a^3 a^2 a^12 a^9 0 a^10 a^14 a^5 a^7 a^1 a^4 a^0 a^8
a^7 |a^7 a^9 a^14 a^12 a^4 a^3 a^13 a^10 0 a^11 a^0 a^6 a^8 a^2 a^5 a^1
a^8 |a^8 a^2 a^10 a^0 a^13 a^5 a^4 a^14 a^11 0 a^12 a^1 a^7 a^9 a^3 a^6
a^9 |a^9 a^7 a^3 a^11 a^1 a^14 a^6 a^5 a^0 a^12 0 a^13 a^2 a^8 a^10 a^4
a^10 |a^10 a^5 a^8 a^4 a^12 a^2 a^0 a^7 a^6 a^1 a^13 0 a^14 a^3 a^9 a^11
a^11 |a^11 a^12 a^6 a^9 a^5 a^13 a^3 a^1 a^8 a^7 a^2 a^14 0 a^0 a^4 a^10
a^12 |a^12 a^11 a^13 a^7 a^10 a^6 a^14 a^4 a^2 a^9 a^8 a^3 a^0 0 a^1 a^5
a^13 |a^13 a^6 a^12 a^14 a^8 a^11 a^7 a^0 a^5 a^3 a^10 a^9 a^4 a^1 0 a^2
a^14 |a^14 a^3 a^7 a^13 a^0 a^9 a^12 a^8 a^1 a^6 a^4 a^11 a^10 a^5 a^2 0

Table 13.Multiplication Table

0 a^0 a^1 a^2 a^3 a^4 a^5 a^6 a^7 a^8 a^9 a^10 a^11 a^12 a^13 a^14
0 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
a^0 |0 a^0 a^1 a^2 a^3 a^4 a^5 a^6 a^7 a^8 a^9 a^10 a^11 a^12 a^13 a^14
a^1 |0 a^1 a^2 a^3 a^4 a^5 a^6 a^7 a^8 a^9 a^10 a^11 a^12 a^13 a^14 a^0
a^2 |0 a^2 a^3 a^4 a^5 a^6 a^7 a^8 a^9 a^10 a^11 a^12 a^13 a^14 a^0 a^1
a^3 |0 a^3 a^4 a^5 a^6 a^7 a^8 a^9 a^10 a^11 a^12 a^13 a^14 a^0 a^1 a^2
a^4 |0 a^4 a^5 a^6 a^7 a^8 a^9 a^10 a^11 a^12 a^13 a^14 a^0 a^1 a^2 a^3
a^5 |0 a^5 a^6 a^7 a^8 a^9 a^10 a^11 a^12 a^13 a^14 a^0 a^1 a^2 a^3 a^4
a^6 |0 a^6 a^7 a^8 a^9 a^10 a^11 a^12 a^13 a^14 a^0 a^1 a^2 a^3 a^4 a^5
a^7 |0 a^7 a^8 a^9 a^10 a^11 a^12 a^13 a^14 a^0 a^1 a^2 a^3 a^4 a^5 a^6
a^8 |0 a^8 a^9 a^10 a^11 a^12 a^13 a^14 a^0 a^1 a^2 a^3 a^4 a^5 a^6 a^7
a^9 |0 a^9 a^10 a^11 a^12 a^13 a^14 a^0 a^1 a^2 a^3 a^4 a^5 a^6 a^7 a^8
a^10 |0 a^10 a^11 a^12 a^13 a^14 a^0 a^1 a^2 a^3 a^4 a^5 a^6 a^7 a^8 a^9
a^11 |0 a^11 a^12 a^13 a^14 a^0 a^1 a^2 a^3 a^4 a^5 a^6 a^7 a^8 a^9 a^10
a^12 |0 a^12 a^13 a^14 a^0 a^1 a^2 a^3 a^4 a^5 a^6 a^7 a^8 a^9 a^10 a^11
a^13 |0 a^13 a^14 a^0 a^1 a^2 a^3 a^4 a^5 a^6 a^7 a^8 a^9 a^10 a^11 a^12
a^14 |0 a^14 a^0 a^1 a^2 a^3 a^4 a^5 a^6 a^7 a^8 a^9 a^10 a^11 a^12 a^13

Henceforth, we compare the results of the generic Reed-Solomon Coding and the proposed fault tolerance technique [Table – 14] based on the initial code.

Table 14.Fault TolleranceResult

Parameter	Generic RS	Proposed Optimized RS
Initial Polynomial	a^1 a^3 a^5	a^1 a^3 a^5
Encoded Data	a^5 a^3 a^1 a^6 a^4 a^2 a^0	0 00 a^6 a^4 a^2 a^0
Fault Tolerance Code	a^5 a^3 a^1 a^6 a^4 a^2 1	a^6 a^4 a^2 1
Optimization Reduction	0%	57%

VII. Conclusion

In this work the commercial cloud storage services are been compared based on the cost and performance factors. The result of the comparative measures provided the understanding of the demand for highly reliable and cost effective fault tolerance system. Henceforth, in this work we study the core Reed - Solomon fault tolerance mechanism based on Erasure codes. The work contributes towards the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

improved performance code for fault tolerance for digital storage devices rather than magnetic. Also the work enhanced the performance of the proposed technique by applying the improvement in terms of optimization. The result of the proposed optimization technique is 57% reduction in the storage cost without negotiating with the fault tolerance reliability.

REFERENCES

[1] J. Kubiatowicz, D. Bindel, Y. Chen, P. Eaton, D. Geels, R. Gummadi, S. Rhea, H. Weatherspoon, W. Weimer, C. Wells, and B. Zhao, “Oceanstore: An Architecture for Global-Scale Persistent Storage,” Proc. Ninth Int’l Conf. Architectural Support for Programming Languages and Operating Systems (ASPLOS), pp. 190-201, 2000.

[2] P. Druschel and A. Rowstron, “PAST: A Large-Scale, Persistent Peer-to-Peer Storage Utility,” Proc. Eighth Workshop Hot Topics in Operating System (HotOS VIII), pp. 75-80, 2001.

[3] A. Adya, W.J. Bolosky, M. Castro, G. Cermak, R. Chaiken, J.R. Douceur, J. Howell, J.R. Lorch, M. Theimer, and R. Wattenhofer, “Farsite: Federated, Available, and Reliable Storage for an Incompletely Trusted Environment,” Proc. Fifth Symp. Operating System Design and Implementation (OSDI), pp. 1-14, 2002.

[4] A. Haeberlen, A. Mislove, and P. Druschel, “Glacier: Highly Durable, Decentralized Storage Despite Massive Correlated Failures,” Proc. Second Symp. Networked Systems Design and Implementation (NSDI), pp. 143-158, 2005.

[5] Q. Tang, “Type-Based Proxy Re-Encryption and Its Construction,” Proc. Ninth Int’l Conf. Cryptology in India: Progress in Cryptology (INDOCRYPT), pp. 130-144, 2008.

[6] G. Ateniese, K. Benson, and S. Hohenberger, “Key-Private Proxy Re-Encryption,” Proc. Topics in Cryptology (CT-RSA), pp. 279-294, 2009.

[7] J. Shao and Z. Cao, “CCA-Secure Proxy Re-Encryption without Pairings,” Proc. 12th Int’l Conf. Practice and Theory in Public Key Cryptography (PKC), pp. 357-376, 2009.

[8] J. Bellorado and A. Kavcic , "Low-complexity soft-decoding algorithms for Reed-Solomon codes: An algebraic soft-in hard-out chase decoder" , IEEE Trans. Inf. Theory. , vol. 56 , no. 3 , pp.945 -959 , 2010

[9] F. Garc-Herrero, J. Valls and P. K. Meher , "High-speed RS (255, 239) decoder based on LCC decoding" , Circuits Syst. Signal Process. , vol. 30 , no. 6 , pp.1643 -1669 , 2011

[10] W. Zhang, H. Wang and B. Pan , "Reduced-complexity LCC Reed-Solomon decoder based on unified syndrome computation" , IEEE Trans. Very Large Scale Integr. (VLSI) Syst. , vol. 21 , no. 5 , pp.974 -978 , 2013

[11] J. Jiang and K. R. Narayanan , "Algebraic soft-decision decoding of Reed-Solomon codes using bit-level soft information" , IEEE Trans. Inf. Theory , vol. 54 , no. 9 , pp.2008 -3928.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

AMSFLO: Optimization Based Efficient Approach For Assosiation Rule Mining

Vikram Rajpoot
Research Scholar
MGCGV Chitrakoot
SATNA, India
rajputvikram022@gmail.com

Dr. Akhilesh Tiwari
Professor
Dept. of CS&IT MITS Gwalior
Gwalior, India
atiwri.mits@gmail.com

Dr. Bharat Mishra
Associate Professor
Faculty of Science & Environment
MGCGV Chitrakoot SATNA, India
bharat.mgcgv@gmail.com

***Abstract*—Rules discovering from various facts keeps on being an extreme association rule mining (ARM) issue Most of the algorithm consideration on discrete priori ability from numerical learning. In addition, in the finding connection system among data, regularly more prominent than one target is require, and in greatest cases, such goals conclusion clashing measures. This paper addresses different ARM optimization algorithms for arithmetical ARM that finds numerical association rules. We propose an improved SFLA way to deal with offer better outcomes. The proposed strategy mine intriguing and comprehensible AR without using minimal support and minimum self-recognition limits in best single analysis. In the trial portion of the paper we produce a gander at administrators results utilized on this study, contrast our approach with Particle Swarm Optimization (PSO) system. The last outcomes and assessment stage display that our proposed AMSFLO take out maximum dependable and valuable knowledge's from data set inside a period.**

Keywords-Apriori Algorithm, ARM, SFLA, Rough set data, Support and Confidence etc.

I. INTRODUCTION

In present years, fast growth in applying credit cards for making purchase has caused a significant data amount. These data can be profitable for analyzing the customer's utilization practices design. The credit card holder has been curious about forecasting default possibility of a credit score card holder. Negative risk emerging from the client practices can prompt a money loss. In this manner, credit card holder required to utilize data mining approach for predicting and arranging client's extra efficiency. Consequently, data mining is an imperative approach for all activity of the credit card procedure. For example, it can be utilized for ordering great/ bad clients in light of their application learning and, additionally, detecting a credit card abuse in view of purchase customer data [1].

The predicting ability the integrity/disagreeableness of a candidate can diminish credit danger of a credit card issuer. Nonetheless, if the card holder makes a wrong choice through issuing Visas to the credit cards, it will result in income and liquidity loss. This credit risk can lead financial emergency of the world economy for instance Tom-Yum-Kung emergency in 1997 and the sub-prime home loan in USA in 2008. As a result of a colossal measure of accessible data, activity analysis in the charge card strategy required to depend on data mining strategies for its efficiency and viability. Predominantly, data mining is separating designs methodology from data. It combines technique which used to factual, machine learning and database with a specific end goal to expel and recognize significant data from huge database.

A. Association Rule Mining Algorithm

Rakesh Agrawal first proposed ARM. It is an approach used to find intriguing rules from metadata and it's miles Data Mining (DM) ponders locales in present years. AR can offer usable data inside the vital significance estimation. For instance, find looking for among disease and blood records in enrolled myocardial localized necrosis case. Khalili makes use of Apriori calculation to watched gadget area approach for industrial intrusion determination through applying use of basic state [2]. Make most of the AR with Database (DB) approaching. Additionally, data mining in administered database is the use of with security protection. Rozenberg considered ARM issue in vertical allocated distributed database.

Definition 1: Transaction: T= t_1, t_2, t_3 … t_n It is a n subsets set called transaction and all transaction in T recognizes a items set ti ⊆ I.
Definition 2: AR: I={ i_1, i_2, i_3 … i_m }It is m factors set which are known as question and rule is portray as a ramifications of the shape: X(antecedent) →Y(consequent), in which X,Y ⊆I,X∩Y = ∅. The left-hand run aspect is predecessor and rule right- side is consequent.
Definition 3: Itemsets Support: The thing set X support, support (X), depict different transactions in T together which incorporate X.

$$\text{Support}\,(x) = |t|, x \in t \wedge t \in T$$

Definition 4: Frequent item set: I = i_1, i_2, i_3 … i_m, T =t_1, t_2, t_3 … t_n,S ⊆I, if bolster (S) ≥ min_support, at that point S is alluded to as regular itemset, where min_support is a limit portray through clients.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Definition 5, Transactions Count: N=|T|, N is finished number of transactions in database.
Definition 6: Length of biggest transaction: E=max (|ti|), E is various things incorporate into vital database transaction.
Definition 7: Ti = {i_1, i_2, i_3 … im} It is an set of object. . K-itemset={i1, i2, i3 … iK}k-itemset⊆I.
Definition 8: Confidence of administer: The rule X → Y has self-belief, it genuinely is transactions rate in T together with itemset X that still incorporate itemset Y. The guidelines that fulfill each guide (X → Y)≥ min_support and self-conviction (X → Y) ≥min_confidence are called robust regulations. The limits are depicted through users.

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support } (X \cup Y)}{|\text{Support } (X)|}$$

B. Shuffled Frog Leaping Algorithm

In SFLA, in step with fitness from massive to the small or from small to huge, individual are assigned to many groups in tum wherein worst individual P_w^thas learned from best individual P_b^tin a subgroup. If there is no progress, will examine from global excellent individual P_g^t. If there is nonetheless no development, P_w^twill be modified thru a random person. The various iterations in algorithm are provide through (t).

$$\text{Dis}^t = R * (P_b^t - P_w^t) \quad \ldots(1)$$

$$P_w^{t+1} = P_w^t + \text{Dis}^t \quad (\text{Dis}_m \geq \text{Dis} \geq -\text{Dis}_m) \quad \ldots(2)$$

Where: $P_w^{t+1} = (P_{w1}^{t+1}, P_{w2}^{t+1}, P_{w3}^{t+1} \ldots \ldots P_{wn}^{t+1})$
is a novel individual created through updating approach, Dis^t is all moving phase length. R is a random number its modify range is 0 to 1, [-Dis_m, Dis_m] is values range of step.
After updating, if the recently created is better than the old, will be replace by; generally, will be supplant through. On the off chance that still show no progress, it will be supplanted randomly through a novel person. This is an iterative system with different emphasess being equivalent to different subgroup people. At the point when subgroup handling is finished every subgroup will be arbitrarily arranged and re-separated into novel subgroups, technique being rehashed until predetermined termination criteria is satisfied. In considering the SFLA literature archives various approach designed recover the algorithm performance, the performance enhancements being commonly addressing updating approach

C. Modified Shuffled Frog Leaping Algorithm (MSFLO)

As mentioned earlier, SFLA is an algorithm for powerful optimization which has present superior performance than various evolutionary algorithms BA [13], KH [14-15], CSA [16-17], HS [18], FA [19-20], TLBO [21-22], CA [23], HBMO [24-25]. Nevertheless, we propose a novel modification technique to enhance complete search SFLA greatly ability. This modification technique is using to the SFLA in order to increase complete search SFLA ability and to avoid the premature convergence. In first phase we use a random walk to raise population diversity. It is known as Lévy flight. It can be expressed as following formulations:

Where t is the iteration number and φ_1 is a random value in the range [0, 1]. The new solution is enhanced than the end one then the system replaces it. Presently in the second period of the change we plan to move the normal of the populace toward the best arrangement. Therefore the mean estimation of the populace column-wise Mp ought to be figured, at that point every arrangement in populace is refreshed as takes after: X_{Gbest} is best frog in populace and TF is an random integer equal equivalent to 1 or 2 The essential SFL algorithm disadvantage is slow merging, closely identified with absence of adaptive acceleration terms in the position refreshing equation. In condition (1), rand characterize development phase frog's sizes through places of X_b and X_W. In the standard SFL, those stage sizes are random numbers among 0 and 1 for each frog. In each cycle, the target work esteem is a foundation that presents relative frog development upgrade as for past one. Along these lines, position changing formulae swings to following structure.

$$D_i = \text{rand} \times C \times (f(X_b) - f(X_w)) \times (X_b - X_w) \quad \ldots(3)$$

$$\text{New position: } X_{i+1} = X_i + D_i \quad \ldots(4)$$

Where $C \in (0, C_{max}]$ is a consistent, C_{max} is a case dependant upper limit, f(Xb) and f(Xw) are ideal and worst fitness capacities. Like unique SFL, if the system creates an enhanced arrangement, the worst frog is change through the better one. Something else, estimations in conditions (3) and (4) are rehashed with perceive to international high-quality frog instead (i.e. Xg and f (Xg) refresh Xb and f(Xb), individually). In the event that no improvement is conceivable, at that point another arrangement is arbitrarily made to most noticeably worst frog supplant Therefore, two distinctive specified alterations are added to standard SFL calculation. This novel form is known as MSFL. The essential MSFL characteristics algorithms are: containing adaptive developments, fast convergence, and better broadening capacity and getting away from local optima. At long last, proposed MSFL is as yet a typical optimization algorithm that can use to any real world continuous optimization problems. Section 2 describes the literature survey to understand the concept of the Shuffled Frog Leaping Algorithm with various techniques. Section 3 describes the problem statement to explain the problem of the existing work which can be overcome by the propose technique. Section 4 describes the proposed work in which we use ARM with MSFLO algorithm. Section 5 describes the result analysis of our proposed methodology in which details of the experiment has been shown. Section 6 presents the conclusion and future work for further research in this field.

II. LITERATURE SURVEY

Luo (2015) et al in propose a power law especially optimization neighborhood seek method intended to increment searching speed. Wang and Fang in present a procedure in

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

which virtual frogs are encrypted as expanded multi-mode movement list and decrypted through multi-mode serial schedule generation strategy. Li propose a stepped forward SFLA which increases leaping rule through broadening the leaping level size and adding a leaping idleness factor to represent social lead. The announced result for inquire about distinguished exhibits SFLA efficacy [4].

Canister Hu (2016) et al introduce which is high dimensional biomedical datasets incorporate various highlights which can be utilized as a part of the molecular diagnosis of disease, in any case, Such datasets include various unimportant or helpless connection highlights which have an impact on predictive analysis accuracy. In this paper, appropriate here present SFLA which exhibits a disorder memory weight issue, an outright strength foundation strategy and a adaptable switch issue. To evaluate proposed method efficiency have used K-nearest neighbor system with a similar analysis in which contrast proposed procedure and genetic algorithms, PSO, and SFLA. Test results show that enhanced algorithm accomplish change in the related subsets recognizing evidence and in characterization accuracy [5].

Israel Edem Agbehadji (2016) et al present that huge data has ended up being one of key sources for useable learning and as data becomes greater it poses couple of computational test in finding a generally perfect. Meta-heuristic calculation when using to mining AR target to discover most perfect principles from data without being stuck in local optimal. For instance, meta-heuristics calculations finish up GA and PSO algorithm.

Discovery appropriate representation of various known about examples applying harsh numerical esteems characteristics is as yet a test in light of the fact that most ARs can't be the utilization of to numerical measurements without discretization which can likewise prompt the know-how loss. Mining numeric AR is a hard optimization inconvenience instead of being a discretization, consequently, this paper proposes a novel meta-heuristic estimation which makes usage of WSA for numeric ARM from troublesome regards inside average extents [6].

Junwan Liu (2011) et al offers that DNA microarray clustering insights that could mine basic examples to help in understanding quality cooperations and direction. This is a standard MOP. Starting late, a couple of researchers have created stochastic search for frameworks which emulate skilled effective lead for instance, ants, honey bees, birds and frogs, as a way to deal with are looking quicker and advance robust solutions for complex optimization bother. The PSO is a heuristics-essentially based optimization system simulating bird flock developments finding food. Exploratory results on two distinctive genuine datasets exhibit that strategy can productively discover imperative high quality bi-clusters [7].

Golda George (2015) et al present that data mining progresses as a promising arrangement in exploring knowledge concealed in clustering and database is one its application. Clustering can be depicted as unsubstantiated examples arrangement into gatherings. Various target functions are utilized to measure partition efficiency through analyzing couple of inherent property groups minimization, remove measures, cluster symmetry and density. Yet, consideration of those targets won't make a commitment to amend sort to clusters. MOO approach is available days utilized as an option technique to yield improved clustering outcomes. In the past procedure, partial cuckoo look wind up plainly utilized for data clustering. In this paper, intended to expand clustering performance through consideration of hybridized optimization technique which will utilize firefly algorithm with GSO. The hybridization is brought out through substituting most exceedingly terrible wellness esteems all GSO emphasis with the refreshed esteems from firefly algorithm. Multiple goal features are utilized for computation of the wellness and individuals utilized fitness are Fuzzy DB-Index, XB-Index and Sym-Index. The proposed approach is executed applying MATLAB and accomplishment of the clustering strategy is assessed applying different files in the CVAP device and contrasted and diverse techniques in CVAP instrument [8].

Wenchuan Yang (2016) et al demonstrate that AR is a essential data analysis and mining technique, and FP-Growth and regular FP-Tree calculation is used as a part of the total rules confidence.. This paper proposes an incremental queue algorithm models in light of the AR, which is enhanced FP4W-Growth calculation. It is proposed and using to the check association message through the incremental queue connection. Its common sense is affirmed through investigation. After calculation and model improvement, it can find concealed and beneficial novel data and new illustration. Besides, those precepts found in the substance can be used as logical decision-making approaches [9].

EsraSaraç (2013) et al show that Upsurge in the data total on Web has caused the required for accurate mechanized classifiers for Web pages to the keep up Web indexes and to development search engines' presentation. As every (HTML/XML) tag and each term on each Web page can be considered as a component, required practical approaches to manage select best features to the lower trademark space of the Web page arrangement bother. In this study, use FA to select features subset, and to the estimate fitness of the chose features J48 classifier of the Weka data mining device is enlisted. Watched that once capacities subset are settled on by means of making utilization of FA, WebKB and Conference datasets have been arranged without accuracy loss, even additional, time required to order novel Web pages [10].

Satpal Singh (2015) et al displays that Traditional data mining technique offers effectively statistical analysis with discovery of frequent patterns and hidden knowledge. It succeeded in relationship discovery among items through measurable significance however couldn't give more parameter to data revelation. As opposed to traditional strategy, profit

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

significance use as a measure to compute novel confidence and support established completely upon " profit " give intriguing examples. To limit this hole, required to encompass couple of techniques that will represent the beneficial rules generation. One approach to do it is through including the benefit bolster and in addition benefit certainty by method for considering the genuine benefit and averaging the whole benefit of each question. Subsequently the rules being supreme in nature will be more profitable than the past ones [11].

Zuleika Nascimento (2013) et al display that impressive exertion has been made through researchers in the zone of community traffic sort, since the Internet develops traffic volume and protocols and packages develops traffic volume and protocols and applications. The task of traffic recognizing verification is a complex task undertaking in light of consistently changing Internet and an development in encoded data. There are different approaches for requesting network traffic for example port-based and Deep Packet Inspection (DPI), yet they are not effective since various applications use unpredictable ports and the payload could be encrypted. This paper proposes an OHM that makes run based model (Apriori) use close by model to deal with the issue of development arrange without making use of the payload or ports. The proposed system similarly permits the AR generation for novel obscure applications and further marking through specialists. Other than that, an optimizer known as Firefly Algorithm was in like manner used to improve comes about through updating both Apriori and SOM parameters and a close study. The OHM accommodates be advanced to a non-updated show for both eMule and Skype applications, finishing ranges progressed to 94% for rightness charge. The OHM offers to be advanced to a non- optimized variant for each eMule and Skype programs, achieving levels advanced to 94% for rightness charge. The OHM was in like manner endorsed against another model in perspective of computational knowledge, named Real-time, and the OHM proposed in this work presented upgraded comes about when attempted continuously [12].

III. PROPOSED METHODOLOGY

Association Rule Mining (ARM) has the major downsides for obtaining the non-interesting rules and huge range of determined regulations or short algorithm overall performance for fixing the complex mining problem. In prior work PSO is used to acquire superior result but there are specific problem with PSO that it is easy to belong into confined optimum in space of high-dimension and has a lesser rate of convergence in the repetitive process. Apply the algorithms of optimization to overcome this problem and discover superior outcomes.

A. Proposed Methodology

Modified Shuffled Frog Lippy Optimization (MSFLO) has applied on spellman.csv database in this research work to investigate association rules and the technique is called ARMSFLO. The subsequent of this segment have several significant components of the algorithm which are clarified: rough set conversion MSFLO encoding, fitness function, and lastly the previous component of the section explain the methodology of ARMSFLO.

B. Proposed Algorithm:

Step1: Apply Apriori algorithm for locating rules

- F = Fetch data file as input
- Generate length of data file (F)
- Form candidate itemset C_m of size m
- Create Frequent itemset L_m of size m
- Put common items in L_1
- For (C_m =1; L_m! = null; m++) do begin
- Candidate created from L_m puts in C_{m+1}
- For every transaction T_i in DB
  do
  Number of every candidates increased which include in T
  C_{m+1} 's candidates with minimum support put in L_{m+1}
  End
  Return $\cup_m$ L m;

Step2: Evaluating the value of support and confidence by formulas shown below:

- Support value (item) = Support number of item/Total number of every items
- Confidence value(A|B)= Support value(AB)/Support value(A)

Step 3: Rule Fitness evaluation for SFLOmigration:

- Fitness_overall(j)=absolute(log(Confidence(j))+log(α*Support(j))/(len(Support)+len(Confidence);
- $Net_{fitness}$=sum(Fitness_overall)/len(data)
- $Net_{fitness}$=abs($Net_{fitness}$*(len(Support)/(min Support *min Confidence*threshold)));

Step 4: Now for the calculation of MSFLO

- Evaluate fitness of rules above evaluated produced less fitted rules by using function of bench mark
- Nmem = fix(size(x,1)/npopmem);
- Npop = nmem*npopmem
- Use the formula below for updating the position (Pos)
  Pos = rand().S.(Z_b-Z_w)
  Where rand() is the random function, S is huge cost at the simulation state, Z_b is global best value, Z_w is worst value
- Evaluate again the global minima and the compare the value

Step5: Exit

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

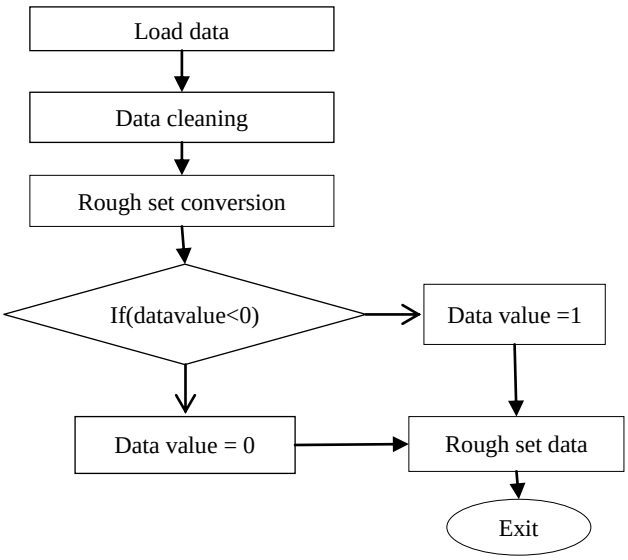


Fig 1: Rough set conversion

The elementary conception following Rough Set Theory is the rough calculation of lower and upper spaces of a set, the estimation of areas being the recognized grouping of expertise concerning the concern field. To demonstrate the irregular investigation we don't forget a straightforward case of choice of applicants to a school. The candidates to the school have suggested their packages of application with certification of secondary school, curriculum vitae and previous school's opinion, for attention via an admission committee based on those documents; the candidates were described the usage of 7 criteria collectively with corresponding scales ordered from the high-quality to the worst cost, specified under.

b_1 - Mathematics Score: $vb_1 = \{6, 5, 4\}$, b_2 - Physics Score: $vb_2 = \{6, 5, 4\}$,
b_3 – English Score: $vb_3 = \{6, 5, 4\}$, b_4 - Other subjects Mean: $vb_4 = \{6, 5, 4\}$,
b_5 - Secondary school Types: $vb_5 = \{2, 3, 4\}$, b_6 -Motivation: $vb_6 = \{2, 3, 4\}$,
b_7 - Previous school Opinion: $vb_7 = \{2, 3, 4\}$, d - Committee Decision: v_d {Acc, Rej}
Then the set of circumstance attributes is $B = \{b_1, b_2, b_3, b_4, b_5, b_6, b_7\}$.

Fifteen candidates with instead exceptional application packages were sorted by means of the committee after due consideration. They create the set of examples. The set of decision attributes is D = {Acc, Rej}, where Acc stands for an admission and Rej for a rejection. The information is signified in Table 1.

Table 1. Decision Table Composed of Example

Condition / Candidate	B1	B2	B3	B4	B5	B6	B7	Decision D
Z1	5	5	5	5	3	3	2	Acc
Z2	4	4	5	4	3	2	2	Rej
Z3	4	5	4	4	2	3	3	Rej
Z4	6	4	6	5	3	2	3	Rej
Z5	5	5	6	5	3	2	3	Acc
Z6	4	5	4	5	3	2	4	Rej
Z7	5	5	6	5	3	3	3	Acc
Z8	5	5	5	5	3	3	3	Acc
Z9	5	5	5	5	2	2	3	Rej
Z10	6	4	6	5	3	2	3	Acc
Z11	6	5	5	5	2	2	3	Acc
Z12	5	4	4	4	4	3	3	Acc
Z13	4	4	5	4	3	4	4	Rej
Z14	5	6	6	5	3	2	2	Acc

There are $Y_{Acc} = \{Z_1, Z_5, Z_7, Z_8, Z_{10}, Z_{11}, Z_{12}, Z_{14}\}$ and $Y_{Rej} = \{Z_2, Z_3, Z_4, Z_6, Z_9, Z_{13}\}$. The lower approximation and upper approximation of Y_{Acc} and Y_{Rej} are as follows $B\text{-}Y_{Acc} = \{Z_1, Z_4, Z_5, Z_7, Z_{10}, Z_{11}, Z_{12}, Z_{15}\}$, $B^{-}Y_{Acc} = \{Z_1, Z_4, Z_5, Z_7, Z_8, Z_9, Z_{10}, Z_{11}, Z_{12}, Z_{15}\}$.

C. Description:

1. Using Apriori Algorithm, we initially find the rules which generate candidate itemset in the first step. At each iteration the frequent itemset are generated from the candidate itemset.
2. Calculate the support and confidence value of the rules in this step.
3. Fitness function of SFLO is calculated in this step which is useful for finding the useful items.
4. In the last step, we calculate the MSFLO method which calculates fitness function, global best and worst value.

For the calculation of support and confidence value mainly ARM used in data mining. It generates the rules for various support and confidence value then we applied this rules in the SFLO which perform the fitness function and generate the value then compare this value and update the database with better results.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

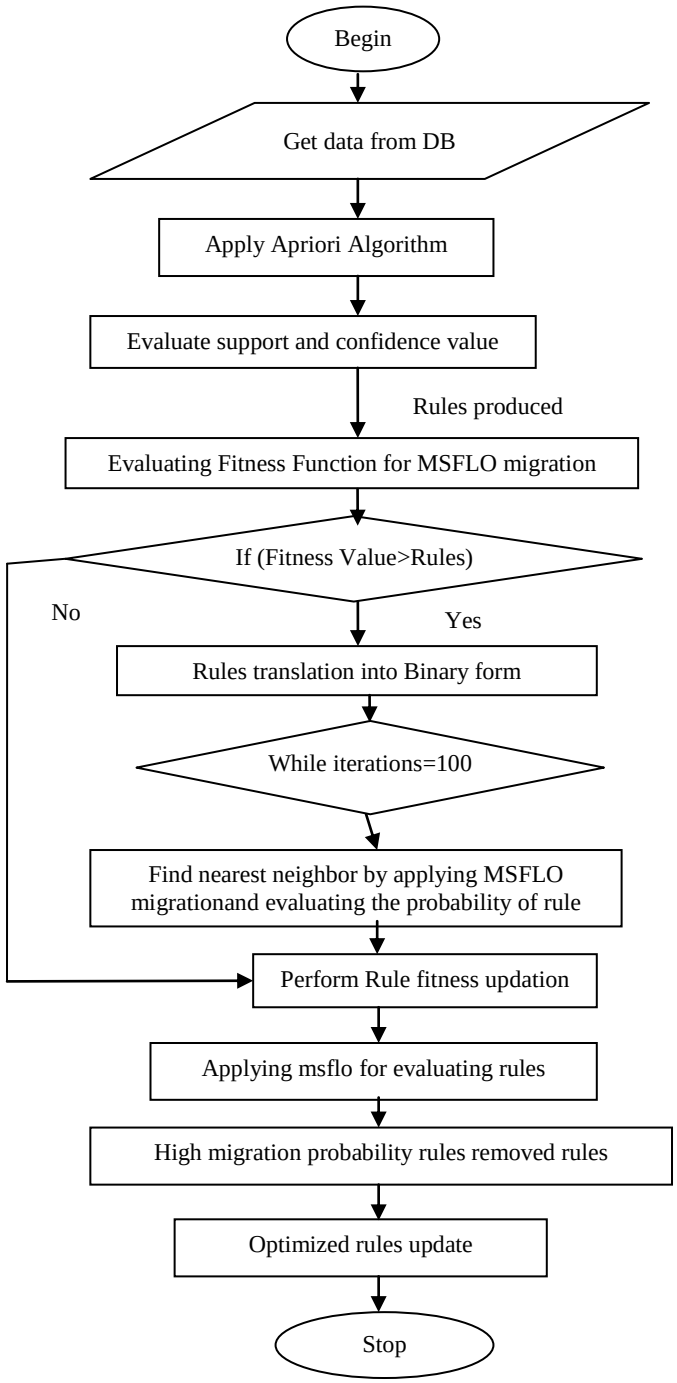


Fig 2. Flow chart of Proposed Algorithm

D. Working example of MSFLO application on rough set data:

1. The above revealed flowchart gives an analysis in broader regarding the proposed algorithm. The utilization of MSFLO in ARM is first calculated using ARM method. Support value and confidence value of calculated and by using these 2 aspects, evaluation of fitness function performed. Calculate the rules which are less than fitness function as these rules will be less fit rules and necessitate to be transferred. The idea of MSFLO is at present functional on to the rules less than fitness value by evaluating their probability of migration. For each case, probability is updated and subsequent location for the motion is estimated. In this approach, those rules which were not as much fit originally will shift to a superior position and will survive. This will enhance their possibility of survival and thus, improved rules can be extracted.

2. Net_fitness evaluation

   The fitness value is evaluated based on the support and confidence that have been resulting in the Apriori algorithm formerly in the proposed methodology. Also, a net fitness value is to be evaluated for the consideration of overall fitness. The fitness examination is used to choose the rules that are to be modified using MSFLO method. The less appropriate rules are searching out by contrasting their fitness value with $Net_{fitness}$. If it is fewer, then they are weaker rules. The formula derived for fitness evaluation is:

$$\text{Fitness\_overall} = absolute\left(\frac{\log\big(confidence(j)\big) + log\big(\alpha * support(j)\big)}{\big(length(support) + length(confidence)\big)}\right)$$

$$Net_{fitness} = sum(\text{Fitness\_overall}))/length(data)$$

$$Net_{fitness} = abs(Net_{fitness} * (length(support)/(minsupport * minconfidence * threshold)\ );$$

Where**,** minsupport and minconfidence are predefined.

IV. RESULT SIMULATION

Numerous testing were accomplished on a 3.3 GHz Intel Processor with the 8 GB main memory. In accumulation, testing were execute on Windows 7 OS. The entire algorithms used in testing were shown in the MATLAB 2014.We performed the simulation for the generation of rules on three different support and confidence value which is valuable to show the presentation of the proposed work. The table below show the parameters used in the proposed work and their values which is useful to understand the implementation easily.

Table 2: Parameter and their Values

Parameter	Value
Tool used	MATLAB
RAM size	512 MB
Hard Disk	1.60 GHz
Dataset	Spellman.csv
Algorithm	Apriori Algorithm
Optimization Techniques	PSO, MSFLO

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Minimum support	Minimum confidence
0.1	0.5

210,240,250 ->130 (10.0205%, 76.2153%)
140,170 ->150,240 (10.0205%, 51.285%)
140,150,170 ->240 (10.0205%, 67.1254%)
140,150,240 ->170 (10.0205%, 62.1813%)
140,170,240 ->150 (10.0205%, 72.562%)
160,170,200 ->240 (10.0205%, 78.2531%)
160,170,240 ->200 (10.0205%, 61.1421%)
170,200,240 ->160 (10.0205%, 57.6115%)

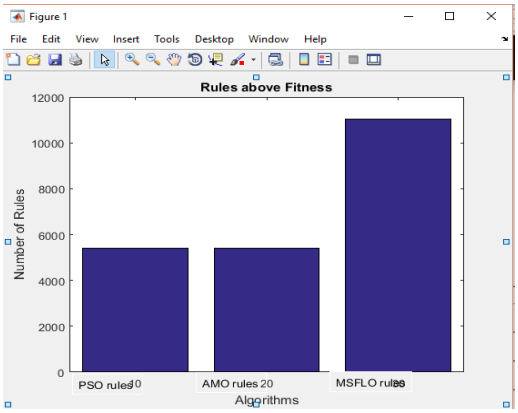


Fig. 3. Rules generated by algorithms PSO, AMO, MSFLO

Above graph shows that MSFLO achieve improved comparison to PSO algorithm by considering above graph it can say that by using MSFLO we get optimize result for this support and confidence value. MSFLO technique achieves all the procedure after and generates the better results as compared to particle swarm optimization.

Minimum support	Minimum confidence
0.2	0.4

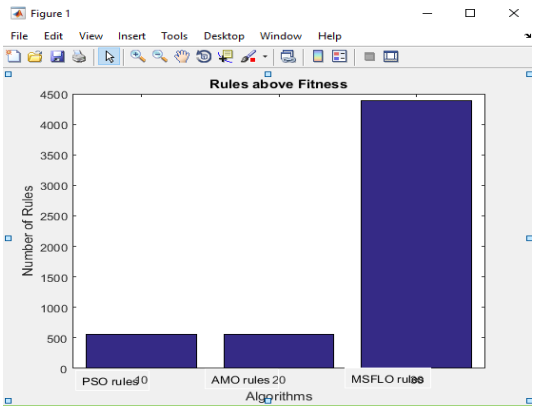


Fig 4. Rules generated by algorithms PSO, AMO, MSFLO

40,130 ->150,170 (20.0183%, 53.443%)
40,150 ->130,170 (20.0183%, 53.4756%)
40,170 ->130,150 (20.0183%, 59.2167%)
130,150 ->40,170 (20.0183%, 43.4373%)
130,170 ->40,150 (20.0183%, 52.1713%)
150,170 ->40,130 (20.0183%, 44.5857%)
40,130,150 ->170 (20.0183%, 70.7829%)
40,130,170 ->150 (20.0183%, 83.6832%)
40,150,170 ->130 (20.0183%, 76.0624%)
130,150,170 ->40 (20.0183%, 62.2427%)

The above graph shows that the fitness value of proposed work is much improved than the existing techniques which mean that we can get a large amount of fitted values from the database for getting more proficient results.

Minimum support	Minimum confidence
0.3	0.5

220,240 -> 200(30.2899%, 73.3149%)
90-> 70,190(30.2442%, 58.3187%)
190-> 70,90(30.2442%, 60.2821%)
70,90-> 190(30.2442%, 76.6782%)
70,190-> 90(30.2442%, 86.1508%)
90,190-> 70(30.2442%, 81.9926%)
250-> 40(30.2214%, 59.4522%)
50-> 250(30.2214%, 56.751%)
250-> 50(30.2214%, 59.4522%)
190-> 110,130 (30.2214%, 60.2366%)
110,13 -> 190(30.2214%, 67.4478%)
110,190-> 130(30.2214%, 81.9307%)
130,190-> 110(30.2214%, 84.4388%)
80-> 70(30.1986%, 59.0361%)
210-> 60(30.1758%, 59.8732%)
180-> 170(30.016%, 56.8772%)

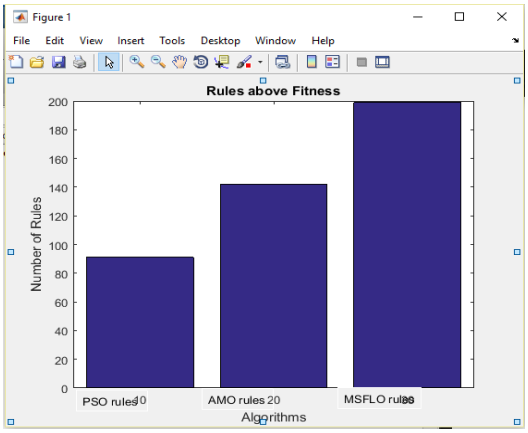


Fig 5. Rules generated by algorithms PSO, AMO, MSFLO

Above tables shows comparison among associate rule created by base and proposed optimization method of both algorithm

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

by seeing result. We are able to declare that the proposed algorithm work enhanced as compare to current one.

Table 3: Comparison between numbers of rules generated by PSO, AMO Techniques

Support	Confidence	PSO rules	AMO rules	MSFLO rules
0.3	0.5	91	142	199
0.2	0.4	559	560	4389
0.1	0.5	5405	5406	11054

V. CONCLUSION

The work specific in this paper offers a viable approach for investigating high quality AR. In this paper we offered the execution of MSFLO DM device have the capability to create the stacks or extra noteworthy rules. So the greater part of the arrangements is not intriguing, just small fraction of the approaches could be enthusiasm to individual. Create most effective appealing rules is the motivation behind optimization algorithms in ARM. In this analysis, the AR produce with the help of Apriori calculation with MSFLO algorithm connected. In this proposed approach, frogs encrypting is the utilization of to from database extricate rules. For mining standard, individual rule fitness value is registered in inclination to minimum useful resource and minimum certainty edges. This has a gain that DB is filtered once best which enhances adequacy of the system in CPU time and memory utilization phrases. To enhance the effectiveness and accurateness of the streamlining there might be requirement of including additional procedures in changed SFLO for such DM issues. The proposed method connected over MATLAB and final product changed into as contrasted and present enhancement method and find that our proposed perform better. In the future work, different datasets and parameters can be taken into consideration to show the efficiency of the proposed work.

References

[1] Pornwatthana Wongchinsri and WerasakKuratach (2016), " A Survey - Data Mining Frameworks in Credit Card Processing", IEEE.

[2] Maocai Cheng, KaiyongXu and Xuerong Gong (2015)" Research on Audit Log Association Rule Mining Based on Improved Apriori Algorithm", IEEE.

[3] TaherehHassanzadeh and Mohammad Reza Meybodi (2012)" A New Hybrid Approach for Data Clustering using Firefly Algorithm and K-means", The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012), pp: 007- 011.

[4] Yongqiang Dai, Bin Ru*, Yun Su, Cheng sheng Mao, ling Chen, Xiaowei Zhang, Philip Moore, LixinXu and RanshuCai (2015)" Feature Selection of High-dimensional Biomedical Data using Improved SFLA for Disease Diagnosis" IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp: 458-463.

[5] Bin Hu, Yongqiang Dai, Yun Su, Philip Moore, Xiaowei Zhang, Chengsheng Mao, Jing Chen and LixinXu (2016)" Feature Selection for Optimized High dimensional Biomedical Data using an Improved Shuffled Frog Leaping Algorithm", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS .

[6] Israel Edem Agbehadji, Simon Fong and Richard Millham (2016)" Wolf Search Algorithm for Numeric Association Rule Mining"IEEE International Conference on Cloud Computing and Big Data Analysis, pp: 146-151.

[7] Junwan Liu, Zhoujun Li, Xiaohua Hu, Junwan Liu and Yiming Chen (2011)"Multiobjective Optimization Shuffled Frog-leaping Bi-clustering", IEEE International Conference on Bioinformatics and Biomedicine Workshops, pp: 151-156.

[8] Golda George and LathaParthiban (2015)" Multi Objective Hybridized Firefly Algorithm with Group Search Optimization for Data Clustering" IEEE.

[9] Wenchuan Yang, Lei Hui, Dong Zhang and Yimin Fu (2016)" An Improved Incremental Queue Association Rules for Mining Mass Text" International Symposium on Computer, Consumer and Control, pp: 447-450.

[10] EsraSaraç and Selma Ayşe Özel (2013)" Web Page Classification Using Firefly Optimization", 2013 IEEE.

[11] Satpal Singh and VivekBadhe (2015)" Profit Association Rule Mining with Inventory Measures" International Conference on Computational Intelligence and Communication Networks, pp: 951- 955.

[12] Zuleika Nascimento, Djamel Sadok and Stˆenio Fernandes (2013)" Comparative Study of a Hybrid Model for Network Traffic Identification and its Optimization using Firefly Algorithm" 2013 IEEE.

[13] A. Kavousi-Fard, A. Khosravi, S. Nahavadi (2015) "A New Fuzzy Based Combined Prediction Interval for Wind Power Forecasting" IEEE Trans. on Power System.

[14] M. Rostami, A. Kavousi-Fard, and T. Niknam (2015) " Expected Cost Minimization of Smart Grids with Plugin Hybrid Electric Vehicles Using Optimal Distribution Feeder Reconfiguration" IEEE Trans. on Industrial Informatics.

[15] A. Kavousi-Fard, A.Abunasri, A. Zare, R. Hoseinzadeh, Impact of Plug-in Hybrid Electric Vehicles Charging Demand on the Optimal Energy Management of Renewable Micro-Grids,78 Energy (Elsevier), 2014, 904-915

[16] A. Kavousi-Fard, T. Niknam (2014) "Optimal Distribution Feeder Reconfiguration for Reliability Improvement Considering Uncertainty" IEEE Trans. On Power Delivery, 29(3) 1344 – 1353.

