IJCSIS Vol. 15 No. 8 (I), August 2017
ISSN 1947-5500

International Journal of Computer Science & Information Security

Pennsylvania, USA

Indexed and technically co-sponsored by :

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 2017 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 Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.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
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 **August 2017 issue** (Volume 15 Number 8 Part I & II) 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 9700 times and the number is quickly increasing. This 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, Academia.edu and EBSCO among others.*

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.

*"We support researchers to succeed by providing high visibility & impact value, prestige and excellence in research publication." 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. 15, No. 8, August 2017 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

Dr. K. V. Prasad (1), Dr. G. Sanjay Gandhi (2)
(1) Professor, Dept of CSE, TKR College of Engineering and Techlogy, Meerpet, Hyderabad,
(2) Professor, Dept of CSE, Visweswaraiah College of Engineering and Technology, Hyderabad

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

7. PaperID 31071710: Roty_Shift: A Proposed Method for Generating Secret Keys (pp. 40-55)

Qusay Mohammed Jafar,
Department of Computer Communication Engineering, AL-Rafidain University College, Iraq-Baghdad

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

8. PaperID 31071711: Ambulance Diversions Reducing and Dispatching Theory for Rescue Operations (pp. 56-62)

(1) Y. Bouhallaf, (1) L. Radoui, (2) O. Malass, (1) H. Belhadaoui, (1) M. Rifi
(1) Laboratory of Network, Computing, Multimedia & Communication, EST de Casablanca, University of Hassan II, BP. 8012, Morocco.
(2) Ecole Nationale Supérieure des Arts et Métiers, 4 rue Augustin Fresnel, 57078 METZ Cedex 3, France.

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

9. PaperID 31071712: The Application of Predictive Analytics in Healthcare Sector (pp. 63-81)

Fatimetou Zahra Mohamed Mahmoud,
Faculty of ICT, International Islamic University Malaysia (IIUM), Kuala Lumpur, Malaysia

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

10. PaperID 31071714: Towards Education Delivery Through E-learning in an Environment of Scarcity of Teachers: A Case of A-level Science Education in Tanzania (pp. 82-92)

(1) Mr. Zakaria Ezekiel Moshi, (2) Dr. Zaipuna O. Yonah, Member IEEE
1 Master's Scholar, School of Computational and Communication Science and Engineering, Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania.
2 Senior Lecturer, School of Computational and Communication Science and Engineering, Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania.

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

11. PaperID 31071716: Comparison Between PSO and HPSO In Image Steganography (pp. 93-100)

Ziyad Tariq Mustafa Al-Ta'i, Department of Computer Science, University of Diyala - College of Science, Baghdad, Iraq
Enaam Rabah Mohammad, Department of Computer Science, University of Diyala - College of Science, Diyala, Iraq

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

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

24. PaperID 31071749: Detection of Microaneurysm in Diabetic Retinopathy (pp. 200-203)

Morium Akter,
Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh

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

25. PaperID 31071755: Study Egyptian Students' Perception of Using Social Media in Learning (pp. 204-214)

Abeer A. Amer, Computer Science & Information System Department, Sadat Academy for management and Sciences, Alexandria - Egypt

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

26. PaperID 31071756: A Comparative Study to Removal Salt & Pepper Noise from Satellite Image (pp. 215-219)

Dr. Salem Saleh Ahmed Al-amri
Department of Engineering Geology Oil & Minerals Faculty, Aden University, Aden, Yemen.