[17] A. Kavousi-Fard, T. Niknam, H. Taherpoor, A. Abbasi, (2015) "Multi-objective Probabilistic Reconfiguration Considering Uncertainty and Multi-Level Load Model" IET SMT, vol 9 (1), 2015, pp.44-55.

[18] A. Kavousi-Fard, T. Niknam, M. Khooban (2014) "An Intelligent Stochastic Framework to Solve the Reconfiguration Problem from the Reliability view" IET SMT, 8(5), p. 245 – 259.

[19] A. Kavousi-Fard, T. Niknam (2014) "Optimal Stochastic Capacitor Placement Problem from the Reliability and Cost Views using Firefly Algorithm" IET SMT, vol. 8(5), pp. 260 – 269.

[20] A. Kavousi-Fard, H. Samet, F. Marzban (2014) "" A New Hybrid Modified Firefly Algorithm and Support Vector Regression Model for Accurate Short Term Load Forecasting, Expert Systems With Applications" 41(13) 6047–6056

[21] A. Kavousi-Fard, T. Niknam, A. Khosravi (2014) "Multi-Objective Probabilistic Distribution Feeder Reconfiguration Considering Wind Power Plants, International Journal of Electrical Power and Energy Systems" (55) 680-691.

[22] A. Kavousi-Fard (2013) "A new fuzzy-based feature selection and hybrid TLA–ANN modeling for short-term load forecasting, Journal of Experimental & Theoretical Artificial Intelligence" 25(4) 2013, 543-557.

[23] A. Kavousi-Fard, F. Kavousi-Fard (2013) "A New Hybrid Correction Method for Short Term Load Forecasting Based on ARIMA, SVR and CSA, Journal of Experimental & Theoretical Artificial Intelligenc" 25(4), 559-574.

[24] A. Kavousi-Fard, T. Niknam (2013) "Considering uncertainty in the multi-objective stochastic capacitor allocation problem using a novel self adaptive modification approach, Electric Power Systems Research" 103, , 16-27.

[25] A. Kavousi-Fard, H. Samet (2013) "Multi-objective Performance Management of the Capacitor Allocation Problem in Distributed System Based on Modified HBMO Evolutionary Algorithm, Electric Power and Component systems" 41 (13) 1223:1247.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Decision Making in Uncertainty:
A Bayesian Network for Plant Disease Diagnoses

Jawed Naseem [1]
SSITU, SARC
Pakistan Agriculture Research Council
Karachi, Pakistan
jawed.naseem@gmail.com

S. M. Aqil Burney [2]
College of Computer Science
& Information System
IoBM University, Karachi, Pakistan
aqil.burney@iobm.edu.pk .I

Abstract: *Decision making in uncertain situation with missing and partial truth is challenging in crop cultivation management like any other domain of application. In crop production attack of disease is a significant risk factor affecting yield and quality of crop. Rusts is one of the economically important fungal diseases of wheat difficult to diagnose in field condition due to ambiguity in classifying factors. Computer based soft computing methods can provide several intelligent solution for disease diagnoses in plant more precisely. In this paper a model for probabilistic decision making system in uncertain situation is discussed. The model is utilized to develop Bayesian Network for diagnoses of leaf, stem and strip rust disease of wheat. Proposed Bayesian network efficiently capture interdependence of classifying factor like color, shape and distribution of disease in different parts of plant along with removing uncertainty by employing conditional dependence. The proposed BN achieve upto 81% accuracy in wheat disease diagnoses.*

Key Words: Decision making, Uncertainty Wheat, Rust disease, Bayesian network,

I. INTRODUCTION

Decision making in uncertain situation with missing and partial truth is challenging in every field of science and agriculture has no exception. Disease attack during cultivation of crop is one of the major risks in crop cultivation management. Timely decision subject to prevailing environmental condition is required to control the disease and reduce risk. Rusts are economically important disease of wheat. Three distinct types of rusts, leaf rust, stripe rust and stem rust occur on wheat. The potential yield loss caused by these diseases depends on host susceptibility and weather conditions, but the loss also is influenced by the timing and severity of disease outbreaks relative to crop growth stage. The greatest yield losses occur when one or more of these diseases occur before the heading stage of development. Early detection and proper identification of disease is critical to disease management and control. Symptoms of various wheat diseases are so common that it is difficult to identify the disease appropriately without detail knowledge. Even the expert with sufficient knowledge can do mistake due to ambiguity in classification.

Computer based technologies can be utilized for decision making with ambiguous information. Artificial Intelligence (AI) is the area of computer science which focuses on developing machines and computer systems requiring intelligence like humans being [Iowa, 2006]. Using AI techniques and methods researchers are creating systems which can mimic human expertise in any field of science. Application of AI ranges from creating robots to soft computing models (softbot) that can reason like human expert and suggest solutions to real life problems. AI can be used for reasoning on the basis of incomplete and uncertain information and delivering predictive knowledge. In AI several machine learning techniques and method can be employed to perform automated task which are difficult to perform manually. Bayesian networks are one of classifying technique which effectively employed in uncertain and ambiguous situation.

In this paper a mechanism is discussed to develop a probabilistic reasoning system for decision making in uncertain situation. The model is used to develop Bayesian Network for diagnoses of rust disease in wheat crop. Section II discuss challenges and issues in wheat disease diagnoses, Section III describes development of Bayesian network. Section VI discuss outcome of the experiment along with efficacy of the proposed system, Section V highlight future work.

II DYNAMICS OF WHEAT DISEASE DIAGNOSES & CONTROL

Wheat cultivation is associated with several risk posed by environment, economic stability and management of crop. One of the economically significant risks is disease attack during cultivation. Wheat is attacked by several diseases during cultivation including rust disease. The wheat rust fungi are obligate parasite as they can grow and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

multiply in nature only on living plant tissue. Rust disease affect crop yield significantly due to their wide distribution, capacity to form new races that can attack previously resistant cultivars, ability to move long distances, and potential to develop rapidly under optimal environmental conditions [Wegulo, 2012]. Stem rust is capable of destroying entire wheat fields over a large area within a period of just a few weeks. The following are the important parameter in wheat rust disease [Table-1]

- Parts of plant infected:
- Shape and distribution of lesions
- Lesion color
- Degree of damage
- Tearing of Tissue

In different rust disease of wheat one or many of the above factors can be used to diagnose the disease.

	Leaf rust	Stem rust	Stripe rust
Pustule location	Leaf, mainly on the upper surface	Stem and leaf, upper and lower surfaces of leaf; occasionally on head and seeds	Leaf, upper surface; occasionally on head and seeds
Pustule arrangement	Single and random	Single and random	Stripes
Pustule shape and size	Round or slightly elongated; small to medium	Oval shaped or elongated; small to large	Round, blister-like; small
Tearing of host epidermis	Rare, visible with magnification	Conspicuous	None
Optimum temperature for infection	59-68 F	59-84 F	45-54 F
Optimum temperature for disease development	68-77 F	79-86 F	50-59 F

Table-1 Comparison of Wheat Rust disease [Wegulo, 2012]

Leaf and stripe rust can be distinguished by the color and shape of pustules and the location of the infection. However the symptom of these three types of disease has very slight variation which makes it difficult to distinguish one from another [Table-1]. Leaf rust pustules are orange brown in color, circular to oval in shape and chiefly found scattered on the upper surface of leaves. Stripe rust pustules are yellow-orange. Initially, the pustules are small and circular, but develop into yellowish stripes on the upper leaf surfaces, leaf sheaths and inside glumes.

III BAYESIAN BELIEF NETWORK

Bayesian network is a probabilistic graphical model used to represent knowledge system about a uncertain domain [Ben-Gal, 2007]. Any system having inherent uncertainty can be represented by Bayesian network. The simplest example of BN is a estimating probability of rain on a given day which is dependent on certain factor like temperature humidity and weather condition on last few days. In BN each node in the graph represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables. These conditional dependencies in the graph are often estimated by using known statistical and computational methods. Hence, BNs combine principles from graph theory, probability theory, computer science, and statistics. Bayesian Network is based on Baye's theorem which explains conditional dependence of one variable on other. The prior probability of event used to estimate posterior probability.

Formally, Bayesian network B is an annotated acyclic graph that represents a JPD over a set of random variables V. The network is defined by a pair

$$B = (G, \Theta)$$

Where G is the DAG (directed Acyclic graph) whose nodes $X_1, X_2, \ldots, X_n$ represents random variables, and whose edges represent the direct dependencies between these variables. The graph G encodes independence assumptions, by which each variable X_i is independent of its non-descendants given its parents in G. The second component Θ denotes the set of parameters of the network. This set contains the parameter $\theta_{xi} \mid \pi_i = PB(xi \mid \pi_i)$ for each realization x_i of X_i conditioned on π_i, the set of parents of X_i in G. Accordingly, B defines a unique JPD over V, namely:

$$P_B(X_1, X_2, \ldots \ldots X_n) = \prod_{i=1}^{n} P(X_i \mid \pi_i) \quad = \prod_{i=1}^{n} \theta_{Xi} \mid \pi_i$$

If Xi has no parents, its local probability distribution is said to be unconditional, otherwise it is conditional. If the variable represented by a node is observed, then the node is said to be an evidence node, otherwise the node is said to be hidden or latent. The conditional independence statement of the BN provides a compact factorization of the JPDs. Instead of factorizing the joint distribution of all the variables by the chain rule is applied. The reduction provides an efficient way to compute the posterior probabilities given the evidence

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Learning Bayesian Network

Bayesian network explicitly define the interdependence among variable of interest. In practical application learning Bayesian network is one of the crucial steps. The process involves learning topology or structure of network to depict causal relationship among variable and secondly estimating the parameter. Different approaches are used for learning the BN. The most common approach is learning through data and using expert knowledge. In this paper a hybrid approach is adopted. Expert knowledge is helpful in defining the structure while learning through data is effective for estimating the parameter. In learning through data a prior probability density function is assigned to each parameter vector and training data is used to compute the posterior parameter distribution and the Bayes estimates.

Probabilistic Reasoning through BN

The ultimate objective of developing BN is to inference the most probable outcome based on available evidence. BN is mathematically represented through JPD in a factored form which can be used to evaluate all possible inference by marginalization, i.e. summing out over "irrelevant" variables. Two types of inference support are often considered: *predictive support* for node *Xi* , based on evidence nodes connected to *Xi* through its parent nodes called *top-down reasoning*, and *diagnostic support* for node *Xi* , based on evidence nodes connected to *Xi* through its children nodes known as *bottom-up reasoning*.

The complexity of JPD increases with increasing number of nodes. Even if the variable have binary outcome JPD has size O(2n), where n is the number of nodes. Hence, summing over the JPD takes exponential time. In general, the full summation (or integration) over discrete (continuous) variables is called exact inference and known to be an NP-hard problem. However, some efficient algorithms exist to solve the exact inference problem in restricted classes of networks. One of the most popular algorithms is the message passing using Junction Tree algorithm.

The junction tree algorithm [Kahle 2008] is a method to extract marginalization in general graphs. In essence, it entails performing belief propagation on a modified graph called a junction tree. The basic premise is to eliminate cycles by clustering them into single nodes. The general problem here is to calculate the conditional probability of a node or a set of nodes, given the observed values of another set of nodes.

The basic concept in junction tree is clustering of predicted attributes. In belief updating instead of approximating joint probability distribution of all targeted variable (cliques) cluster attributes are formed and potential of clusters are used to approximate probability. So basically junction tree is the graphical representation of potential cluster nodes or cliques and a suitable algorithm to update this potential. Junction tree algorithm involve several steps as moralizing the graph, triangulation junction tree formulation, assigning probabilities to cliques, message passing and reading cliques marginal potentials from junction tree. Consistency in junction tree is a requirement which ensure that potential of a particular node with in two different cliques marginal probability of the node of interest is same .

IV BN RUST DISEASE DIAGNOSES

The development of Bayesian Network of rust disease diagnoses is carried out through six tire processes as below;

i. Identification of parameter/ variable of interest
ii. Identifying relationship, interdependence among variable
iii. Representing structure/topology of network through Directed Acyclic Graph(DAG)
iv. Estimating Conditional probabilities and joint probability distribution(JPD)
v. Belief updating using junction tree algorithm by marginalizing/ factoring JPD
vi. Inference BN through message passing algorithm

We have used a hybrid approach for learning the network. In the first step expert knowledge together with technical detail of occurrence of disease is used to identify the variable of interest define the interdependence of various factors [Fig. 2] and their expected probability. The following factors are identified significant in diagnoses of disease [Table1]

- Parts of plant infected,
- Shape and distribution of lesions,
- Lesion color,
- Degree of damage of tissue
- Visibility of damage
- Occurrence of Disease(Common, Occasional, Rare)

- In the second steps parameter learning of conditional probability dependence of variable is determined using data. The collected data divided in two parts as learning and test data set. Individual record selected randomly in two data set. However data contain replicate of all possible outcomes of

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

identified variable. Open source tool BNsoft used for structure learning through data. In third step model generated reviewed by the expert.

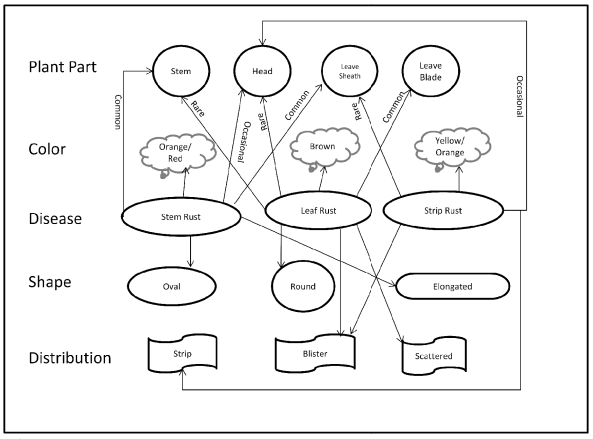
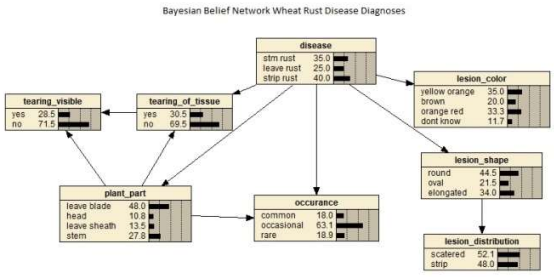


Fig. 2 Dynamics of Rust Disease Diagnoses

The developed Bayesian Belief network of rust (BBNRust) disease diagnoses depicted in Fig 2. The network efficiently estimates probability of occurrence of respective disease subject to instantiation of dependent variable. The network is capable of diagnosing the disease in case of missing instant of particular variable. The system can update the probability as soon as more information is available about variable.



The JPD of the BN is given as under

P (D) = P(LC) x P(Occ) x P(LS/B) x P(ToT/TV)
x P(PP/TV,TOT)

Where
D = Disease, LC = Lesion Color, Occ = Occurrence
LS = Lesion Shape, LD = Lesion distribution
ToT = Tearing of Tissue, TV = Tearing visible
PP = Plant Part

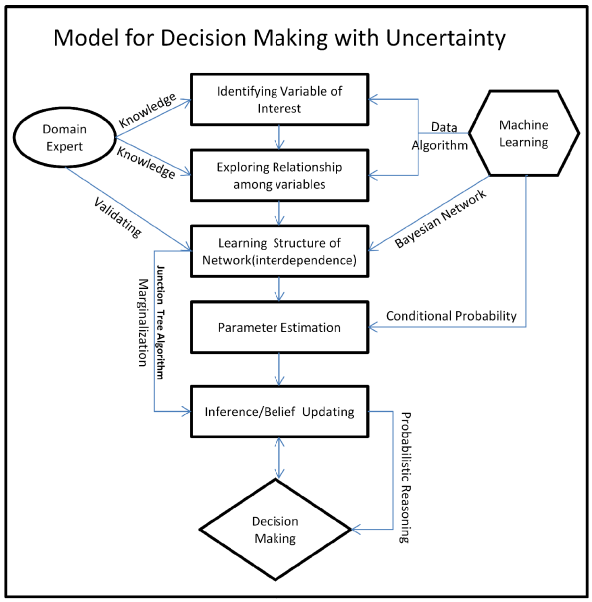
Simplifying joint probability distribution to marginalize the require probability is carried out Using junction tree algorithm. The following sets of clique are formed;

Clique [Joined To]	Member nodes (* means home)
0 [1]	(disease, plant_part, *occurance)
1 [0 2 3]	(*tearing_of_tissue, disease, *plant_part)
2 [1] plant_part	(*tearing_visible, tearing_of_tissue,
3 [1 4 5]	(*lesion_shape, *disease)
4 [3]	(*lesion_distribution, lesion_shape)
5 [3]	(*lesion_color, disease)

II RESULT AND DISCUSSION

Decision making in uncertain situation is a challenge particularly in plant disease diagnoses. Bayesian belief network proved to be an effective method for diagnoses of rust disease in wheat. The BN efficiently estimated conditional dependence of diagnostic parameters by capturing causative relationship between variable. Expert knowledge along with learning through data successfully identified underlying structure of the system.

We proposed a model (Fig 2) for developing a system for decision making in uncertain situation



The mechanism is multi facet process involving domain expert as well as state of art computer based machine learning methods to develop the system.

We have proposed a hybrid system as in many situations it is difficult use expert knowledge alone or purely learning the structure through data. The hybrid

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

approach ensures to capture the relationship which can not be distinguished with data. The model involves identifying variable of interest, exploring relationship and estimating parameter.

Employing the model BN for rust disease diagnoses is developed (Fig. 1). The BN diagnose the disease up to 81% accuracy. However variation exists among different kind of disease. The diagnoses of Stem Rust disease is more accurate as compare to other disease (Table 1)

Disease Diagnosed	**Accuracy Rate (%)**
Stem Rust	87.5
Leave Rust	78.3
Strip Rust	76
Over All	**81.3**

The proposed system is flexible as well as scalable. Bayesian network ensure inclusion of more variable of interest in rust diagnoses over period of time. Further the network can be extended for diagnoses of other plant diseases.

The overall accuracy of 81 % is not optimum main reason is the fact that shape and distribution of lesion is still posing confusion as human inspection may contribute to inaccuracy. The possible option is to use images recognition for distribution of lesion.

VI FUTURE WORK

The limiting factor of the proposed network, as mentioned, is the more precise recognition of shape of lesion which can be achieved by image processing. Authors have plan to undertake research for incorporation of automated image recognition component in the proposed system.

REFRENCES

1. Ben-Gal, I, 2007, “Bayesian Networks”, Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons
2. Burney, S. M Aqil, Nadeem Mehmood, 2012 “Generic Temporal and Fuzzy Ontological Framework, (GTFOF) for Developing Temporal-Fuzzy Database Model for Managing Patient’s Data, Journal of Universal Computer Science, vol. 18, no. 2 (2012), 177-193
3. Burney, S. M Aqil,, Jawed Naseem, 2010, “Efficient Probabilistic Classification Methods for NIDS”, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. 8, pp 168-172
4. Iowa, Ames, 2006,“Arterial Intelligence: An Overview” Iowa State University, https:// pdfs.semanticscholar.org
5. Kahle, David et al, 2008, Junction Tree Algorithm Rice University”, https://www.cs.helsinki.fi/u/bmmalone/probabilistic-models-spring- 2014/JunctionTreeKahle.pdf
6. Kolhe Savita, Raj Kamal, Harvinder S. Saini, G.K. Gupta, 2011, “A web-based intelligent disease-diagnosis system using a new fuzzy-logic based approach for drawing the inferences in crops”, Computers and Electronics in Agriculture Volume 76, Issue 1, Pages 16-27
7. Magarey, R.D.; Travis, J.W.; Russo, J.M.; Seem, R.C. & Magarey, P.A. 2002. Decision Support Systems: Quenching the Thirst. Plant Disease, Vol. 86, No. 1, pp. 4-14,
8. Murphy, K. (1998). “A brief introduction to graphical models and Bayesian networks”. http://www.cs.ubc.ca/~murphyk/Bayes/ Inintro. html.
9. Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, San Francisco.
10. Wegulo, Stephen N. et al, 2012, “Rust Diseases of Wheat” Institute of Agriculture and Natural Resources, University of Nebraska–Lincoln, http://extensionpublications.unl.edu/assets/pdf/g 2180.pdf
11. Zhi Ping Ding, 2011, “The Development of Ontology Information System Based on Bayesian Network and Learning,” Advances in Intelligent and Soft Computing”, Volume 129, , Pages 401-406

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

AUTHOR'S PROFILE

Dr. S. M. Aqil Burney[2] is Professor at College of Computer Science and Information Systems (CCSIS) at Institute of Business Management (IoBM) Karachi, one of the leading Business School in Pakistan. Dr. Burney was a Meritorious Professor (R.E.) and approved supervisor in Computer Science and Statistics by the HEC, Govt. of Pakistan. He was also the founder Project Director (UBIT) & Chairman of the Department of Computer Science, University of Karachi. He is also member of various higher academic boards of different universities of Pakistan. His research interest includes artificial intelligence, soft computing, neural networks, fuzzy logic, data science, statistics, simulation and stochastic modeling of mobile communication system and networks and network security, currently heading the detp. of actuarial science and risk management at CSCIS - IoBM. Teaching mostly MS(CS) Ph.D(CS))courses such as Data Warehousing, data mining & ML and information retrieval systems, fuzzy systems ,advanced theory of statistics, Markov chains and Financial Time Series.

Jawed Naseem[1] is Principal Scientific Officer (RE) in Pakistan Agricultural Research Council. He has MCS & M.Sc (Statistics) and currently a Ph D. scholar department of Computer Science, University of Karachi, Pakistan. His research interest includes data modeling, machine learning, probabilistic reasoning, Information Management & Security and Decision Support System particularly in health care and agricultural research. He has experience in research & education at national regional (SAARC) and international level.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The Sharp Increase in Unmasking of Obtrusion into Internet of Things (IoT) IPV6 and IPV6 Low –power Wireless Personal Area Network (6LoWPAN), a Lead Way to Secure Internet of Things Services

Modesta .E. Ezema
Computer Science Department
University of Nigeria Nsukka
Enugu , Nigeria
modesta.ezema@unn.edu.ng

Anthony .O. Okwori
Computer Science Department
Federal University Wukari
Wukari , Nigeria
okwori@fuwukari.edu.ng

Francis . A. Okoye
Computer Engineering Department
Enugu State University of Science and Technology Enugu
Enugu , Nigeria
francisce@esut.edu.ng

Christopher .C. Ezema
info@infodata.com.ng
Victoria Island, Lagos Nigeria
chrisraph4@gmail.com

Abstract*— The Internet of Things (IoT) is thriving network of smart objects where one physical object can exchange information with another physical object. In today's Internet of Things (IoT) the interest is the concealment and security of data in a network. The obtrusion into Internet of Things (IoT) exposes the extent with which the internet of things is vulnerable to attacks and how such attack can be detected to prevent extreme damage. It emphasises on threats, vulnerability, attacks and possible methods of detecting intruders to stop the system from further destruction, this paper proposes a way out of the impending security situation of Internet of things using IPV6 Low -power wireless personal Area Network.

Keywords **security; data; threat; network;**

I. INTRODUCTION

The connection in Internet of Things hardware, communication and software implementations are always connected through low –power IPV6 which is untrusted and unreliable, these has led to the increase in the attacks and threats on these devices[1] , Encryption and authentication could have helped but for the exposure to wireless attack from IPv6 Low-power Wireless Personal Area Network and the internet.[2] using wireless sensor network could have been better as IPv6 Low-power Wireless Personal Area Network are connected to the distrusted internet and the cybercriminal can get access to the hardware and software resources from any where on the internet. Access to the internet from any where in the world helps obtrusion into Internet and makes vulnerable even for the attackers targeting 6LoWPAN networks. These vulnerabilities have been showing up targeting the physical interfaces of IoT devices, wireless protocols, and user interfaces [3].Providing security in IoT is difficult as the channels for exchange of information are not stable , and the devices uses a set of unique IoT mechanizations such as Routing Protocol for low power and lossynetwork(RPL) [4]. Therefore, to have a safe implementation of information exchange in IoT an Intrution detection system (IDS) must be implemented to guard against cyber-attacks. [5]. A more-realistic approach to develop a security system for IoT comprises of the following; Analyzing an attack to guard against it in the future, avoiding occasions that can lead to an attack in the future , detecting an attack before it is carried out by the attacker and finally identifying security breaches .Also there must be measures to identify misuse of the computer system restricted access and abuse of computer resources [6]. It can be a software or hardware tools that inspect and investigate machines and user actions, detect signatures of well-known attacks and identify

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

malicious network activity. It aims at observing the networks and nodes, detect various intrusions in the network, and alert the users after intrusions had been detected. It works as an alarm or network observer to avoid damage to the systems by generating an alert before the attackers cause any harm to the system. The IDSs for internet of things monitors several devices connected by a network.

A ***Threats Associated to IoT***

Increase in connections of equipments to the internet attract more cybercriminals this is only because IoT devices have less security protections against cyber threats and they are easy to exploit [7] Cyber criminals take advantage of poorly protected IoT devices to spy on people ,cause physical damage, and to project massive denial of service attacks. Some IoT devices that can control threats to the internet of Things network are as follows;

•**Smart Grid:** Recent discoveries in Homeland Security Department flawed the hardened grid and router provider products of RuggedCom. This was achieved by reducing the traffic between an end user and the RuggedCom products, which could led an attacker to launch an attack to expediently accept standards that are lower than the energy grid[8].

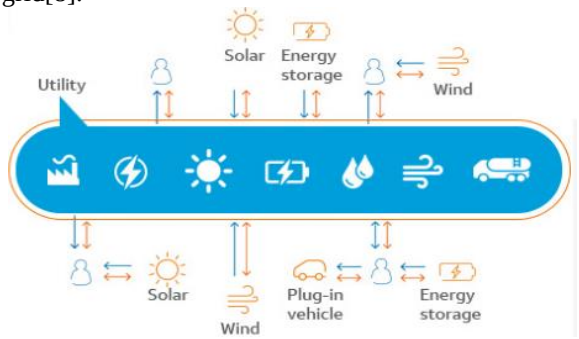


Fig .1, Smart grid

•**Home Network Routers:** Of all the Internet connected devices in homes these days, the network router continues to be by far the most targeted in attacks [9]. "Most Internet routers that are keystone to our home network are riddled with security issues, which make them easy picking for hackers. Most routers are locked with authentication code worldwide had default or basic username and password combinations, like "admin" and while others some use other relevant information like office or home address and birthday of the user to authenticate access to the router . Because of this, most routers are vulnerable to simple password attacks, which is basically an open invitation to malicious hackers.Not surprisingly, attackers have begun taking advantage of vulnerable home routers to create botnets for relaying spam and launching DDoS attacks.

•**Digital Video Recorders (DVRs)**: The near ubiquitous set-top boxes, which people use in their homes to record TVs shows, have become another favorite target for attackers. Recently most of the massive DDoS attacks are connected to Compromised DVRs .as discovered by investigation they do this by creating big botnets of such devices for use in various malign ways. The security controls of DVRs is next to none most of them are connected to the internet with weak user name and password . Often DVRs from multiple manufacturers integrate components from the same supplier. As a result, a security flaw in one product is likely to exist in another vendor's product as well. Security vendor Flashpoint recently analyzed malicious code that was used in DDoS attacks involving IoT devices. The company discovered that a large number of DVRs being exploitedby the malware [10] were preloaded with management software from a single vendor. The supplier sold DVR, network video recorder (NVR), and IP camera boards to numerous vendors who then used the parts in their own products. Flashpoint estimated that more than 500,000 network-connected DVRs, NVRs, and IP cameras were vulnerable to the attack code because of a vulnerable component from a single vendor.

• Smart Fridges/Smart Home Products: In January 2014, a researcher at security vendor Proof point who was analyzing spam and other e-mail borne threats discovered an Internet-connected refrigerator being used to relay spam. The incident was used to offer proof of what analysts have for some time been stressing: the startling vulnerability of many network-enabled devices being installed in homes these days such as smart fridges, TVs, digital assistants, and smart heating and lighting systems.Refrigerators, personal assistants, and TVs have enough processing power to be used in botnets or to be used as access points to the rest of the network Lamar Bailey in [11]. Such devices pose a threat in the enterprise context as well. For instance, a connected fridge in an office break room could provide an unexpected gateway to systems containing corporate data. This isn't about hacking the fridge, it's about hacking through it to gain network access. Since the connected fridge is on the corporate network, which also connects to enterprise applications, it can be leveraged and exploited by hackers to gain valuable corporate and customer data.[12]

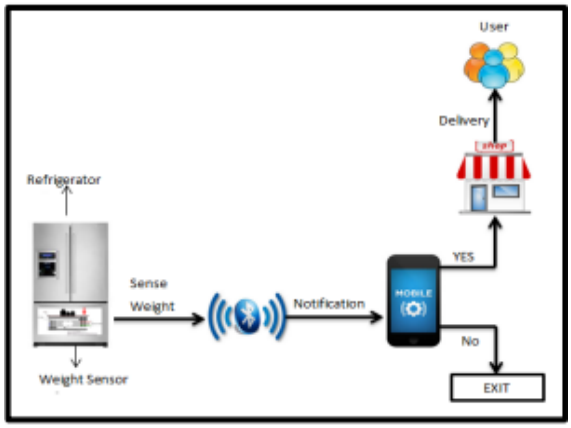


Fig .2, Block diagram of Smart Refrigerator

•**Implantable Medical Devices:** Vulnerabilities in wireless-enabled implantable medical devices such as insulin pumps, pacemakers, and defibrillators make them tempting targets for malicious attacks. In recent years, security researchers have shown how attackers can take advantage of unencrypted and

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

generally weak communications protocols in such devices to gain remote control of them and to get them to behave in potentially lethal ways. An attacker could take advantage of weaknesses in the wireless management protocol and pairing protocols of devices like insulin pumps to gain remote access to it and get it to release lethal doses of insulin to the wearers of this device.[13]

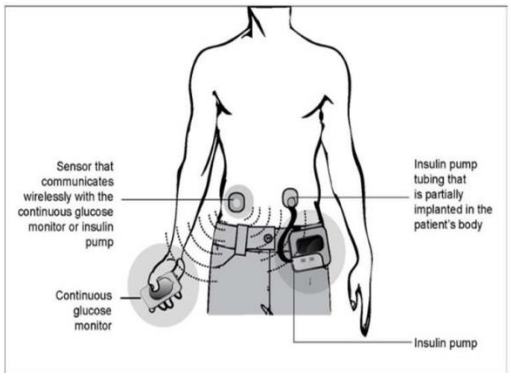


Fig .3, Continuous Glucose Monitoring System and Insulin Pump

•**Supervisory Control and Data Acquisition (SCADA) Systems:** The Supervisory Control and Data Acquisition (SCADA) systems that are used to manage industrial control equipment and critical infrastructure are part of the IoT devices that are vulnerable as many SCADA systems are now network-enabled but lack efficient security controls by using hard-coded passwords and poor patching processes. Also, Industrial controllers (SCADA) systems that have been in place and are difficult to update are specifically vulnerable for attacks. Attackers could use compromised SCADA systems in DDoS attacks or in ransom ware attacks.

•**Baby Monitors**: Consumer products that are used to monitor babies are another category of IoT devices that are vulnerable to attacks and compromise. Some vulnerabilities associated with baby monitors includes: hard-coded passwords, unencrypted communications, privilege escalation, easily guessable passwords, backdoor accounts, and flaws that would have let an attacker alter device functions[14].These vulnerabilities let attackers hijack video sessions, or view video stored in the cloud, or gain complete administrative control of the baby monitor. All of the flaws were easy to exploit and would have given attackers varying degrees of remote control over compromised devices. This vulnerable device could pose a threat to any computer connected to the home network, including those used by remote workers. An infected IoT device could be used to pivot to other devices and traditional computers by taking advantage of the unsegmented, fully trusted nature of a typ+ical home network. [15]

Fig .4, Baby monitor

•**Connected Cars:** Most Modern cars are part of the IoT devices as the irnumerous components are network-accessible and exposed to network-borne threats. The weaknesses in the controller area network of a Jeep Cherokee could be exploited to gain remote control of the vehicle's accelerator, braking, and steering systems. Other threat in connected cars includes proof-of-concept attacks on Toyota and Ford models[16]

Fig. 5, Internet of Things connected vehicles

B ***Vulnerabilities***

- **Insecure Web Interface:** to break into IoT systems the cyber criminal apply some of these tactics like trusting default passwords, shaky passwords, or offering a "forgot password" functionality [17]. The following can also lead to insecure web interface they includes XSS attacks(cross- site) this happens when a cyber-criminal uses a web application to send malevolent code, generally in the form of a browser side script, to a many end users, cross-site request forgery as well as sequential query langauge injection.
- **Insufficient Authorization /Authentication:** the truth is that to access web interface has a lot of security implications **.** when an uncertified user gain access to the web interface it is usually disastrous [18]. It is therefore necessary to improve authentication and authorization to adequately protect credentials.Such authentication and authorization should equally be revoked when necessary. It is important to ensure that application, device, and server authentication are required. A unique session keys are required with authentication token.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- **Threatened Network Services:** For a reliable and secure network services device's open ports should be reviewed and tested for incursions as DoS attacks.
- **Deficiency in Integrity verification/Transport Encryption** To reduce this vulnerability, the network traffic, mobile applications as well as could connections should not pass any clear text along the transport layer. The encryption protocols should be ensuredand the use of SSL and TLS should be up to date.
- **Insecure Cloud Interface:** To be certain of the cloud interface it is vital to keep to default username and passwords, block user accounts that fail to login after a defined number of attempts should be monitored and also review all cloud interfaces for possible attack.
- **Insecure Mobile Interface:** Ensure to determine if the authorization are mistakenly exposed, when hooked up to wireless networks and to offer two-factor authentication options.
- **Poor Physical Security:** Physical security can equally enhance vulnerability in the IoT network. Storage medium should be secure from easy removal, stored data be encrypted, prevent bad actors from gaining access to the ports as well as ensuring that the device cannot be easily disassembled

This template, created in MS Word 2000 and saved as "Word 97-2000 & 6.0/95 – RTF" for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

C *Attacks Associated with Internet of Things*

Different types of vulnerabilities exist but the attack on Internet of things are overwhelming . Internet of things connects millions of equipment that are potential victims to traditional style cyber-attacks. At its core, the Internet of things continue to connect and network devices that up until now have not necessarily been connected. The implication is that those equipment whether new or old creates another entry point there by posing another security risk to the system. When the cyber-criminal attacks a network its effect varies depending on the ecosystem, the equipment and the environment ,e.t.c. , the available protection level and many more. Some cyber attacks and its effect on the IoT are discussed below;

Botnets: this is another form of malware distribution where systems are connected with the purpose to distribute malicious code. The inter connected systems may include personal computers , servers, mobile devices and IoT devices [19]These systems may be used by the cyber criminal to hack private information from the bank, from online operations , e.t.c. there is another form of botnet called thingbots which gathers all connected objects like mobile phones, personal computers and other smart devices which are internet enabled . Botnet and thingbot have many things in common including transferring data via a network .

Man-In-The-Middle Concept: As the name implies here a cyber-criminal seeks to breach communication between two systems, he quietly intercepts the communication of two parties when they believe they are communicating with each other. The recipients is played to believe that he is getting the right message, cases of hacked vehicle, smart refrigerators are documented for IoT in a specified threat area [20].These attacks can be alarming in the internet of things as the nature of what is being hacked matters, these includes industrial tools, machinery, vehicles, smart televisions or garage door openers.

Data and Identity Theft: This is brought about by the laxity of the user for mishandling his devices there by giving way for an opportunistic user to have access to it. When this happens the cyber-criminals can access your bank account and useful information they may get from your internet connected devices like smart watches, phones, e.t.c., their target is to amass data**.** They can go extra mile by seeking information about their victim in the social media . this will give them idea of personal identity. When they get detailed information about the owner , the easier and the more dangerous a purposeful attack aimed at identity theft can be.

Social Engineering: this is the act of maneuvering people in order to obtain their confidential information such as passwords, bank information or by accessing a computer so that they can quietly install malware codes that will give them access to their private data of their victim and they will take over the computer. They do this in so many ways like sending phishing emails, to divulge information or send it to websites of some financial institution and business sites that look legitimate, enticing users to enter their details.

Denial of Service (DoS): In this type of attack a needed service is declared unavailable to frustrate the user, there is a situation where many systems are involved and it is called Distributed Denial of Service (DDoS) attack, here a larger number of systems malevolently attack one target. In a botnet attack, many devices are programmed to ask for a service at the same time . This is often done through a botnet. For instance in 21stOctober 2016, there wj as an attack on internet activity because of (DDoS) attack in the US [21]. The cause of the attack was that there were unsecure connections of numerous devices , these devices includes the under listed as home routers, surveillance cameras among others. The cybercriminal use many of such devices that had been infected with malevolent code to form a botnet.[21]

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Fig. 6, a network of system showing (DDoS) attack in IoT

D ***IntrusionDetection***

We have software and hardware devices that monitor the activities of malevolent applications and policy violations in a network . when such activities are noticed or detected report is sent to a management station these software and hardware devices are called Intrusion Detection System. An IDS audit categories of network activity and pinpoints distrust patterns that may reveal a network or system of attack from someone planning to break into or accord a system. Some categories of detection system are as following;

Barbaric detection/ Exception detection: In barbaric detection, the Intrusion Detection System analyses the information it gathers and compares it to large databases of attacker signatures to search for a specific attack that has already been documented. Barbaric detection software is only as good as the database of attack signatures that it uses while in barbaric detection, the system administrator defines the baseline state of the network's traffic load, protocol, and typical packet size.
The exception detector monitors network segments to compare their state to the normal default of the system and look for changes .

Hook-up-based/Manager-based systems: In a hook-up Intrusion Detection System based system (HIDS) the individual packets flowing through a network are analyzed. Thehook-up Intrusion Detection System (HIDS) can detect malevolent packets that are designed to be disregarded by a firewall's filtering rules while in a manager-based system, the Intrusion Detection System (IDS) understudies the activity on each individual system or host such as monitoring of the systems ' s configuration files to discover unsuitable settings; also it checks the file containing the password for wrong passwords, and monitor other system areas to detect policy violations.

A hook-up based Intrusion Detection System (HIDS) sensor has two interfaces [22]. They are manageable interface and the listening interface which is in dissolute mode. This dissolute interface can not be accessed over the internet. and it is not manageable. The monitoring interface which is manageable is connected to the network segment, that is being monitored. The sensor accesses every packet that crosses the hooked up segment. Hooked up based sensors apply predefined attack signatures to each frame to identify strange traffic. If it finds a match against any signature, it notifies the Intrusion Detection System management console see fig.7, below,[21].

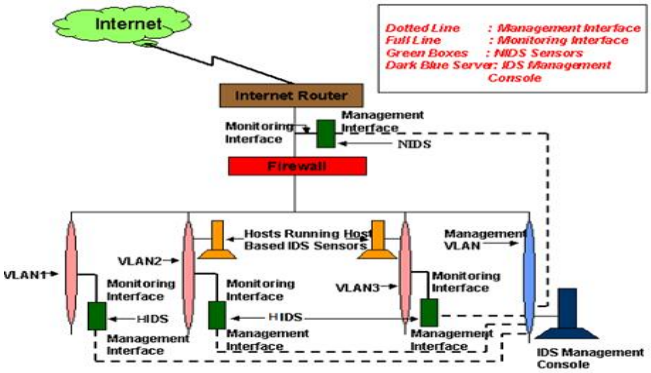


Fig .7, Implementation of Intrusion Detection System (IDS) Sensors and in a network

5.3 Idle vs Sensitive systems: In an idle system, the intrusion detection system (IDS) sensor reveals a potential security breach, sends the information and flags an alert on the system the owner while In a sensitive system, the IDS reacts to the distrustful activity by logging off a user or by changing the firewall to block network traffic from the suspected malevolent source.

Obtrusion detection systems came up as a result of increasing cases of attacks on major sites and networks, those at Pentagon, the White House, e.t.c were all inclusive. The protecting of our systems from cyber criminals are becoming increasingly difficult, this is because the technologies they attack even though it is becoming ever more sophisticated; at the same time, less technical ability is required for the novice attacker, because proven past methods are easily it can be accessed through the Web. The work of Intrusion detection system IDS includes: watching over and analysing both user and system activities, analysing system configurations and vulnerabilities ,assessing system and file integrity,ability to recognize patterns typical of attacks,analysis of abnormal activity patterns as well astracking user policy violations

E ***Proposed Obtrusion detection methods for IoT***

Lean obtrusion detection system: This was among the earliest Intrusion detection system designed for Internet of Things (IoT) [23]. It is made of an integrated firewall, which consists of 6LoWPAN Mapper that extract information about the network and construct it using IPv6 Routing Protocol for Low-Power and Lossy Networks

It recognizes obtrusion by analyzing the mapped data.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

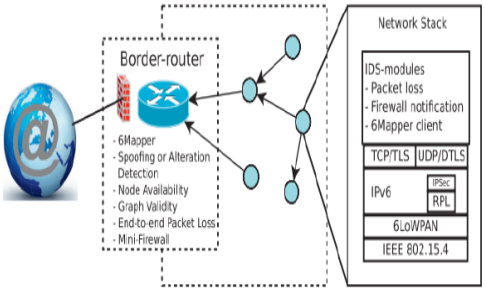


Fig .8, Lean Intrusion Detection system in IoT

An Engine Based Obtrusion Detection Technique : This is a similar obtrusion detection method for the vast diverse IoT networks based on an automata model. This method can signal and report the possible IoT attacks with these three methods of attack namely; jam-attack, false-attack, and reply-attack automatically [23]

Outcross Obtrusion Detection method: This method of obtrusion detection in internet of things was discovered by Sedjelmaci et al In [23] using Game Theory. This method mixed the usage of signature and aberration ways for IoT obtrusion detection. It accomplish this by creating the game model of infiltrator and normal user.

Complicated Event-Processing IDS: this method was invented by J. Chen and C. Chen. It is which is a Actual time pattern matching system for IoT devices .This method uses the Complex Event Processing (CEP) that focuses on the use of the features of the events flows to determine the intrusions, which can minimize the fake alarm rate comparing with the conventionall intrusion detection methods.

Fake Neural Network (FNN) Intrusion Detection System: Here supervised fake neural network is tracked using internet packet traces and was assessed on its ability to circumvent Distributed Denial of Service (DDoS/DoS) attacks on IoT devices [24].The detection was based on classifying normal and threat patterns. It was able to identify successfully different types of attacks and showed good performances in terms of authentic and untrue rates.

F *Conclusion*

This research presents an overview of obtrusion detection in internet of things as well as detail knowledge of various threats, vulnerabilities, attacks and available methods of detecting a cyber criminal in our internet of things(IoT) .It exposed various ways cyber criminals thrive to delude users in other to hijack their authentic data to manipulate it and cause them pain. It also enlisted ways of protecting your data to avert such mayhem in our thriving Internet of things technology(IoT).

References

[1] Mohamed Abomhara and Geir M. Køien, (2015), "Cyber Security and the Internet of Things: Vulnerabilities, Threats, Intruders and Attacks" Department of Information and Communication Technology,University of Agder, Norway

[2] ShahidRazaa,_, Linus Wallgrena, ThiemoVoigta,baSwedish, "SVELTE: Real-time Intrusion Detection in the Internet of Things" http://www.cs.umanitoba.ca/~comp7570/assets/media/0404Singh_M.pdf accessed on 15th/05/2017

[3] Adam Kliarsky, (2017), "Detecting Attacks Against The 'Internet of Things"

[4] PavanPongle, GurunathChavan , (2015), "Real Time Intrusion and Wormhole Attack Detection in Internet of Things" International Journal of Computer Applications, Volume 12.

[5] Tariqahmad Sherasiya1, Hardik Upadhyay2 &Hiren B Patel3, (2016), "A Survey: Intrusion Detection System For Internet Of Things", International Journal of Computer Science and Engineering (IJCSE), Vol. 5, Issue 2, page: 91-98

[6] Nicholas J. Puketza, Kui Zhang, Mandy Chung, Biswanath Mukherjee and Ronald A. Olsson, (1996), "A Methodology for Testing Intrusion Detection Systems" Department of Computer Science University of California, Davis Davis, CA 95616 Second revisionM. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[7] JAI Vijayan, (2016), "7 Imminent IoT Threats" available at http://www.darkreading.com/endpoint/7-imminent-iot-threats/d/d-id/1327233accessed on 12th/05/2017

[8] Smart Grid Solution available at https://www.business.att.com/enterprise/Service/internet-of-things/smart-cities/iot-smart-grid/ accessed on 14th /05/2017

[9] Http://www.darkreading.com/endpoint/7-imminent-iot-threats/d/d-id/1327233?image_number=2Accessed on 15th /05/2017

[10]]http://www.darkreading.com/endpoint/7-imminent-iot-threats/d/d-id/1327233?image_number=3Accessed on 13th/05/2017

[11] http://www.darkreading.com/endpoint/7-imminent-iot-threats/d/d-id/1327233?image_number=4 Accessedon 14/05/2017

[12] Rishabh S. Khosla, Pranul S. Chheda, Smith R. Dedhia, Dr.BhaveshPatel , (2016), International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 4 Issue: 1, Shah & Anchor Kutchhi Polytechnic, Mumbai, Indias"

[11] PavanPongle, GurunathChavan , (2015), "Real Time Intrusion and Wormhole Attack Detection in Internet of Things" International Journal of Computer Applications, Volume 12.

[12] Tariqahmad Sherasiya1, Hardik Upadhyay2 &Hiren B Patel3, (2016), "A Survey: Intrusion Detection System For Internet Of Things", International Journal of Computer Science and Engineering (IJCSE), Vol. 5, Issue 2, page: 91-98

[13]] Hacking Implantable Medical Devices, (2014), available at http://resources.infosecinstitute.com/hcking-implantable-medical-devices/ Accessed on 15th/05/17

[14] http://www.darkreading.com/endpoint/7-imminent-iot-threats/d/d-id/1327233?image_number=7Accessed on 13th/05/17

[15] Aamir Lakhani , Hacking, (2016), "When Baby Monitors are a Model For Iot Security", available at http://www.drchaos.com/when-baby-monitors-are-a-model-for-iot-security/Accessed on 15/05/17

[16] Mr. Shashank Dhaneshwar, (2015), "Internet Of Things Application For Connected Vehicles And Intelligent Transport Systems", Available at https://www.slideshare.net/shashankdhaneshwar/iot-applications-for-connected-vehicle-and-its accessed on 13th/05/2015

[17] http://www.ioti.com/security/10-biggest-iot-security-vulnerabilities/gallery?slide=1 ACCESSED ON 15TH/05/2017

[18] http://www.ioti.com/security/10-biggest-iot-security vulnerabilities /gallery?slide=2 ACCESSED ON 13TH/05/2017

[19] http://searchsecurity.techtarget.com/definition/botnet

[20] "5 Common Cyber Attacks in the IoT - Threat Alert on a Grand Scale" (2016),

[21] STEPHEN COBB, (2016), "10 things to know about the October 21 IoTDDoS attacks" Available at https://www.welivesecurity.com/2016/10/24/10-things-know-october-21-iot-ddos-attacks/ accessed on 15th/05/2017

[22] SANS Institute 2001, Intrusion Detection Systems: Definition, Need and Challenges, available on https://www.sans.org/reading-room/whitepapers/detection/intrusion-detection-systems-definition-challenges-343

[23] Yulong Fu, Zheng Yan, Jin Cao, Ousmane Koné and Xuefei, (2017), "An Automata Based Intrusion Detection Method for Internet of Things" available at https://www.hindawi.com/journals/misy/2017/1750637/ accessed on 13th/05/2017

[24] Elike Hodo, Xavier Bellekens, Andrew Hamilton, Pierre-Louis Dubouilh, Ephraim Iorkyase, Christos Tachtatzis and Robert Atkinson, "Threat analysis of IoT networks Using Artificial Neural Network Intrusion Detection System", available at https://arxiv.org/ftp/arxiv/papers/1704/1704.02286.pdf accessed on 15th/05/2017

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

CROWD CONSCIOUS INTERNET OF THINGS ENABLED SMART BUS NAVIGATION SYSTEM

Devi Mukundan
PG Student ,
Department of Computer Science and Engineering
R.M.D Engineering College
Kavaraipettai,Tamil Nadu, India
devibe281@gmail.com

Dr.P.Ezhumalai
Professor &Head of Department
Department of Computer Science and Engineering
R.M.D Engineering College
Kavaraipettai,Tamil Nadu, India.
hodcse@rmd.ac.in

***Abstract*— Public transport service is one of the most preferred modes of transportation in today's smart cities. People prefer public transport mainly for the cost benefit reasons. The problems faced by the people while using the public transport can be overcome by the technology such as Internet of Things (IOT). In this paper, we present how this technology can be applied to eliminate the problems faced by the passengers of the public bus transport service. The Internet of Things technology is used to provide the passengers waiting at the bus stop with real time information of the arriving buses. Information such as arrival time, crowd density and traffic information of the arriving buses are predetermined and provided to the passengers waiting at the bus stop. The display boards fitted at the bus stops provide the real time bus navigation information to the waiting passengers. This Smart Bus Navigation system enables the passengers to make smart decisions regarding their bus journey. This system reduces the anxiety and the waiting time of the passenger's at the bus stop. The smart bus navigation system creates a positive impact and increases the number of people who prefer to use the public mode of transportation.**

Keywords-Public Transport; smart cities; Bus Navigation System; Internet of Things; display boards.

1. INTRODUCTION

People in smart cities prefer smart modes of transportation where they can reach their destination in a faster and efficient manner. An intelligent public transport service is one of the essential needs for the fast growing cities of today to satisfy the requirements of the urban mobility. People prefer the public bus transport for the social, economic and environmental reasons. Though the public bus transport system have their own advantages they suffer from several drawbacks .Passengers are seen waiting at the bus stop without knowing the exact arrival time of the bus. One of the important requirements of the modern traveller information system is the provision of arrival time predictions of the next available bus or train [1]. The crowd density in the arriving bus is not known to the passengers which results in long waiting time at the bus stops. The traffic information of the arriving buses is not known to the passengers. The technologies play an important role to overcome the problems faced by the public transport system. [2],[3],[4].

The Internet of Things (IOT) technology provides the way to overcome the drawbacks in the existing public transport system. The IOT in transportation can be used for control, communication and information processing across various transportation systems. The use of smart devices and powerful enabling technology improves the data collection, automation and operations. Passenger satisfaction levels among public transport users were found to decline for those who travel on crowded or unreliable bus services and those who have long wait-times at the bus stop [5].

In this paper we show how the Internet of Things Technology can be used to improve the satisfaction of the passengers by reduced waiting time at the bus stop. The smart bus navigation system reduces the anxiety of the passengers at the bus stop by providing the 1) arrival time information 2) crowd density in the arriving bus 3) traffic information of the arriving bus. Altogether Smart Bus navigation System provides the passengers with all the information that is required for a pleasant journey.

The remainder of the paper is structured as follows. In section 2 we review related work. Section 3 will be problem analysis. Section 4 will be proposed work and finally section 5 will be conclusion.

2. RELATED WORK

One of the key problems that is found by the cities in today's world is finding trustworthy public transport services that understand the needs and the demands of the passengers [6].Marcus Handle et al., propose the Urban Bus Navigator (UBN). The main feature of the urban bus navigator is that it provides the micro navigation and the crowd aware route recommendation to the passengers. The passengers are guided along their journey and the route recommendations enable the passengers to take better decisions along their bus travel. Though this system has many useful features it does not consider the traffic conditions along the road which may affect the arrival time of the buses at the bus stop. Wenping Liu et al., discusses about WiLocator which is a powerful tool that was implemented to tackle the problems faced during the arrival time prediction of the bus[7]. This tool partition the radio frequency signal space of the Wi-Fi access points. It

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

tackles the problem of noisy signal and AP dynamics. Smartphones are used to collect the navigation information that is predicted using the surrounding Wi-Fi signal information.In this paper Signal Vornoi Diagram(SVD) is used to tackle the noisy received signal strength and uses tool called WiLocator for navigation. It provides accurate and real-time traffic map and predicted travel time on each road segment. This system does predict the crowd in the bus and it is Smartphone based.

GPS technology can be used to gather information regarding the location of the vehicles. It can be used for single vehicle or a group of vehicles. M.B.M Kamel proposes a vehicle tracking system that is based on GPS and GPRS [8]. It uses traffic modified coding method to encode and compress data about the location before it is transmitted to the destination. It uses a simple security mechanism that guarantee the privacy of the transmitted data. . A protected web interface is used by the authorized user to track the vehicle. This system involves cost effective usage of network traffic. It also includes the drawbacks which are seen commonly in GPRS systems like the distance factors. Cemil Sungur et al., propose a smart bus station passenger information system that provides the passengers waiting at the bus stops with the current location and the status of the vehicle [9]. Embedded mini-computers and digital monitors are used to provide the location information. The passengers are provided with information such as bus status information, remote bus information and status management. Micro-Navigation discussed by S.Foell et al proposes a tool called the Urban Bus Navigator (UBN) which is a reality aware navigation system [10].Micro-Navigation is done by using the Internet of Things Technology. The proposed system provided end to end route guidance to the bus riders. Though micro navigation improves the satisfaction of the passengers it does not predict the crowd in the arriving buses and it is smart phone based. P.Zhou et al., proposes a system for predicting the bus arrival time by using a mobile phone based participatory sensing [11].This system gathers the navigation information from the cell tower signals, movement statuses, audio recordings etc., rather than from the GPS. This system provides accurate travelling route and arrival time estimates than GPS operated solutions. It results in loss of information when disruptions occur in cell tower signals.

A.Thiagarajan et al., discusses about a crowd sourced technique for transit tracking [12]. This system makes use of built in sensors, GPS modules, Wi-Fi and accelerometer to detect the user's activity. It determines whether the user is driving in a transit vehicle or not. A central tracking server is used to send periodic and anonym zed location updates The underground vehicle tracking can also be done using this system. J.Zimmerman et al., proposes a transit information system called Tiramisu where the commuters share GPS traces and also submit problem details [13]. The incoming traces of information from the commuters are processed by the Tiramisu and it generates real-time arrival time for the buses. The proposed system was also fielded trailed with 28 participants. This paper mainly discusses how crowd sourcing can be used to generate cooperative production between the commuters and the public transport services

3. EXISTING TECHNOLOGIES

There are various technologies that are used for real time bus navigation. Researchers have increasingly turned their attention to digital technologies that can overcome the inherent drawbacks that is observed in the present bus transport system and can lead to an efficient bus transport system. Technologies such as Zigbee, RFID, GPS, GLONASS are used for real time bus tracking.

3.1 Satellite Navigation System

3.1.1 GPS

The Global Positioning System (GPS) is United States Government owned space based radio navigation. It is also known as Navstar(Navigation System for Training and Ranging).It was initially developed with 24 satellites. It currently comprises of 31 satellites orbiting the earth every 12 hours at 12,000 miles in altitude.The first generation of GPS was developed by the US Department of Defence in 1973 for military purposes. GPS has global coverage and it works in all weather conditions.

The GPS system comprises of three segments: space segment, user segment, and control segment. The space segment consists of 24 to 32 satellites. It helps to locate the position of the object by broadcasting the signal used by the receiver. The signals of four satellites are needed to calculate the position. The user segment includes military and civilian users. This segment comprises of a receiver which can detect signals and it consists of a computer to convert the data that is received to required information. The GPS receiver locates the position and it consists of security measures that disallow the person from being tracked by someone else. The control segment is required to work efficiently. It is that the transmission signals are kept updated and the satellite should be maintained in appropriate orbits. Using input of GPS we can identify the current location of the bus. With the help of built-in sensors, such as GPS, the application will automatically detect when the user is riding in the vehicle. The arrival time of the bus can be predicted with extreme accuracy, since estimates are constantly being updated in real time. GPS is used in many cities for bus navigation that improves the efficiency of city bus operation. Passengers could use in co creation of value by using the GPS equipped mobile phones they carry to generate real-time bus arrival information.

3.1.2 GLONASS

The Global Navigation Satellite System is a navigation system based on satellites developed in the Soviet Union and is operated by the Russian Aerospace Defense Forces. It was built to overcome the problems faced by Tsikada system. The Tsikada system required several hours of observation to provide accurate position. It can be used as a

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Technologies	Coverage	Frequency	Precision	Coding	Advantages	Disadvantages
GPS	Global	1.57542 GHz (L1 signal) 1.2276 GHz (L2 signal)	15m	CDMA	Global coverage Easy navigation Low cost	Does not pierce through solid walls and structures, accuracy depends on signal quality
GLONASS	Global	Around 1.602 GHz (SP) Around 1.246 GHz (SP)	4.5m-7.4m	FDMA	Global coverage Better accuracy than GPS at high latitudes	Satellite errors,atmosphererro rs (ionosphere, troposphere)
GALILEO	Global	1.164–1.215 GHz (E5a and E5b) 1.260–1.300 GHz (E6) 1.559–1.592 GHz (E2-L1-E11)	1m	CDMA	Global coverage, Better accuracy than GPS, GLONASS at high latitudes	Atmospheric errors, receiver noise
ZIGBEE	10-100 meters	2.4 to 2.4835 GHz (worldwide)	10m-20m approx	CSMA/CA	Low cost Low power Wireless technology	Short distance coverage, high replacement cost, less secure.
RFID	1-500 meters	120-150KHz (LF) 3.1-10GHz (microwave)	3feet(passive tags) 20-25feet(UHF) 300feet(Active tags)	TDMA(Aloha/slotted Aloha)	Easy to install No line of sight limitation RFID tags can store lot of information	Expensive, signal frequencies are non-standardised, privacy concerns

Table1 Comparison between different navigation technologies

Substitute for the GPS and it is the second navigational system which is used widely for accurate navigation information. It is composed of 24 satellites that provide the navigation details with precision. GLONASS is suited for usage in high latitudes where receiving a GPS signal can be problematic. It provides horizontal positioning accuracy within 5–10 meters. GLONASS is supported by devices like smartphones and tablets that provided that speed and accuracy in difficult conditions. some modern receivers combine both GLONASS and GPS together which provides improved coverage and efficiency

3.1.3 GALILEO

Galileo is a satellite based navigation system created by the European Union. It provides high precision in higher latitudes than other navigation system such GPS and GLONASS. It consists of a total of 30 satellites out of which 22 are operational. The first satellite was launched in 2011. The fully operational Galileo system will consist of 24 operational satellite and 6 active spares and it is scheduled for completion in2020. It is independent navigation system but it is compatible and interoperable with GPS. Galileo's higher orbit coupled with inclination increase enables it to have better Coverage at high latitudes compared with its counterparts. Galileo's signal design is expected to have improved acquisition and tracking, improved multipath performance, and improved building penetration. Their dual civil frequency mitigates the ionospheres uncertainties.

3.2 ZIGBEE

Zigbee is a wireless technology that is mainly aimed at remote control and secures applications. It is a low-cost, low power wireless network. It is best suited for several embedded applications, industrial control and home automation. It covers 10-100 meters within the range. It is less expensive and simpler than Bluetooth and wifi. Zigbee networks are extendable with the use of routers and many nodes to interconnect with each other for building a wider area network.

Zigbee structure consists of coordinator, router and end device. Coordinator is one of the essential devices in the zigbee network. It acts as the root and bridge of the network. The handling and storing of information is done by the coordinator. Zigbee routers are responsible for transmitting the data to and from other devices. End devices have limited functionality to transmit and receive data from the parent nodes. Zigbee

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

follows different types of network topology such as star, mesh and cluster tree.

3.3 RFID

The Radio Frequency Identification (RFID) system makes use of RFID tags for tracking the objects. It makes use of the radio signals to detect the presence of an object. RFID electronic tags do not require a viewable scan and it can carry essentially more information. The normal method used for identification is to store a serial number that identifies a person or object on a microchip which is fixed to an antenna. The chip transmits the identification information to the reader. The chip transmits the information with the help of the antenna .The radio waves that are reflected back from the RFID tag is converted into digital information that can be passes on to computers for using it efficiently.

RFID system consists of an unique serial number for every object. The serial number is used to identify the object. It transmits the identity of an object using radio waves. A RFID system consists of three components:RF tags (transponder),An antenna (coil),A transceiver. An RFID tag is made up of a microchip containing information for identifying the object. The chip contains a serialized identifier or bus stop identifier.

4 PROPOSED SYSTEM

Location Technologies that are based only upon GPS are defenceless and it needs to be supported by additional sources of information to obtain the desired availability, accuracy, integrity and uninterrupted service [14]. In the proposed system we are using Internet of Things (IOT) to implement an efficient real time bus navigation system. Providing additional information about the expected number of passengers can be very useful since it enables the passengers to travel in comfort [15].IOT can be used to determine the accurate location and the arrival time of the bus and also the crowd in the bus which leads to less crowded bus routes to the bus riders.

The Internet of Things (IOT) technology consists of a network of physical objects that require sensors and API's to communicate and exchange data over the internet. It requires sensors to collect the details about the object. It collects data by various technologies such as GPS, RFID and then autonomously flow the data between other devices. It consists of unique identifiers to transfer data without requiring human-to-human or human-to-computer interaction.

The proposed system provides the passengers waiting at the bus stop with the information such as arrival time of the bus, crowd density in the arriving bus and traffic information. The provision of this information to the waiting passengers enables them to make smart decisions regarding their journey.

4.1Location Tracking

The Buses are fitted with GPS modems to obtain the real time information about the location of the buses. The GPS modem receives the signal from at least three satellites. It displays the latitude and the longitude of the location based on the received signal. This information is used to predict the arrival time of the bus. The estimated arrival time of the bus can be obtained with extreme precision since estimates are constantly being updated in real time. The position values obtained from the GPS modem are sent to the microcontroller unit through serial communication. An IOT modem is used to transfer the location details to the cloud server. The information from the cloud server is retrieved at the bus stops. The arrival time information is displayed to the passengers through the LCD Display boards fitted at the bus stop.

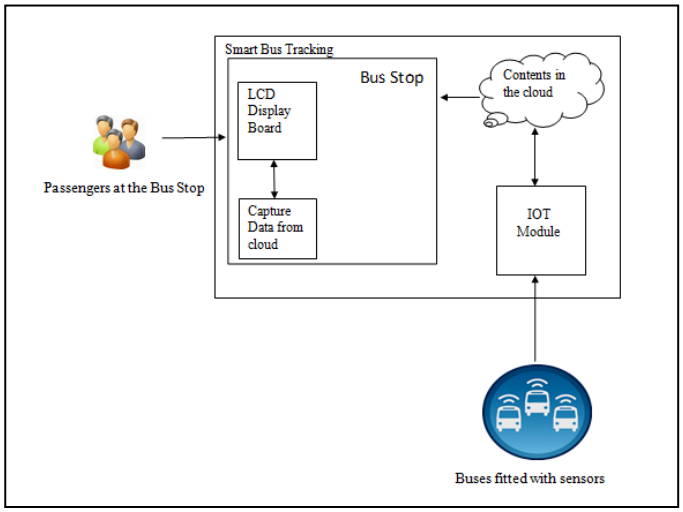


Fig 1:System Architecture

4.2Crowd Prediction

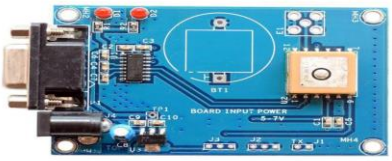
We often see that passengers are waiting at the bus stop without knowing the crowd in the arriving bus that is whether the arriving bus has enough seating capacity for the passenger. The IOT technology can be used to predict the number of passengers in the arriving bus. This reduces the waiting time of the passengers at the bus stop. This enables the passengers to make decisions as to whether wait for the bus or move on.

The buses are fitted with IR sensors to find the number of passengers in the bus. Each bus is fitted with two IR sensors. IR1 that is used to determine the in-count or the people entering the bus and IR2 that is used to determine the out-count or the people leaving the bus. The data received from the IR sensors is used to calculate the number of passengers in the bus. This information is collected by the IOT modem from the sensors and transferred to the cloud server. The crowd information is retrieved from the cloud at the bus stops. The people waiting at the bus stop are able to view the crowd in the arriving bus through the LCD Display Boards fitted at the bus stop.

4.3Traffic Analysis

The provision of traffic related information of the arriving bus to the passengers reduces the anxiety and the waiting time of the passengers at the bus stop. When the passenger is aware that the bus is stuck in traffic he is able to make better

https://sites.google.com/site/ijcsis/
ISSN 1947-5500



GPS with internal Antenna

IOT

decisions regarding his travel. The sensors are placed at the points on the road where traffic may occur such as near the traffic signals. The information from the sensor is collected by the IOT modem and transferred to the cloud server. This information is retrieved from the cloud at the bus stops. The people waiting at the bus stop are able to view the traffic information of the arriving bus through the LCD Display Boards fitted at the bus stop.

The IOT module plays a major role in retrieving all the information regarding the location, people count and traffic related information from the sensors fitted in the bus. IOT module consists of UART, controller that captures the information and stores it in a cloud server. The IOT module acts as an interface between the buses related information and the cloud server.

5 CONCLUSION

In this paper, a crowd conscious smart bus navigation system can able to enhance the passenger bus journey. The passenger is able to know the arrival time information and the crowd density of the arriving bus. The passengers are also provided with the traffic conditions on the road that enable them to make correct decisions regarding their bus journey whether to wait for the bus or not. When the crowd in the bus is beyond the seating capacity alternative bus options are provided to the passengers. Thus the system reduces the anxiety and the waiting time of the passengers at the bus stop. The bus information is stored in the cloud which is retrieved and displayed through the LCD Display Boards fitted at the bus stops. The Internet of Things (IOT) devices can be monitored and controlled by easy to use applications available thus improving the performance of the system. Thus the crowd conscious smart bus system enables the passenger to make smart decisions regarding their bus journeys.

REFERENCES

[1] Aidan O'Sullivan, Francisco C. Pereira, Jinhua Zhao, and Harilaos N. Koutsopoulos " Uncertainty in Bus Arrival Time Predictions: Treating Heteroscedasticity With a Metamodel Approachputing Systems "IEEE Transactions On Intelligent Transportation Systems, November 2016.

[2] T. Camacho, M. Foth, and A. Rakotonirainy, "Pervasive Technology and Public Transport: Opportunities Beyond Telematics," IEEE Pervaisive, pp. 18–25, 2013.

[3] J. Hare, L. Hartung, and S. Banerjee, "Beyond deployments and testbeds: Experiences with public usage on vehicular wifi hotspots," in Proc. of the 10th Intl. Conf. on Mobile Systems, Applications, and Services (MobiSys), 2012.

[4] S. Foell, R. Rawassizadeh, and Gerd, "Informing the design of future transport information services with travel behaviour data," in Proc. of the Workshop on SenCity: Uncovering the Hidden Pulse of a City, 2013

[5] M. Cantwell, B. Caulfield, and M. O'Mahony, "Examining the factors that impact public transport commuting satisfaction," *J. Pub. Transp.*, vol. 12, no. 2, pp. 1–21, 2009.

[6] Marcus Handte, Stefan Foell, Stephan Wagner, Gerd Kortuem, and Pedro Jose Marr ´ ," An Internet-of-Things Enabled Connected Navigation System for Urban Bus Riders", IEEE Internet Of Things Journal,2016

[7] Wenping Liu, Jiangchuan Liu, Hongbo Jiang, Bicheng Xu, Hongzhi Lin, Guoyin Jiang,Jing Xing, Huazhong University of Science and Technology, China, "WiLocator: WiFi-sensing based Real-time Bus Tracking and Arrival Time Prediction in Urban Environments", 2016 IEEE 36th International Conference on Distributed Computing Systems.

[8] M. B. M. Kamel, "Real-time GPS/GPRS based vehicle tracking system," International Journal Of Engineering And Computer Science, Aug. 2015

[9] CemilSungur,Ismail Babaoglu, AysegulSungur. 2015. Smart Bus Station-Passenger Information System. 2015 2nd International Conference on Information Science and Control Engineering

[10] S. Foell *et al.*, "Micro-navigation for urban bus passengers: Using theInternet of Things to improve the public transport experience," in *Proc.*

[11] P. Zhou, Y. Zheng, and M. Li, "How long to wait? Predicting bus arrival time with mobile phone based participatory sensing," *IEEE Trans.Mobile Comput.*, vol. 13, no. 6, pp. 1228–1241, Jun. 2014.

[12] A. Thiagarajan, J. Biagioni, T. Gerlich, and J. Eriksson, "Cooperative transit tracking using smart-phones," in *Proc. 8th ACM Conf.2010*

[13] J. Zimmerman et al., "Field trial of tiramisu: Crowd-sourcing bus arrivaltimes to spur co-design," in Proc. Annu. Conf. Human Factors Comput.S

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[14] I. Skog and P. Händel, “In-car positioning and navigation technologies—A survey,” IEEE Trans. Intell. Transp. Syst., vol.10, no. 1, pp. 4–21, Mar. 2009

[15]M. Handte et al., “Crowd density estimation for public Transport vehicles,” in Proc. Workshop Min. Urban Data MUD) Join Conf EDBT/ICDT, Mar. 2014, pp. 315–322.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Knowledge Extraction by Applying Data Mining Technique to Use in Decision Making for Development Policy Making

Md. Humayun Kabir
Dept. of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka-1342, Bangladesh

Abstract— **Data mining technique has a key role in knowledge extraction from databases to promote efficient decision making. This paper presents an approach for knowledge extraction from a sample database of some school dropped students using association rule generation and classification algorithms to demonstrate how knowledge-based development policy making decisions can be processed from the extracted knowledge. A system architecture is proposed considering mobile computing devices as user interface to the system connecting mass people database with cloud computing environment resources. The causes of education termination are investigated by analyzing the sample database in terms of attribute value relationship in the form of association rules to reason about the causes based on the computed support and confidence. It is observed that if the affected family had no service holders, the dropped student had to stop his education because of financial problem. Classification is applied to classify the dropped students in different groups based on their level of education.**

Keywords-database; data mining; knowledge extraction; decision making; development policy making

I. INTRODUCTION

In recent days, massive data are collected through customized application software operating various organizations. It is infeasible to extract knowledge from millions of data records which are stored using various RDBMS tools, e.g., Oracle, MySQL etc. manually for using in decision making. Various data mining tools e.g., Weka, DBMiner, Oracle Miner are available for mining knowledge from databases for easier and efficient decision making. Android-based mobile devices are massively used to access web-based applications. Data about the personal quality and activities of the mass people can be collected through web applications. Intelligent applications can be developed to be executed on application servers to access the mass people database to analyze data and extract knowledge to assist in decision making on their doings to improve their life standard. To build a knowledge-based developed society, the people's activities may be monitored and guided based on their personal information and daily activities to help and suggest accordingly as required to ensure development [1].

Data mining is the extraction of hidden knowledge or interesting patterns from databases or data warehouses. Various methods e.g., classification, association rule analysis and clustering can be applied to the database for the extraction of hidden knowledge or interesting patterns. In this research, the application of association rules [2]-[7] and classification technique [8]-[10] have been applied on a sample database to extract knowledge to aid decision making [11]. The use of association rules in classification purpose is presented in [9], [10]. One of the objectives of this research is to identify and define the real world problem using data mining concepts so that domain knowledge can be extracted using data mining technique to aid decision making. The other objective is to investigate the design of a framework integrating mass people database, mobile computing devices and data mining system within cloud computing environment [12]-[14] for knowledge-based development policy making [1]. This paper presents the steps toward development policy making by providing simple examples. The details of the development policy making is out of scope of this paper.

The paper is organized as follows. Section II describes development policy making. Section III explains how data mining technique can be applied to extract knowledge for decision making defining an example problem. Section IV presents the methodology defining the various stages of knowledge extraction and decision making for development policy making. Section V presents the proposed system architecture. Section VI provides experimentation and results using a data mining tool called Weka on a sample database to extract knowledge for decision making. Finally, section VII concludes and gives a guideline for future work.

II. DEVELOPMENT POLICY MAKING

Any development in personal or organizational level may be achieved by conducting some development activities for a period of time systematically. The deviation from the expected development in personal and organizational level may be traced using peoples and organizational databases. This may be monitored to try to construct successful life and organizational success by guiding the activities using intelligent human guidance and automatic intelligent devices executing intelligent software systems [1]. Android-based mobile devices can be used as an interface to access web-based applications connecting cloud database and application

The research work presented in [1] was funded by the University Grants Commission (UGC) of Bangladesh in 2014-2015 in the Faculty of Mathematical and Physical Sciences, Jahangirnagar University, Savar, Dhaka, Bangladesh.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

servers to get big database storage and computing support. These cloud database servers can be used to store data about the personal quality and activities of the people collected through various applications. For the proper running and faster development of any organization, better policies must be formulated, and need to be employed in its operation. Domain knowledge, operational data as well as contextual information play vital role in making successful policies for organizations. Business data collection and automatic processing of data using intelligent software systems may enable knowledge extraction from the collected data to aid decision making. Appropriate decisions may be taken and policies may be formulated based on the extracted knowledge considering the contextual factors. In this paper, three terms: i) domain knowledge ii) decision making and iii) development policy are key terms. The development policy making involves: i) Specifying the problem describing the intended development ii) Finding a solution involving decision making using the domain knowledge with identifying the development activities for policy formulation to solve the problem to achieve the expected development, and iii) Building an overall plan to implement the policy. Tracking progress in the domain is also required to monitor the success of the development policy.

III. DEFINING DATA MINING PROBLEM FOR DECISION MAKING

A social problem is stated below to clarify a real world system problem for automation and knowledge extraction for decision making to fix and formulate the appropriate development policy.

Example 1. A school going boy of a rural area is not going to his primary school for a month. An undergraduate student of that village noticed this. The fact is due to the shortage of money because of being a family member of a very low income group. So, he was unable to continue his school education because of being busy with money earning job to bear the family expenses. Such students can be identified of an area and their family properties can be analyzed using data mining techniques to identify the reason of termination of school education of the dropped students to provide any form of financial help to continue their school education.

In Example 1, a problem is stated which clarifies that a school going boy has to stop his school education because of being busy with money earning job to support his family expenses. In a region, various data about such school going boys who have to stop their school education can be collected through web-based applications using mobile devices or any computers where Internet facility is available. The collected data can be stored in database servers for further processing to extract knowledge about the problem to reason about the cause of termination to help decision making to support their further school education. Finding and allocating some sort of financial help to the affected students is urgent to continue their education. Any development policy to support funding the school education of the affected students may include granting any form of financial support to the students who are unable to continue their school education.

The solution to the above problem requires data collection about the school dropped students as described in the problem statement of Example 1 of a region, data analysis and knowledge extraction from the collected data of the affected students. A solution may be provided concerning decisions and a policy formulation based on the extracted knowledge. The design of a framework for making any development policy need to include a data mining system for knowledge extraction from the domain database for using in decision making [1]. Data mining technique can be applied to extract knowledge from this data for using in decision making [11]. Millions of data records are collected from the daily operations in various organizations, e.g., super shops, education sector, hospitals, business organizations, and many other sectors. The Management Information System (MIS) and Decision Support System (DSS) personnel of the organizations make use of the knowledge or patterns hidden in this massive data in decision making. The real problem is to identify the patterns, rules and models, and the extraction of decision making knowledge from the extracted patterns, rules and models. This requires the data mining software systems to incorporate intelligent algorithms to see insight into the data records, and discover patterns from the data records to extract knowledge to use by the decision making authority in making successful decisions to advice new policy to make changes for improvement.

IV. METHODOLOGY

The decision making and policy formulation using the extracted knowledge are two main activities of the knowledge-based development policy making process. Some of the main steps are explained below.

A. Identifying and Defining the Data Mining Problem for Knowledge Extraction

In this step, the real world problem is identified and the data mining problem statement is expressed to solve this problem as described in Example 1. The problem statement clarifies the domain problem and specifies what sort of knowledge is to be extracted from the domain data for particular decision making. It should also provide a hints in the formulation of the development policy. Considering Example 1, this step should specify the frequent causes which force the students to leave their schools.

B. Preparing Sample Data

This step includes the sample data preparation activities. Data may be collected through customized application software directly into databases stored in database servers using database management system (DBMS) tools, e.g., MySQL, Oracle or data can be manually collected to store into spreadsheets or databases using DBMS tools for further processing. In this research work, a sample database of some school dropped students is stored into a spreadsheet for processing using Weka data mining tool. Most of them are of current age 15 to 30 years.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE I represents the properties about the school dropped students to reason about the termination causes of their school education at an early age. The sample database can be used to extract knowledge applying data mining technique for using in decision making.

C. Choosing the Knowledge Extraction Methods

Various data mining methods e.g., classification, association rule analysis and clustering are used for knowledge extraction from databases and data warehouses. A database created for a real world domain, e.g., Banking System, Super Shop Sales System may contain various patterns of data. Data may need to be organized in groups to apply the appropriate data mining method to extract knowledge. For example, classification method may be applicable to a particular group of data while association rule mining may not be applicable to that group of data.

D. Specifying the Decision Making Knowledge to be Extracted

The output obtained by applying the various intelligent algorithms employed in data mining methods on the training data set can be represented in the form of association rules, decision trees and neural networks in terms of existing attribute values, test conditions, predicted values, and constraints. Data mining is usually performed on a single relation, though multi-relational data mining methods can also be employed. In Example 1, the knowledge required for decision making can be of the following forms:

Sub-Problem 1: To know about the termination reasons of school education of the dropped students.

Sub-Problem 2: To classify the school dropped students based on their last education at which they had to stop their education.

The knowledge required for Sub-Problem 1 can be specified by mining association rules to represent the association relation among the attribute values using the sample data by applying the association rule generation algorithm and computing their support and confidence. The knowledge required for Sub-Problem 2 can be specified by building a classification model from the sample data using classification algorithms [8]-[10] by constructing decision trees.

E. Extraction of Knowledge

Appropriate data mining methods implementing intelligent data mining algorithms [2]-[5], [8]-[10] can be applied on the database for knowledge extraction. Various data mining tools e.g., Weka, Neuralware, DBMiner, Rapid miner are available which can be applied on the database to extract knowledge. Customized applications can be developed implementing intelligent algorithms to mine knowledge and patterns from personal and organizational databases, texts and web pages in the form of decision trees, neural networks, association rules with their support and confidence, if-then rules, clusters and so on.

F. Decision Making Using the Extracted Knowledge

Making appropriate decisions at the right time in organizational policy making is crucial for organizational success. Interesting patterns and knowledge extracted from the organizational database storing data in countrywide various sectors e.g., agriculture, education, law and discipline, mass people activities, environment and business organizations can be used by the decision making body of the organizations to make strategic, managerial and operational decisions [1] for organizational development. Knowledge extraction from personal data using data mining technique may help development policy making at personal level. Decisions should be made based on the extracted knowledge to decide what should be done or not. In this paper, two approaches-association rule mining [2]-[5] and classification [8]-[10] have been applied on sample database to generate association rules and build decision trees. The support and confidence of each of the association rules and the decision tree model can be used in decision making [11], [15]-[17]. The extracted knowledge helps optimization in decision making [17] process.

To clarify decision making, we consider a super shop sales

TABLE I. ATTRIBUTE DESCRIPTION [1]

Attribute	Meaning
Person	Person Identifier
Gender	Represents whether Male or Female
Age	Current age
LastclassStudied	The last class in which the student left his school
ReasonofStudyTermination	The main cause of study termination
FamilyEducation	Represents that the dropped student's family has any other educated member or not
ServiceHolders	Represents that the dropped student's family has any service holder or not
Other_Income_Source	Represents that the dropped student's family has any other income source except service or not
Last_Education	Level of the last education

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

system. If the support and confidence both are above 50% for the current purchase of the customer Z, which contains item Y or items (X,Y), then the customer Z is a frequent buyer of item Y or items (X,Y), and he may get a discount on item Y or items (X,Y). This may inspire the customers to buy these items at a regular basis, which may increase the sales of the super shop to earn more profit. This may cause higher amount of purchase orders to the suppliers to increase the stock.

G. Formulating the Development Policy

A development policy needs to execute many decisions during the implementation of the policy where the decisions are made based on the extracted knowledge. The formulation of a policy involves building a plan consisting of a set of actions to reflect the decisions to achieve some form of development. Upon executing the actions during the life span of the policy, the specific policy is implemented and it is expected that some development can be achieved.

H. Justifying the Correctness of the Proposed Approach

The success of an organization depends on the use of domain knowledge in efficient decision making which is actually done by applying human skill, labor, expertise and intelligence in most of the cases. The knowledge required for decision making can be extracted by applying intelligent data mining algorithms on the domain database for easier and faster decision making which will speed up the decision making process. Millions of records are collected through using automated software systems and stored using DBMS tools, which can't be manually processed to extract knowledge. Data mining algorithms are applied on the database after arranging the attributes in a relation properly. The mined output is evaluated using interestingness measures, e.g., support and confidence for association rule analysis and hence the extracted patterns and knowledge hopefully will be correct and relevant for decision making. For the decision tree model constructed using classification method, the model is built using valid domain data, so the constructed model hopefully will function correctly in decision making.

V. SYSTEM ARCHITECTURE

A system architecture shown in Fig. 1 is proposed connecting target group of people using mobile computing devices with cloud computing environment resources.

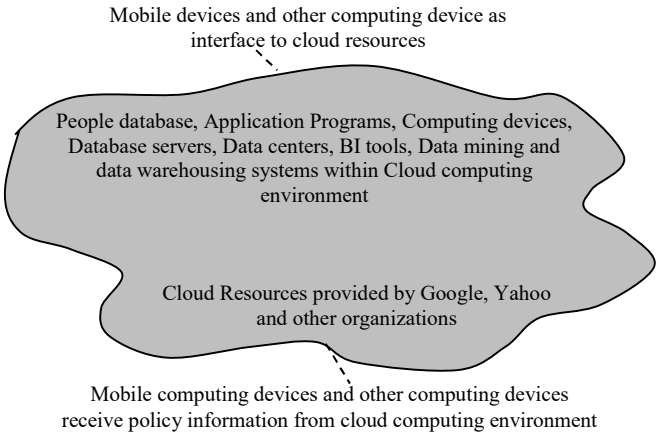


Fig. 1: Proposed System Architecture Connecting Computing Resources within Cloud Computing Environment [1].

Each of the mobile devices and other computing devices in Fig. 1 can be used to request a remote cloud server for a service, and may get a service automatically being processed within the cloud environment. The data mining system can be applied on the database to extract knowledge for decision making to reason about the domain problem. Web-based applications can be developed to collect sample data from a target group of people or any organization to store it directly into the database servers. Cloud computing environment [12]-[14] may provide the required resources for the solution of massive data collection and processing by providing sharable computing power and powerful database servers while running intelligent software systems on application servers. The security issues of the proposed system architecture are considered of the standard security measures usually available in cloud computing environment.

VI. EXPERIMENTATION AND RESULTS

Some sample data about some school dropped students are analyzed using a data mining tool called Weka. The family properties of the school dropped students are stored in a spreadsheet database. TABLE I defines the meaning of 9 attributes of the sample database, and TABLE II summarizes the attributes and their corresponding values. Two methods-association rule mining and classification have been applied on the sample database which are described below. The data items must be organized emphasizing the causes of

TABLE II. ATTRIBUTE VALUES [1]

Attribute	Values
Person	{M1, M2, M3, M4, M5, M6, M7, M8, M9, M10, M11, M12, M13, F1, F2, F3, F4, F5, F6}
Gender	{M, F}
LastclassStudied	{III, IV, V, VI, VIII, IX}
ReasonofStudyTermination	{Early Marriage, Financial Problem}
FamilyEducation	{Yes, No}
ServiceHolders	{Yes, No}
Other_Income_Source	{Poor Agriculture, Middle class Agriculture, No Fixed Source, Poor Business}
Last_Education	{Primary School, High School}

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

termination of the school education of the dropped students in organizing data relationships. Each data mining method is applied to the sample database and the mined output is analyzed for knowledge extraction to verify its use in decision making.

A. Application of Association Rules in Decision Making for Development Policy Making

An association rule [2]-[5] is an implication expression of the form $X \Rightarrow Y$ where X and Y are antecedents and consequents which are subsets of an item set W respectively, and $X \cap Y = \varnothing$ where $\Rightarrow$ is the implication operator. The Weka 3.4.3 Associator[1] is used to generate association rules from the sample database containing 19 instances. The *Age* attribute is removed from analysis as this attribute contains continuous numeric values. Apriori is a famous algorithm [2] for generating association rules from transactional databases. A number of research works [3]-[5] have been carried out on this algorithm in various applications. In this research work, the data mining tool Weka is used to generate association rules by applying Apriori algorithm using the attribute values of the sample database. It is assumed that the Weka Associator will be able to find association relationship among the attribute values in the form of association rules to reason about the causes of the termination. The strong rules consisting of the most occurring attribute values can be used to reason about the related domain facts.

The support s of an association rule $X \Rightarrow Y$ can be defined as follows [2]-[5]:

Support, $s(X \Rightarrow Y) = \sigma(X \cup Y) / N \times 100$

where N is the number of records in the database.

The confidence c of an association rule $X \Rightarrow Y$ can be defined as follows [2]-[5]:

Confidence, $c(X \Rightarrow Y) = \sigma(X \cup Y) / \sigma(X) \times 100$.

The best association rules generated using Weka 3.4.3 Associator using the sample database with minimum support = 0.4 and minimum confidence = 0.9 are shown in TABLE III. Among the generated association rules, Rule 7 and Rule 9 are rejected from analysis as these rules have no relevance to reasoning. Other rules have also weak relevance to reasoning though these rules need simplification by eliminating some antecedent's attribute-value relationships. Rule 1 is justified as the most relevant rule for Example 1 and has the highest support with $sup = 14/19 \times 100 = 73.68\%$, and confidence, $conf = 14/14 \times 100 = 100\%$ as shown in TABLE IV, which is the most frequent association rule contained in TABLE III with 14 occurrences of both antecedents and consequents within the sample database. Rule 2 has also a good relevance to the problem as stated in Example 1 with $sup = 10/19 \times 100 = 52.63\%$ with $conf = 10/10 \times 100 = 100\%$.

TABLE IV. THE BEST ASSOCIATION RULE SELECTED FOR REASONING

Rule No.	sup (*s*%)	conf (*c*%)
1	73.68	100.00

In Rule 1 as defined in TABLE III, the rule consequent is the attribute *ReasonofStudyTermination* with the only value *Financial Problem*. Rule 1 expresses that the reason of their termination of the school education is *Financial Problem* with no service holders in the family to earn money. A decision can be made to provide any sort of financial support to the affected students by the proper authority if the problem is identified at the right time. Hence, a development policy by the proper authority may need to include the decision to support the continuation of such school dropped students by providing any financial support if the termination causes can be identified when the problem occur.

TABLE III. THE BEST ASSOCIATION RULES [1] GENERATED USING WEKA 3.4.3

Rule No.	Rule
1.	ServiceHolders=No 14 ==> ReasonofStudyTermination=Financial_Problem 14 conf:(1)
2.	FamilyEducation=No ServiceHolders=No 10 ==> ReasonofStudyTermination=Financial_Problem 10 conf:(1)
3.	ServiceHolders=No Last_Education=Primary_School 10 ReasonofStudyTermination=Financial_Problem 10 conf:(1)
4.	Other_Income_Source=Poor_Agriculture 9 ==> ReasonofStudyTermination=Financial_Problem 9 conf:(1)
5.	Gender=M ServiceHolders=No 9 ==> ReasonofStudyTermination=Financial_Problem 9 conf:(1)
6.	FamilyEducation=No ServiceHolders=No Last_Education=Primary_School 8 ==> ReasonofStudyTermination=Financial_Problem 8 conf:(1)
7.	Gender=M 13 ==> ReasonofStudyTermination=Financial_Problem 12 conf:(0.92)
8.	FamilyEducation=No 13 ==> ReasonofStudyTermination=Financial_Problem 12 conf:(0.92)
9.	Last_Education=Primary_School 13 ==> ReasonofStudyTermination=Financial_Problem 12 conf:(0.92)
10.	FamilyEducation=No Last_Education=Primary_School 11 ==> ReasonofStudyTermination=Financial_Problem 10 conf:(0.91)

[1] https://www.cs.waikato.ac.nz/ml/weka/

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

B. *Application of Classification Method in Decision Making for Development Policy Making*

In data mining, classification method [8]-[10], [18], [19] can be applied on a set of training data records using several algorithms e.g, ID3, C4.5 to build a decision tree model. The decision tree model is applied on a set of test data to verify its accuracy, and later it can be used for classification. Fig. 3 shows the decision tree obtained by using Weka 3.7.12 [2] classify module by applying J48 algorithm on the sample database. By replacing the attribute value High School with HS and Primary School with PS for the column *Last_Education* within the rows of sample database, the application of Weka 3.7.12 classify module with J48 classification algorithm on the resulting database constructed the decision tree shown in Fig. 4. By summarizing the data shown in Fig. 3 and Fig. 4, a more meaningful information can be provided as shown in TABLE V, which is graphically plotted in Fig. 5.

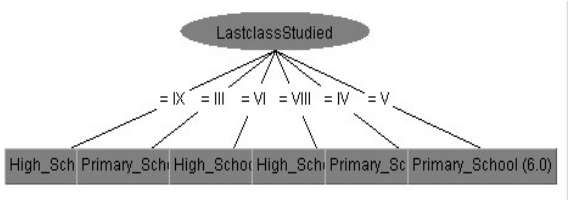


Fig. 3: Classification Using Weka 3.7.12 Classify Module Using the Sample Database.

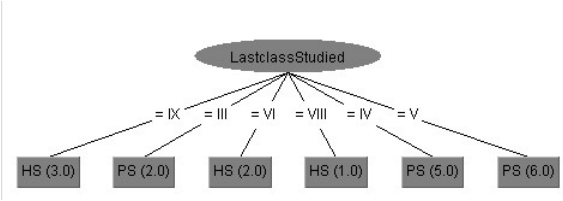


Fig. 4: Classification Using Weka 3.7.12 Classifier Using the Modified Database Obtained from the Sample Database.

TABLE V. NUMBER OF DROPPED STUDENTS IN EACH CLASS MOST ARE OF CURRENT AGE 15-30 YEARS

LastclassStudied	Last_ Education	Number of Dropped Students
III	Primary School	2
IV	Primary School	5
V	Primary School	6
VI	High School	2
VIII	High School	1
IX	High School	3

[2] https://www.cs.waikato.ac.nz/ml/weka/

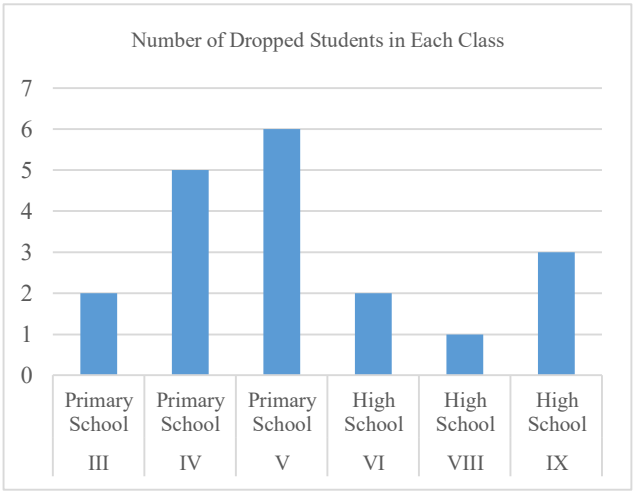


Fig. 5: Number of Dropped Students in Each Class Most are of Current Age 15-30 Years.

The information and knowledge provided in TABLE V and Fig. 5 can be used in the formulation of a development policy for improving child education and should emphasis on reducing the drop of students at primary school education level more than high school education level as more number of students are dropped at primary school level. The affected students can be supported by developing an education policy called *Child Education Policy for the Poor Children* to provide any form of financial support.

VII. CONCLUSION AND FUTURE WORK

In this paper, the application of two data mining methods has been investigated to extract knowledge from the sample database of some school dropped students to use in decision making for development policy making. An architecture for knowledge extraction to aid development policy making has been presented. Attribute value relationship has been analyzed using association rules generated from the sample database using Weka data mining tool. The attribute value relationships are analyzed by computing the support and confidence of the rules to reason about the causes of termination of the school education using the extracted knowledge. It is a novel application of association rules in reasoning about facts. Classification technique has been applied on the sample database to construct a decision tree model which represents the category of the dropped students based on their last education at which they had to stop their school education. To make more efficient use of information and knowledge represented by the decision tree model, the data and information are extracted and summarized using Table and chart for efficient decision making. The data collection about the dropped students through online systems at real time may help to achieve effective solution using the extracted knowledge by the proper authority to formulate a development policy to provide financial assistance to these students so that they can continue their school education. In future, data mining application software may be developed to integrate with web-based on-line customized applications to aid knowledge-based efficient decision making in order to speed up successful

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

development policy formulation to automate social problem solution.

ACKNOWLEDGMENT

The author would like to thank the experts who made valuable comments to this research work. He also thanks all of the people who helped him in any matter in relation to this research work.

REFERENCES

[1] Md. Humayun Kabir, Study on the Architectural Design of ICT-based Development Policy Making System (DPMS) to Inspire Child Education Using Data Mining System. Research Project Report 2015. Faculty of Mathematical and Physical Sciences, Jahangirnagar University, Savar, Dhaka, Bangladesh.

[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," in Proceedings of the 20th VLDB Conference, 1994, Santiago, Chile, pp. 487-499.

[3] R. Srikant and R. Agrawal, "Mining Generalized Association Rules," in Proceedings of the 21st VLDB Conference, 1995, Zurich, Swizerland, pp. 407-419.

[4] Z. Zhu and J.-Y. Wang, "Book Recommendation Service by Improved Association Rule Mining Algorithm," in Proceedings of the 6th International Conference on Machine Learning and Cybernetics. Hong Kong, 2007. Pages 3864-3869. IEEE Conference Publications.

[5] R. Agrawal, T. Imielimski and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases," in Proceedings of the ACM SIGMOD Conference on Management of Data, Washington DC, 1993, pp. 207-216.

[6] Md. Humayun Kabir, Development of a Data Mining Software for Data Warehouse Construction and Mining Interesting Patterns from Organizational Databases. Research Project Report 2010. Faculty of Mathematical and Physical Sciences, Jahangirnagar University, Bangladesh.

[7] Md. Humayun Kabir, Integration and Testing of Association Rule Discovery Algorithm for Developing Masranga Data Mining Software System. Research Project Report 2012. Faculty of Mathematical and Physical Sciences, Jahangirnagar University, Savar, Dhaka, Bangladesh.

[8] L. Zhang, Y. Chen, Y. Liang and N. Li, "Application of Data Mining Classification Algorithms in Customer Membership Card Classification Model," International Conference on Information Management, Innovation Management and Industrial Engineering 2008. IEEE Computer Society.

[9] R. H. Rajon and J. P. Malar Dhas, "A Method for Classification Based on Association Rules Using Ontology in Web Data", International Journal of Computer Applications (IJCA), Volume 49- No, 8 July 2012.

[10] B. Liu, W. Hsu and Y. Ma, "Integrating Classification and Association Rule Mining," KDD Proceedings-1998, Conference of AAAI 1998.

[11] A. Kusiak, "Data Mining and Decision Making," in Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery 2002.

[12] Y. Jadeja, and K. Modi, "Cloud Computing- concepts, architecture and challenges," International Conference on Computing, Electronics and Electrical Technologies (ICCEET), 2012, pp. 877-880. IEEE Conference Publications.

[13] W. Zhao, Y. Peng, F. Xie and Z. Dai, "Modeling and simulation of cloud computing- A review", IEEE Asia Pacific Cloud Computing Congress (APCloudCC), 2012, pp. 20-24. IEEE Conference Publications.

[14] S. Secci and S. Murugesan, "Cloud Networks: Enhancing Performance and Resiliency", IEEE Computer Society, 2014.

[15] N. A. Haris, Munaisyah Abdullah, Abu Talib Othman, and Fauziah Abdul Rahman, "Optimization and Data Mining for Decision Making," WCCAIS'2014 Congress. IEEE Publication.

[16] Mohd. M. Ali and L. Rajamani, "Automation of Decision Making Process for Selection of Talented Manpower Considering Risk Factor: A Data Mining approach," International Conference on Information Retrieval & Knowledge Management (CAMP) 2012. IEEE Publication.

[17] M. Gramatikov, "Data Mining Techniques and the Decision Making Process in the Bulgarian Public Administration," NISPAcee Annual Conference, 2003.

[18] LIU Qin and WANG Yongquan, "Improved ID3 Algorithm Using Ontology in Computer Forensics," 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), 2010, pp VII 494-497.

[19] R. Cao and L. Xu, "Improved C4.5 Algorithm for the Analysis of Sales," 2009 Sixth Web Information Systems and Applications Conference, IEEE Computer Society, 2009, pp 173-176.

AUTHOR PROFILE

Md. Humayun Kabir received his Doctor of Philosophy degree in the area of formal software development from the School of Computing, Dublin City University in 2007, Ireland. He received Master of Science degree in Computer Science from Dhaka University in 1992, and Bachelor of Science (Hons.) degree in Applied Physics and Electronics from Dhaka University in 1991, Bangladesh. He is currently working as a Professor in the Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka. His research interests include Data mining, program transformation, automatic program construction, software architecture and so on.

Parallel Evolutionary Algorithms for Feature Selection in High Dimensional Datasets

Safa Adi [#1], Mohammed Aldasht [*2]

[#] *Computer and IT Department, Palestine Polytechnic University*
Palestine
[1] safa_adi@ppu.edu

[*] *Computer Engineering Department, Palestine Polytechnic University*
Palestine
[2] mohammed@ppu.edu

***Abstract*—Feature selection in high-dimensional datasets is considered to be a complex and time-consuming problem. To enhance the accuracy of classification and reduce the execution time, Parallel Evolutionary Algorithms (PEAs) can be used. In this paper, we make a review for the most recent works which handle the use of PEAs for feature selection in large datasets. We have classified the algorithms in these papers into four main classes (Genetic Algorithms (GA), Particle Swarm Optimization (PSO), Scattered Search (SS), and Ant Colony Optimization (ACO)). The accuracy is adopted as a measure to compare the efficiency of these PEAs. It is noticeable that the Parallel Genetic Algorithms (PGAs) are the most suitable algorithms for feature selection in large datasets; since they achieve the highest accuracy. On the other hand, we found that the Parallel ACO is time-consuming and less accurate comparing with other PEA.**

Index Terms: Evolutionary algorithms, parallel computing, classification, feature selection, high dimensional dataset.

I. Introduction

Nowadays many disciplines have to deal with high dimensional datasets which involve a huge number of features. So we need data preprocessing methods and data reduction models in order to simplify input data.

There are two main types of data reduction models [1]. The first is: instance selection and instance generation processes are focused on the instance level. (i.e. select a representative portion of data that can fulfill a data mining task as if the whole data is used) [14]. The second is: feature selection and feature extraction models which work at the level of characteristics. These models attempt to reduce a dataset by removing noisy, irrelevant, or redundant features. Feature selection is a necessary preprocessing step in analyzing big datasets. It often leads to smaller data that will make the classifier training better and faster [3].

Feature selection is a problem with big datasets. In order to make classification faster and more accurate, we need to select the subset of features that are discriminative. Evolutionary algorithms like Genetic algorithms, Swarm intelligence optimization, Ant colony optimization, etc. These methods can be effective for this problem, but they require a huge amount of computation (long execution time), also memory consumption. In order to overcome these weaknesses, parallel computing can be used.

In this survey, we will review a set of papers about parallel evolutionary algorithms that used for feature selection in large datasets. Furthermore, we will compare the performance of different algorithms and environment.

The rest of the paper is organized as follow: Section2 is Background about feature selection approaches and parallel architecture in general. Section3 talk about parallel evolutionary algorithms. Section 4 will discuss and review many papers which talk about the feature selection problem by using parallel computing. Section5 contains the summary of the survey, the last section is the conclusion and future work.

II. Background

In general, there are three classes of feature selection: filter-based, wrapper, and embedded. The filter approach analyzes the features statistically and ignores the classifier [18]. Most of filter-based methods perform two operations, ranking and subset selection. In some cases, these two operations are performed sequentially, first the ranking, then the selection, in other cases only the selection is carried out. These methods are effective in terms of execution time. However, filter methods sometimes select redundant variables; since they don't consider the relationships between variables. Therefore, they are mainly used as a pre-processing method. In the wrapper model [15], the process of feature selection is depending on the performance of a specific classifier. But its disadvantages are time-consuming and over fitting. The last

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

method for feature selection is the embedded. In this method, the feature selection process and the learning algorithm (tuning the parameters) are combined to each other[6, 15].

The selection of optimal feature subset is an optimization problem that proved to be NP-hard, complex, and time-consuming problem [13]. Two major approaches are traditionally used to tackle NP-hard problems, as seen in Figure1: exact methods and metaheuristics. Exact methods allow exact solution to be found, but this approach is impractical since it is extremely time consuming for real world problems. On the other hand, metaheuristics are used for solving complex and real world problems. Because metaheuristics provide suboptimal (sometimes optimal) solution in reasonable time [2, 11, 13].

As seen in Figure1, Metaheuristics are divided into two categories [13]:

- Trajectory-based (exploitation-oriented methods): the well-known metaheuristics families based on the manipulation of a single solution. Include Simulated Annealing (SA), Tabu Search (TS), Iterated Local Search (ILS), Variable Local Search (VNS), and Greedy Randomized Adaptive Search Procedures (GRASP).
- Population-based (exploration-oriented methods): the well-known metaheuristics families based on the manipulation of a population of solutions. Include PSO, ACO, SS, Evolutionary Algorithms (EAs), Differential Evolution (DE), Evolutionary Strategies (ES), and Estimation Distribution Algorithms (EDA).

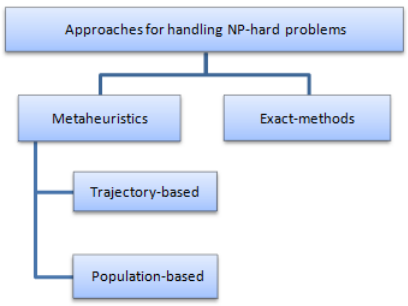


Fig. 1. Approaches for handling NP-hard problems

Metaheuristics algorithms have proved to be suitable tools for solving the feature selection accurately and efficiently for large dimensions in big datasets [2]. The main problems when dealing with big datasets are: The first is execution time because the complexity of the metaheuristics methods for feature selection is at least $O(n^2D)$, where n is the number of instances and D is the number of features. The second is memory consumption since most methods for feature selection need to store the whole dataset in memory. Therefore, the researchers try to parallelize the sequential metaheuristics to improve their efficiency for feature selection on large datasets. There are many programming models and paradigms, such as MapReduce (Hadoop, spark), MPI, OpenMP, CUDA [1, 6, 13]. Parallel computing can be process interaction (shared memory, message passing) or problem decomposition (task or data parallelization) [6].

Parallel computing is a good solution for these problems since many calculations are carried out simultaneously in the task and/or data [6]. Population-based metaheuristics are naturally prone to parallelize since most of their variation operators can be easily undertaken in parallel [2, 13].

Parallel implementations of metaheuristics are an effective alternative to speed up sequential metaheuristics; by reducing the search time for solutions of optimization problems. Furthermore, they lead to the more precise random algorithm and improve the quality of solutions [11]. As seen in Figure2, the implementation of parallel metaheuristics is divided into two categories [13].

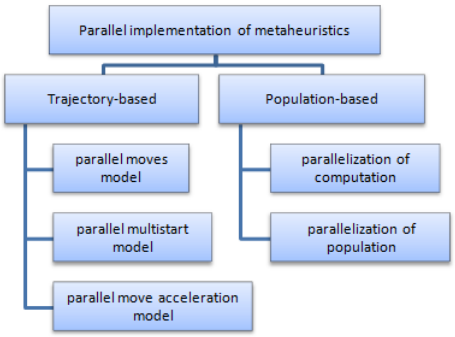


Fig. 2. Parallel implementation of metaheuristics

Parallel evolutionary algorithms are used in many works rather than feature selection, such as inferring phylogenies, traffic prediction. In [9] Santander et al., used MPI/OpenMP with a hybrid multiobjective evolutionary algorithm (fast non-dominated sorting genetic algorithms and firefly algorithm); for phylogenetic reconstruction (Inferring evolutionary trees). In [10] Jiri at al., used parallel multiobjective GA with OpenMP. In order to make traffic prediction more accurate. Master-Slave scheme of GA was implemented on multi-core parallel architecture. They reduced the computational time, but it was successful for short-term traffic prediction.

III. Overview of Parallel Evolutionary Algorithms for Feature Selection

Feature selection algorithms are used to find an optimal subset of relevant features in the data. In this section we will talk about parallel evolutionary algorithms that are used for feature selection problem in large datasets. We will illustrate the steps of six algorithms (PGA, PCHC, PPSO, PGPSO, PSS, and PACO).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A. Parallel Genetic algorithm (PGA)

In order to increase the efficiency and reduce the execution time of the genetic algorithm (GA); the researchers used parallel GA. Algorithm 1 presents the parallel GA methodology, with the master-slave model of parallel GA.

Algorithm 1 Parallel genetic algorithm [10]

```
Create initial population
Evaluate initial population
Create slaves
while not done do
  Start slave
  Wait for slave to finish
  Run mutation operator
end while
for i=1 to slave iterations do
  Select individuals
  Run crossover operator
  Evaluate offsprings
  if solution found then
    set done=True
  end if
end for
```

B. Parallel CHC algorithm (PCHC)

A CHC is a non-traditional GA, which combines a conservative selection strategy (that always preserves the best individuals found so far), that produces offsprings that are at the maximum hamming distance from their parent. The main processes of CHC algorithm are [1]:

- Half-Uniform Crossover (HUX): This will produce two offsprings, which are maximally different from their two parents.
- Elitist selection: this will keep the best solutions in each generation.
- Incest prevention: this step prevents two individuals to mate if the similarity between them greater than a threshold.
- The Restarting process: if the specified population stagnated, then this step generated a new population by choosing the best individuals.

C. Particle Swarm Optimization (PSO)

This subsection handles the geometric particle swarm optimization (GPSO) and shows the algorithm that used to parallelize PSO or GPSO.

1) Geometric Particle Swarm Optimization (GPSO): GPSO is a recent version of PSO. The key issue in GPSO is the using a multi-parental recombination of solutions (particles). In the first phase, a random initialization of particles created. Then the algorithm evaluates these particles to update the historical and social positions. Finally, the three parents (3PMBCX) move the particles, as shown in Algorithm 2:

Algorithm 2 GPSO algorithm [2]

```
S:SwarmInitialization()
while not stop condition do do
  for each particle i of the swarm S do do
    evaluate(solution(xi))
    update(velocity equation (hi))
    update(global best solution (gi))
  end for
  for each particle i of the swarm S do do
    xi:3PMBCX ((xi, wa), (gi, wb), (hi, wc))
    mutate(xi)
  end for
end while
Output: best solution found
```

2) Parallel Multi Swarm Optimization (PMSO): Parallel multi swarm optimization presented in [2], it was defined in analogy with parallel GA as a pair of (S, M), where *S* is a collection swarm, and *M* is a migration policy. Algorithm 3 depicts the parallel PSO methodology.

Algorithm 3 Multi swarm optimization [2]

```
DO IN PARALLEL for each i  1,...,m
initialize(Si)
while not stop condition do do
  iterate Si for n steps /* PSO evolution */
  for each Sj  (Si) do  do
    send  particles in s(Si) to Sj
  end for
  for each Sj such that Si  (Sj ) do do
    receive  particles from Sj
    replace  particles in Si according to r
  end for
end while
Output: best solution ever found in the multi-swarm
```

D. Parallel Scatter Search (PSS)

Scatter search is an evolutionary method that was successfully applied to hard optimization problems. It uses strategies for search diversification and intensification that have proved effective in a variety of optimization problems, see Algorithm 4.

E. Parallel Ant Colony Optimization (PACO)

When dealing with huge search space, parallel computing techniques usually applied to improve the efficiency. Parallel ACO algorithms can achieve high-quality solutions in reasonable execution times comparing with sequential ACO [18]. In Algorithm 5, the methodology of PACO is presented.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Algorithm 4 Parallel scatter search methodology [11]

```
Create Population (Pop, PopSize)
Generate ReferenceSet (RefSet, RefSetSize)
while Stopping Criterion1 do
  while Stopping Criterion2 do
    Select Subset (Subset, SubsetSize)
    for each processor r=1 to n do in parallel do
      Combine Solutions (SubSet, CurSol)
      Improve Solution (CurSol, ImpSol)
    end for
  end while
  Update ReferenceSet (RefSet)
end while
```

Algorithm 5 Parallel ant colony optimization methodology [18]

```
Generate Ants
Initialize N processors
Multicast to all slaves processors N and the task ids of all slaves
for each slave do do
  Send a number between 0 and N that identifies the task inside the program
end for
while not all slaves have sent back solution do
  Wait for solution
  if a slave returns a solution that is better than any solution received then
    Multicast this solution to all slaves
  end if
end while
Return the best solution
```

IV. Parallel evolutionary algorithms for feature selection

We reviewed a set of research papers, which were dealing with feature selection problem for high dimensional datasets in a parallel environment and using parallel evolutionary algorithms. Let us discuss these studies in the following subsections.

A. Parallel GA

Liu et al., [5] used parallel GA for selecting informative genes (features) in tissue classification, using wrapper approach. The main purpose was to find the subset of features with fewer elements and higher accuracy. The parallelization of GA performed by dividing the population into sub-populations, and then the GA run on each sub-population. Therefore, the searching for the optimal subset of genes can be on several CPUs/computers at the same time.

For evaluation, the Golub classifier was used. This classifier introduced by the authors and it depend on the sign of the results for classification; if the sign is positive the sample x belongs to class 1, else if it negative the sample x belongs to class 2. This classifier used only if the datasets have two classes. The accuracy of the classifier tested by using the LOOCV (leave one out cross validation) method. The results showed that using the parallel GA increased the accuracy, and reduced the number of genes that used for classification.

In [8] Zheng et al., analyzed the execution speed and solution quality of many parallel GA schemes theoretically. Furthermore, they pointed to the best scheme of parallel GA that used on multi-core architecture. This paper considered the relationship between speed and parallel architecture along with solution quality.

They analyzed (Master-Slave, Synchronous Island, Asynchronous Island, Cellular, and hybrid scheme of Master-Slave and Island) schemes of parallel GA, with Pthread library on multi-core parallel architecture.

To validate their theoretical analyzing an experiments performed. The hybrid scheme of (Master-Slave and Asynchronous Island) was the best scheme in performance using multi-core architecture. The Island scheme has the best execution time, but the worst solution quality. To improve the solution quality when using Island model it is better to decrease the number of islands. The Asynchronous Island is faster than the Synchronous. The Master-Slave scheme has the best solution quality and the worst execution time.

Soufan at el., [15] developed a web-based tool called DWFS, which used for feature selection for different problems. This tool followed a hybrid approach of wrapper and filter. First, the filter used as preprocessing and select the top ranked features based on tunable and a predefined threshold. In the next step, parallel GA based on wrapper approach applied to the selected features to search for subset features that increase the classifier accuracy. The scheme of parallel GA was Master-Slave; the master node used to create initial population and GA steps. While the slave (worker) nodes used for fitness evaluation of each chromosome, this implementation is performed on 64 core.

For evaluation, they used three different classifiers (Bayesian classifier, K-nearest neighbor, and a combination of them). The results of the experiments show that DWFS tool provided many options to enhance the feature selection problem in different biological and biomedical problems.

In [7] Pinho et al., presented a framework called ParJEColi (java-based library) for a parallel evolutionary algorithm in bioinformatics applications. The aim of this platform was to make the parallel environment (multi-core, cluster, and grid) easy and transparent to the users. This library adapted itself to the problem and the target parallel architecture. The user can easily configure the parallel model and the target architecture; since, ParJEColi encapsulated the parallelization

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

concerns as features. The explicit steps implemented by a simple GUI.

The experiments for validation this framework were done on 2 biological dataset and many bioinformatics scenarios. The results indicate that the proposed framework improves the computational performance (decreases execution time) also the solution quality.

B. Parallel CHC

In [1] Peralta et al., presented a parallel evolutionary algorithm called CHC algorithm by using the MapReduce paradigm for selecting features in high dimensional datasets to improve the classification. The parallelization of CHC algorithm is done by using MapReduce procedure (Hadoop implementation).

A cluster of computers of 20 computing nodes were used. Each dataset split into 512-map task. For evaluating their work, three classifiers where used SVM (support vector machine), logistic regression, and Bayesian classifier.

The results showed that the run time for classification increased as the number of features decreased, except for Bayesian classifier. They explained this result as follow: if the number of blocks less than the number of computing machines; this leads to have some machines remain idle. In addition, if the number of blocks greater than the number of computing machines, the blocks maybe will not distributed in efficient way.

They compared parallel CHC with the serial version, and they concluded that the accuracy of classification increased by using parallel CHC. Furthermore, the parallel version of CHC reduced the run time when the datasets is high dimensional.

C. Parallel PSO

PSO is an efficient optimization technique, it used to solve the problem of feature selection in high dimensional datasets. In [4] Chen et al., used the parallel PSO algorithm for solving two problems at the same time. By creating an objective function that takes into account three variables at the same time (the selected features, the number of support vectors, and average accuracy of SVM). In order to maximize the capability of SVM classifier in generalization.

The proposed method called PTVPSO-SVM (parallel time variant particle swarm optimization support vector machine), it had two phase: 1) the parameter settings of SVM and feature selection work together. 2) the accuracy of SVM evaluated using the set of features and the optimal parameters from the first phase.

They used parallel virtual machine (PVM) with 8 machines; and 10-fold cross validation. The results showed that they could achieve the following aims: increasing the accuracy classification, reducing the execution time comparing with sequential PSO, producing an appropriate model of parameters, and selecting the most discriminative subset of features.

Feature selection can be carried out based on rough set theory with searching algorithm as in [3, 6]. In [6] Qian et al., proposed three parallel attribute reduction (feature selection) algorithms based on MapReduce on Hadoop. The first algorithm was built by constructing the proper (key, value) by rough set theory and implementing MapReduce functions. The second algorithms were done by realizing the parallel computation of equivalence classes and attribute significances. The last parallel algorithm was designed to acquire the core attributes and a reduce in both data and parallel task.

The experiments are performed on a cluster of computers (17 computing node). They considered the performance of the parallel algorithms, but they did not focus on the classification accuracy; since the sequential and parallel algorithms gave the same results. The results showed that the proposed parallel attribute reduction algorithms could deal with high dimensional datasets in an efficient way and better than the sequential algorithms.

In [3] Adamczyk, use rough set theory for attribute reduction, to increase the efficiency he implemented parallel Asynchronous PSO for this problem. The parallelization was done by assigning the complex function computations in slave cores and the main core make the updating particle and checking the convergence of the algorithm.

From their experiments it was noticeable that the efficiency and speedup of parallel PSO algorithm were raising as the size of dataset increased. The achievable accuracy was not astonishing, but it was better than the classical algorithms.

D. Parallel GPSO

In [2] Garcia-Nieto et al., parallelized a version of PSO called GPSO which is suitable for feature selection problem in high dimensional datasets. The proposed method was called PMOS (Parallel multi-swarm optimizer). Which was done by running a set of parallel sub PSOs algorithms, which forming an island model. Migration operation exchanged solutions between islands based on a certain frequency The aim of the fitness function increasing the classification accuracy and reduce the number of selected genes (features).

They used the SVM classifier (Support Vector Machine) to prove the accuracy of the selected subset of features. In their experiments, they used a cluster of computers as a

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

parallel architecture. They found that 8-swarm PMSO was the best choice for parallelization. The results pointed out that this algorithm was better than the sequential version and other methods in term of performance and accuracy while it selected few genes for each subset.

E. Parallel SS

In [11] Lopez et al., present a parallel SS metaheuristics for solving feature selection problem in classification. They proposed two methods for combining solutions in SS.

The first method is called GC (greedy combination): in this strategy, the common features of the combined solutions are added, then at each iteration one of the remaining features is added to any new solution.

The second strategy is called RGC (reduced greedy combination), it has the same start as GC, but in the next step, it considers only the features that appear in solutions with good quality. Then the parallelization of SS is obtained by running these two methods (GC, RGC) at the same time on two processors. Using different combination methods and parameters settings at each processor.

They compared the proposed parallel SS with sequential SS and GA. The results show that the quality of solution in parallel SS is better than solutions which was obtained from the sequential SS and GA. Also, the parallel SS use a smaller set of features for classification. The run time is the same for parallel and sequential SS.

F. Parallel ACO

This subsection shows how the parallel ACO is used to solve feature selection problem for classification in high dimensional datasets.

In [17] Meena et al., implemented a parallel ACO to solve the feature selection problem for long documents. The parallelization was done using MapReduce programming model (Hadoop) that automatically parallelize the code and data then run them on a cluster of computing nodes. The wrapper approachis used as evaluation criteria that used Bayesian classifier. Furthermore, the accuracy of the classifier was based on these metrics: precision, recall, accuracy and F-measure.

The enhanced algorithm (parallel ACO) was compared with ACO, enhanced ACO, and two feature selection methods, CHI (Statistical technique) and IG (Information Gain). They used Bayesian classifier in evaluation process. The results showed that for a given fixed quality of the solutions the proposed algorithm could reduce the execution time but without considered the solution quality. On the other hand, the accuracy of the classifier was increased using parallel ACO comparing with sequential ACO and feature selection methods.

TABLE I
SUMMARY OF ALGORITHMS AND PROGRAMMING MODELS

Paper	**Used evolutionary algorithm**	**Parallel Programming model**
Peralta et al. [1]	CHC (Type of GA)	MapReduce
Garcia-Nieto et al. [2]	GPSO	MALLBA
Adamczyk [3]	PSO	Unknown
Chen et al. [4]	PSO	PVM
Liu et al. [5]	GA	Unknown
Lopez et al. [11]	SS	Unknown
Soufan et al. [15]	GA	MPI
Meena et al. [17]	ACO	MapReduce

In [12] Cano et al., parallelized an existing multi-objective ant programming model that used as the classifier. This algorithm was used for rule mining in high dimensional datasets. The parallelization was done on data and each ant encoded a rule. This was achieved by let each processor perform the same task on a different subset of the data at the same time. In the implementation, they used GPUs, which are multi-core and parallel processor units architecture. This parallel model Followed CUDA method.

For evaluation they used these metrics: true positive, false positive, true negative, false negative, sensitivity, and specificity. The results indicate that the efficiency of this model was increased as the size of datasets increased.

V. SUMMARY AND DISCUSSION

The summary of the papers that implemented the parallel EA for solving the classification problem in high dimensional datasets is reported in Table 1 and Table 2.

Many research papers [2, 3, 7, 8, 9, 10, 12], stated that we can reduce the execution time and achieve acceptable speed ups, when applying parallel evolutionary algorithms on multiple processors. We noticed that they achieved a reasonable speed up in many cases.

In the next table (Table 2), when comparing the accuracy of parallel EA it is important to notice how many classifiers were used to measure the accuracy. Furthermore, we should consider the metrics that were used to evaluate the classifier. For example, the parallel PSO and its variants have the higher accuracy; but they used only one metric which is the success rate. This means that the parallel PSO is not the most accurate parallel EA based on Table 2.

On the other hand, the parallel GA and its variant has the least accuracy, but they used from two to five metrics for evaluation purpose. Based on these metrics, we can say that

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE II
SUMMARY OF DATASETS, CLASSIFIERS, AND ACCURACY RESULTS

Paper	dataset	Classifiers	Metrics for classification	Accuracy
Peralta et al. [1]	Epsilon	Bayesian	AUC= (TPR+TNR)/2	0.71
		SVM		0.68
		Logistic Regression		0.70
	ECBDL 14-ROS	Bayesian		0.67
		SVM		0.63
		Logistic Regression		0.63
Garcia-Nieto et al. [2]	Colon	SVM	Success Rate	0.85
	Lymp			0.97
	Leuk			0.98
	Lung			0.97
Adamczyk [3]	15 Data Set	—	Success rate	0.70 (Avg)
Chen et al. [4]	30 Data Set	SVM	Success rate	0.87 (Avg)
Liu et al. [5]	Leukemia	Golub	Success rate	0.88
	Colon			N/A
Lopez et al. [11]	12 Data Set	Nearest Neighbor	Success rate	0.86 (Avg)
		Bayesian		0.87 (Avg)
		Decision Tree		0.86 (Avg)
Soufan et al. [15]	9 Data Set	K- Nearest Neighbor	F1, PPV, GMean,...	0.81(Avg) (GMean)
		Bayesian		0.79(Avg) (GMean)
Meena et al. [17]	2 Data Sets	Bayesian	F-measure, recall,....	0.64 (Avg)

the parallel GA is the best parallel EA for feature selection in high dimensional datasets

VI. CONCLUSION

After the review of different parallel EA that are used to solve the feature selection problem in high dimensional datasets. We adopted the accuracy as a measure to compare the algorithms performance.

The following points show our conclusion about the performance of the mentioned algorithms in this chapter for feature selection:

- GA and its variants: based on the papers we reviewed, the parallel GA has the higher accuracy.
- PSO and its variants: the parallel PSO has the same accuracy as sequential PSO.
- SS: parallel SS gives better results in case of accuracy than GA and sequential SS.
- ACO: parallel ACO has the less accurate results than the other parallel EA.

It is noticeable that PGAs are the most suitable algorithms for feature selection in large datasets; since they achieved the highest accuracy. On the other hand, the PACO is time-consuming and less accurate comparing with other PEA.

References

[1] D. Peralta, S. RA.O, S. Rama.rez-Gallego, I.Triguero, J. Benitez, F. Herrera. "Evolutionary Feature Selection for Big Data Classification: A MapReduce Approach". Hindawi Publishing Corporation, Mathematical Problems in Engineering, Volume 11, Article ID 246139, (2015).

[2] J. Garca-Nieto, E. Alba. "Parallel multi-swarm optimizer for gene selection in DNA microarrays". DOI 10.1007/s10489-011-0325-9, Appl Intell (2012) 37:255266.

[3] M. Adamczyk. "Parallel Feature Selection Algorithm based on Rough Sets and Particle Swarm Optimization". DOI:.10.15439/2014F389, ACSIS, VOL.2. IEEE, (2014).

[4] H. Ling Chen, B. Yang, S. Wang, G. Wang, D. Liu, H. Zhong Li, W. Liu. "Towards an optimal support vector machine classifier using a parallel particle swarm optimization startegy". Applied Mathematics and Computation 239, 180-197. Elsevier, (2014).

[5] J. Liu, H. Iba. "Selecting Informative Genes with Parallel Genetic Algorithms in Tissue Classification". Genome Informatics 12: 14-23, (2001).

[6] J. Qian, D. Miao, Z. Zhang, X. Yue. "Parallel attribute reduction algorithms using MapReduce". Elsevier, Information Sciences (2014).

[7] J. Pinho, L. Sobral, M. Rocha. "Parallel evolutionary computation in bioinformatics applications". Elsevier. Computer methods and programs in biomedicine 110, 183191, (2013).

[8] L. Zheng, Y. Lu, M. Ding, Y.Shen, M. Guo, S. Guo. "Architecture-based performance evaluation of genetic algorithms on multi/many core systems". The 14th IEEE International Conference on Computational Science and Engineering. 978-0-7695-4477-9/11, (2011) IEEE DOI 10.1109/CSE.2011.65

[9] S. Santander-Jimenez, M. Vega-Rodrguez. "Parallel Multiobjective Metaheuristics for Inferring Phylogenies on Multicore Clusters". DOI 10.1109/TPDS.2014.2325828, IEEE (2014).

[10] J. Petrlik, L. Sekanina. "Towards Robust and Accurate Traffic Prediction Using Parallel Multiobjective Genetic Algorithms and Support Vector Regression". (2015) IEEE 18th International Conference on Intelligent Transportation Systems.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[11] F. Lopez, M. Torres, B. Batista, J. Perez, J. Vega. "Solving feature selection problem by a parallel Scatter Search". Elsevier, European Journal of Operational Research 169, 477-489, (2006).

[12] A. Cano, J. Olmo, S. Ventura. "Parallel Multi-Objective Ant Programming for Classification Using GPUs". Journal of Parallel and Distributed Computing, November (2012).

[13] E. Albaa, G. Luquea, S. Nesmachnowb. "Parallel metaheuristics: recent advances and new trends" Intl. Trans. in Op. Res. 20 (2013) 148, DOI: 10.1111/j.1475-3995.2012.00862.x

[14] H. Liu, M. Hiroshi. "Instance Selection and Construction for Data mining". Springer ,Eds (2001).

[15] O. Soufan, D. Kleftogiannis, P. Kalnis, V. Bajic. "DWFS: A wrapper Feature Selection Tool Based on a Parallel Genetic Algorithm". PLOS ONE, DOI: 10.1371/journal.pone.0117988, February (2015).

[16] A. Gottlieb, G.S. Almasi. "Highly Parallel Computing". Benjamin-Cummings Publishing Co., Inc. Red word city, CA, USA (1989).

[17] M. Meena, K.R. Chandran, A. Kathik, A. Samuel. "A parallel ACO algorithm to select terms to categorise longer documents". Int. J. Computational Science and Engineering, Vol 6 , No.4, (2001).

[18] A. Sameh, A. Ayman, N. Hasan. "Parallel Ant Colony Optimization". International Journal of research and reviews in Computer Science (IJRRCS), Vol.1, No. 2, June (2010).

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Attribute-based Permission Model for Android Smartphones

Abdulrahman A. Muthana[#1], Abdulraqeb I. Al-Samei[*2]

[#]*Faculty of Computer Science and Information Systems*
Thamar University, Yemen
[1]ab.muthana@smartsecurity-y.com

[*]*Computer Networks, Faculty of Computing & Information Technology*
University of Science & Technology, Yemen
[2]abdu.alsamee@gmail.com

***Abstract*- The dependence of users on smartphones to accomplish their daily works is growing increasingly. Every day many mobile applications are downloaded and installed by the users to perform different desirable tasks for them. Before it can be installed in the smartphone, the mobile application requests from the user granting some sort of permissions, which may include the access right to users' sensitive resources. In absence of a security mechanism that can enforce fine-grained permission control, the application may abuse the granted permissions and thus violates the security of sensitive resources. This paper proposes an attribute-based permission model ABP for Android smartphones to control how the mobile application can exercise the granted permissions. The finer granularity of the permission language used by ABP model ensures that the mobile application cannot violate the user's security. By using ABP model, the users can enjoy the useful tasks the mobile applications provide while protecting sensitive resources from unauthorized use.**

Keywords: android smartphone; attribute-based permission; fine-grained permission; mobile application;

I. INTRODUCTION

Modern mobile systems such as Android and iOS implement permission-based access control model to protect sensitive resources from unauthorized use. In this model, the accesses to protected resources without granted permissions would be denied by the permission enforcement system. Ideally, the Android permission model should prevent malicious applications from abusing sensitive resources. However, due to some features of the Android ecosystem, malicious entities could easily abuse permissions, leading to the explosion of Android malware and the numerous reported application vulnerabilities in the past few years [1].

Given this problem, a number of extensions have been proposed to refine the Android permission model. Dr. Android and Mr. Hide framework [2] provides fine-grained semantics for serval permissions by adding a mediation layer. SEAndroid [3] hardens the permission enforcement system by introducing SELinux extensions to the Android middleware. FlaskDroid [4] extends the scope of current permission system by regulating resource accesses in Linux kernel and Android framework together within a unified policy language. Context aware permission models [5], [6] are proposed to support different permission policies according to external contexts, such as location, time of the day. However, these works still could not address the two limitations described above. There are also some work dedicated to reducing the risk of inter-application communication [6], [4] or to isolate untrusted components inside an application [7], [8]. However, none could achieve unified and flexible control according to the system-wide application context.

In this paper we present an attribute-based permission model ABP to protect personal data and sensitive resources in Android platform. The finer granularity of the permission language used by model ABP ensures that the mobile application cannot violate the user's security.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The remainder of this paper is organized as follows: section 2 describes security model in android platform focusing mainly on permissions. A summary of related work is given in section 3. Section 4 and 5 present the proposed attribute-based permission model and its security analysis and section 6 concludes.

II. ANDROID SECURITY MODEL

Security of smartphones is extensively affected by user behavior, as every potentially dangerous application requires permissions when being installed. Malicious software usually requires inadequate set of permissions according to its purpose. If users paid a proper attention to these permissions, the risk of threats to their devices would be minimized. However, according to many studies, only around 20% of users pay attention to permissions when installing applications to their smartphones [9]. In other words, the applications that are installed form Android store may compromise personal security, user, and mobile privacy by misusing sensitive information such as documents, SMS, e-mails, contact list, calling services, location (GPS) network /data, camera, and battery [10].

Android enforces permission-based mechanisms to provide a fine-grained access control to system resources and third-party applications. Specifically, sensitive system APIs are protected by system permissions, and third-party applications can make use of these APIs by first requesting the corresponding permissions in its manifest file. At the beginning of installation process, all requested permissions are presented to the user. If the user agrees to complete the installation, all those requested permissions are to be granted. Applications may also define and enforce their own permissions, which is called custom permissions. All custom permissions can specify one of the four protection levels: normal, dangerous, signature, signatureOrSystem. The custom permissions, as well as system permissions, can be used to protect third-party applications. An application can specify a certain permission that client applications must have for interaction, by setting the android permission attribute of the application element (for all components) or of a component in the manifest file. It is also possible for an application to check caller's permissions during runtime, which is embedded in its source code.

For accessing sensitive resources, however, users should grant the requested permissions to applications. These permissions are of two types in Android: signature and system permissions, which both sound privileged services, content providers, and regular permissions which are all available to all applications such as Android Manifest file (Android-Manifest.xml). Whenever the application tries to access a privileged system resource, the Android framework requires permission management system PMS to check the state of application of whether it has the necessary permission to do so or not [11].

As the capability-based security model found in the Android operating system proves to fall short at protecting the users' privacy, the need arises to find another solution for this problem to make users free and have control over their personal data. This problem evokes the researcher to think deeply for the sake of addressing this problem. In fact, a number of studies have never covered this part of inquiry due to the copious interests in other critical areas ignoring the most important tenets in the current technologies devices such as android systems as will show later.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

III. RELATED WORK

Several security researches have discussed permission-based systems [1, 2, 5, 11, 12, 13, 16, 20, 21, 22, 23, 24, 29, 32, 33] and on Android security [9, 14-19, 25-28, 31] and the permission model of Android has been well described in [21]. Here we introduce some most relevant related works.

The authors in [20] investigated Android OS system to find how the permission methods are implemented. To this end, they analyzed around 1100 Android-based applications. The analysis results, showed that the applications have excessive usage of permissions that negatively impact users. The researchers suggested development access control model to control the usage of permission.

Stowaway, [21] developed a tool "Stowaway" to help in checking permissions files and the source code of the application to reveal the application that has API calls on the source-code, which was not specified in the Androidmanifest.xml file. According to the study results, 35% of the applications used unnecessary permissions. Additionally, the study analyzed the causes behind why the applications behave this way and investigated suspicious behaviors, unnecessary permission usage and method calls.

In [22], the researchers investigated the popular and the more used permissions, and how many of these permissions were actually used by application and how they affected the users. To this end, researchers analyzed 10,000 applications by using data mining techniques. According to the research analysis results, 40% of applications used unnecessary permissions and the permissions that are more popular were misused more. Although the researchers provide a deep analysis of Android applications, but the researchers don't provide any solution for the problem that they referred to.

The authors in [2] Dr. Android and Mr. Hide framework provides fine-grained semantics for several permissions by adding a mediation layer to protect security and privacy issues in the Android systems. The research was conducted on only 19 applications from various categories. According to the results of study, the system rebuilt the amended applications successfully. However, Mr. Hide causes an extra 10-50% overhead on the android OS, and it takes around one minute to rebuild the application. It causes overload, in addition to that the modified application runs slower than their original ones. Furthermore, the proposed system wasn't released to the public that makes it unusable.

Like the other studies [23] developed a software that analyzed permissions of the android-based applications and generates risk signals accordingly. The study analyzed only 121 malicious apps and 150,000 harmless apps during the generation of the risk signals. The research only conducted on 26 permissions out of 122 Android permissions, which are high critical. However, the results of the study limited on the calculation of the risk scores according to specific data that were in hand.

FineDroid [1] is used for providing a fine-grained permission system. It covers both intra-inter-application in addition to systematizing the context sensitive permission rules. TaintDroid [16] is used for detaining information flows on Android smartphones because it is assumed that all applications installed by users cannot be trusted. However, the study doesn't consider control flows and is limited to flow of data

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

tracking. Moreover, it notifies the user when discovered an illegal data flows, but it doesn't enforce permissions to prevent that. PasDroid [24] is a real-time security scheme based on TaintDroid which is used for informing users about the state of permissions of whether they should be allowed or not.

The authors in [5] developed a tool "CRePE" on Android OS to enable/disable functionalities "objects" and enforce fine-grained policies of security taking into consideration both time and location features. There are interception and enforcement for the policies when starting the activities by CRePE. Policies in CRePE are composed of propositional conditions concerning the allowance and denying of actions.

As shown by this state of the art, a solution that provides a fine-grained access control mechanism is required. Module proposed in this work is an authentication language that can serve as a policy for a fine-grained access control to protect personal data in android operating system.

IV. THE SOLUTION

To protect the security of sensitive resources in Android smartphones against unauthorized access, it is required to control how the mobile application makes an access to these resources. To meet these requirements, our solution provides an attribute-based permission model ABP for a fine-grained permission enforcement in Android smartphones.

```
Permissions := <Permission-Object> [, <Permissions>] | <Permission-Object>
Permission-Object  := {<Action>, <App>, <Permission>, <Object>, <Context>}
Action := grant | deny
```

Figure 1. Permission language

The permission language is a declarative language to express the rules for handling permission requests in a context-sensitive manner. Fig.1 shows the general structure of our permission language while the details of permission rules are given in Appendix A. Basically, the language specifies the action <Action> to perform when an application <App> requests a permission <Permission> under the application context of <Context>. To ease the expression of permissions, each language rule is structured in JSON format, with the following main keys:

- **Action key**. It is the key for specifying a policy action. Tow values are required to designate the action value (grant or deny) when the user grant or revoke the permission <Permission> to/from an application <App>.
- **App key**. It describes the subject or applications, which user grant or revoke the permission <Permission> to/from. Package name can also be used as the identity of the application.
- **Permission key**. It describes permission information for a single application participated.
- **Object key**. It describes the object name, which user grant or revoke the permission <Permission> on.
- **Context key**. It describes the constrictions on application through running in the calling context.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The Android platform has a wide range of permissions that provide access to different kinds of resources and objects. However, the current Android security model cannot provide the required fine-grained permission control. For example, an application can be granted a permission to the whole SD-Card, while to perform its task it needs only to access some files on the SD-Card. This way, a malicious mobile application can abuse such permission and make unauthorized access to SD-Card content. To protect against unauthorized access to sensitive resources like SD-Card, the application should be given only a permission to files that are necessary for its functionality, which complies with the least privilege security principle.

The policy that Android security model follows when granting permissions to an application to access the resources is "Everything or Nothing". Meaning that, when it is granted a permission to an object the application can make access to the whole object even if it is not necessary as mentioned in the above example. This policy can lead to security issues as the mobile application can abuse the granted permissions. The solution consists in improving Android permission control model to ensure a finer control over mobile applications and ensure that least privilege security principle is always maintained.

This paper suggests an improvement of Android permission model by introducing an-attribute-based permission concept where more attributes for the resources objects to be accessed are considered and the granted permission can be parametrized by the object attribute(s) or part(s) granted to the requester application. The permission parameterizing ensures a fine-grained permission control enforcement. It worthy to note that, attribute-based concept is not applied to all objects. Only certain objects are refined, which represent the most important resource objects in Android smartphone. The object refinement results in two types of permissions: Multi-Attributes Objects and Single-Attribute Objects. Table 1 shows the types of resource objects.

TABLE 1
Types of Objects in Android Permission Control Model

Multi-Attributes Objects		Single-Attribute Objects	
SD_Card	Phone_address	Credential	PACKAGE_SIZE
SMS	Phone_call	STATUS_BAR	BATTERY_STATS
PHONE	Sites_Zone	Tasks	KEYGUARD
Wi-Fi / Network	Internet	STICKY	Alarm
Bluetooth	Site	BOOT_COMPLETED	WAP_Push
Camera	System_settings	MOCK_LOCATION	TIME_ZONE
Microphone	Downloads	BACKGROUD_PROCESSES	WALLPAPER
Contact_List	Accounts	ANIMATION_SCALE	APN_SETTINGS
Social_Information	System_tools	PERSISTENT_PROCESSES	FILESYSTEMS
Calendar	USB	HISTORY_BOOKMARKS	NFC
Location	SYNC	User_Dictionary	SIP
		Social_Stream	Lock

A. Multi-Attributes Objects

This type contains 22 Android resource objects. In this type, granting access to the whole object without specifying the object's attribute for which access is granted may lead to compromise of personal security and exposure sensitive resources to dangers. This type includes objects like documents, SMS, e-mails, contact list, calling services, location (Global Positioning System GPS), network/data, camera, and battery [10]. By restricting the permission scope, a finer-grained permission control could be achieved and results

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

in better security. Finally, it is up to the user to decide whether the permission can be granted to the whole object or to some attributes of the object. The user may grant an access to the whole object if he decides that the application may need to. To make permission granting process more flexible, the presented permission language supports using the wild card character "*". The use of * indicates that permission on the while object is granted to the application.

B. Single-Attribute Objects

This type contains 24 Android resource objects that do not fit into the first type, either because they already are sufficiently fine-grained or because they would not benefit from finer granularity. For example, RECEIVE BOOT_COMPLETED has only one purpose that does not seem useful to subdivide, and while KILL BACKGROUND_PROCESSES could potentially be fragmented (e.g., by restricting the processes that could be killed), doing so seems unlikely to add much practical security.

V. SECURITY ANALYSIS

As the number of Android-based smartphones increases, more data are used by these devices. Due to the enormous amount of personal data on these devices, they pose a threat and present an inviting environment within which cyber criminals can attack. To defeat attacks, we propose a fine-grained permission language that can serve as a policy for fine-grained permission enforcement. The key idea is to include more resource object attributes and minimize the object surface exposure to attack. The object surface is the part/attribute of the object that is granted to the application. For each permission request a detailed permission rule including attributes or parts of the resource object is constructed when suitable and make granting decisions based on this permission language. Since the permission enforcement system could keep the access rights to the sensitive resource objects at the minimum, it can mitigate the impact of any security breach that caused by a malicious mobile application. In the following, we discuss some scenarios that demonstrate where Android permission control model pitfalls and demonstrate how our permission language can protect sensitive resource objects.

A. SD Card Permission

The SD Card READ permission in Android permission control model allows an application to read the whole content of SD Card, so any application may read the other applications data. Our fine-grained permission language parameterizes the permission (Fig. 2) with the folder(s) that can be accessed by the application and thus minimizing the surface exposure to attack.

```
Sd_Card   :=  Files  |  Folders  | *
          Files    : =  File |  File  [ , Files]
          File   : = FileName
          Folders  :=  Folder  | Folder [ , Folders ]
          Folder  : = FolderName
```

Figure 2. SD Card permission

B. INTERNET Permission

The Internet access "Wi-Fi/Network" permission in Android permission control model allows an application to access Internet indiscriminately. However, the application may contact malicious websites,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

and thus breaching the security of user's privacy or leaking sensitive data. It is wise to request the application to specify the Internet domain(s) or website(s) that it will communicate with to perform its tasks, so the user made be aware of such website(s). Our fine-grained permission language parameterizes the permission (Fig. 3) with the website(s) that can be communicated by the application and thus minimizing the surface exposure to attack.

```
Wi-Fi / Network     : =  Wi-Fi / network_connectivity | Wi-Fi/ network _connectivity | Full_network | Sites_Zone  | *
      Sites_Zone    : = Internet  |  Local intranet  |  Trusted_sites  |  Restricted_sites
      Internet      : = Site |  Internet
      Site          : = Location  |  URI  | IP
```

Figure 3. Wi-Fi/Network permission

C. SMS Permission

The SMS access "SMS" permission in Android permission control model allows an application to access SMS service for send/receive or read/write permissions. whenever it is granted sending SMS messages, the application may send SMS to specific mobile numbers that are intended for advertisements and provided with the applications and thus consuming the user's credit without his knowledge. It would better for the user if he knows the destination mobile numbers distained by the application before granting SMS permission to the application. Our fine-grained permission language parameterizes the permission (Fig. 4) with the mobile number(s) that can be distained by the application and thus protecting the user against such attacks.

```
SMS  :=  SMS-TYPE  to  Contacts |  *
     SMS-TYPE : = SMS  |  MMS
     Contacts : = Local  |  International  |  *
     Local  := +967  Contact_List [, Local] | *
     International  : = + Number Contact_List [,International]
    Contact_List := Number  |  Numbers
     Numbers  : =  Number  |  Number  [ , Numbers  ]
     Number := 0|1|2|3|4|5|6|7|8|9
```

Figure 4. SMS permission

6. CONCLUSION

This paper presents an attribute-based permission model for fine-grained permission enforcement in Android smartphones. By using the proposed permission model, mobile applications cannot breach the security of the sensitive resources. The fine-grained permission rules provide the users with a flexible method to control the access to their sensitive resources and ensure that unauthorized access to sensitive resources will not occur.

APPENDIX A. PERMISSION LANGUAGE

The main result of this appendix is the rules of fine-grained permission language to protect sensitive resources from unauthorized use in Android mobile system.

Permissions := <Permission-Object> [, <Permissions>] | <Permission-Object>

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Permission-Object := {<Action>, <App>, <Permission>, <Object>, <Context>}
Action := grant | deny
Object := SD_Card | Phone_address | SMS | Phone_call | PHONE | Sites_Zone | Wi-Fi / Network | internet | Bluetooth | Site | Camera | SYNC | USB | Calendar | System_settings | Microphone | Downloads | Contact_List | Accounts | Social_Information | System_tools | Location
Permission = read | open | send | receive | pair | write | call | connect | edit | view | modify | close | access | active | take | install | uninstall | draw | allow | add | locate | remove | measure | record | get | create | subscribed_feeds| authenticate | manage | set | process | find | google | admin | use | wake | reorder | disable | expand | format | mount | unmount | broadcast| change | control

admin := [discover, Requst_pair, Replay_pair, unpaired, accept_connection, visible, invisible, enable, disable, change]
control := [send, receive, show_files]

Sd_Card := Files | Folders | *

Files : = File | File [, Files]

File : = filename

Folders := Folder | Folder [, Folders]

Folder : = foldername

SMS := SMS-TYPE to Contacts | *

SMS-TYPE : = SMS | MMS

Contacts : = Local | International | *

Local := +967 Contact_List [, Local] | *

International : = + Number Contact_List [,International]

Contact_List := Number | Numbers

Numbers : = Number | Number [, Numbers]

Number := 0|1|2|3|4|5|6|7|8|9
Phone := phone_status | phone_identity | phone_address | call_log | SIP

outgoning_call | voicemail | None | * | device_ID | phone_call |

phone_address := Contact_list | Non_Contact_list | Number | Numbers | *

Numbers : = Number | Number [, Numbers]

Phone_call := call_phone | call_Emergency | call_Privileged | *

Wi-Fi / Network : = Wi-Fi / network_connectivity | Wi-Fi/ network _connectivity | Full_network | Sites_Zone | *

Sites_Zone : = Internet | Local intranet | Trusted_sites | Restricted_sites

Internet : = Site | Internet

Site : = Location | URI | IP

Bluetooth := Bluetooth _settings | Bluetooth_ Data | audio_channel | Internet_access | Paired_list | Non_ Paired_list | *

Camera := Pictures | Videos | *

Pictures : = Picture | Picture [, Pictures]

Videos : = Video | Video [, Videos]

Microphone := Audios | *

Audios : = Audio_settings | Audio | Audio [, Audios]

Identify the applications that used Camera / Microphone to Record:

Camera / Microphone := Videos | Audios | Video_Audio | None | *

Social_Informations := Contacts | call_log

Contact_List := Full | Non_ Contact_List | Number | numbers | *

Numbers : = Number | Number [, Numbers]

Calendar := Calendar_Date | Event _settings | Calendar_settings

Location := Fine_location | Coarse_location | None | *

Accounts := account | Google_accounts | accounts_password |

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Google_Service_Configure | Local_accounts | accounts_settings | accounts_data | *

System_tools := Shortcut | System_settings | Home_setting | App_Storage_Space | Tool [, Tools]

System_settings := Volume_Control_Widgets | notification_Widgets | GPS_Utilities | settings_Widgets | Wi-Fi_Utilities |

System_setting |

System_setting [, System_settings]

USB := USB_settings

Downloads := Content_URI | Location |

SYSC := SYSC _settings | SYSC _stats |

JSON Format for permission:

{ "app" : "com.masdroidapp","action" : "allow", "permission" : ["receive "," send"], "object" : " WAP_Push ", "context" : "None", "comment" : "None"}
{ "app" : "com.masdroidapp", "action" : "allow", "permission" : ["read "," write "] , "object" : " history_bookmarks",
"context" : "None", "comment" : "None"}
{ "app" : "com.masdroidapp", "action" : "allow", "permission" : "use" , "object" : " credential", "context" : "None", "comment" : "None"}
{ "app" : "com.masdroidapp", "action" : "allow", "permission" : " set " , "object" : " alarm ", "context" : "None", "comment" : "None"}
{ "app" : "com.masdroidapp", "action" : "allow", "permission" : [" get " ," reorder "], "object" : " task",
"context" : "None", "comment" : "None"}
{ "app" : "com.masdroidapp", "action" : "allow", "permission" : "wake " , "object" : " Lock ", "context" : "None", "comment" : "None"}
{ "app" : "com.masdroidapp", "action" : "allow", "permission" : " receive" , "object" : " Boot_Completed ", "context" : "None", "comment" : "None"}
{ "app" : "com.masdroidapp", "action" : "allow", "permission" : " get " , "object" : " package_size", "context" : "None", "comment" : "None"}
{ "app" : "com.masdroidapp", "action" : "allow", "permission" : " access " , "object" : " mock_location ", "context" : "None", "comment" : "None"}
{ "app" : "com.masdroidapp", "action" : "allow", "permission" : " set" , "object" : " time_Zone ", "context" : "None", "comment" : "None"}
{ "app" : "com.masdroidapp", "action" : "allow", "permission" : " set" , "object" : " time_Zone ", "context" : "None", "comment" : "None"}
{ "app" : "com.masdroidapp", "action" : "allow", "permission" : " disable " , "object" : " KeyGuard ", "context" : "None", "comment" : "None"}
{ "app" : "com.masdroidapp", "action" : "allow", "permission" : " expand " , "object" : " status_bar", "context" : "None", "comment" : "None"}

REFERENCES

[1] Zhang, Y., Yang, M., Gu, G., & Chen, H. (2016). Rethinking Permission Enforcement Mechanism on Mobile Systems. IEEE Transactions on Information Forensics and Security, 11(10), 2227-2240.
[2] Jeon, J., Micinski, K. K., Vaughan, J. A., Fogel, A., Reddy, N., Foster, J. S., & Millstein, T. (2012). Dr. Android and Mr. Hide: fine-grained permissions in android applications. Paper presented at the Proceedings of the second ACM workshop on Security and privacy in smartphones and mobile devices.
[3] Smalley, S., & Craig, R. (2013). Security Enhanced (SE) Android: Bringing Flexible MAC to Android. Paper presented at the NDSS.
[4] Bugiel, S., Heuser, S., & Sadeghi, A.-R. (2013). Flexible and Fine-grained Mandatory Access Control on Android for Diverse Security and Privacy Policies. Paper presented at the USENIX Security Symposium.
[5] Conti, M., Nguyen, V. T. N., & Crispo, B. (2010a). CRePE: Context-related policy enforcement for Android. Paper presented at the International Conference on Information Security.
[6] Ongtang, M., McLaughlin, S., Enck, W., & McDaniel, P. (2012). Semantically rich application-centric security in Android. Security and Communication Networks, 5(6), 658-673.
[7] Shekhar, S., Dietz, M., & Wallach, D. S. (2012). AdSplit: Separating Smartphone Advertising from Applications. Paper presented at the USENIX Security Symposium.
[8] Wang, Y., Hariharan, S., Zhao, C., Liu, J., & Du, W. (2014). Compac: Enforce component-level access control in Android. Paper presented at the Proceedings of the 4th ACM Conference on Data and Application Security and Privacy.
[9] Burguera, I., Zurutuza, U., & Nadjm-Tehrani, S. (2011). Crowdroid: behavior-based malware detection system for android. Paper presented at the Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices.
[10] Yuksel, A. S., Zaim, A. H., & Aydin, M. A. (2014). A Comprehensive Analysis of Android Security and Proposed Solutions. International Journal of Computer Network and Information Security, 6(12), 9.

[11] Au, K. W. Y., Zhou, Y. F., Huang, Z., & Lie, D. (2012). Pscout: analyzing the android permission specification. Paper presented at the Proceedings of the 2012 ACM conference on Computer and communications security.
[12] Felt, A. P., Greenwood, K., & Wagner, D. (2011). The effectiveness of application permissions. Paper presented at the Proceedings of the 2nd USENIX conference on Web application development.
[13] Felt, A. P., Wang, H. J., Moshchuk, A., Hanna, S., & Chin, E. (2011). Permission Re-Delegation: Attacks and Defenses. Paper presented at the USENIX Security Symposium.
[14] Bose, A., Hu, X., Shin, K. G., & Park, T. (2008). Behavioral detection of malware on mobile handsets. Paper presented at the Proceedings of the 6th international conference on Mobile systems, applications, and services.
[15] Chaudhuri, A. (2009). Language-based security on Android. Paper presented at the Proceedings of the ACM SIGPLAN fourth workshop on programming languages and analysis for security.
[16] Enck, W., Gilbert, P., Han, S., Tendulkar, V., Chun, B.-G., Cox, L. P., Sheth, A. N. (2014). TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones. ACM Transactions on Computer Systems (TOCS), 32(2), 5.
[17] Enck, W., Octeau, D., McDaniel, P., & Chaudhuri, S. (2011). A Study of Android Application Security. Paper presented at the USENIX security symposium.
[18] Enck, W., Ongtang, M., & McDaniel, P. (2009). On lightweight mobile phone application certification. Paper presented at the Proceedings of the 16th ACM conference on Computer and communications security.
[19] Hornyack, P., Han, S., Jung, J., Schechter, S., & Wetherall, D. (2011). These aren't the droids you're looking for: Retrofitting android to protect data from imperious applications. Paper presented at the Proceedings of the 18th ACM conference on Computer and communications security.
[20] Barrera, D., Kayacik, H. G., van Oorschot, P. C., & Somayaji, A. (2010). A methodology for empirical analysis of permission-based security models and its application to android. Paper presented at the Proceedings of the 17th ACM conference on Computer and communications security.
[21] Felt, A. P., Chin, E., Hanna, S., Song, D., & Wagner, D. (2011). Android permissions demystified. Paper presented at the Proceedings of the 18th ACM conference on Computer and communications security.
[22] Stevens, R., Ganz, J., Filkov, V., Devanbu, P., & Chen, H. (2013). Asking for (and about) permissions used by Android apps. Paper presented at the Proceedings of the 10th Working Conference on Mining Software Repositories.
[23] Sarma, B. P., Li, N., Gates, C., Potharaju, R., Nita-Rotaru, C., & Molloy, I. (2012). Android permissions: a perspective combining risks and benefits. Paper presented at the Proceedings of the 17th ACM symposium on Access Control Models and Technologies.
[24] Hsiao, S. W., Hung, S. H., Chien, R., & Yeh, C. W. (2014). PasDroid: Real-Time Security Enhancement for Android. Paper presented at the Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2014 Eighth International Conference on.
[25] Baláž, A., Madoš, B., & Ambróz, M. (2015). Android Access Control Extension. Acta Informatica Pragensia, 4(3), 310-317.
[26] Bugiel, S., Davi, L., Dmitrienko, A., Fischer, T., Sadeghi, A.-R., & Shastry, B. (2012). Towards Taming Privilege-Escalation Attacks on Android. Paper presented at the NDSS.
[27] Conti, M., Nguyen, V. T. N., & Crispo, B. (2010b). CRePE: Context-Related Policy Enforcement for Android. Paper presented at the ISC.
[28] Hoopes, J. (2009). Virtualization for security: including sandboxing, disaster recovery, high availability, forensic analysis, and honeypotting: Syngress.
[29] Nauman, M., Khan, S., & Zhang, X. (2010). Apex: extending android permission model and enforcement with user-defined runtime constraints. Paper presented at the Proceedings of the 5th ACM symposium on information, computer and communications security.
[30] Shan, Z., Wang, X., Chiueh, T.-c., & Meng, X. (2012). Facilitating inter-application interactions for OS-level virtualization. Paper presented at the ACM SIGPLAN Notices.
[31] Vargas, R. J. G., Huerta, R. G., Anaya, E. A., & Hernandez, A. F. M. (2012). Security controls for Android. Paper presented at the Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on.
[32] Wei, X., Gomez, L., Neamtiu, I., & Faloutsos, M. (2012). Permission evolution in the android ecosystem. Paper presented at the Proceedings of the 28th Annual Computer Security Applications Conference.
[33] Wu, L., Du, X., & Zhang, H. (2015). An effective access control scheme for preventing permission leak in Android. Paper presented at the International Conference on Computing, Networking and Communications (ICNC), 2015.

BADEL WEB APPLICATION- An online web Portal for donation and Exchangeable Resources

Ms. Raheemunnisa Begum
Lecturer, Department of Information System,
King Khalid University,
Abha, K.S.A,
rhem@kku.edu.sa

Manal Ali, Yasmin Hasan, Arwa Saad , Shahad Saeed,
Atheer Khalid, Malak Saeed , bashayer Mohammed
Scholar, Department of Information System
King Khalid University
Abha, KSA.

Abstract - **Badel is a web application available via the internet. To help King Khalid university students in Saudi Arabia to exchange their collectibles or ownership with their university friends. It is a special auctions system while student shows their ownership details and pictures in order to exchange with other university friend's ownership, also he /she can lend it or donate to another friend. Badel provides a way to save money and resources.**

Keywords - Badal; Exchange; resources; sharing; auction.

I. INTRODUCTION

The main aim is of this project is to design and implement an online website like a big virtual marketplace where university members can gathering to exchange, lend, borrow and donate their collectibles and ownerships easily to save their time and money

This system facilitates users in their exchange or donates process by offering them automated methods rather than Traditional barter system to save money, and without allocating part of university staff to enter the information into a database. The responses are processed automatically, and the results are accessible at any time

II. EXISTING SYSTEM

If you've ever swapped one of your book with a friend in return for one of their book, you have badel. The badel system is influenced by the olden system called "Barter System". Bartering is trading services or goods with another person when there is no money involved. This type of exchange was depended upon by early civilizations.

University members use the traditional methods like search in the internet, social communication website and other contacts applications like whatsApp, skype, messenger ….etc. to contact with other university members or to exchange or donate their ownership to them.

A. The disadvantage of the current system

- Sometimes people have to keep their collectibles which they do not need for a long time because they do not know how can they get benefit from it or how can they help another person to get benefit from it?
- Searching for beneficiaries is a time-consuming process.
- Causing this collectibles and ownerships damage or lost and that increase money loss and environmental pollution
- Also, advertising collectibles in the newspaper is expensive and less benefit because it does not reach to a lot of people or for a specific group of people**.**

III. PROPOSED SYSTEM

This system facilitates users in their exchange or donates process by offering them automated methods rather than Traditional barter system to save money, and without allocating part of university staff to enter the information into a database. The responses are processed automatically, and the results are accessible anywhere and at any time

A. System Objectives

- No upfront costs in advertising
- No staff or distributors
- No paperwork that means no population
- Add friendship atmosphere between the University members
- Help needy member to get some products free.
- Save the time and effort of university members in exchange items process

B. Benefits of proposed systems

- Badel system is a web application in which every university members of KKU could log in any time and use our website to do an online exchange or donate, in a new and modern way.
- It helps university members to save money with the great and easy user interface.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- It helps the university members view their ownership through our website.
- The good design of interface provides accurate information and details will increase university members to do exchange process for their ownership.
- Help the environment by decrease the waste.
- System help reduces the workforce by handle task and documentation based on electronically rather than on paperwork, which will be more costly.
- Badel is more than a website that only allows the user to donate and exchange the product, for example, such as portable tablets in exchange for laptops.
- It also contains functions such as friend list, profile editing, advertisement of the product and categories of product etc.

IV. METHODOLOGY

A. *System Requirement*

The system requirements for Badel Web Application

- Database: Microsoft SQL Server
- Operating System: Window 7
- Integrated Development Environment: visual studio 2010, Microsoft SQL server management studio
- Programming Language : ASP.NET
  c# , JavaScript , HTML , CSS
- Web server: iis7

B. *System Design*

Software design and implementation is the stage in the software engineering process at which an executable software system is developed for realizing the design as a program.

1) Admin Module

a) Can login/logout
b) Add, delete or modify the user account
c) Add university member information
d) Add or view the advertisement
e) Delete or stop advertisement
f) Add category
g) Add department

2) User module

a) Can login or logout user
b) Update profile
c) Search product
d) Browse the category of products.
e) Receive(exchange/lend/gift) request
f) Send (exchange/lend/gift) request
g) Stop advertisement
h) Add auction

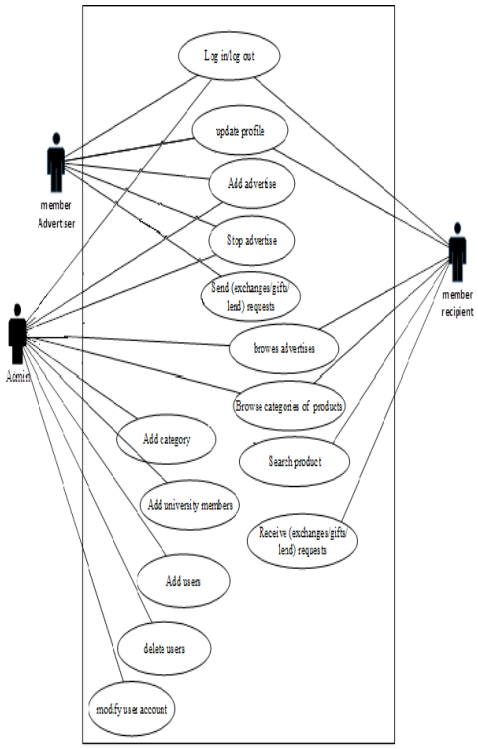


Figure 1 Usecase of Badel

C. *System implementation*

Software design and implementation is the stage in the software engineering process at which an executable software system is developed and accepted by the user.

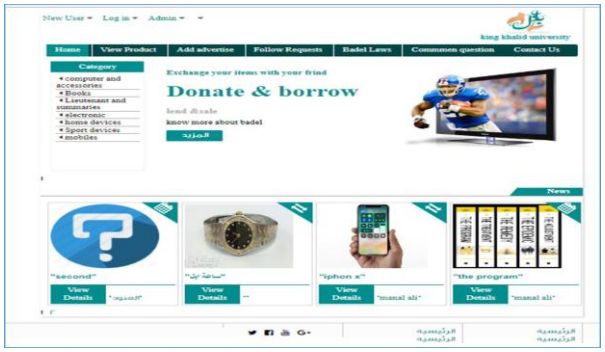


Figure 2 home page

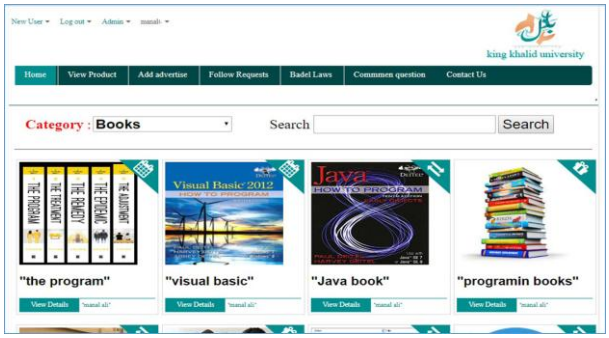


Figure 3 search category interface

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

D. System testing

System testing is a level of software testing where complete and integrated software is tested. The purpose of this test is to evaluate the system's compliance with the specified requirements.

E. Conclusion

Badel is a desktop application, which helps and guides the students to exchange their Collectibles or ownership with their university friends. It is a special Auctions system while student shows their ownership details and pictures in order to exchange with other campus friend's ownership, also he or she can lend it or donate to another friend.

This project goes through several steps starting from gathering information and studies that information in order to reach the final objectives and the final solution which required to be implemented. our own software. Analysis for all of these data, requirements, and methodologies also detailed then.

ACKNOWLEDGMENT

This acknowledgment transcends the reality of formality when we would like to express deep gratitude and respect to all those people behind the screen who guided, inspired and helped me for the completion of our research paper. Special thanks to my parents who motivated me to do this work.

REFERENCES

[1] https://www.mint.com/barter-system-history-the-past-and-present
[2] https://en.m.wikipedia.org/wiki/Microsoft_SQL_Server
[3] http://softwaretestingfundamentals.com/system-testing/

AUTHORS PROFILE

Mrs.Raheemunnisabegum, received her Master in Computer Applications from osmania university in the yer 2005, I have 12years of teaching experience as lecturer in the Depatment of Information system, King Khalid unversity, Ministry of Higher Eduction, Saudi Arabia. My area of intrest include System Analysis and design, Programming and Networking.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Hybrid Cluster Based Congestion Aware (HCBCA) Approaches in Wireless Sensor Network

V.Perumal
Ph.D Research Scholar,
Department of Computer Science,
Erode Arts & Science College(Autonomous),
Erode, Tamilnadu, India.
Email:rishiperumal89@gmail.com

Dr.K.Meenakshi Sundaram
Associate professor,
Department of Computer Science,
Erode Arts & Science College (Autonomous),
Erode,Tamilnadu,India.
Email: lecturerkms@yahoo.com

ABSTRACT

In most efficient clustering technique for WSN has been proved as a congestion control and hierarchical based cluster head selection process. The cluster head reduces the energy wastage and additionally that improves the receiving of data and collection of data from their member sensor nodes. Also transmitting the collected data to the base station (BS). In proposed method hybrid cluster based congestion aware (HCBCA) is mostly focused on traffic that affects the continuous flow of data, Arrival of data from the source to destination delay time, Avoid packet losses and energy consumption process. Mainly congestion happens in the intra cluster to do the process of transmitting the destination of packets in many to one manner form sensor node to CH. The main reason for occurrence of congestion is communication path, nodes energy level and nodes buffer size. When these above it are successful done the congestion, does not exist or otherwise congestion will occur. The purpose of WSN congestion control is to improve the packet delivery ratio and energy consumption.

Keywords: Sensor node, Hybrid Cluster, Congestion Avoidance, WSN.

I.INTRODUCTION

In Wireless Sensor Network (WSN) the sensor nodes are usually scattered over a sensor field and are capable of sensing, processing and transmitting to the base station, based on the requirement application. The major constraints of WSNs are the limited power sources of the sensor nodes. The battery operated sensors are often deployed in an unattended hostile environment, so replacement of their battery is almost impossible which make the sensor node energy constraint.

Clustering sensor node is one of the most effective techniques which is employed to conserve energy of sensor node. In the process of clustering the network is divided into many groups, called Cluster Head (CH). CH responsible for collecting the data from their members sensor nodes within the clusters, aggregate them and send it to a remote base station (BS) directly or through other CHs. The base station is connected to a public network such as internet for public notification of the event. The congestion generally whilst a sensor node utilized as a relay node for multiple flows. Another possible reason of congestion is the unfair distribution of data traffic in the network. The possible effect of unfair traffic utilization will result in unstable paths that can overload the nodes and soon deplete the energy of some sensor nodes, which consequently partition the network [1][2].

In several aspects based on congestion possibly will happen, such as contention due to concurrent transmission, overflow in buffers and time varying wireless channel condition [2]. The congestion can occur while collecting the data and sending it towards the central location over the WSN. Congestion happens mainly in the sensor to base station direction. When packets are transported is a many to one manner. It has negative impact of on network performance and application objective indiscriminate packet losses, increased packet delay, wastage node energy and severe fidelity degradation [3].

The congestion organizes technique in WSN are classified under two categories: Link level congestion and node level congestion. Node level congestion arises from buffer overflow in the node, which results in packet loss. The link level congestion is related to wireless channels shared by several nodes through competitive MAC layer protocol. Link level congestion control can achieve by using multiple access technique such as CSMA, FDMA, TDMA and CDMA to prevent congestion by exercising light degree buffer management [4].

The most challenging congestion mechanisms are congestion Avoidance, detection and alleviation. The congestion avoidance is referring to as proactively routing protocol plays an important role to select best nodes and to route the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

data traffic from the source to destination. Congestion detection in a timely manner during data forwarding, sensor nodes monitor the buffer occupancy and the channel utilization. On the other hand, congestion alleviation schemes control congestion reactively either by adjusting the source traffic rate or by re-discovering the new route. All these three mechanisms are able to increase the performance and to balance the traffic load in multi-hop WSNs.

Congestion avoidance when a source node is triggered by the application; the first step that is performed by the sensor node is to check the availability of route to the desired location through a check route availability process. Congestion avoidance is measuring the data accuracy and data redundancy. Congestion detection process monitors the state of the node and the link between the nodes in initiated in order to detect congestion. If the node or link between the nodes that it will be congested in the near future, then the process to notify the source or precursor node is triggered. The congestion notification process is invoked by the sensor node when congestion or low energy is detected. It measures by an aggregated of two matrices: buffer occupancy and channel utilization. Congestion alleviation is activated in ripple search based when a sensor node receiving a notification message. In this process of congested node or link is bypassed in order to maintain a route. Another procedure to alleviate congestion is to Re-route the traffic to an alter route congestion aware and energy efficient route. It measured, unlike the resource control and traffic control that it will alleviate congestion by adjusting the traffic rate at the source node or intermediate nodes [5][6].

II.RELATED WORKS

Azlan Awang et al [1]. Congestion-aware energy efficient and traffic Load Balancing Scheme (CLS) for routing in *WSNs* is proposed. This scheme utilizes the ignored information during the route discovery process and considers a composite metric that incorporates the consumed energy *E*, participation level *P* of the node and signal strength *S* of the link between the nodes. In addition, a separate field is maintained in the packet for each routing metric in the case of multiple metrics that might overload the node. In this paper, a new congestion aware, energy efficient and traffic load balancing scheme (CLS) for routing has been designed. The proposed scheme compares the proposed routing metric over a new route discovery mechanism, using weighted additive composition approach and lexical approach. The optimum next hop is selected based on a combination of three different metrics such as energy *E*, participation level *P* and signal strength *S* during forward route formation. Using this approach, a least congested and an energy efficient route is discovered that maintains the minimum routing information. Furthermore, this approach increases the PDR, decreases the energy consumption and an ETE delay of the entire network.

Srinivasan et al [2]. proposed an energy efficient cluster head selection algorithm which is based on particle swarm optimization (PSO) called PSO-ECHS. The algorithm is developed with an efficient scheme of particle encoding and fitness function. For the energy efficiency of the proposed PSO approach, we consider various parameters such as intra-cluster distance, sink distance and residual energy of sensor nodes. We also present cluster formation in which non cluster sensor nodes join their CHs based on derived weight function. The algorithm is tested extensively on various scenarios of WSNs, varying number of sensor nodes and the CHs.

Raheleh Hashemzehi et al [3]. The Suggested The unique characteristics of WSN such as coherent nature of traffic to base station that occurs through its many-to-one topology and collision in physical channel are main reasons of congestion in wireless sensor networks. Also, when sensor nodes inject sensory data into network the congestion is possible. Congestion affects the continuous flow of data, loss of information, delay in the arrival of data to the destination and unwanted consumption of significant amount of the very limited amount of energy in the nodes**.** Therefore, Congestion in wireless sensor networks (WSN) needs to be controlled in order to prolong system lifetime improve fairness, high energy-efficiency, and improve quality of service (QoS). It has mainly described the characteristic and the content of congestion control in wireless sensor network and surveys the research related to the Congestion control protocols for WSNs.

Chia-Hsu Kuo et al [4]. Proposed a distributed congestion control protocol called traffic aware congestion control protocol (TACCP). Through the buffer management mechanism TACCP for adaptively allocating an appropriate forwarding rate to potentially jammed sensors for mitigating the congestion load. TACCP can be used to avoid packet loss caused by traffic congestion, reduce the power consumption of nodes, and improve the throughput of the entire network.

Omer chughtai et al [5]. The developed CTLS protocol avoids congestion proactively by modifying the traditional route discovery mechanism in order to select the best node during the forward route formation. It detects congestion in a timely manner by monitoring either the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

remaining space of the buffer, the interval between the consecutive packets and the link utilization based on the number of times a node goes into the Back off stage of CSMA/CA. The CTLS either bypasses the congested node/link through a local repair technique or deviates the traffic to the detour path in order to alleviate congestion. The simulation results show that the CTLS performs better as compared to the congestion avoidance, detection and alleviation and no congestion control schemes in terms of packet delivery ratio, ETE delay, throughput, and energy consumption per data packet in a resource constraint wireless network.

Ji-ming CHEN et al [6]. Proposed a congestion control scheme CADA for congestion avoidance, detection and alleviation in wireless sensor networks. The key objective is to provide high transmission quality for the data traffic under conditions of congestion. The scheme comprises three main mechanisms. Firstly, it attempts to suppress the source traffic from event area by carefully selecting a set of representative nodes to be data sources. Secondly, the onset of congestion is indicated in a timely way by jointly checking buffer occupancy and channel utilization. Lastly, the network attempts to alleviate congestion in the traffic hotspot by either resource control or traffic control, which is dependent on the specific congestion Condition.

Vaibhav Eknath Narawade et al [7]. The Survey of the congestion control and avoidance mechanisms are investigated in terms of their appropriateness in congestion detection and inform the related nodes with the intention that a proper control can be taken. Based on the usage, several methods are applied to manage the congestion. To satisfy the application requirements, either traffic control by throttling the node rates or resource control by utilizing the unused resources are used Different issues and challenges regarding the congestion control protocols were studied which will be useful for further research in this field.

Venugopal K R [8]. Proposed MCDR techniques is effectively mitigates congestion by considering the parameters such as minimum Queue Length, the depth, the distance and maximum residual energy of each node while scattering the traffic towards the sink from the congested area. Improved network throughput is achieved by maintaining the minimum congestion rate due to fair queue length at each node in the network. The looping problem has been drastically reduced by selection of each node that is based on the minimum distance to scatter the traffic towards the sink. The reduction in looping results in lower latency and minimizes energy utilization. The results of our proposed algorithm show that improved network throughput and packet delivery rate for both high and low load conditions and also fulfill the fidelity requirement of different applications.

Majid Gholipour et al [9]. Proposed a hop-by-hop gradient-based routing scheme to evenly distribute traffic in WSNs with non-equivalent sink. The key concept herein is to utilize the number of hops and the current traffic loading of neighbors to make routing decisions reduces the number of packet retransmissions and packets dropped by preventing nodes with overloaded buffers from joining in routing calculation. Simulation results are indicate improves network performance such as end-to-end packet delay, packet delivery ratio, and average energy consumption in comparison to other routing schemes including SPF, CODA, ESRT, and GRATA. To address practical concerns, the proposed routing algorithm can be easily implemented on existing devices without major changes. The limitation of the new method is that the values of traffic factors (α, β, and φ) are chosen based on simulation experiments. Moreover, overhead is a common drawback of proposed algorithms.

Buddha Singh et al [10]. Suggested by a Particle Swarm Optimization (PSO) approach for generating energy aware clusters by optimal selection of cluster heads. The PSO eventually reduces the cost of locating optimal position for the head nodes in a cluster. In addition, it has implemented the PSO based approach with in the cluster rather than base station, which makes it a semi distributed method. The selection criteria of objective functions are based on the remaining energy, intra cluster distance, node degree and head count of the portable cluster head. Furthermore, influence of the expected number of packet transmission along the estimated path towards the cluster head is also reflected in our PSO energy consumption model.

III. PROPOSED METHOD

3.1 Congestion Occurrences in WSN

The WSN is randomly deployed in particular area with base station(BS), positioned at coordinate number of sensor nodes are distributed in particular region of x,y (meters) based on distance. It is presumed that there are total numbers of cluster in the sensor network. Further, using to hybrid node deployment strategy, that means combination of equal and unequal cluster process. We assume the network model Fig.1 represents that intra cluster communication that sends the data from source to destination while in this process for an example in the 11^{th} node the congestion is occurred.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

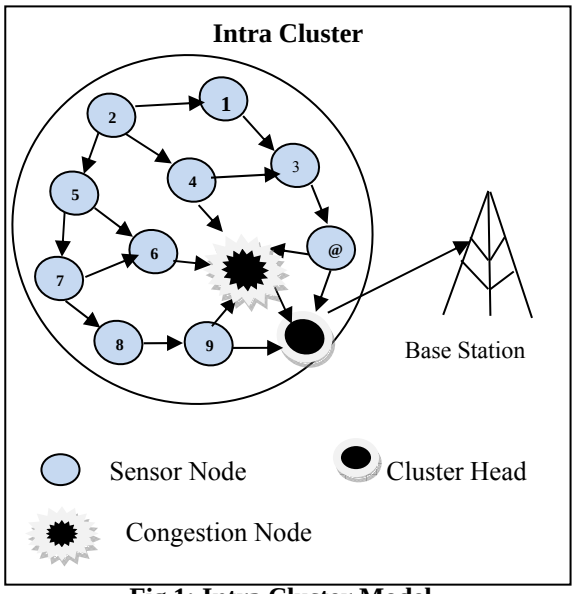


Fig.1: Intra Cluster Model

During the congestion, this is the right to check the nodes energy level and the buffer size. After this process got over nodes energy level and buffer size are high means the data will send through the CH node (or) neighboring CH node via the BS. Suppose any one the nodes energy level (or) buffer size is low in that time repeat request process is activated.

A sensor node is allowed to use different level of transmission power depending upon its distance from the target node. The distance can be estimated from the strength of the signal received from the destination node. The based station periodically sends a request to the cluster head to upload samples collected by the sensors (fig.1) on receiving the request the cluster head broadcast data collecting signal to its entire cluster member. The cluster member nodes are their packets to the CH, after which the CH processes and aggregates the collected packet and finally forwards the information to the Base Station (BS).

In this model summarized as follows

- Calculate the Intra cluster node to base station of node to congestion node distance.
- Determination of the number of communication links between the member nodes to CH nodes.
- Derivation of total retransmission of the collided packets in particular simulation time period.

3.2 Congestion Aware Architecture

The proposed Hybrid Cluster Based Congestion Aware (HCBCA) Algorithm is distributed in Hierarchical clustering communication between sources to destination. It acquires the hierarchical clustering structure in order to achieve the congestion Avoidance.

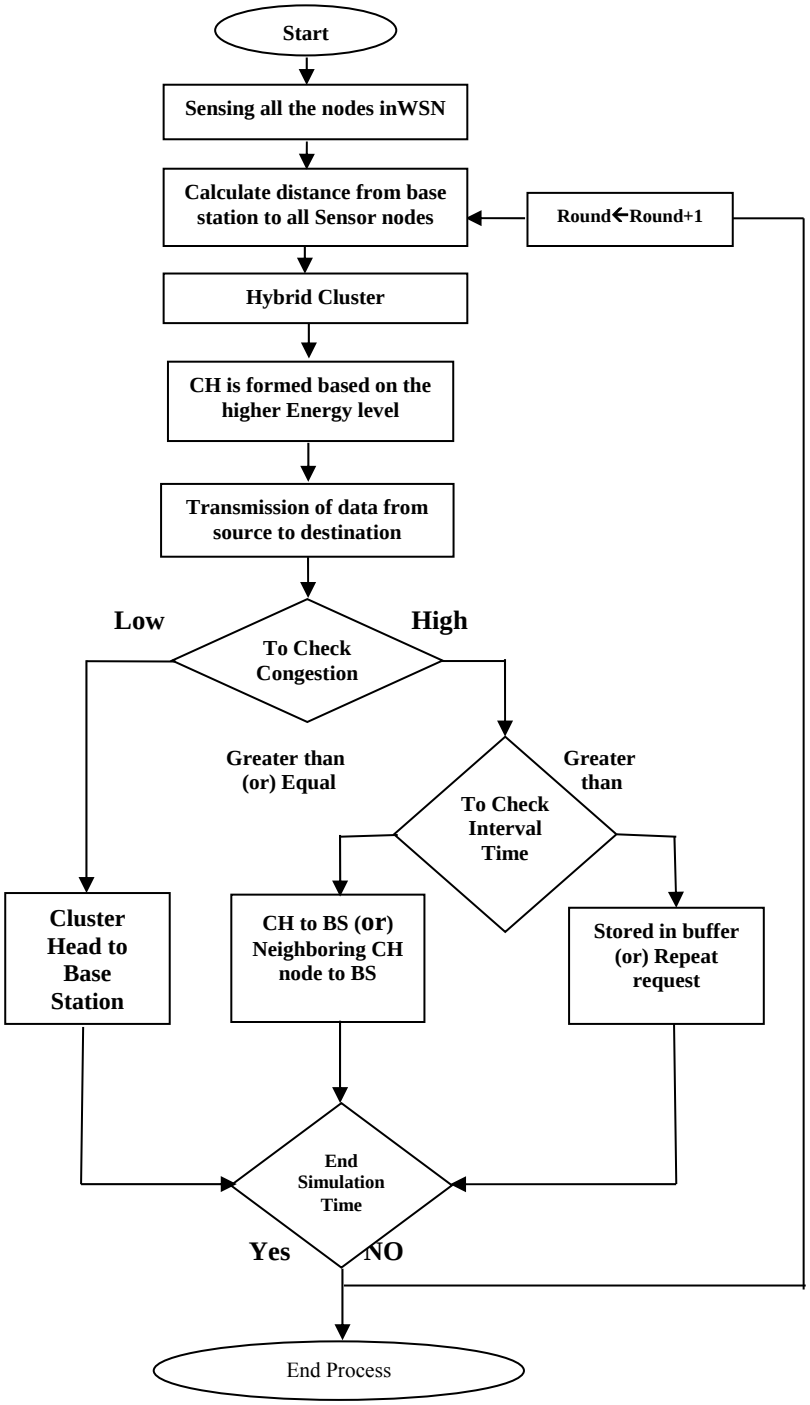


Fig.2: Congestion Aware Architecture

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

The Architecture of hybrid cluster based congestion describes the level of congestion and the performance of congestion in each level. In HCBCA check the condition in two levels. They are low and high level, when the congestion level is low then the packets are directly sends to base station from CH. In order to describe that congestion level is high, and then condition occurs to check the interval time. In mean while condition falls in less than or equal position then the packets are transmitted from CH to BS or in another way it will transmit the data from neighboring CH node to BS. When the condition is greater than the maximum interval time, then the packets are stored in the buffer or repeat request process will be active. At the last all these processes will meet the end simulation time, while the simulation ends then the process will end or otherwise data transmission will repeat its process from the start to end.

3.3 Determine Congestion Metrics

The clarification of these congestion metrics and determination of composite congestion metric are explained in the following subsection.

Distance Calculation

Distance between the nodes source to destination can be calculating using formula

$$D(S,R) = \sqrt{(x1 - x2)2 + (y1 - y2)2} \quad ....(1)$$

Where d(S,R) is the distance between node S and Base Station R, (X1,X2) is the X coordinate of node S and Base Station R and (y1-y2) is the coordinate of node R and Base Station R.

Find Queues length

The Queue length Q_1 is defined as the ratio of number of packets in the buffer to the maximum buffer size of node. It can be calculated as

$$Q_L(i) = \frac{NP}{BS(i)} \quad ..... \quad (2)$$

Where $Q_l(i)$ is the queue length of node i, N_P is number of packet in the buffer, BS(i) is the maximum buffer size of the node i.

Find the flow of Data

Contribution level P is calculated based on the total number of flows passing through a node as

$$P = \frac{\text{Current number of flows}}{\text{Number of sources}} \quad .... (3)$$

A node with more number of flows represents a high level that is more prone of the congestion as compared to the node with less member's flows.

The step for our Proposed Algorithm are described in Table 1

Table 1: Hybrid Clustering Based Congestion Aware (HCBCA) ALGORITHM

Initialization:

Min: Minimum Interval, Max: Maximum Interval

RRQS: Repeat Request

Sensor Nodes: $\{SN_1, SN_2, \ldots\ldots SN_n\}$

CH: The set of CHs based on Energy level $\{CH_1, CH_2 \ldots CH_n\}$

Step 1: Start

Step 2: To Form Sensor nodes with Wireless Sensor Network (WSN).

Step 3: cluster formation is generated based on sensor Nodes distance.

Step 4: cluster head (CH) is formed based on higher Energy node.

Step 5: Then, to collect higher energy level of Neighboring CH node.

Step 6: Transmission of packets between sources to Destination based on Hybrid Model.

Step 7: Congestion Checking

Step 7(a): Congestion level is low

Then, packets are directly sending from CH to BS.

Step 7(b): Congestion level is high

Then, Check interval time from source node to congestion node.

Step 8: If Interval time is less than (or) Equal

Then, packets are directly sending from CH to BS (or) another way Packets are sending From Neighboring CH node to base station.

Else

Interval time is greater than maximum

Then, packets are stored in Buffer (or) Repeat Request Process.

Step 9: To check Simulation Time

If Simulation time is End

Then, end process.

Else

Round→Round + 1

Then, following step 2 to 8.

Step 10: End Process

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

IV. SIMULATION RESULTS

The suggested hybrid congestion avoidance methodology is improved by Network simulator ($NS_{2.34}$) Environments.

Table 2: Simulation Parameter

Parameters	Values
Number of Node	50
Area Dimension	400 * 400(Meter)
Routing Protocol	DSDV
Total Energy	150 Joule
Initial Energy	0.5 Joule
Packet Size	4000 bits
Number Of Round	500
Type of the MAC	802.11
Simulation Tool	NS2.34

The proposed algorithm of EEBHC is highlighting on the network energy with new developed method HCBCA is provide a good output with respect to the packet delivery ratio, End to End delay time, Dead node occurrences in rounds, packet losses and Energy Savings.

Performance of packet delivery ratio

The representation of fig: 4.1 denote that the existing method was overcome by HCBCA. In this packet delivery ratio, the packets are transmitted from the sources to destination by proper routing path to evaluate the number of packets that are delivered in WSN like hybrid cluster head approaches in WSN.

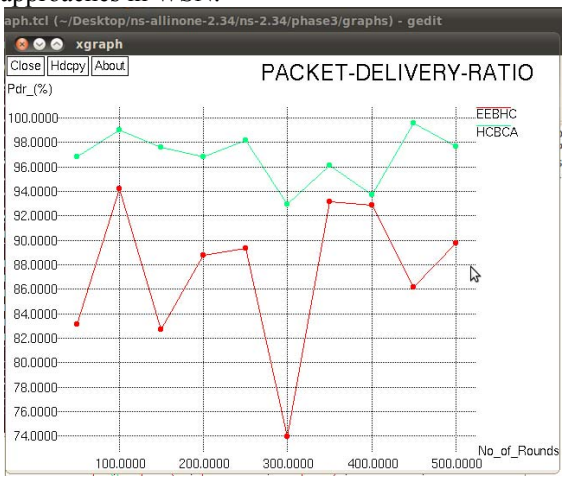


Fig 4.1 packet delivery ratio

While comparing the result of existing method EEBHC denotes 87.41% and the proposed method HCBCA denotes 96.85%.

Performance of End to End delay

The packet delay was average Maximum time to arrive in the destination. It take time maximum delay when congestion is occur for that time packets are stored in buffer or Repeat Request process is activated.

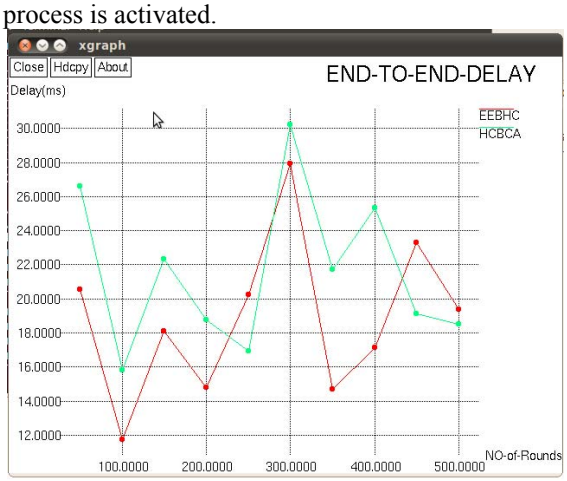


Fig: 4.2 End to End delay

When comparing the existing method EEBHC the result takes maximum time to transmit the packet now the proposed work is better for transmit the packets.

Performance of Dead Node Occurrences

During the packet transmission while the dead node occurred packet losses are reduced in proposed work. In the proposed work when dead node occurs reducing we use buffer management and retransmission concept.

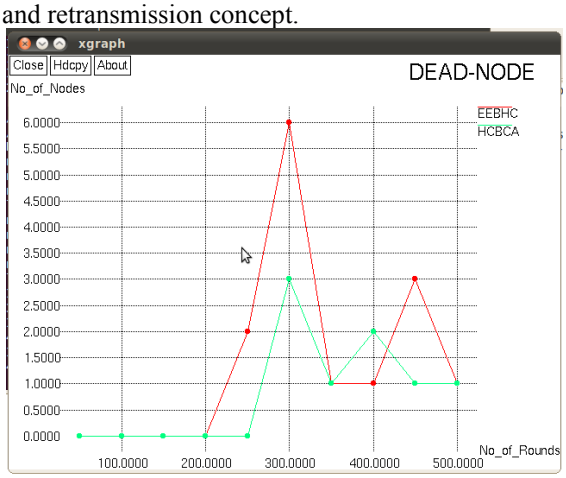


Fig: 4.3 Dead Node Occurrences

When compared to existing method EEBHC the dead node occurs during 237th round and average packet losses is 12.59%. The proposed HCBCA method dead node occurred in 273 round and average packet losses is 3.15%. During the packet transmission while the dead node delay occurred means packet losses will be reduced for that proposed method is used in the effective manner.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Performance of Remaining Energy

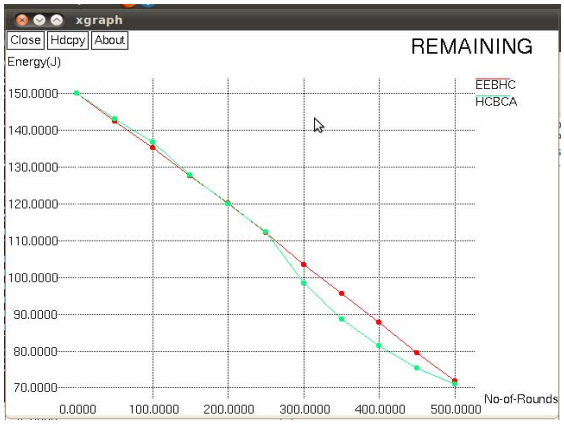


Fig: 4.4 Remaining Energy

Comparing the remaining energy by taking 100% and 150 joules as sample for that, we get the following results existing and proposed work. The result for the existing method EEBHC 47.8% and corresponding joule is 71.760J. So, the proposed works proves that better for saving the remaining energy.

V. CONCLUSION

In this paper, we proposed Hybrid Cluster Congestion Aware method is concentrated on the buffer management and Packet retransmission in WSN. The objective is providing high transmission of packet delivery ratio has been improved the network lifetime performance with respect to time, at the same time packet losses have reduced by packet retransmission.

The initially energy level of the node is being lower than 0.5J the node in noted as a dead node. So, the algorithm HCBCA is used. In that time data losses or data aggregation can be failed during the energy level is low. Finally, the simulator has considered the advantages of HCBCA method and demonstrated for the significant performance in an improvement over existing Scheme.

VI. REFERENCES

[1] Azlan Awang, Omer Chugtai, Nasreen Badruddin "A Congestion-Aware and Energy efficient Traffic Load balancing Scheme for Routing in WSNs" Conference paper in Telecommunication System, IEEE oct-2014.

[2] Haider Banka, Srinivasa Rao p.s, Prasanta, Jana.K "A Particle Swarm optimization based energy efficient cluster head selection algorithm for wireless sensor networks" Wireless Netw-Springer, pp(1270-07), April-2016.

[3] Rheleh Hashemzehi, Reza Nourmandipour, Farokh Koroupi "Congestion in Wireless Sensor Networks and Mechanisms for Controlling Congestion" Indian Journal of Computer Science and Engineering(IJCSE) vol.4, pp(204-207), jun-2013.

[4] Chia-Hsu Kuo, Tzung-Shi Chen, Zheng-Xin Wu "Congestion Control under Traffic Awareness in Wireless Sensor Networks" International Computer Symposium, pp(429-4343), IEEE-2016.

[5] Omer Chghutai, Nasreen Badruddin, Azlan Awang, Maaz Rehan "congestion-aware and traffic load balancing scheme for routing in WSNs" Telecommun Syst-Springer, January-2016.

[6] Ji-ming CHEN ,Wei-wei FANG , Lei SHU ,De-pei QIAN " Congestion avoidance, detection and alleviation in wireless sensor networks" Journal of Zhejiang University – SCIENCE C(Computer & Electronics), pp(63-73)-2010.

[7] Vaibhav Eknath Narawade , Uttm D.Kolekar "Congestion Avoidance and Control in Wireless Sensor Networks: A Survey" IEEE- 2016.

[8] B M Thippeswamy , Raghvendr M, Reshma S "MCDR: Mitigating Congestion using Distance based Routing in Wireless Sensor Networks"IOSR Journal of Computer Eineeringng (IOSR_JCE) , vol-16, Issue 1, pp(17-29), Jan-2014.

[9] Majid Gholipour, Abolfazl Toroghi and Mohammad Reza Meybodi "Hop-by-Hop traffic-aware routing to congestion control in wireless sensor networks" EURASIP Journal on wireless Communication and Networking, pp(1-13), Springer-2015.

[10] Buddha Singh and Daya Krishan Lobiyal "A novel energy-aware cluster head selection based on particle swarm optimization for wireless sensor networks" Human- Centric Computing and Information Sciences, pp(1-18), Springer-2012.

[11] Ji-Yeon Min, Sang-Hwa Chencyung, and Yuvin Ha "An Energy and Latency Aware WSN MAC Protocol for Bidirectional Traffic in Data Collection" International Journal of Distributed Sensor Networks (IJDSN) – 2015.

[12] Younas Khan, Sheeraz Ahmed, Imran Ahmad "ETEEM-Extended Traffic Aware Energy Efficient MAC Scheme" International journal of Advanced Computer Science and Applications (IJACSA), vol-7, pp(272-277),2016.

[13] Muzamil Basha S, Vinodh K, Raghu Veer "Enhanced Congestion Aware Routing for Congestion Control in Wireless Sensor networks" International Journal of Scientific and Research Publications(IJSRP) vol-3, Issue-2, Feb-2013.

Implementation of OFDM System Using Various Channel Modulation Schemes

Muhammad Zunnurain Hussain

Sr. Lecturer CS & IT Bahria University Lahore Campus

zunnurain.bulc@bahria.edu.pk

Muhammad Zulkifl Hasan

Sr. Lecturer CS & IT Lahore Garrison University, Lahore

zulkifl.hasan@lgu.edu.pk

Abstract

This report discusses the planning Associate in nursing the implementation of an OFDM system in several information module schemes like M-QPSK, M-QAM. First, a short introduction is provided by explaining the background and the specification of the project. Then the report deals with the system model, every block of the OFDM system is represented (IFFT, FFT, Cyclic prefix, modulation / reception, Channel estimation, bit error rate). System design is analyzed. The transmission techniques, further because the system parameters for transmission and reception are explained well. Finally, the results are provided.

1 Introduction

In orthogonal frequency division multiplexing (OFDM) the essential principle is to separate a high rate stream into type of lower-rate streams that area unit transmitted at an equivalent time over type of subcarriers (SCs). Each of that's modulated on a separate subcarriers (FDM).So, the knowledge live of the subcarriers becomes smaller the knowledge live of the channel therefore, all subcarriers area unit entirely filled with flat weakening, that produces the trouble technique easier and straightforward. That the image quantity of the individual subcarrier streams is made long compared to the delay unfold of the time-dispersive radio channel. The dispersion area unit about to be reduced attributable to the image amount that's inflated for lower rate parallel subcarriers this will be caused by multipath delay unfold. Lay image interface are eliminated by the introduction of the guard interval inside the sub carrier stream. By selecting a special set of (orthogonal) carrier frequencies, high spectral efficiency is obtained due to the spectra of the subcarriers overlap, whereas mutual influence among the subcarriers are avoided. The system model shows that by introducing a cyclic prefix (the GI), the orthogonality are maintained over a dispersive channel. OFDM are enforced victimization utterly completely different parameters however; we have used DVB-T standards 2k, 4k, 4k modes. We have jointly used utterly completely different modulation schemes for comparison in our coding/ implementation section[1].

2 Previous work

People square measure performing on OFDM because it is currently a awfully helpful technique to send knowledge at high rate with less ISI and delay spreads. Previous work embrace following paper and researches listed below:

- A MATLAB program was written to analyze Orthogonal Frequency Division Multiplexing (OFDM) communication systems. This program is effective for future researchers simulating systems that square measure in theory complicated to research. Single carrier QAM and multicarrier OFDM were compared.
- To demonstrate the strength of OFDM in multipath channels. 2 graphical interface demonstrations show a number of the fundamental ideas of OFDM. [2]
- Orthogonal frequency division multiplexing (OFDM) could be a promising technique for the high rate wireless communications as a result of it

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

may be combat inter-symbol interference (ISI) caused by the dispersive attenuation of wireless channels. The planned analysis focuses on techniques that improve the performance of OFDM based mostly wireless communications and its business and military applications. In particulars the paper addresses the subsequent aspects of OFDM: inter-carrier interference (ICI) suppression, co-channel interference suppression for clustered OFDM, clustered OFDM based mostly anti-jamming modulation, channel estimation for MIMO-OFDM, and precoding for MIMI-OFDM with channel feedback.[1], [3]

• This paper proposes a MIMO OFDM baseband transceiver style for future generation high output wireless LAN mistreatment 2 transmission antennas and 2 receiver antennas. A MIMO OFDM receiver with algorithmic rule for temporal arrangement and frequency synchronization, tracking, channel estimation, and MIMO detection is intended and enforced in software system. Simulation results shows that the planned receiver is capable of transmission with a knowledge rate that's doubly that of the present IEEE 802.1a wireless LAN customary.[4]

• One of the proposals for the physical layer of this technique was entitled innovative modulation for the Brazilian Digital TV System (MI-SBTVD). The MI-SBTVD Project includes high performance error correcting codes; transmit spacial diversity and multi carrier modulation. The focus of this paper is twofold. First, we glance at the transmit diversity theme, which mixes Alamouti committal to writing and OFDM modulation. We tend to then discuss the channel estimation algorithmic rule that has been enforced within the planned system. Pilot subcarriers square measure inserted among knowledge subcarriers, and each uni-dimensional and Bi dimensional linear interpolation at the receiver square measure thought of. Theoretical account results, mistreatment typical digital TV channels, show that the planned theme is in a position to perform on the brink of the case of a superbly acknowledged channel at the receiver.[2]

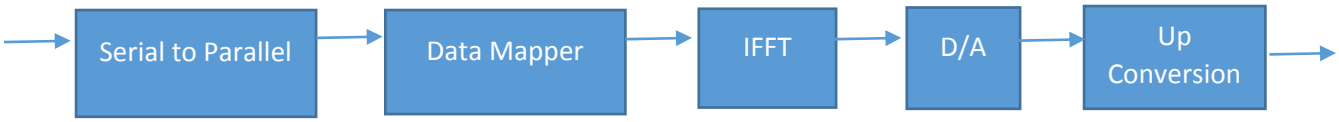


Figure: Basic OFDM Transmitter

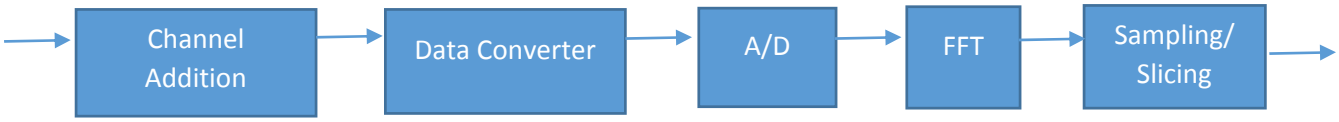


Figure: Basic OFDM Receiver

3 System Design Explanation

3.1 Serial to Parallel

The data input to an OFDM transmitter is in the form of binary bit stream, consisting of [0 , 1]. For constellation mapping, using any useful

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

data scheme, we first need to convert this serial data into parallel data. So this block provides us with parallel data ready for constellation mapping.

3.2 Constellation Mapper

At this stage we mapped our bit stream data, in useful manner as per requirement of OFDM, Orthogonal to each other, using modulation schemes i.e. PSK and QAM. Each bit from random binary stream is picked up and placed according to the modulation scheme at orthogonal frequency to avoid ISI and a graph is being made between real and complex values in XY-plane to see the constellation map.

QPSK, QAM: these are the two techniques for data mapping which we have tested in lab. Moreover, their efficiency changes have also been recorded.

3.3 IFFT

On the transmitter aspect the IFFT of a symptom X(k), wherever k denotes the frequency parts, and x(l) is that the ensuing sampled signal, that is made by the total of the modulated frequency parts X(k) (at their corresponding digital frequency k=K). To retrieve once more the digital frequency parts, the inverse equation should be used.

3.4 Parallel to Serial

In this stage for adding the cyclic prefix and zero padding, we have to convert it to serial data from parallel.

3.5 Digital to Analog Conversion

In this block, we have converted digital subcarriers into analog baseband signal. For this process, we used stream of pulses and convolved it over our subcarriers, getting digitized pulse. Further, by the help of pulse shaping filter, of high order, we converted it into continuous time baseband signal.

3.6 Up converter

We multiply the signal with high frequency to increase the power of signal. [5]

3.7 Channel Addition

In this stage, we have a tendency to check the behavior of our system by introducing some channels like Rician, Lord Rayleigh and additional AWGN noise.

In Lord Rayleigh channel, there is no main path. Instead, the received signal is mirrored into many tiny power signals. Therefore, it is tough to synchronize. The Rician issue K is that the magnitude relation of the ability of the direct path to mirrored ways.

Fix Reception Rician:

$$Y(t)= \frac{\rho o x(t) + \sum_{i=1}^{N} \rho i \; e^{-j\theta i} \; x(t-\pi i)}{\sum_{i=0}^{N} \rho^2}$$

Portable Reception Rayleigh:

$$Y(t)= \frac{1}{\sqrt{\sum_{i=0}^{N} \rho^2}} \sum_{i=1}^{N} \rho i \; e^{-j\theta i} \; x^{(t-\tau i)}$$

Where N is the number of echoes and equals to 20; θi is the phase shift from scattering of the ith path; ρ^i is the attenuation of the ith path and t^i is the relative delay of the ith path. The Rician channel contains a sturdy main path, therefore it is easier to try to on synchronization and channel estimation. Therefore, the system has higher BER performance than the Rician channel. as a result of the trail delays area unit continual, we have a tendency to translate those into separate sample index, therefore we have a tendency to solely select one path of those with a similar separate sample index.[4]

3.8 Down Conversion

In this stage, we multiply the signal with same frequency to get the same signal back.[6]

3.9 Serial to Parallel

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

To remove cyclic prefix (if used in transmission) we have to convert our data from serial to parallel. An OFDM receiver includes a demodulator unit being tried to a receiver signal for demodulating both an in-phase (I) component and a quadrature-phase (Q) component of the receiver signal, a serial to parallel unit for converting the output of the demodulator to a plurality of parallel paths. [7]

3.10 Analog to Digital Conversion

In this block, we have converted analog sub band signals into digital subcarriers. For this process, we used low order pulse shaping filter (Butterworth).[8]

3.10 FFT

At the receiver, ignoring channel affects, time wave form is digitized so born-again back to a Symbol victimization AN FFT. The FFT may be a important a part of the receiver as a result of before reception it converts continuous sign into carriers. once over one carrier is gift, it's the sole sensible technique for sick the info from overlapping carriers. it's uphill, for instance, to use a single-carrier sixty four QAM receiver to pull up a sixty four QAM carrier in AN OFDM system. [9]

3.11 De mapping and sampling

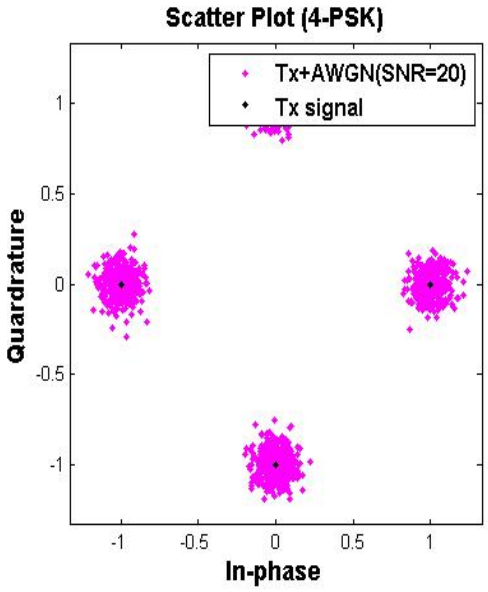
In this block, we have a tendency to take away the zeros from our knowledge to induce the first one. The equalization (symbol Diamond State mapping) needed for police work the information constellation is a component wise multiplication of the DFT output by the inverse of the calculable channel TF (Channel Estimation). For PM schemes, multiplication by the advanced conjugate of the channel estimate will do the equalization. After all, that we have a tendency to square measure with our knowledge streams back.[10]

3.12 Parallel to Serial

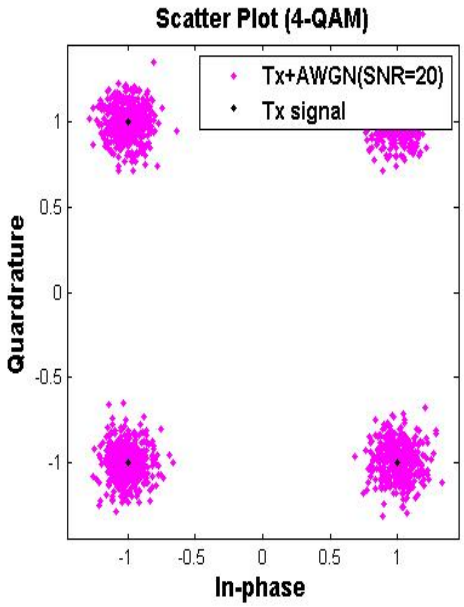
To get the output we have to convert our data back to original form. [11]

4 Results showing effects of noise and modulation on data

4.1 4-PSK Modulated Data & AWGN Noise Added Data

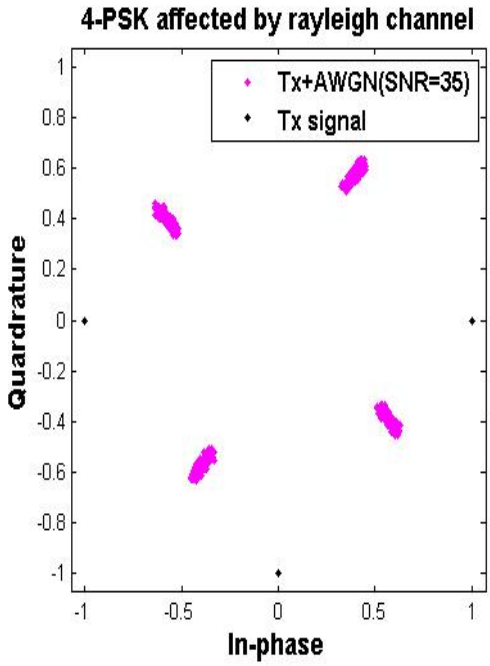


4.2 M-QAM Modulated Data & AWGN Noise Added Data

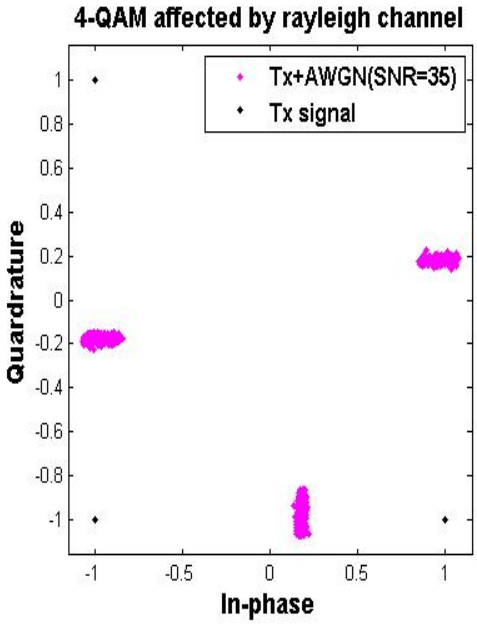


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

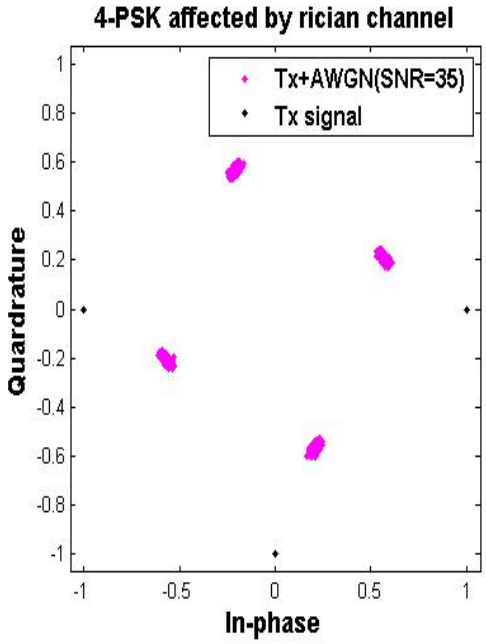
4.3 Rayleigh Channel Addition to 4-PSK Modulated Data



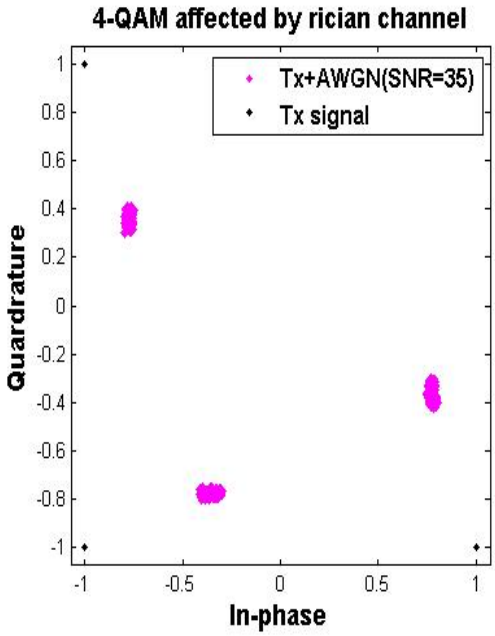
4.4 Rayleigh Channel Addition to 4-QAM Modulated Data



4.5 Rician Channel Addition to 4-PSK Modulated Data



4.6 Rician Channel Addition to 4-QAM Modulated Data

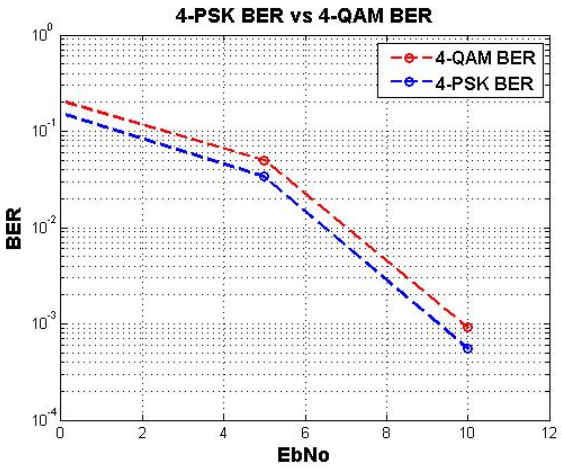


https://sites.google.com/site/ijcsis/
ISSN 1947-5500

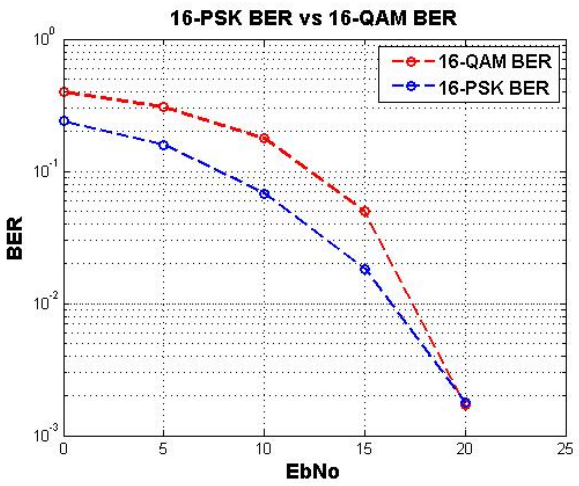
4.7 BER Comparison Graph

These graph show comparison between bit error rates of M- QAM & M-PSK modulated data.

4.7.1 BER Curve Comparison for 4-PSK & 4-QAM



4.7.2 BER Curve Comparison for 16-PSK & 16-QAM



Conclusion & Future Work

The above research is about giving common users, an opportunity, to observe data transmission & reception in a step manner, with OFDM being implemented automatically. The user can do this by “Easy to use software”. This GUI or software allows the user to give the wanted DVB-T Parameters and the user has a choice to choose a modulation type, between QAM & PSK. The software has also a modulation scheme order changing option. Then a channel effect can also added along with AWGN noise at any SNR, the user wants then all the graphs are displayed, stepwise, on the side panel, showing the user all the step results in graphical form, like scatterplot of simple modulated and channel effected data. After seeing, the complete transmission and reception one can compare its bit error rate of any other data, which is transmitted using different configuration of parameters or modulation schemes, as the software has the ability to save two results of different configurations and compare them, by making a comparison BER graph with respect to increase SNR. From these comparison graphs, it is clear that higher order modulation scheme have a better efficiency in even low SNR conditions. But energy per bit increases significantly and the modulation scheme like 4-PSK or 4-QAM which are more vulnerable in low SNR condition have a low energy per bit so it is a tradeoff between accuracy and energy also we can never say that QAM is better technique than PSK or vice versa. Because there is, efficiency varies with changing SNR or Eb/ No

Due to shortage of time, we could not show the transmission and reception of a video file. Instead, we used random bits to observe transmission and reception. In future, we would like to transmit and receive different video formats and other types of data like audio and text. In addition, we are planning to enhance the functionality of our GUI by adding video and image blocks.

References

[1] R. Prasad, *OFDM for Wireless Communications Systems.* Artech House, 2004.

[2] G. C. Lima *et al*, "Design, simulation and hardware implementation of a digital television

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

system: STC-OFDM and channel estimation," in *Spread Spectrum Techniques and Applications, 2006 IEEE Ninth International Symposium On,* 2006, pp. 198-202.

[3] J. R. Oliva, "Underwater Wireless Video Transmission using Acoustic OFDM: An experimental approach to the transmission and reception of OFDM signals for real-time video communication," 2011.

[4] C. Yu, Z. Ding and T. Chiueh, "Design and simulation of a MIMO OFDM baseband transceiver for high throughput wireless LAN," in *Circuits and Systems, 2004. Proceedings. the 2004 IEEE Asia-Pacific Conference On,* 2004, pp. 205-208.

[5] J. Yazdani *et al*, "Modelling of an OFDM transmission technique for power line using UNIX BasedCOSSAP'and power line channel simulator," in *International Symposium on Power-Line Communications and its Applications. T. Coffey and S. McGrath (Editors). University of Limerick. Limerick, Ireland. 5th-7th April,* 2000, pp. 1-874653.

[6] H. Zhang, *Orthogonal Frequency Division Multiplexing for Wireless Communications,* 2004.

[7] H. Zou, B. McNair and B. Daneshrad, "An integrated OFDM receiver for high-speed mobile data communications," in *Global Telecommunications Conference, 2001. GLOBECOM'01. IEEE,* 2001, pp. 3090-3094.

[8] T. Shahana *et al*, "RRNS-convolutional concatenated code for OFDM based wireless communication with direct analog-to-residue converter," *World Academy of Science, Engineering and Technology,* vol. 2, *(2),* pp. 86-93, 2008.

[9] M. Al-Akaidi, O. Daoud and S. Linfoot, "A new turbo coding approach to reduce the Peak-to-Average Power Ratio of a multi-antenna-OFDM system," *International Journal of Mobile Communications,* vol. 5, *(3),* pp. 357-369, 2007.

[10] A. C. Brooks *et al*, "Design and Simulation of Orthogonal Frequency Division Multiplexing (OFDM) Signaling," *Electronic Publication: Digital Object Identifiers (DOIs),* pp. 1-4, 2001.

[11] J. Liu, *System Design and Signal Processing for OFDM-Based Wireless Communication Systems,* 2004.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

https://sites.google.com/site/ijcsis/
ISSN 1947-5500
International Journal of Computer Science and Information Security (IJCSIS),
Vol. 16, No. 3, March 2018

Information Technology in Animal Health Care System

Omankwu, Obinnaya Chinecherem, Nwagu, Chikezie Kenneth, and Inyiama, Hycient

[1] Computer Science Department, Michael Okpara University of Agriculture, Umudike Umuahia, Abia State, Nigeria
saintbeloved@yahoo.com

[2] Computer Science Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria, Nwaguchikeziekenneth@hotmail.com

[3] Electronics & Computer Engineering Department, Nnamdi Azikiwe University, Awka Anambra State, Nigeria.

ABSTRACT

The information technology played an important role in information and knowledge dissemination in the last decade. The usage of IT to transfer information and knowledge in the animal health care domain using expert systems is one of the areas investigated by many institutions. The current era is witnessing a vast development in all fields of animal health care. Therefore there is a need for an unconventional method to transfer the knowledge of experts in this domain to the general public of livestock holders, especially that the number of experts in new technologies is lesser than their demand in a certain domain. The transfer of knowledge from veterinary consultants & scientists to livestock holders represents a bottleneck for the development of animal health care in any country. Expert systems are simply computer software programs that mimic the behaviour of human experts. They are one of the successful applications of the Artificial Intelligence field, a branch in Computer Science that investigates how to make the machine think like human or do tasks that humans do. Expert Systems are very helpful to ensure an effective and nationally coordinated approach in response to emergency incidents and in routine bio-security activities. Such systems enable better management of the information and resources used to manage animal's diseases and emergency responses to incursions**.**

***Keywords*— Artificial Intelligence, Expert System, IITV, Ultrasound, Tomography, MRI, DSA , Endoscopy**

INTRODUCTION

Livestock wealth is very precious for a developing country like India. In India, animal husbandry is no longer a subsidiary to agriculture or a backyard vocation. Animal husbandry has metamorphosed into an industry and the latest reports suggest that the contribution of animal husbandry sector to the GDP of the nation is substantially higher despite the meager input. Animal husbandry offers a better scope for marginal farmers whose income from agriculture is dwindling fast due to vagaries of monsoon, fragmentation of landholdings, pest problems, poor pricing etc. Though the growth of livestock industry is very promising, in order to make India a global leader in animal husbandry, it is imperative to integrate it with developments in other fields. The developments in Information Technology over the past few decades are tremendous and offer great potential in improving animal health through various measures like effective disease forecasting, rapid and accurate disease diagnosis, modern therapeutic measures etc.

INFORMATION TECHNOLOGY IN ANIMAL HEALTH CARE

Medical diagnostic technology has made rapid strides with the advent of the computer. Many of the advances in human diagnostic technologies are translated into veterinary medicine in developed countries. Newer branches like Imaging, Radio diagnosis; Telemedicine, Telesonography and Teleradiology have emerged. Broadly, the instrumentation/devices which have been created with modern technology in the present digital age are listed below.

1) Image Intensifier TV system (IITV): Generally used in orthopedic surgery. IITV helps in X-ray imaging of the intra-operative site for orthopedic manipulations, and the same can be stored for future reference.

2) Ultrasound: In small animal ultrasound is routinely used as a diagnostic aid. Ultrasonography seems to have a promising future in veterinary medicine, particularly for the assessment of intra/peri-abdominal disease. Ultrasonography is non-invasive and non-surgical armamentarium of the veterinary clinician since the advent of the fiber optic endoscope.

3) Computerized Tomography (CT): CT has been an extremely significant development which has a unique cross-sectional imaging ability useful for the diagnosis of tumors, malformations, inflammation, degenerative and vascular diseases and trauma.

4) Magnetic Resonance Imaging (MRI): MRI is a highly sensitive and non-invasive technique providing accurate and detailed anatomic images with good contrast and spatial resolution. MRI is still in its infancy and its use is infrequent. To date, MRI has been used in developed countries in clinical cases as well as a research tool especially for diseases in small animals.

5) Digital Subtraction Angiography (DSA): DSA is a radiographic modality which allows dynamic imaging of the vascular system following intravascular injection of iodinated X-ray contrast media, through the use of image

intensification, enhancement of the iodine signal and digital processing of the image data.

6) Laparoscopy: Only in the last 15 years, its use has been extensively in various animal species for research and clinical diagnostic and therapeutic purposes. The most advantageous characteristic of laparoscopy is that it allows direct examination of abdominal cavity with only minimal and superficial surgical intervention.

7) Endoscopy: It is a minimal invasive diagnostic modality which aids in documenting mucosal inflammation, hyperemia, active bleeding, irregular mucosal surface etc. and facilitates biopsy in tubular organs like GI tract, respiratory and the urogenital systems.

As all the above techniques require human experts to analyse the results of IT application, there is a need for an unconventional method to transfer the knowledge of experts in this domain to the general public of livestock holders through the expert system. Expert system is one of the successful applications of the Artificial Intelligence field. Artificial intelligence may be defined by comparing computer and human functions. If the computer performs a task that seems intelligent when it is done by humans it can be said to be exhibiting artificial intelligence. In medicine, most artificial intelligence research has been devoted to creating computer systems that contain detailed information about a specific medical subject. By focusing relevant knowledge on the problems facing the physician, these programs are designed to act like consultants and thereby have the potential of expanding the practitioner's expertise.

Expert systems are computer programs that typically contain large amounts of knowledge for making decisions about specific problem domains such as an area of medicine. In medicine, several important experimental expert systems have been developed. For example: INTERNIST - Diagnosis in internal medicine, PIP - Renal disease, VM - Ventilator Management, PUFF - Pulmonary function and ATTENDING - Anesthetic Management. Similarly researchers have gone through following different researches based upon the various expert systems in animal health care domain.

REVIEW OF LITERATURE

Jeffrey C. Mariner1, Dirk U. Pfeiffer2(2011)

Through their research they suggested the Participatory Epidemiology Network for Animal and Public Health (PENAPH) seeks to facilitate research and information-sharing among professionals interested in participatory approaches to epidemiology and risk-based surveillance. As part of this process, the network supports innovation in institutional capacity by promoting minimum training guidelines, good practice and continued advancement of methods through action research.

Graeme Garner(2011)

This research indicates that trade and market access is a major focus of surveillance in Australia. The animal health surveillance system in Australia has evolved to meet a range of regional, state/territory, national and industry needs including Notifiable disease reporting, Trade and market access, Regional and national animal disease management, Monitoring endemic diseases and Early detection of exotic and emerging diseases

Jampour,M (2011)

In his research, he concludes that animal health and domestic products health undoubtedly are the most basic health factors, although, there are complete and correct information in the disease of animal with neurological involvement, however, generally defined neurological diseases only on the basis of clinical symptoms is not so simple as so proximity neurological signs and in most instance veterinarians will doubt in diagnose. In this research researcher use the fuzzy logic model approach to determine and calculate lack or involvement of each the possible disease with neurological signs and sufficiently reduced natural Uncertainty regarding the diagnosis of disease.

Gustavo Sotomayor(2011)

He commented that the Animal Protection Division of the Agriculture and Livestock Service of Chile (SAG) has moved from using file-based information and local databases – in other words a nonstandard, non-interconnected system – to a centralized database with which users connect via a WAN (Wide Area Network). Until 2004 the recording, storage and analysis of data (information management) was mainly carried out using local, spreadsheet-type files compiled by those responsible for the different programmes. These were sent to the SAG operational offices and then bound as management reports or epidemiological analysis.

Hosein Alizadeh, Alireza Hasani-Bafarani, Hamid Parvin, Behrouz Minaei, Mohammad R. Kangavari(2008)

Through their research, researchers highlighted the possibility of developing of an expert system for replacing human expert investigated. Also, the knowledge extraction methods are scribed. Fuzzy logic is used for dealing with uncertainty. Finally, the Knowledge representation methods are discussed and fuzzy rule base is proposed for representing this knowledge.

Soegiarto(2011)

His research is based on Indonesian animal health service using computerized information systems to assist in managing animal and zoonotic disease for almost 20 years.

Initially these were adaptations of programs developed internationally, but in the past ten years these have been replaced by three nationally developed systems: SIKHNAS for managing surveillance data, InfoLab used by regional veterinary laboratories, and the HPAI Information System for monitoring HPAI surveillance and control. These applications are all standalone, which can lead to data integration problems at a national level.

P.L. Nuthall, G.J. Bishop-Hurley(1999)

Their research is a section of a wider study involving expert systems for feed management which covers the

development of a successful interface for expert systems and the farmers attitudes to the expert systems themselves. Alternative forms of the interface were created and presented to both professionals and farmers for evaluation and use. Their responses were used to conclude on a number of interface design questions. A clear preference for data input through as few screens as possible using pick lists and a mouse is evident, as is the benefit from providing on-call pictures to visually depict alternatives where the user has a choice.

Van Dang Ky(2011)
Through his research, he indicated that Viet Nam's disease information and surveillance system has been in place since the 1960s. However, before the year 2000 the system showed limitations, such as slow outbreak detection and delayed information transmission. Many outbreaks, therefore, could not be detected early on and the implementation of control measures was delayed, causing diseases to spread. At pesent, many diseases are under intensive surveillance and monitoring. Rapid response to outbreaks is performed well at different levels of the veterinary system.

Dickens M Chibeu(2011)
The researcher has explained the role of the Animal Resources Information System (ARIS) in decision-making, planning and monitoring cannot be overstated. Specifically, ARIS is useful in early warning and rapid response, allocating resources, assessing the level of livestock contribution to livelihoods and GDP, and formulating policy. About a decade ago, there was no comprehensive information system at IBAR or in most Member States (MS) capable of contributing efficiently to these surveillance and decision-making activities. The focus then was on disease reports for international organizations, with no systematic data collection, analysis and information dissemination. Data from different sections of Animal Resources was fragmented, with a majority of MS using paper-based data management rather than databases.

Kellaway, RC(1988)
His research is a design of CAMDAIRY, a computer model containing a package of programs designed to help advisers, farmers, students and research workers who are involved in the feeding of dairy cows. Details of the model are given by Hulme .The core program incorporates functions to predict nutrient requirements, feed intake, substitution effects when feeding concentrates, tissue mobilisation and partition of nutrient utilisation between milk production and growth. Nutrient partitioning is described by a series of asymptotic curves relating energy intake to milk production, such that energy requirements per litre increase progressively with level of milk production.

Mokganedi Mokopasetso(2011)
He concludes that within the Southern African Development Community (SADC) member states, livestock farming is considered one of the main pillars for developing rural livelihoods. In particular, there is a critical need to strengthen national epidemic surveillance systems to enable timely collection, reporting and analysis of animal disease data. The overall project objective was to strengthen regional preparedness against the spread of trans-boundary animal diseases, and its main undertaking was to strengthen animal disease surveillance through improving disease data collection and processing for decision-making. This is the context in which Digital Pen Technology (DPT) was introduced to the region as an innovative way to collect and send animal disease surveillance data from remote areas in the field to Central Epidemiology Units for analysis and decision-making.

Lawrence R. Jones(1990)
Technologies outlined in this research represent the foundation of the next generation of computer applications for dairy herd management. If adopted, these technologies will allow the development of systems that are more intuitive to use, are easier to learn to use, and provide more complete access to management information. Integrated decision support systems have the potential to supply dairy herd managers and their consultants with a complete computerized system to address many farm problems. As these systems are augmented with more intelligent user interfaces, they should eliminate many of the problems facing dairy herd managers in selecting and using software. The result of adopting such technologies will be better informed management.

Mat Yamage and Mahabub Ahmed(2011)
Developed for Avian Influenza Technical Unit, Food and Agriculture Organization of the United Nation Department of Livestock Services, Dhaka, Bangladesh, his research described the SMS gateway system which is a tool for transmitting a large amount of information from the grassroots level via a mobile phone to a central Internet server and consolidating this information automatically for handling by a single database manager. The flow of information is bi-directional and timely instructions can be given in response to a particular situation. The system is suitable for the surveillance of HPAI H5N1 in Bangladesh where the majority of poultry farms are in rural areas and not readily accessible to the national veterinary services owing to a shortage of human and material resources.

C. H. Burton, H. Menzi, P.J. Thorne, and P. Gerber, Cemagref(2008)
Their research commented that dissemination and knowledge transfer remain a challenge in many fields of research. This is especially the case for the application of livestock waste management in developing countries where there an overwhelming volume of material is already available to the farm advisor and the real need is often the transfer of such knowledge to the local level. The object of this project is to package up suitable techniques as an expert and design system that can be applied directly to farm situations across South East Asia. The software will comprise both calculation models (e.g. nutrient excretion of animals, nutrient balance, design and costs of manure treatment facilities), and decision tree elements (e.g. structured analysis of the present situation at a given farm). Outputs will include summary reports providing specific recommendations, specifications, case study examples and supporting multimedia background information.

A.J. Mendes da Silva1, E. Brasil(2011)
Their research indicates that the Second Inter-American Meeting on foot-and-mouth disease and Zoonoses Control

(RICAZ), under Resolution I, took the first steps towards establishing a Continental Epidemiological Information and Surveillance System (SCIV). The proposal was put forward by the Pan-American foot- and- mouth disease Centre (PANAFTOSA), which had at that time already established procedures by which member countries were urged to submit periodically, epidemiologic information on the occurrence of foot-and-mouth disease (FMD) and vesicular stomatitis, as well as other diagnosed types and subtypes of virus.

Sanjay S. Chellapilla(2003)
This research describes the design and implementation of DairyMAP, a Web-based benchmarking analysis and Expert System for Dairy Herd Producers, as part of the Dairy Management Analysis Program undertaken by the Edgar L. Rhodes Center for Animal and Dairy Science. The system consists of two major components – a preliminary statistical benchmarking analysis (based on Dairy Herd Information reports provided by the Dairy Records Management Systems, Inc., in Raleigh, NC) and, a detailed expert evaluation of the four major areas of dairy herd management, viz., Somatic Cell Count and Mastitis, Reproduction, Genetics, and Milk Production. The preliminary analysis provides information to the producer about the areas of concern within each component of dairy management, and suggests further evaluation and diagnosis by the Expert System, concluding with comments and recommendations for improving the producer's herd.

T. Rousing, M. Bonde & J. T. Sorensen(2001)
Their research suggests a welfare assessment protocol for loose housing systems for dairy caws based on four sources of information being the system, management, animal behavior and animal health. The animal behavior indicators refer to social behavior, man-animal relationship and resting/rising behavior. Health indicators focus on causes of pain and discomfort to the animal:
Extreme body condition, skin injuries and disorders, udder and teat lesions, lameness, hoof disorders and systemic diseases with general affection of the animal. The listed indicators were included in a protocol, which will be tested in ten commercial dairy herds. The herds will be visited regularly during a one-year period. System and management will be described and the behavioral and health indicators will be measured on a sample of the animals. The evaluation of the indicators will include statistical analyses, expert opinion and interviews with the articipating farmers.

A. Dagnino, J. I. Allen, M. N. Moore, K. Broeg, L. Canesi and A. Viarengo(2007)
Through their research they developed an expert system which is based on a set of rules derived from available data on responses to natural and contaminant-induced stress of marine mussels. Integration of parameters includes: level of biological organization; biological significance; mutual inter-relationship; and qualitative trends in a stress gradient. The system was tested on a set of biomarker data obtained from the field and subsequently validated with data from previous studies. The results demonstrate that the expert system can effectively quantify the biological effects of different levels of pollution. The system represents a simple tool for risk assessment of the harmful impact of contaminants by providing a clear indication of the degree of stress syndrome induced by pollutants in mussels.

J. Enting, R.B.M. Huirne, A.A. Dijkhuizen, M.J.M. Tielen(1999)
For constructing a knowledge-based system in the field of animal health management a documentation methodology has been developed and is reported in their research. The methodology was based on, among other things, the CommonKADS technique.
It includes three subsequent phases: documenting concepts and facts in hierarchies, documenting separate inferences which integrate knowledge documented in hierarchies, and documenting the strategy or sequence of the inferences to be made. The method supports the full pathway of the documentation process and addresses both declarative and procedural knowledge. Also, the method provides a quick insight into knowledge of a knowledge source (e.g. experts) and comprehensible transcripts for the expert. The latter facilitates the process of knowledge verification.

Michele Ruta, Floriano Scioscia, Eugenio Di Sciascio(2009)
Their research is based on an innovative Decision Support System for healthcare applications which is based on a semantic enhancement of RFID standard protocols. Semantically annotated descriptions of both medications and animals, or person case history are stored in RFID tags and used to help doctors in providing the correct therapy. The proposed system allows discovering possible incompatibilities in a therapy suggesting alternative treatments.
From above reviews, it clear that most of them are based on Animal Disease Surveillance to improve disease analysis, early warning and predicting disease emergence and spread. As a preventive measure, disease surveillance is aimed at reducing animal health-related risks and major consequences of disease outbreaks on food production and livelihoods. Early warning systems are dependent on the quality of animal disease information collected at all levels via effective surveillance; therefore, data gathering and sharing is essential to understand the dynamics of animal diseases. Through the proposed expert system researchers will utilize the experts knowledge for the best management practices for developing rules in variety of animal health care issues with special reference to lactating animals.

REFERENCES

[1] Jeffrey C. Mariner1, Dirk U. Pfeiffer2, Surveillance for the present and the future Challenges Of Animal Health Information Systems And Surveillance For Animal Diseases And Zoonoses By Food And Agriculture Organization Of The United Nations Rome, 2011
[2] Graeme Garner, Surveillance For Animal Diseases And Animal Health Information Management In Australia, Challenges Of Animal Health Information Systems And Surveillance For Animal Diseases And Zoonoses By Food And Agriculture Organization Of The United Nations Rome, 2011

[3] Kate Elizabeth Sawford, A Thesis Submitted To The Faculty Of Graduate Studies In Partial Fulfilment Of The Requirements For The Degree Of Doctor Of Philosophy

[4] Jampour,M, A Fuzzy Expert System to Diagnose Diseases with Neurological Signs in Domestic Animal,Information Technology: New Generations (ITNG), 2011 Eighth International Conference on Digital Object Identifier: 10.1109/ITNG.2011.220 Publication Year: 2011 , Page(s): C1 IEEE Conference Publications

[5] Gustavo Sotomayor, Progress and challenges in official information systems for disease surveillance in Chile , Challenges Of Animal Health Information Systems And Surveillance For Animal Diseases And Zoonoses By Food And Agriculture Organization Of The United Nations Rome, 2011

[6] Hosein Alizadeh, Alireza Hasani-Bafarani,Dairy Cattle Judging: An Innovative Application for Fuzzy Expert System www.iaeng.org/publication/WCECS2008/WCECS2008_pp247-250.pdf

[7] Soegiarto, Animal health information systems in Indonesia, Challenges Of Animal Health Information Systems And Surveillance For Animal Diseases And Zoonoses By Food And Agriculture Organization Of The United Nations Rome, 2011

[8] P.L. Nuthall , G.J. Bishop-Hurley, Expert systems for animal feeding management PartI: Presentation aspects, *Computers* and electronics in agriculture ISSN 0168-1699 CODEN CEAGE6. Source / Source. 1996, vol. *14*, no*1*

[9] Van Dang Ky, Animal health information system in Viet Nam , Challenges Of Animal Health Information Systems And Surveillance For Animal Diseases And Zoonoses By Food And Agriculture Organization Of The United Nations Rome, 2011

Dickens M Chibeu AU-IBAR , Animal Resources Information System (ARIS) of the AU-IBAR, Challenges Of Animal Health Information Systems And Surveillance For Animal Diseases And Zoonoses By Food And Agriculture Organization Of The United Nations Rome, 2011.

[5] Kellaway, RC, CamDairy : a computer program for performance prediction, ration analysis and ration formulation to maximise profit from dairy cows. http: //www.asap.asn.au /livestocklibrary /1988 /Kellaway88.PDF

[6] Mokganedi Mokopasetso, Digital pen technology for animal disease surveillance in Southern Africa, Challenges Of Animal Health Information Systems And Surveillance For Animal Diseases And Zoonoses By Food And Agriculture Organization Of The United Nations Rome, 2011

[7] LAWRENCE R. JONES, The Current State of Human-Computer Interface Technologies for Use in Dairy Herd Management',"Emerging Agricultural Technology: Issues for the 1990s" by the Office of Technology Assessment, Congress of the United States, Washington, DC.

[8] Mat Yamage and Mahabub Ahmed, Active surveillance of H5N HPAI using sms gateway in Bangladesh1 Challenges Of Animal Healt Information Systems And Surveillance For Animal Diseases And Zoonose By Food And Agriculture Organization Of The United Nations Rome, 2011

C. H. Burton, H. Menzi, P.J. Thorne, A computer-based Expert system to support the selection and implementation of systems for sustainable livestock waste management in South-East Asia, by Food and Agriculture Organization of the United Nations (FAO), Rome, Italy, www.ramiran.net/doc08/RAMIRAN_2008/Burton.pdf

[30] A. J. Mendes da Silva1*, E. Brasil1, SIVCONT epidemiological information and surveillance system , Challenges Of Animal Health Information Systems And Surveillance For Animal Diseases And Zoonoses By Food And Agriculture Organization Of The United Nations Rome, 2011

[31] Sanjay S. Chellapilla, Dairymap: A Web-Based Expert System For Dairy Herd Management , www.Ai.Uga.Edu/Iai/ Theses /Chellapilla_Sanjay.Pdf

[55] T. Rousing, M. Bonde & J. T. Sorensen; Indicators for the assessment of animal welfare in a dairy cattle herd with a cubicle housing system www.velferdsprotokoller.org, 2001

[56] A. Dagnino, J. I. Allen, M. N. Moore, Development of an expert system for the integration of biomarker responses in mussels into an animal health index 2007, Vol. 12, No. 2 ,

[57] J. Enting, R.B.M. Huirne, A knowledge documentation methodology for knowledge-based system development: an example in animal health management ,Computers and Electronics in Agriculture,Volume 22, Issues 2–3, April 1999, Pages 117–129

[58] Michele Ruta, Floriano Scioscia, Eugenio Di Sciascio, A knowledge-based framework enabling decision support in RFID solutions for healthcare International Journal of Recent Trends in Engineering, vol. 1, no. 4, pp. 68–71,2009

Implementation of Secure Cloud data Storage –Data Transaction by Using an Orthogonal Hand Shaking Authentication Mechanism (OHSAM)

Dr.K.Subramanian [#1] and M.Mohamed Sirajudeen[#2(*)]

[#1] Department of Computer Science, H.H Raja College, Pudukottai, Tamil Nadu, India. {subjjcit@gmail.com}
[#2] (*) Corresponding Author, Department of Computer Science, J.J College of Arts and Science, Pudukottai, Tamil Nadu, India. {mdsirajudeen1@gmail.com}

ABSTRACT: *In general, the Cloud computing utilization becomes unavoidable in each and every data communication as well as the service sharing center with various applications. Based on the requirements, the clients are performing service selection (such as infrastructure, software or platforms) towards to fulfil their needs in optimized manner. Whenever the data is going to be stored in a third party network, it automatically brings a question mark for secure access along with its storage infrastructure. The security for data transaction between cloud service providers and cloud clients is forever carry with the help of crypto graphic algorithms either it may be an symmetric or asymmetric key generation mechanism with certain limitations . In this research paper implement a new approach for ensure the secure data transaction by using Orthogonal Handshaking Authentication Mechanism under cloud along with a proposed storage authentication protocol. It creates a roadmap for data retrieval progress of authenticated cloud users among the service access in cloud computing. Data on Cloud storage will get an encrypted format by using a symmetric key helps to maintain its security and authentication will provide the cloud data utilization with appropriate cloud users.*

Key words: *Security, Orthogonal, Key, Data and Authentication.*

I. INTRODUCTION

In most of the circumstances, the data always resides in the cloud data servers (CDS). It secures with the help of cryptographic algorithms and authentication mechanism for sharing among different users through communication channels. The functional architecture for each and every security algorithm is based on its key (K) management with different authentication mechanism [1]. If the cloud data storage is combined with symmetric key crypto algorithms maintain, the same key for both encryptions (E_K) and decryptions (D_K) in order to maintain its secrecy over the communication channel. In contrast, the asymmetric key algorithm maintains a separate encryption (E_K) and decryption key (D_K) as well as to make it any one of the key as public. In most of the cases these secure mechanisms are not providing any significant impact on its secure access among the cloud data users [2]. In spite of , all the existing security algorithms (Julius Ceaser Cipher, Transposition Cipher, RSA, DES, MD5), the RSA (Rivest, Shamir, Aldimer) is used in many occurrences regarding to ensure the security of cloud data storage in effective and efficient manner [3]. Security for the data storage as well as the utilization of secure cloud storage is a challenging task either in the public /private cloud environments for the reason of the service consumer/clients attempt themselves to do the data transactions.

In private cloud, the service providers are always ensure themselves a secure data storage management or service allocation for the clients by using any kind of authentication mechanism. Anyhow, there will be numerous possibilities for intruders attack on the data transaction by the third party agents/intruders over the internet [4].

The major issues for the security in existing cloud data in the cloud server is utilize the service via the cloud may cause the problem of data loss. In general, the extracted features for cloud computing components are as follows:

- ***Software as a Service (SaaS):*** Instead of dedicated software applications for an organization, everyone tries to utilize the leased software product in cloud service provisions. Growing demands in industry push the clients to move towards the software as a Service (SaaS) over the communication channel [5].

- ***Platform as a Service (PaaS):*** The cloud computing progress take account of the leased platforms for its concerns regarding to utilize the internet based services over the cloud service provision in the communication channel. The PaaS service model creates all of the conveniences required to maintain the complete web applications and services consumption [5].
- ***Infrastructure as a Service (IaaS):*** The competence makes available to the clients in the provision of cluster servers (Cloud Service Providers) , processing units ,

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

storage infrastructures , communication channel (Intranet or Internet) , and other fundamental computing resources related with build an Infrastructures residues under Infrastructure as a Service (IaaS) [5].

The general layered architecture for cloud data storage comprises two major components are: Application Interface and Access Layer. The access layer acts an interface between the security algorithms and cloud data storage. Simultaneously, the application interface creates a bridge between the data as well as the crypto-key mechanism with the security algorithms. In general, the aspect of data storage into the cloud servers or service providers follow certain storage mechanism such as sequential number generations , block allocations and the corresponding link address to indicate the succeeding memory locations [3]. Whenever the data transfer from data owner into the cloud server is always resides in the encryption format. If it is required by cloud users/clients need to confirm its authentication by using authentication mechanism. The following diagram (Figure 1), illustrate cloud memory storage architectural framework layout for 3x3 and every block is assigned by sequential number in order to retrieve the required content among the clients.

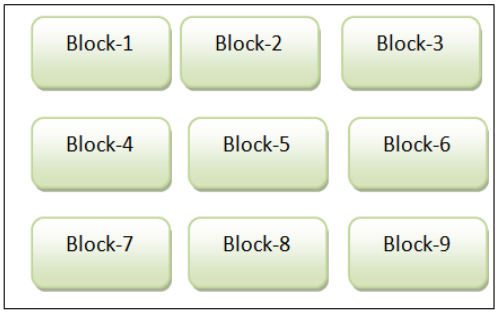


Figure 1 Cloud memory storage frame work for 3x3 CSP

Enterprise cloud data storage (Figure 2) is comprised with the components of Data Processor, Data Verifier and Pseudo-Random Number generator between the client and cloud service provider [6].

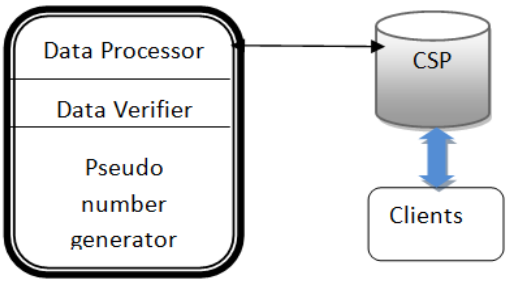


Figure 2 an enterprise Cloud Storage Architecture.

II. RELATED WORK

The cloud service users or clients send their request to the cloud service providers (CSP) in order to get service for either one of the service offered by cloud such as: IaaS, PaaS and SaaS. The service approval is required to get an authentication from the Data owners in private cloud regarding to ensure its secure data communication over the network with the help of authentication protocols [6]. The specified architectural framework (figure 3), the third party component (Authentication Protocols) is act an interface between the cloud service providers (CSP) and the data owners in order to maintain the Block Allocation Sequential Link Table (BASLT) and security control flow towards the CSP. The BALST comprising the following components:

- **Physical address:** Used to specify the location identification for cloud servers.
- **Index** **:** An indication for type of services is required by the cloud clients.
- **Data** **:** User original data in textual format.
- **Link Address** **:** The location for succeeding or proceeding data location in Cloud servers.

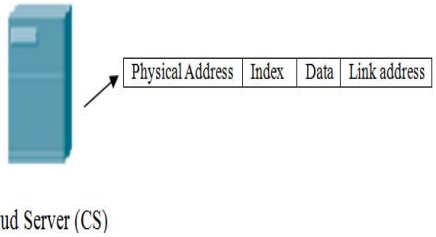


Figure 3 CSP Internal Storage Architecture with BALST

In most of the cloud storage architectural specification with crypto algorithms includes the major components such as: Storage Attached Network (SAN), Network Attached Storage (NAS) and Direct Attached Storage (DAS) [7]. All the three technologies focus on only the secure concern and not for cloud storage allocation mechanism [8]. Most of the storage allocation designed to maintain the crypto form data on the cloud with different geographical location and it's utilized by varieties of end users or cloud clients with the help of any kind of authentication mechanism [9][10]. In order to make a proper allotment of data store on cloud server is a significant factor to determine the secure cloud data transmission over the communication channel and it includes the following characteristics,

- ***Error free data on cloud storage:*** It provides the appropriate data block

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

allocation and carry on that error free data service among all types of cloud service users such as private, public and hybrid.

- ***Easiest way to localize the misbehave on cloud server:*** It powerfully modifies the cloud server even if the cloud service is not provoked in a proper manner by the client's infrastructure over the communication channel. The link address and physical location of cloud data is must be a location transparency for cloud data.
- ***Data Dependency:*** This attribute helps to maintain a healthier data link among the existing cloud data in the cloud service provider.

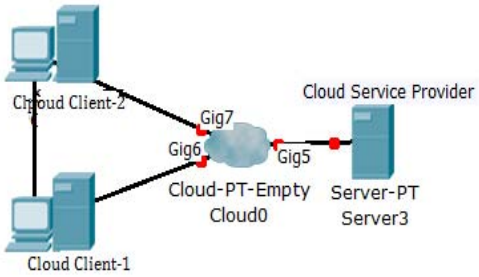


Figure 4 Data Access via Clients from Cloud Service Provider

In addition with the above researchers quote, the following design goals are play an important role to establish secure cloud data storage on CSP [9][10].

- ***Encryption:*** The Encryption process is used to secure the cloud data transmit over the open communication channel. The encryption is carried with any one of the symmetric or asymmetric crypto algorithms.
- ***Updations on web server:*** In this component relate with the web server make proper updations brings an uninterrupted cloud service over the network.
- ***Decryption:*** The Decryption is the reverse process foe the data encryption work with the help of key. The fetching key is done by using a procedure for Data Verification.
- ***Data Verification:*** This procedure helps to perform data store or retrieval on cloud storage by using authentication or data verification mechanism.

In general, most of the researchers contribute their work regarding to secure the data storage on cloud server or cloud service provider by using different cryptographic algorithms (Symmetric Key and Asymmetric Key). The principle of these cloud data storage mechanism focus encrypted data (E_D) along with its key (E_K) [7][8] is stored **on same server**. Whenever the data is required by another client or cloud user, it **may access via authentication** [9]. The following figure 5 illustrates, the way of key authentication with the server.

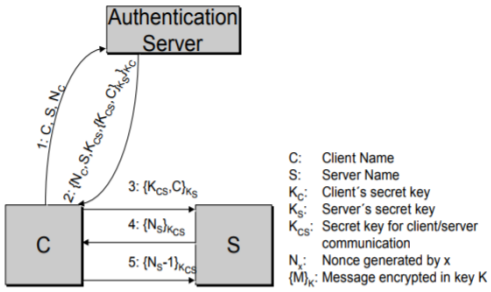


Figure 5 Key Authentications

Few research articles propose, a **one-time session key** is generated by a Key Distribution mechanism (KDM) for use in symmetric key encryption of a single session between two parties. By using the one-time session keys from the KDM, a user is freed from having to establish *a priori* its own shared key for each and every network entity with whom it wishes to communicate. Instead, a user need only have one shared secret key for communicating with the KDM, and will receive one-time session keys from the KDC for all of its communication with other network entities [10].

Initially, the entire message or data is divided into a number of discrete blocks (B). Instead of authenticating each and every individual data or message packets over the communication channel, this approach is used to perform a group of certain bit length messages or data together in to a block (B).

At the moment, every block using certain authentication procedure in order to provide a cloud service as secure manner through the network (Figure 6). The outline of authentication algorithm by using block structure is given below [14][15][16],

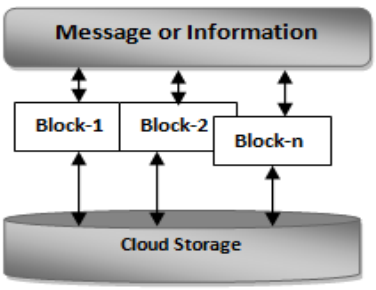


Figure 6 Existing Block storage in Cloud server

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

A. Auditing Algorithm Shell (AAS) :

Step 1 : Start with the initial message or information.
Step 2 : Fake stream generation.
Step 3 : Merge the fake stream with original message.
Step 4 : Store in to the cloud server.
Step 5 : Just reverse it to get original text by using Decryption.

B. Matrix Encryption Algorithm(MEA):

Step 1 : Count the No. of character (N) in the plain text without space.
Step 2 : Convert the plain text into equivalent ASCII code. And form a square matrix (S X S >=N).
Step 3 : Apply the converted ASCII code value from left to right in the matrix. Divide matrix into three part namely upper, diagonal and lower matrix.
Step 4 : Read the value from right to left in each matrix.
Step 5 : Each matrix uses three different key K=K1, K2, K3 for encryption. Do the encryption.
Step 6 : Apply the encrypted value into the matrix in the same order of upper, diagonal and lower.
Step 7 : Read the message b y column by column. Here the order in the columns read from the matrix is the key K4.
Step 8 : Convert the ASCII code into character value.

C. Procedure Block_Authetication Algorithm(PBAA) :

Step 1: Divide the original data or message (M) into fixed size of Packets (S).
Step 2: Group the specified number of packets into a Block (B) as a fixed size or variable –length size.
Step 3: Assign the authentication code from any one of cryptographic algorithms.
Step 4: Do the authentication process by using the key (K) at the time of receive any request from cloud clients or users.
Step 5: If the Authentication process is success, then provide the service, otherwise to terminate from the request.

Limitations :

The followings major limitations are identified from existing works related with secure the cloud data storage and listed as follows,

- The encrypted text and key is stored in same cloud server [11].
- There is no standardized mechanism for key storage in existing cloud servers [13].
- Every cloud storage concentrate on cloud crypto format data storage only, not focus on the storage mechanism [14].
- The way of cloud encrypted data and its appropriate will always stored on consecutive locations over the communication channel [13].
- The time consumption for authentication mechanism will take longer time in order to provide required data to the cloud users [15][16].

III. IMPLEMENTATION

The cloud data storage and secure mechanism for cloud service utilization is carrying with the help of Orthogonal Handshaking Authentication Mechanism (OHSAM) (figure 5). In general, the term “orthoganality” is generally referred to as “Perpendicular with each other” working principles on cloud data storage set.

Whenever the cloud services (SaaS, PaaS or IaaS) is required by the cloud clients or user is initiated with the steps for registration on the trusted third party network. The registration process is illustrated with the following diagram (figure 6). The new cloud client or user before to initiate the service utilization, must be register on the relevant cloud service provider and get an ID for authentication purpose through the random ID creation module.
The proposed work includes the following working Steps:

Step 1: Registration in the cloud service providers (CSP).
Step 2: Encryption or Decryption Mechanism
Step 3: Key Distribution
Step 4: Authentication – Handshaking
Step 5: Retrieval or Response

Most of the cloud data encryption security algorithm’s key (E_K) is stored in the same cloud server (Figure 7, 8). But, in the proposed mechanism encrypted data is stored into one cloud server and its relevant key will store in another cloud server based on the orthogonal selection in the clustered cloud server infrastructure over the communication channel. The content or service (SaaS, PaaS or IaaS) availability during the retrieval process is carried out with the help of handshaking –authentication mechanism. In order to secure and create a cloud data storage block allocation for encrypted user’s data is linked with data store in the cloud service provider by symmetric key generation.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

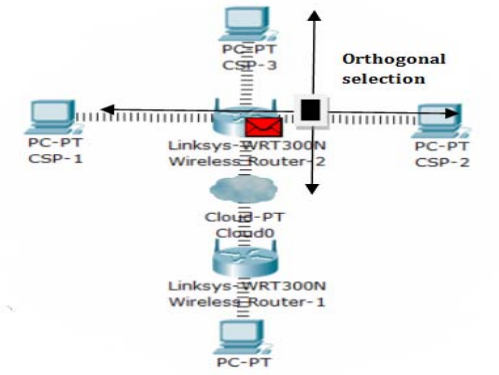


Figure 5 Cloud Storage by using Orthogonal Principle

The data or information in textual format initiated by cloud users or clients interacts with the modules of encryption or decryption process along with registering in the cloud service providers (CSP).

The registration process provides a sequential number for verification regarding the cloud clients over the internet. At the same moment, the encryption or decryption process is carried out with the help of crypto graphic algorithms such as symmetric or asymmetric mechanism.

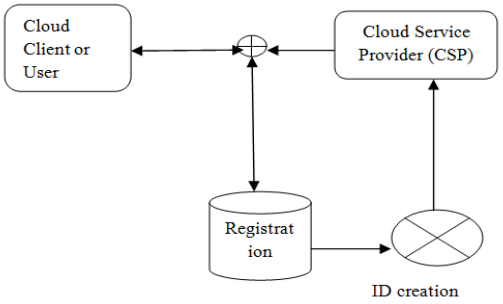


Figure 6 Registration in the cloud service providers (CSP)

The encrypted key for original data (For example , if consider Cloud client-1 send its own data to the cloud service provider based on the orthogonal selection, the relevant key will store another CSP) will occupy in one CSP and its data will available in another CSP based on the orthogonal selection mechanism. One of the fields in original data segment includes the link address for its key distribution cloud server over the network.

The authentication for registered users will ensure with the help of orthogonal handshaking mechanism and it will carry out as a continuation of this research work in consecutive publication. The source code for data store relevant with secure cloud data storage in order to generate key is specified as follows,

```
Function OHSAP _plaintext ()
 {
  m = InputString.length;
  // Skip blanks, and comments within { } and additional blanks afterwards.
  Ch = InputString.charAt (InputIndex);
  While ((ch == ' ' || ch == '\t' || ch == '\n' || ch == '\f' || ch == '\r') && InputIndex < m)
  {    Ch = InputString.charAt (++InputIndex);
    If (ch == '\n' || ch == '\f' || ch == '\r') lineno++;
  }
  While (ch == '{')
 {
    Ch = InputString.charAt (++InputIndex);
    While (ch! = '}' && InputIndex < m) ch = InputString.charAt (++InputIndex);
    If (ch == '}')
     {
     Ch = InputString.charAt (++InputIndex);
     While ((ch == ' ' || ch == '\t' || ch == '\n' || ch == '\f' || ch == '\r') && InputIndex < m)
       {
          Ch = InputString.charAt (++InputIndex);
          If (ch == '\n' || ch == '\f' || ch == '\r') lineno++;
       }
     }
   }
```

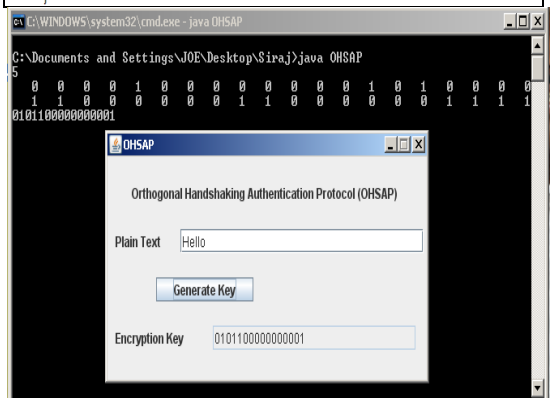


Figure 7 Source code for Orthogonal Key generation

The basic principle of the proposed architectural framework mainly focuses on the storage of encrypt and decrypt message in private cloud. There are different mechanism of secure data transmission is proposed by different cryptographic algorithms for example Auditing Algorithm Shell (AAS), Matrix Encryption Algorithm (MEA) and procedure Block Authentication Algorithm(PBAA) .

The cloud client's registration ensures the secure cloud service utilization over the communication channel by using authentication process. Without registering the cloud service access brings in secure data access and data store on the cloud servers. The given chart analysis (figure 9) depicts the analysis of performance among the existing algorithms. The relevant procedure in order to perform the authentication /handshaking on secure cloud data storage access mechanism by using orthogonal handshaking mechanism continues as a future work. Based on the following factors, the performance evaluation will be listed as follows with the implemented OHASM (Table 2) and the figure 10.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

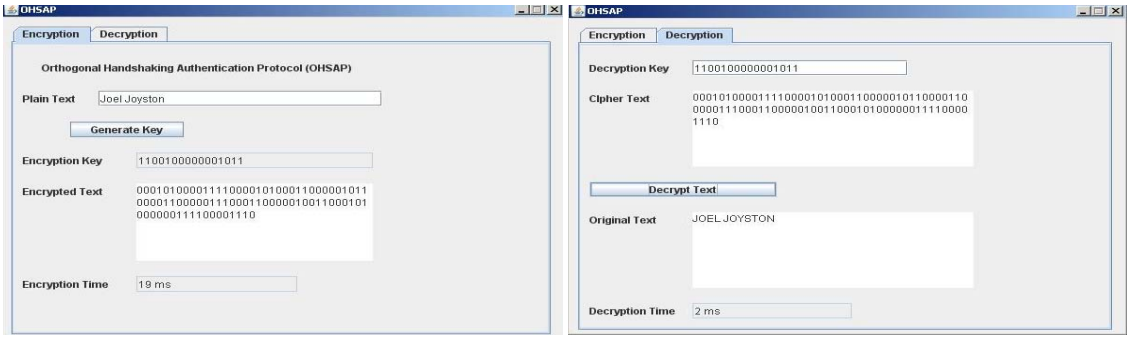


Figure 8 Encryption and Decryption by using OHSAM

Table 2. Performance comparison

Evaluation Factors	AAS	MEA	PBAA	OHASM
Storage Space(KB)	10.2	15.0	13.3	7.5
Retrieval time(mS)	72	56	81	34
Key generation Speed(mS)	8.2	4.5	4.1	2.0
Encryption Speed (mS)	36	23	19	16
Decryption Speed(mS)	12	10.5	8.4	2.0
Authentication (%)	56	45	67	72.3

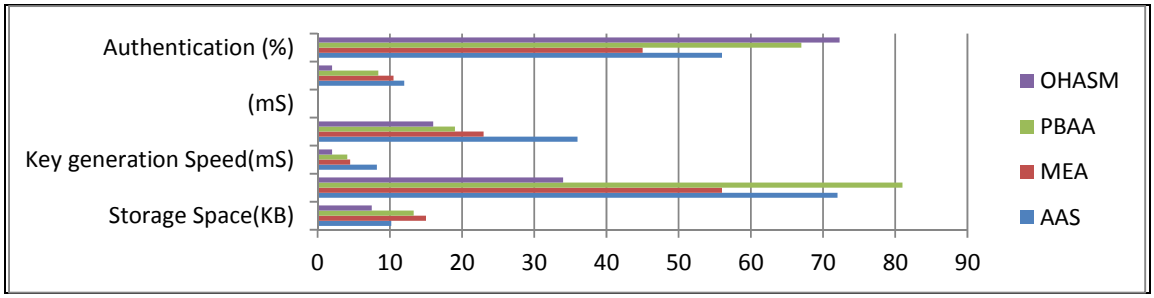


Figure 9. Comparison with Existing mechanism

IV.CONCLUSION AND FUTURE WORK

In general, most of the cloud computing secure data transaction or cloud data storage is always carried out by using a standalone crypto graphic algorithm either in symmetric or asymmetric mechanism. It never focuses on its internal storage infrastructure reading to ensure its secure cloud data access over the communication channel. In this research work and its previous relevant publications are concentrating to eliminate such drawbacks with a novelty approach on secure cloud data storage management by using “Orthogonal Handshaking Authentication mechanism”. In reality, the practical implementation of this research work includes different modules and herewith to show the relevancy of empirical analysis shows except the key distribution and authentication mechanism.

REFERENCES

[1] R.Vaishali, M.Menaka"Attribute based Encryption and Key Distribution for Secure Storage in Clouds ", © 2014 IJEDR | Conference Proceeding (NCISECT 2015) ISSN: 2321-9939.

[2]. Prof. Dr.K.Subramanian, M. Mohamed Sirajudeen, “Enhancement of the Private Cloud Data Transaction by using an Orthogonal

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Handshaking Authentication Protocol (OHSAP)", International Journal of Computer Applications (0975 – 8887) Volume 96 – No.23, June 2014.

[3] Guigang Zhang, Chao Li, Yong Zhang, Chunxiao Xing. Research on Cloud Storage Model based on Mass Information Processing. Beijing: Computer Research and Development, vol. 49(suppl), pp.32-36, 2012.
[4]. National Institute of Science and Technology. "The NIST Definition of Cloud Computing", Page.7, Retrieved July 24, 2011.
[5] Ngongang Guy Mollet, "Cloud Computing Security" Thesis, p. 34 + 2 appendices Published April 11, 2011.
[6]. EMC, and EMC Education Services. Information Storage and Management: Storing, Managing, and Protecting Digital Information. Libre Digital, 2010.
[7]. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., &Zaharia, M. (2010). A view of cloud computing. Communications of the ACM,53(4), 50-58.
[8]. Chunhua, ZHOU Ke WANG Hua LI. "Cloud Storage Technology and Its Application" ZTE Communications 4 (2010): 013.
[9].Wu, Jiyi, et al. "Recent Advances in Cloud Storage." Proceedings of the Third International Symposium on Computer Science and Computational Technology (ISCSCT'10). 2010.
[10] De Li, Report of the Cloud Computing Technology Development (2012). Beijing: Science Press, 2012.
[11] Liang Liu. Research on Cloud Storage based on Virtualization and Distributed Technology. Anhui: Computer Knowledge and Technology, vol. 8, pp. 2641-2642.2012.
[12] Gantz John, Reinsel David. Digital Universe Study -"Extracting Value from Chaos". 2011. http://idcdocserv.com/1142.
[13] Jiyi Wu, Lingdi Ping, Xiaoping Ge, Ya Wang, Jianqing Fu. Cloud Storage as the Infrastructure of Cloud Computing. International Conference on Intelligent Computing and Cognitive Informatics, pp. 380-383. 2010.
[14]. An Efficient Framework for Information Security in Cloud Computing Using Auditing Algorithm Shell (AAS), M. Omer Mushtaq, Furrakh Shahzad, M. Owais Tariq, Mahina Riaz,, Bushra Majeed, "International Journal of Computer Science and Information Security (IJCSIS, Volume 4,Novembe 7, 2016), ISSN:1947-5500.
[15]. "Encryption algorithm for Data Security and privacy cloud storage", Manish a R. Shinde and Rahul D. Taur, "American Journal of Computer Science and Engineering Survey, ISSN 2349 – 7238.
[16]. Access Control for Cloud Computing Through Secure OTP Logging as Services Priyanka Patel Nirmal Gaud, "International Journal of Computer Applications" (0975 – 8887), Volume 141 – No.14, May 2016.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Analysis of Key Transmission Issues in Optical Wireless Communication for Indoor and Outdoor Applications

Muhammad Waqas Khan[1], Mehr-e-Munir[2] ,Umar Farooq[3], Amjad Amer[4]
Departmentof Electrical Engineering , Iqra National University ,Peshawar Pakistan[1,2,3,4]
Email: m.waqas@inu.edu.pk[1], mehre.munir@cusit.edu.pk[2], umer.farooq@inu.edu.pk[3] , amjadaamer008@gmail.com[4]

Abstract- **Optical Wireless Communication (OWC) has attracted the researchers as an alternative broadband technology for wireless communication. In OWC optical beams are used to transport data through atmosphere or even vacuum. We have proposed an OWC model and analyze the transmission performance of OW channel for indoor/ outdoor application. The performance has been judged on the basis of key parameters like BER and OSNR. A theoretical model has also been presented and validated by the simulation results. The proposed OWC channel was simulated in Optisystem which is a powerful tool of Optical communication System**

Keywords- OOW Model, Laser, OWC Model

I. INTRODUCTION

The OWC provides optical bandwidth connections using lasers. 2is an optical wireless communication technique in which light spread in the space, free space, i.e. space, air, transmission of data wirelessly for computer and telecommunication networking. At present optical wireless communication has a capacity of transmitting around 2.5 Gega byte /s, voice, video and other forms of data transmission through space permits optical connectivity without the need for optic fibre cable or getting spectrum licenses. Optical wireless communication operates b/w the 780 to 1600 nm bands by using converters i.e. electrical to optical and Optical to electrical. OWC needs light, which can be focused by using lasers or LEDs. Using the lasers is very similar to using fibre optic cables for transmission difference is the medium .

OWC connectivity doesn't require any optic fibre cable, or security license for the RF (radio frequency) solution. Digging is not popular in metropolitan cities and also prohibited by local administration. Also cost may be increase for digging specially for river and railway tracks.. So OWC can provide cost effective connectivity. OWC provides low BER, high SNR, low cost, power efficient, easy installation and maintenance

Personal communication system (PCS) is a major area of application of OW. The progress in optic technology has enabled the mass with optical components that are fast and available at low costs are suitable for short ranged OW. In the 1990 Optical Wireless becomes an emerged technology for data communication transfer for Personal Computers as the IDA (Infrared data association)[2] has developed the required protocols which enabled the standardization and commercialization the OW ports which are very popular and are now found on mobile phones and PCs.

Almost 30 years ago OWC (Optical wireless communications) a new broadband technology for wireless transmission to use as was suggested. [1]. OW has a very simple basic concept: utilization of Laser to carry data through vacuum and free spacev. This means that the architecture of OW link is very similar to that of point to point fiber optic links, except for the fact that optical fibres are not used as a transmission medium. It is also like to Radio Frequency links but light waves are used instead of radio waves and an antenna with an optical trans-receiver for free space media. In spite of the apparent resemblance b/w the two(RF links and OW). OW has many better characteristics compared to RF. Optical components are very power and cost efficient in comparison to RF components. Also they don't undergo interference or multipath fading and operate under strict safety protocols. This does not at all mean that OW can replace RF completely. RF technology is of no match when it comes to area coverage and user mobility compared to OW which is quite limited. But because their photo-electric conversion mechanism impacts light noise sources, incoherent OW receivers present lower sensitivity compared to RF receivers.

Artificial magnetic conductor is used for miniaturization of an antenna and it reduces an antenna size but results in lower gain [1]. Complementary split ring resonators are used for miniaturization but size reduction is only 10% [2]. Size reduction of 21% is presented by using Koch fractal shape but after few iterations gain starts to decrease [3]. In short circuited technique, patch is shorted to the ground plane and this technique reduces size up to great extent but gain also decreases [4]. .Another main problem of the smaller size antenna is its narrow impedance bandwidth and lower gain

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[5]. In multiband response of the microstrip patch antenna is reported but gain is very smaller for most of bands [6]. Miniaturization of microstrip patch antenna with multiband performance is presented but impedance bandwidth is very narrow for all the desired bands [7]. Meta materials are used as ground plane to reduce antenna size [8].

With the high permittivity substrates size of antenna can be reduced upto great extent but this technique reduces radiation efficiency of antenna and impedance bandwidth of antenna also reduces [9]. Magnetic substrates can be used for this purpose but pure magnetic substrates are unlikely to obtain [10].

Therefore in the present study we used technique for miniaturization of double patch antenna with a good gain and satisfactory impedance bandwidth for each band. We used combination of U-Shape and L-Shape slots on the ground plane and H-Shape slot on the fractal patch. We also employed shorting pin between fractal patch and ground plane. By the combination of all these proposed techniques size of antenna reduced upto 69.29% and it produced multiband response in the frequency range of 1-8GHz and impedance bandwidth and gain are satisfactory for each band. We can adjust different bands by changing position of shorting pin.

II. Network architecture of the proposed OWC model

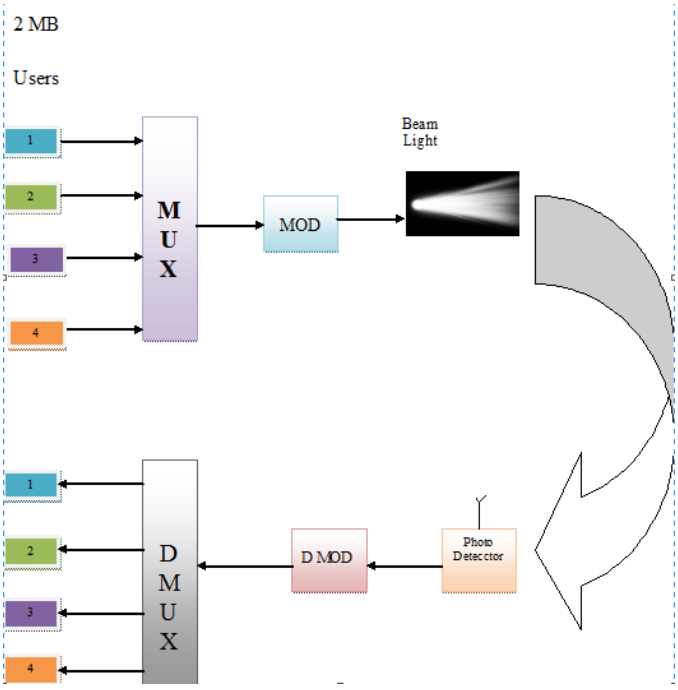


Fig1. Shows Network Architecture of Proposed OWC Model

In figure 3 for user are using OWC. they are mux together at MUX showing above. Then there combined data is modulated by using a modulator. The modulating data is then transmitted through a light beam / laser transmitter. The light beam coming out from transmitter coming through the straight beam through some distance. Some of data however scattered by the way .

At receiver end photo detector receive the OWC signal and collect the information. This signal then pass through the demodulator which demodulate the data of four users. The signal then passed through the Demux where data of four users is separated and four user the data and required information.

The laser emit the light which is ristricted to a narrow cone. But when Laser propagates the beam out ward it fans out slowly or it diverge. For an electromagnetic beam, beam divergence is the angular measure of the increase in the radius with distance from the optical aperture as the beam emerges.

The laser beam divergence can be calculate if the beam diameter d_1 and d_2 at two separate distances are known. Let $z_1$and z_2 are the distances along the laser axis, from the end of the laser to points "1" and "2".

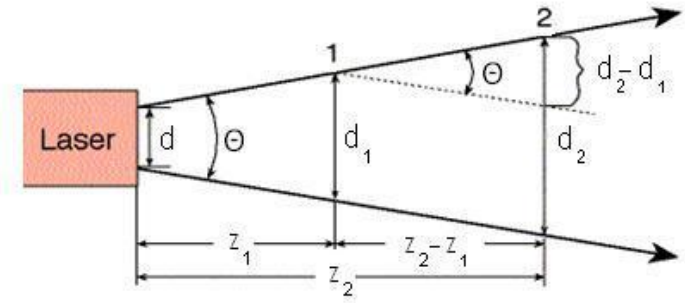


Fig 2 . laser beam divergence

The divergence angle is taken as the full angle of opening of the beam. Then,

$$\Theta = \frac{d_2 - d_1}{z_2 - z_1}$$

Half of the divergence angle can be calculated as

$$\Theta = \frac{w_2 - w_1}{z_2 - z_1}$$

Where w_1 and w_2 are the radii of the beam at z_1 and z_2.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Like all other electromagnetic beams, the lasers beam are subject to divergence, which is measured in mill radians (mrad) or degrees. Lower divergence beam is preferable for many applications.

Atmospheric attenuation effects OWC, limiting the reliability and performance. Scintillation, rain, fog and haze cause atmospheric attenuation which has a harmful effect on OWC. Mie scattering contributes the most in the scattering on the laser beam. The aerosol that existed in the atmosphere due to fog and haze causes this scattering and visibility can be used to calculate it whose value can go up to 100s of decibels in thick fog which reduces visibility lower than 50 meters and can effect the performance of OWC. Rain (non-selective) scattering does not affect considerable attenuation in wireless IR links as it does not depend on wavelength, it affects mainly on radio system and microwave that transmit energy at longer wavelengths. Scintillation and laser beam spreading and wander are the three main effects on turbulence. Change in refractive index of air causes scintillation which causes the light intensity to be non-uniform. The OWC components such as divergence of the beam, diameter of the aperture of both the transmitter and receiver are the responsible values for geometric attenuation. The sum of geometric and atmospheric attenuation is the total attenuation.

To design OWC system the effect of geometric loss and atmospheric attenuation is small, to reduce the total attenuation.

III. Results and Discussion

Visibility depend on the climatic condition, as a quantity measured by a human observer which is defined as (Kruse model) the distance where an optical signal is reduced from 550nm to 0.02 of its original. But there are many objective and physical factors effect this estimation. The essential meteorological quantity, namely the transparency of the atmosphere, can be measured objectively and it is called the Runway Visual Range (RVR) or the meteorological optical range. Some values of atmospheric attenuation due to scattering based on visibility are presented in Table 1.

Visibility S (Line of Sight)(km)	λ=800 nm (db/km)	λ=2500 nm (db/km)
0.5	32.5	30.8
0.7	23	21
0.9	18	16
1.1	14.5	12.5
1.3	12	10
1.5	10	8.33
Source		

The visibility depends on the degree of coherence of the source, on the distance between the paths as well as on the location of the detector with regard to the source. The coherence between different beams reaching the detector depends on the crossed media. For an example, the diffusing medium can decrease the coherence. For links referred to as "in direct sight" links, coherent sources can be used, given that parasitic reflections do not interfere with the principal beam, inducing modulations of the detected signal.

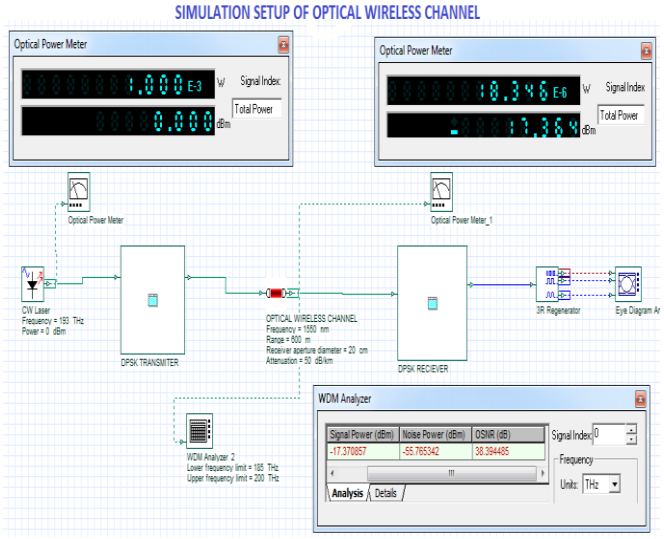


Fig 3.Simulation Setup for Optical Wireless Channel

The Network Architecture of the OWC link show in figure 3.1 is simulated in Optisys Software as shown in Fig.3.4 .The transmitter consist of a DPSK transmitter. A CW Laser modulates the basic information at DPSK Transmitter. The Launch power is measured by Power meter, Then it is passed through a optical wireless channel working as same as a beam light. The Parameters of OWC channel is shown in Fig. 3.5.In the receiving end a DPSK Receiver is used to receive the signal. Which regenerate the signal. We have analyzed the transmission performance of the proposed OWC channel using parameters of BER, SNR, Range, Beam Divergence and Receiver Aperture. Next section will cover the simulation results of various parameters.

Different results taken from simulation as shown in graphical interpretation we compare different parameters to check the link performance. As given below.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

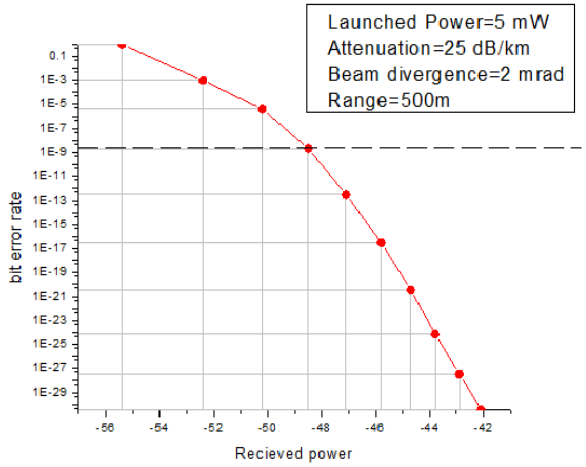


Fig 4. Received Power

In figure 4 we can see that as the bit error rate is low received power is high. As u can see in graph bit error rate at -9 is acceptable in OWC communication. In this graph we use some constant parameters i-e launched power=5mW, Attenuation=25 dB/km, Beam divergence=2mrad,range=500 m and we vary receiver aperture.

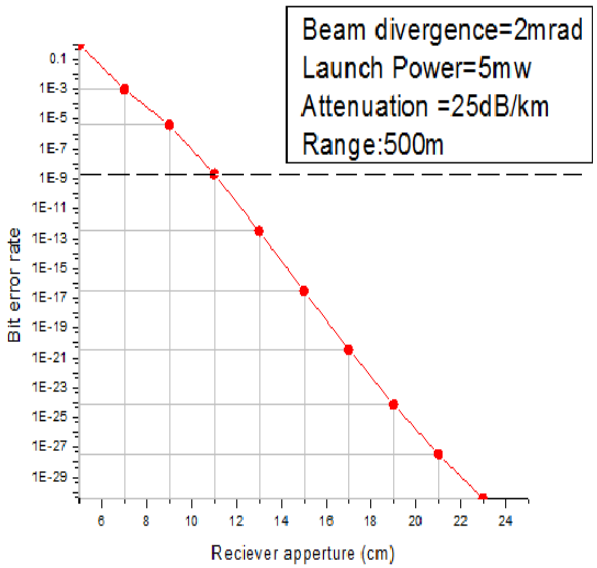


Fig 5. Received Apperture

In fig 5 we can see relation of bit error rate and receiver aperture if we increases receiver aperture bit error rate is low.

In this graph we vary receiver aperture and beam divergence, launch power, attenuation and range is constant.

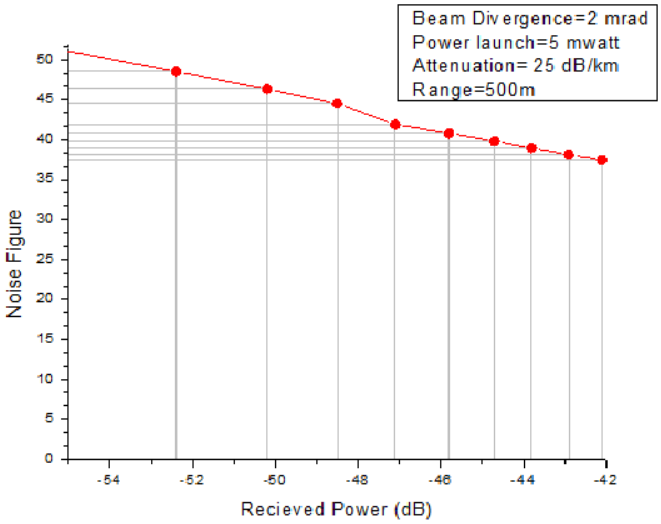


Fig 6. Beam Divergence

In figure 6 we see the effect of noise figure on receiver power there is not much effect on it as u can see in figure. As received power is increases noise figure decreases as not too much effect on it.

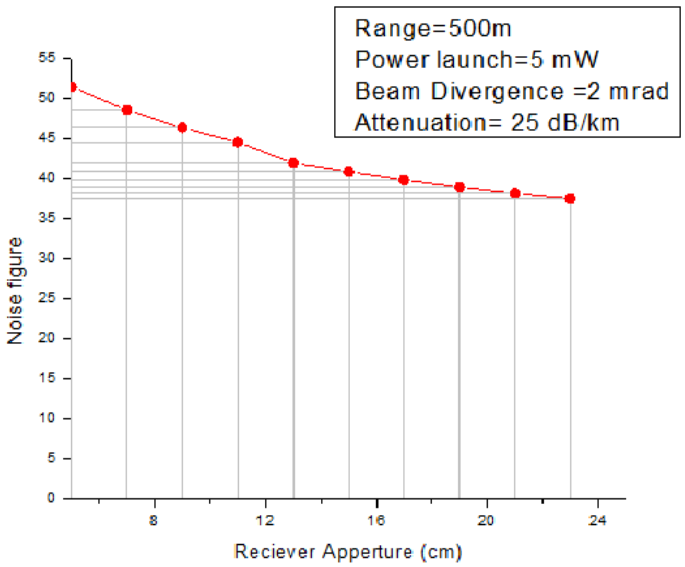


Fig 7. Noise Figure

In figure 7 we can see the effect of noise figure on receiver aperture it is not much effect on it. As we increase receiver

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

aperture the noise figure is low. In this case we vary receiver aperture.

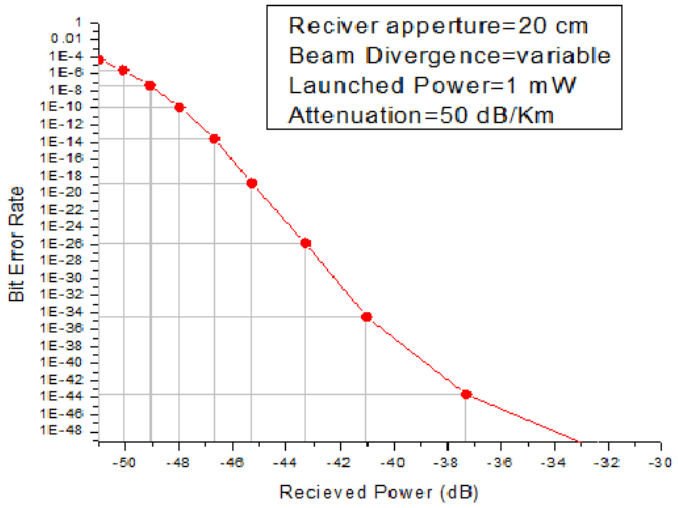


Fig 8. Bit Error Range

In this figure 8 we see the effect of bit error rate on received power u can see in figure if we increased received power bit error rate is low and performance is good in this case we vary beam divergence and received aperture, launched power and attenuation is constant.

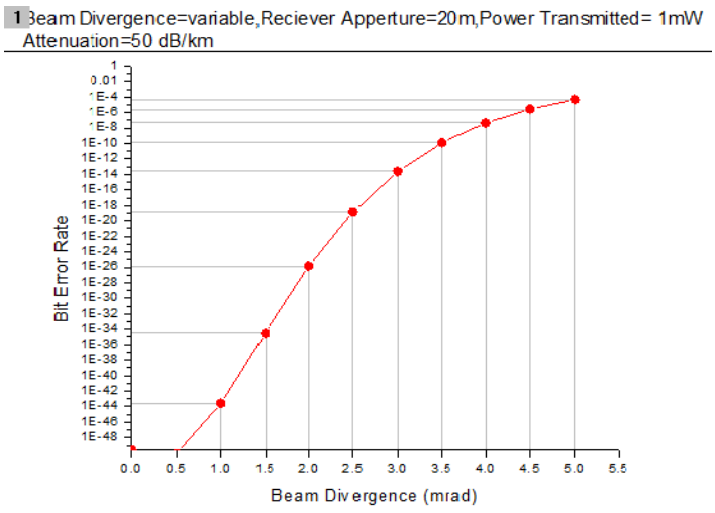


Fig 9. Beam Divergence on Bit Error Rate

In figure 9 we can see the effect of beam divergence on BER as we can see as beam divergence is increase bit error rate also increased. We vary beam divergence to see the effect of BER.

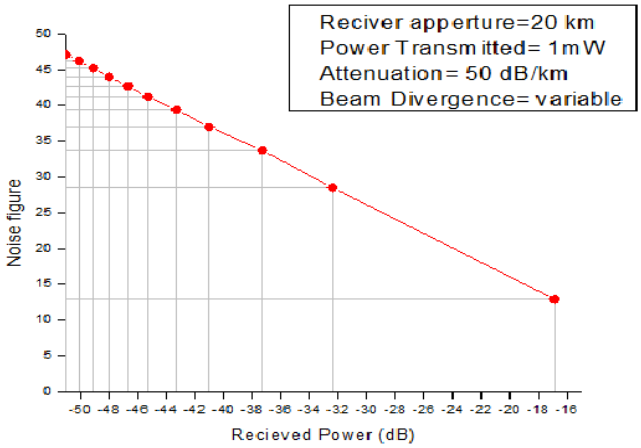


Fig 10. Effect on Noise Figure in Received Power

In figure 10 we can see the effect on noise figure on received power. As received power increased noise figure decrease in this case beam divergence is variable and receiver aperture, transmitted power, attenuation is constant.

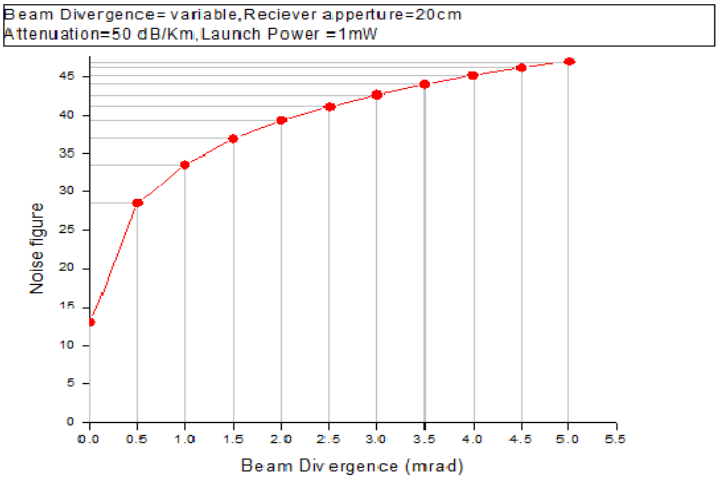


Fig 11. Comprise between Noise Figure and Beam divergence

In this figure 11 we comprise between noise figure and beam divergence. As we increase beam divergence noise figure also increased in this case we vary beam divergence to see effect on noise figure under some constant parameters i-e receiver aperture, attenuation, launch power.

IV. CONCLUSION

We have done analysis on Optical wireless communication parameters and we conclude that Optical wireless communication is considered a promising technology for the long range distance communication we proposed and analytically demonstrate a transmission link based on Optical

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

wireless communication. An analytical model has been presented and validated by simulation. .

In our experiments we have selected the DPSK as transmission modulators. DPSK is investigated on the basis of the key issues related to range, attenuation and data rate has been addressed.

Bit error rate (BER) and optical signal to noise ratio shows the good transmission performance with the 500m range, 1Gbps data rate and up to 150 dB/km of attenuation in analyzed optical wireless communication.

We have simulate OWC link in optisystem and we have done some analysis on some parameter like Beam divergence, receiver aperture, transmitter aperture, BER, noise figure by keeping some parameters constant and some variable and also we show our analyses on graphical interpretation and finally we implement practically a transmitter which transmit a audio data signal through free space optics and receiver to receive that signal.

.

REFERENCES

1. Gfeller F, Bapst U: Wireless in-house data communication via diffuse infrared radiation. Proc IEEE 1979, 67(11):1474-1486.
2. Infrared Data Association (IrDA)
3. Visible Light Communications Consortium
4. Chan VWS: Optical satellite networks. IEEE/OSA J Lightw Technol 2003, 21(11):2811-2827. 10.1109/JLT.2003.819534
5. IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. 1997.
6. IEEE 802.15 Working Group for WPAN]
7. IrDA: Advanced Infrared (AIr) MAC Draft Protocol Specification--Version 1.0, Infrared Data Association. 1999.
8. IrDA: Advanced Infrared Logical Link Control (AIrLC) Specification--Version 0.1, Infrared Data Association. 1999.
9. IrDA: Advanced Infrared Physical Layer Specification (AIr-PHY) --Version 1.0, Infrared Data Association. 1998.
10. IrDA: Point and Shoot Profile--Version 1.1, Infrared Data Association. 2000.
11. Kahn J, Barry J: Wireless infrared communications. Proc IEEE 1997, 85(2):265-298. 10.1109/5.55422
12. Chu T-S, Gans M: High speed infrared local wireless communication. IEEE Commun Mag 1987, 25(8):4-10.
13. Heatley DJT, Wisely DR, Neild I, Cochrane P: Optical wireless: the story so far. IEEE Commun Mag 1998, 36(12):72-74. 79-82 10.1109/35.735881
14. Pang G, Kwan T, Liu H, Chan C-H: LED wireless. IEEE Ind Appl Mag 2002, 8(1):21-28. 10.1109/2943.974354
15. Hranilovic S: Wireless Optical Communication Systems. Springer, New York, NY; 2004.
16. Barry JR, Kahn JM: Link design for nondirected wireless infrared communications. OSA Appl Opt 1995, 34: 3764-3776.
17. Wang K, Nirmalathas A, Lim C, Skafidas E: 4 × 12.5 Gb/s WDM optical wireless communication system for indoor applications. IEEE/OSA J Lightw Technol 2011, 29(13):1988-1996.
18. Pakravan MR, Simova E, Kavehrad M: Holographic diffusers for indoor infrared communication systems. Int J Wirel Inf Netw 1997, 4(4):259-274. 10.1023/A:1018876326494
19. Elgala H, Mesleh R, Haas H: Indoor broadcasting via white LEDs and OFDM. IEEE Trans Consumer Electron 2009, 55(3):1127-1134.
20. Liu Q, Qiao C, Mitchell G, Stanton S: Optical wireless communication networks for first- and last-mile broadband access [Invited]. OSA J Opt Netw 2005, 4: 807-828. 10.1364/JON.4.000807
21. Chan VWS: Free-space optical communications. IEEE/OSA J Lightw Technol 2006, 24(12):4750-4762.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Sentiment analysis of Autistic Child by Emotional advisor

Muhammad Zunnurain Hussain

Senior Lecturer Bahria University Lahore Campus
engrrhusain@gmail.com

Muhammad Zulkifl Hasan

Senior Lecturer, Lahore Garrison University

engrrhasan@gmail.com

Dr. Khadeeja Anjum

Demonstrator, Biochemistry Dept. CMH Lahore Medical College Lahore

Khadeeja.anjumz@gmail.com

Dr. Arha Anwar

Registrar, Oral Medicine Dept. Rashid Latif Dental College Lahore

arhaanwar201@gmail.com

Abstract:

The deficiency or hindrance in the capacity to think passionate states is known as mind visual deficiency. This condition is seen to be the key inhibitor of social and enthusiastic insight for mentally unbalanced individuals. A mental imbalance is a range of neuro-formative conditions which influences one's social working, correspondence what's more, is frequently went with redundant practices and over the top interests. Failures coming about because of mind-visual deficiency incorporate measuring the enthusiasm of different gatherings amid discussions, withdrawal from social contact, obscurity to social signals, in distinction to individuals' conclusions and inconceivable non-verbal correspondence. The current assistive gadgets and instruments generally fill in as healing apparatuses that give a learning condition for mentally unbalanced youngsters to find out about the standards of social conduct. In any case, these instruments do not have the capacity to work in conjunction with certifiable circumstances. A thought is recommended that means to satisfy this need. We propose a compact gadget which can help extremely introverted individuals in correspondence in genuine circumstances. We trust that this versatile gadget can help to limit the hole amongst us and the universe of extreme introvertedness through helped correspondence. In this paper, we introduce one a player in this gadget, which is called Emotional Advisor to help extremely introverted youngsters in taking part in significant discussions where individuals can learn how they are feeling amid correspondence.

Keyword: autism spectrum disorders, complex genetics, copy number variation, disconnection syndrome, neuroimaging, neuropathology, diagnosis, assessment, diagnostic instruments, Risk factors, Perinatology, Mental retardation

1. Introduction

A mental imbalance Spectrum Disorder (ASD) is a complex neurodevelopmental scatter that is portrayed by impedances in social cooperation, for example, dialect aptitudes, specifically social correspondence. Not at all like the majority of us, mentally unbalanced individuals confront colossal challenges in understanding meaningful gestures and traditions; they cannot appropriately express non-verbal correspondence and non-verbal communication. These failures thwart them from understanding verbal and non-verbal correspondences, and perusing human outward appearances adequately. They could not distinguish and comprehend the feelings that they are presently encountering. Without this understanding, they will stay unaware of other individuals' goals, feelings and consequently influences their basic leadership. The absence of such critical earlier learning of the condition, they barely settle on an

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

educated choice. The result of such postures challenge for extremely introverted individuals to associate in the very mind boggling social condition. As of late, investigate on feeling acknowledgment has quickly expanded [1-4]. This marvel proposes that the capacity to distinguish and decide one's feelings can fill in as a strengthening for the field of computerized reasoning and give ascend to more quick witted, even more intense machines that comprehends the goal of clients. A keen machine with enthusiastic mindfulness can accomplish the weaknesses of mentally unbalanced individual. With that enthusiastic mindfulness, the machine is prepared to do educating and managing extremely introverted individuals on the best way to react fittingly when the individual that he or she is imparting with is communicating different feelings. Such machine has the potential to connect the correspondence abyss between the society and those determined to have extreme introvertedness. The focal point of this exploration is on the improvement of a real time feeling response consultant for mentally unbalanced kids that goes about as a counselor showing them how they can act appropriately based on how the other party is feeling amid verbal correspondence. The framework creates recommendations for the suitable reaction base on the feeling of that individual as anticipated by the framework. The proposed framework envelops two essential parts: a feeling acknowledgment module what is more, a passionate guide module. The proposed framework comprises of a couple of glasses with smaller than normal camera interface with a PDA or a portable PC. The feeling acknowledgment module working on the compact gadget perceives the facial feeling of the other party and articulates the feeling through an earpiece to the mentally unbalanced client. Other than criticism of the enthusiastic state to the client, the enthusiastic counsel module will show a suitable counsel or proposal on how the client can react as indicated by the sentiment the other party. In this paper, advancement of this enthusiastic counsel is principally introduced. The improvement of the feeling acknowledgment module can be alluded to our distributed papers in subtle elements [6, 7]. The enthusiastic consultant module is basically a fluffy administer based framework which shapes the database of various sorts of prompts, each comparing to one or a blend of feelings that can be shown by the communicator. The passionate consultant module will figure out which counsel to yield to the extremely introverted client relying upon the info caught by the passionate acknowledgment framework. The entire acknowledgment and prompting process is constant and the preparation is intuitive, i.e. the learning (database) of the framework is refreshed ceaselessly. The paper is sorted out as takes after: Section II gives a diagram of the computational system for the passionate counsel utilizing fluffy tenets. A preparatory testing is performed that intends to decide the practicality of actualizing fluffy rationale in the passionate counsel. Segment III exhibits the outcome of this preparatory testing, which traces the benchmarking process and gives an examination of the test deduction. Segment IV wraps up, exhibits the finish of the undertaking, and proposes some conceivable developments for future looks into.

2. Computational Structure

This exploration depends on a blend of fluffy rationale, learning and example acknowledgment together with a neuroscience comprehension of intellectual and visual flag exchange in crossing over the correspondence gorge between mentally unbalanced kids what's more, the world. The philosophy of fluffy sets and participation, response to-feeling relationship by fuzzification what's more, defuzzification and fluffy IF-THEN principles are talked about. The exploration is centered on

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

looking at the attainability of applying fluffy standards into the passionate consultant. Results from the trials led have demonstrated that the passionate consultant, which frames as one of the modules our proposed smart feeling framework as appeared in Fig. 1, has satisfied the fundamental criteria of portable application: speed and effectiveness. The fluffy framework is fit for ordering and summing up with a precision that outperforms different famous classifiers like Naïve Bayes. This one of a kind normal for fluffy rationale is fundamental in handling true situations in light of the fact that the world is loaded with vulnerability. In any case, debates still stay among control builds whose inclinations influence towards two-esteemed rationale and analysts who just acknowledge Bayesian rationale. By and by, fluffy rationale has been effectively fused in a significant number of the specific fields today and it has additionally been a investigate point which is widely considered in the course of the last few decades. In view of its capacity to tame vulnerability, fluffy rationale is a rationale hypothesis that suits the idea of our venture and can be embraced to form the system of our enthusiastic guide[8-9].

A. Input

The three data sources got from the outward appearance recognizer, enthusiastic indexer and expectation are sustained to the fluffy enthusiastic counselors and put away in various content records to be specific output.txt. ei.txt and predict.txt separately. Four conceivable results can be gone after every individual record. Table I demonstrates these conceivable yields that can be produced from every content records. An aggregate of 64 novel blends can be shaped in view of the diverse yields. In this examination, the goal is to test the practicality of executing fluffy rationale in the enthusiastic consultant. To accomplish this, four unique esteems are conveyed for each passionate input. It is to streamline and diminish the extent of the prototyping process. The testing is executed progressively utilizing MATLAB programs. The execution of utilizing fluffy rationale in the enthusiastic consultant is looked at and benchmarked against other mainstream classifiers. Itemized examination of the yield comes about are examined and compressed in the later segment[10--12].

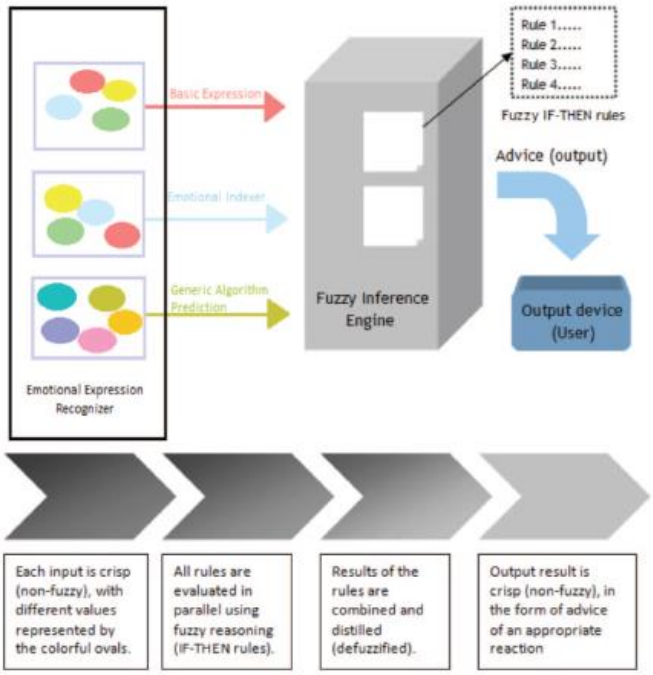


Figure 1: working model for generating advises for the autistic children which is advised by experienced advisor after collecting various factors (facial expression, emotions)

Table 1: Output predicted by using each.txtfile

Type of output/ txt.file	Type 1	Type 2	Type 3	Type 4
Output.txt	1 (neutral)	2 (happy)	3 (Sad)	4(Surprise)
Ei.txt	NeutralAverage	HappyAverage	SadAverage	SurpriseAverage
Predict.txt	1(encouraging)	2(Interesting)	3(Discouraging)	4(unsure)

B. Classes & Datasets

As featured before, there are 64 one of a kind blends that can be framed by the 3 diverse enthusiastic sources of info. To decide the quantity of classes (i.e. the quantity of conceivable advices that can be produced) and the quantity of information blend that the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

framework ought to convey, tests are led with various number of classes and informational indexes utilizing 279 a few classifiers gave by WEKA. From the given outcomes, the quantity of classes and informational indexes to be conveyed in the real test set is resolved. This test set will be utilized for the testing of the fluffy principles and utilized for the examination between all the diverse classifiers. The outcomes are classified and displayed in the later area [13-15].

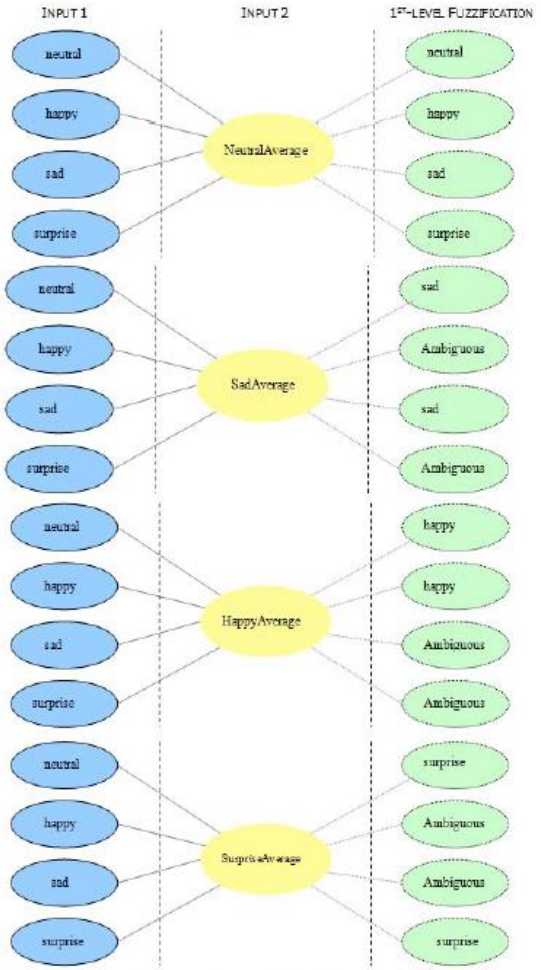


Fig.2. The first-level collections of the three emotional inputs, and the subsequent truth values of each sentiment combination

C. Fuzzification

Fuzzification is the way toward changing fresh esteems into evaluations of enrollment for fluffy sets. The enrollment work relates a review to each term characterized in the sets. Feelings are mind boggling and dubious; consequently, there is a need to relate each enthusiastic contribution to a fluffy set, to precisely pinpoint the generally speaking, predominant feeling that the client is encountering. Individual may show diverse responses towards specific feelings. Hence, this feeling to-response affiliation is not a coordinated capacity. For example, when one is cheerful, one may begin to sing. On the other hand, others may express bliss by purchasing a dessert for themselves. Both are coherent and subjected to the person's inclinations. The accompanying representations diagram the relationship between different enthusiastic information sources. Fuzzification is done in two Moderate advances. The yield of the primary accumulation is encouraged into the other three sources of info, which gives us the last resultant passionate states. Subtle elements of the fluffy thinking will be expounded in the later areas. There are two levels of totals as appeared in Fig. 2 and 3. Fig. 2 demonstrates the first level Accumulations of the three enthusiastic information sources, and the resultant truth estimations of every blend of feeling. Fig. 3 demonstrates the second-level accumulations of the three enthusiastic inputs, and the resultant truth esteems for every feeling mix [16-17].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

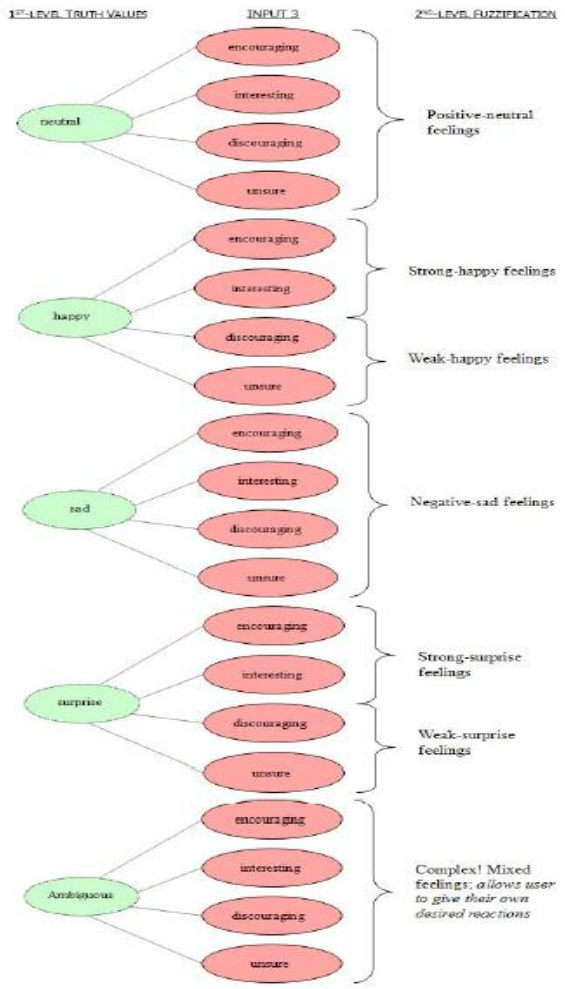


Fig. 3. The second-level collections of the three emotional inputs, and the resultant truth-values for each sentiment combination

D. Fuzzy IF then guidelines

Numerous down to earth applications utilize a moderately confined yet essential piece of fluffy rationale, which fixates on the utilization of IFTHEN rules. This part of fluffy rationale involves accumulation of ideas and techniques for taking care of a decent variety of learning which can be spoken to as an arrangement of fluffy IF-THEN principles whereby the precursors, results, or then again both, are fluffy as opposed to fresh esteems. The term 'fresh' alludes to precision of a substance. Fluffy esteems are characterized as fluffy in light of the fact that they in part have a place with at least one sets. A set, which comprise of fluffy esteems is known as fluffy sets. In embodiment, the IF-THEN standards change over contributions to yields, one fluffy set into another. Fluffy rationale permits the transformation of etymological control methodology in light of master learning, into a computerized control methodology. The principle excellence is that fluffiness of the predecessor's disposes of the requirement for a correct match with the info, subsequently giving space for equivocalness, which is inescapable in relatively every circumstance. Given the collections, eleven fluffy guidelines are composed, fit for foreseeing the general feeling condition of the client.

Table II indicates parts of the fluffy IF-THEN principles composed for the enthusiastic guide. These standards are completed and reenacted in the MATLAB program to test the attainability of ongoing age of response advices. For speedier Calculation, choice tree structure is directed. By encouraging in the example input information, the tests demonstrate that the program can create advices easily utilizing one-moment interims[18-20].

Table II: Partially designed Fuzzy IF-THEN for the advisors

Rule		Conditions		Consequences
1	IF	Basic expression IS a.Strong OR a.Average OR a.Mild	AND	
		Emotional advisor is a.Average	OR	
		Basic expression IS a.strong	AND	
		Emotional indexer IS NeutralAverage	THEN	Generate Very.a advice
2	IF	Basic expression IS NeutralMild	AND	
		Emotional indexer is a.Average	THEN	Generate Very.a advice
3	IF	Basic expression IS NeutralStrong	THEN	Generate Neutral advice
4	IF	Basic expression IS NeutralAverage AND emotional indexer is HappyAverage	OR	
		Basic expression IS HappyMild OR HappyAverage AND emotional indexer is NeutralAVerage	AND	

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

		Prediction IS Encouraging OR Interesting	THEN	Generate LittleHappy advice

3. Investigational consequences and deliberations

Since the quantity of classes utilized as a part of the testing sets contrasts, portrayals of each passionate state are balanced as per the quantity of classes that were made accessible. For each test set, the feeling to-response affiliations are refined so as the classes influenced accessible to can satisfactorily separate and well-spoken the resultant passionate state. The resultant passionate state is reality esteem got by conglomeration of the three enthusiastic data sources. Table III demonstrates the prescient execution utilizing the different mixes of classes and information examples picked. For occurrence, Naïve Bayes is appeared to have the best outcomes moderately to alternate classifiers amid the testing of informational collection A, with 5 out of 40 restorative grouped examples, yield a effective rate of 12.5%. In the wake of learning the quantity of classes and experiments that delivers the ideal outcomes, these classes and experiments are absorbed into the fluffy deduction motor and frame the real informational collection that examines the exhibitions of fluffy rationale and different classifiers in their prescient competency. 50 cases were haphazardly chosen what's more, extricated from the completely accessible blend and shaped 8 distinct arrangements of informational collections to test the prescient exhibitions of fluffy tenets and other well-known classifiers. The comes about are outlined and demonstrated as follows [21].

TABLE III
PREDICTIVE PERFORMANCE WITH VARIOUS COMBINATIONS OF CLASSES AND DATA INSTANCES

	Naïve Bayes	AdaBoost	SimpleCart	DecisionTable
Correctly Classified Instance	7, 17.5%	6, 15%	5, 12.5%	6, 15%
Incorrectly Classified Instance	33, 82.5%	34, 85%	35, 87.5%	34, 85%
Root Mean Square Error	0.2996	0.301	0.3179	0.3029
Weighted Avg for TP Rate	0.175	0.15	0.125	0.15
Weighted Avg for FP Rate	0.106	0.121	0.124	0.111

(a) Results for 40 Data, 10 Classes, 3 fold

	Naïve Bayes	AdaBoost	SimpleCart	DecisionTable
Correctly Classified Instance	4, 10%	3, 7.5%	2, 5%	5, 12.5%
Incorrectly Classified Instance	36, 90%	37, 92.5%	38, 95%	35, 87.5%
Root Mean Square Error	0.3037	0.3078	0.3141	0.3023
Weighted Avg for TP Rate	0.1	0.075	0.05	0.125
Weighted Avg for FP Rate	0.114	0.127	0.13	0.121

(b) Results for 40 Data, 10 Classes, 7 fold

	Naïve Bayes	AdaBoost	SimpleCart	DecisionTable
Correctly Classified Instance	26, 65%	10, 25%	19, 47.5%	10, 25%
Incorrectly Classified Instance	14, 35%	30, 75%	21, 52.5%	30, 75%
Root Mean Square Error	0.3125	0.3841	0.3947	0.3862
Weighted Avg for TP Rate	0.65	0.25	0.475	0.25
Weighted Avg for FP Rate	0.133	0.311	0.187	0.183

(c) Results for 40 Data, 5 Classes, 3 fold

	Naïve Bayes	AdaBoost	SimpleCart	DecisionTable
Correctly Classified Instance	31, 77.5%	11, 27.5%	21, 52.5%	16, 40%
Incorrectly Classified Instance	9, 22.5%	29, 72.5%	19. 47.5%	24, 60%
Root Mean Square Error	0.2824	0.3676	0.3612	0.3743
Weighted Avg for TP Rate	0.775	0.275	0.525	0.4
Weighted Avg for FP Rate	0.09	0.224	0.183	0.138

(d) Results for 40 Data, 5 Classes, 7 fold

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

	Naive Bayes	AdaBoost	SimpleCart	DecisionTable
Correctly Classified Instance	22, 55%	17, 42.5%	17, 42.5%	17, 42.5%
Incorrectly Classified Instance	18, 45%	23, 57.5%	23, 57.5%	23, 57.5%
Root Mean Square Error	0.3811	0.4099	0.4358	0.4161
Weighted Avg for TP Rate	0.55	0.425	0.425	0.425
Weighted Avg for FP Rate	0.293	0.44	0.38	0.396

(e) Results for 40 Data, 4 Classes, 3 fold

	Naive Bayes	AdaBoost	SimpleCart	DecisionTable
Correctly Classified Instance	16, 40%	16, 40%	18, 45%	16, 40%
Incorrectly Classified Instance	24, 60%	24, 60%	22, 55%	24, 60%
Root Mean Square Error	0.3952	0.4138	0.4301	0.4119
Weighted Avg for TP Rate	0.4	0.4	0.45	0.4
Weighted Avg for FP Rate	0.403	0.43	0.347	0.13

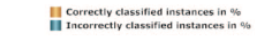
(f) Results for 40 Data, 4 Classes, 7 fold

	Naive Bayes	AdaBoost	SimpleCart	DecisionTable
Correctly Classified Instance	33, 66%	19, 38%	23, 46%	27, 54%
Incorrectly Classified Instance	17, 34%	31, 62%	27, 54%	23, 46%
Root Mean Square Error	0.3274	0.3768	0.3804	0.3632
Weighted Avg for TP Rate	0.66	0.38	0.46	0.54
Weighted Avg for FP Rate	0.156	0.395	0.301	0.069

(g) Results for 50 Data, 5 Classes, 3 fold

	Naive Bayes	AdaBoost	SimpleCart	DecisionTable
Correctly Classified Instance	35, 70%	19, 38%	28, 56%	24, 48%
Incorrectly Classified Instance	15, 30%	31, 62%	22, 44%	26, 52%
Root Mean Square Error	0.311	0.3819	0.3663	0.3699
Weighted Avg for TP Rate	0.7	0.38	0.56	0.48
Weighted Avg for FP Rate	0.148	0.389	0.168	0.12

(h) Results for 50 Data, 5 Classes, 7 fold



4. CONCLUSION

We can reason that consolidating fluffy rationale in the enthusiastic counselor is an achievable strategy that can yield a moderately high precision in foreseeing the right passionate state and producing fitting guidance for the end client. As fluffy master frameworks are displayed observationally, they have the potential to catalyze better execution and are responsive to changes and upgrades. The examination we display here altogether propels the beginning capacity of machines to gather subjective full of feeling passionate states progressively from nonverbal articulations of individuals. By utilizing fluffy rationale in building up a constant framework for the derivation of a wide scope of feeling states past the fundamental feelings, we have augmented the extent of human-PC collaboration situations in which this innovation can be coordinated. This is a critical advance towards building socially and sincerely astute machines. Keeping in mind the end goal to actualize the passionate guide in the certifiable circumstance, the principles of the fluffy standards ought to be dictated by specialists in the field of brain research, feelings what's more, extreme introvertedness range issue to accurately speak to and build up the connection be tween's different feelings, and furthermore to supply appropriate response reactions which extremely introverted kids is ready to comprehend and perform when created by the passionate counselor.

Fig. 5. Results obtained by different classifiers for Dataset 3, 4 & 5

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

5. REFERENCES

[1] Chen L. S. Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction. PhD thesis, University of Illinois at Urbana-Champaign, Dept. of Electrical Engineering, 2000.
[2] Wong J-J, Cho S-Y. Facial emotion recognition by adaptive processing of tree structure; 2006; Dijon, France. ACM press. P23-30.
[3] Jia-Jun Wong and Siu-Yeung Cho, "A Brain-inspired model for recognizing human emotional states from facial expression",Neurodynamics of Cognition and Consciousness, Edited by: Leonid I.Perlovsky and Robert Kozma, Springer-Verlag, 2007.
[4] Yok-Yen Nguwi and Siu-Yeung Cho, "Support Vector based Emergent Self-Organizing Approach for Emotional Understanding" Connection Sciences (*ISI impact factor: 0.806*), vol. 22, iss. 4, pp. 355-371, 2010.
[5] Teik-Toe Teoh, Yok-Yen Nguwi and Siu-Yeung Cho, "Towards a Portable Intelligent Facial Expression Recognizer", Intelligent Decision Technologies, vol. 3, no. 3, pp. 181-92, 2009.
[6] Siu-Yeung Cho, Teik-Toe Teoh and Yok-Yen Nguwi, "Advanced Feature Selection and Classification methods", Advances in Face Image Analysis: Techniques and Technologies, Edited by Yu-Jin Zhang (Tsing Hua University), IGI Global Publishing, 2010.
[7] Carlos Busso, Zhigang Deng , Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, Shrikanth Narayanan , Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information, Emotion Research Group, Speech Analysis and Interpretation Lab Integrated Media Systems Center, University of Southern California, Los Angeles.
[8] Ce Zhan, Wanqing Li, Philip Ogunbona, and Farzad Safaei. 2007 Real-Time Facial Feature Point Extraction, University of Wollongong. Pacific-Rim Conference on Multimedia (pp. 88-97). Germany: Springer.
[9] Faten Bellakhdhar, Kais Loukil, Mohamed ABID, computer embedded system, University of Sfax 2012. SVM classification for face recognition, Journal of intelligent computing volume 3 Number 4 December.
[10] G. U. Kharat, S. V. Dudul, 2009 Emotion Recognition from facial expression using neural networks, Human-computer systems interaction advances in intelligent and soft computimg.
[11] Hua Gu Guangda Su Cheng Du Department of Electronic Engineering, Feature Points Extraction from Faces Research Institute of Image and Graphics, Tsinghua University, Beijing, China. Image and vision computing NZ.
[12] Ira Cohen, Ashuto,sh Garg, Thomas S. Huang, Emotion Recognition from Facial Expressions using Multilevel HMM, Beckman Institute for Advanced Science and TechnologyThe University of Illinois at Urbana-Champaign.
[13] Jui-Chen Wu, Yung-Sheng Chen, and I-Cheng Chang 2007. An Automatic Approach to Facial Feature extraction for 3-D Face Modeling, IAENG International Journal of Computer Science, 33:2, IJCS_33_2_1, 24 May 2007.
[14] L. S. Chen. Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction. PhD thesis, University of Illinois at Urbana-Champaign, Dept. of Electrical Engineering, 2000.
[15] Lee, C. M. , Yildirim, S. , Bulut, M. , Kazemzadeh A. , Busso,C. , Deng, Z. , Lee, S. , Narayanan, S. S. Emotion Recognition based on Phoneme Classes. To appear in Proc. ICSLP'04, 2004.
[16] Mase K. Recognition of facial expression from optical flow. IEICE Transc. , E. 74(10):3474–3483, 0ctober 1991.
[16] P. Ekman and W. V. Friesen, Facial action coding system: Investigator's Guide. Consulting PsychologistsPress, Palo Alto, CA, 1978.
[17] Priya Metri1, Jayshree Ghorpade and Ayesha Butalia,Department of Computer Engineering, MIT-COE,"Facial Emotion Recognition Using Context Based Multimodal Approach", Int. J. Emerg. Sci. , 2(1), 171-182, March 2012 ISSN: 2222-4254 © IJES 171, Pune.
[18] Qiuxia wu, Zhiyong Wang, member, IEEE, Feiqi Deng, member, IEEE, Zheru Chi, member, IEEE, and David Dagan Feng, Fellow IEEE 2013. Realistic human action recognition with multimodal feature selection and fusion. IEEE transactions on systems, man, and cybernetics: systems, VOL. 43, NO, 4, July 2013.
[19] T. Kanade,T. Kanade, J. F. Cohn, and Y. Tian. Comprehesive database for facial expression analysis. In Proc. Of 4rd Intl Conf. Automatic Face and Gesture Rec. , pages 46–53, 2000.
[20] V Soroosh Mariooryad, Student Member, IEEE, and Carlos Busso, Member, IEEE. Ding, W. and Marchionini, G. 2013 Exploring Cross-Modality Affective Reactions for Audiovisual Emotion Recognition IEEE Transactions On Affective Computing, Vol. 4, No. 2, April-June.
[21] Yoshitomi, Y. , Sung-Ill Kim, Kawano, T. , Kilazoe, T. Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. Robot and Human Interactive Communication, 2000. RO-MAN 2000. Proceedings. 9th IEEE International Workshop on, 27-29.

Fig. 6. Results obtained by different classifiers for Dataset 6, 7 & 8

IJCSIS REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
e Engineering & Management Institute Shamli (U.P), Indiaa
Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Assoc. Prof. N. Jaisankar, VIT University, Vellore,Tamilnadu, India
Dr. Amogh Kavimandan, The Mathworks Inc., USA
Dr. Ramasamy Mariappan, Vinayaka Missions University, India
Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China
Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA
Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico
Dr. Neeraj Kumar, SMVD University, Katra (J&K), India
Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania
Dr. Junjie Peng, Shanghai University, P. R. China
Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia
Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India
Dr. Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain
Prof. Dr.C.Suresh Gnana Dhas, Anna University, India
Dr Li Fang, Nanyang Technological University, Singapore
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia
Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India
Dr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand
Dr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.)/ Dimat Raipur, India
Dr. Hayder N. Jasem, University Putra Malaysia, Malaysia
Dr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India
Dr. R. S. Karthik, C. M. S. College of Science and Commerce, India
Dr. P. Vasant, University Technology Petronas, Malaysia
Dr. Wong Kok Seng, Soongsil University, Seoul, South Korea
Dr. Praveen Ranjan Srivastava, BITS PILANI, India
Dr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong
Dr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia
Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan
Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria
Dr. Riktesh Srivastava, Skyline University, UAE
Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia
Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt
and Department of Computer science, Taif University, Saudi Arabia
Dr. Tirthankar Gayen, IIT Kharagpur, India
Dr. Huei-Ru Tseng, National Chiao Tung University, Taiwan
Prof. Ning Xu, Wuhan University of Technology, China
Dr Mohammed Salem Binwahlan, Hadhramout University of Science and Technology, Yemen
& Universiti Teknologi Malaysia, Malaysia.
Dr. Aruna Ranganath, Bhoj Reddy Engineering College for Women, India
Dr. Hafeezullah Amin, Institute of Information Technology, KUST, Kohat, Pakistan

CALL FOR PAPERS
International Journal of Computer Science and Information Security

IJCSIS 2018-2019
ISSN: 1947-5500
http://sites.google.com/site/ijcsis/

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc,), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on
its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embeded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://sites.google.com/site/ijcsis/authors-notes .

ISSN 1947 5500
http://sites.google.com/site/ijcsis/