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

27. PaperID 31071757: Linked List vs. AVL Tree in Modulo Ten Search (pp. 220-223)

Soheir Noori, Department of Computer Science, University of Karbala, Karbala, Iraq

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

28. PaperID 31071758: Optimised Kd-Tree Approach with Dimension Reduction for Efficient Indexing and Retrieval from Multibiometric Database (pp. 224-231)

Revathi B, Department of ECE, Pondicherry Engineering College, Puducherry, India
Gnanakumar D., Department of ECE, National Institute of Technology, Trichirappalli, India
Dr. Gnanou Florence Sudha, Professor, Department of ECE, Pondicherry Engineering College, Puducherry, India

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

29. PaperID 31071759: Object Oriented Modeling of Space Efficient LSB based Steganography including Compression in Transmission of e-Learning Documents (pp. 232-239)

[1] Soumendu Banerjee, [2] Sunil Karforma
[1] Research Scholar, Department of Computer science, The University of Burdwan, West Bengal, India
[2] Associate Professor, Department of Computer science, The University of Burdwan, West Bengal, India

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

Sidi Mohamed Ben Abedellah University (USMBA), Fez Morocco

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

36. PaperID 31071767: A Modified GA-based Workflow Scheduling for Cloud Computing Environment (pp. 276-283)

Safwat A. Hamad, Department of Computer Science, Faculty of Computers & Information, Cairo University, Cairo, Egypt
Fatma A. Omara, Department of Computer Science, Faculty of Computers & Information, Cairo University, Cairo, Egypt

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

37. PaperID 31071768: Towards a Semantic-based Context-as-a-Service for Internet of Things (pp. 284-293)

Manar Elkady, Dept. of Computer Science, Faculty of Computers and Information, Cairo University, Cairo, Egypt
Abeer M. El-Korany, Dept. of Computer Science, Faculty of Computers and Information, Cairo University, Cairo, Egypt
Reem Bahgat, Dept. of Computer Science, Faculty of Computers and Information, Cairo University, Cairo, Egypt

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

38. PaperID 31071769: Improvement and Enhancement Point Search Algorithm (pp. 294-299)

Sundus Khaleel Ebraheem, Dept. of Computer Sciences, College of Computer Sciences and Mathematics, Mosul University, Mosul, Iraq
Eman Abdulaziz, Dept. of Computer Sciences, College of Computer Sciences and Mathematics, Mosul University, Mosul, Iraq

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

39. PaperID 31071772: Simulation and Experimental Analysis of an AC-DC Converter with D.C Load (pp. 300-303)

Kalsoom Bhagat (1), Mohsin A. Mari (1), Asif A. Solangi (1), M. Zubair Bhayo (2), M. Haroon Nadeem (3)
(1) Department of Electrical Engineering, Mehran UET SZAB Campus Khairpur Mir's
(2) Department of Electrical Engineering, The Benazir Bhutto Shaheed University of Technology and Skill Development Khairpur Mir's
(3) Department of Electrical Engineering, Islamia University of Bahawalpur

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

40. PaperID 31071774: Heterogeneous Networks of Remote Monitoring with High Availability and Resilience Application to Wireless Sensor Networks (pp. 304-314)

Ayoub Marzak, Mohamed Hamraoui, Hicham Belhadaoui
RITM Laboratory, CED Engineering Sciences, Ecole Supérieure de Technologie, Hassan II University of Casablanca, Morocco

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

New approach for keys distribution througth publish/subscribe protocols

Abdessamad MEKTOUBI*, Olaf MALASS#,Hicham BELHADAOUI* and Mounir RIFI*
* CED Engineering Science, ENSEM, Lab. RITM/ESTC Hassan II University Casablanca, Morocco
# ENSAM/ParisTech, Metz,France
amektoubi@gmail.com
olaf.malasse@ensam.eu

***Abstract*— The Internet of things (IoT) is the interconnection of physical objects with the Internet network, using existing technologies and communication protocols. The speed up and massive uses of this network's type in a multitude domain, from large industrial applications to small every day's uses, raises new serious problematic related to security. Indeed, recent reports on cyber security highlighted the Internet of things' vulnerability, and the risks continue to increase exponentially, especially with the deployment of intelligent networks [1][2][3].**

Indeed, the user is able, in such an environment, to have access to the information, such information passes through different networks using different protocols, with a wide range of services delivered by different providers. In addition, end users ask to be constantly connected anyhow to the information. They also wish to have secure access to this information through different protocols.

Faced with this heterogeneity of support and protocol, several security challenges arise. The environment is increasingly open, multiple protocols are not known in advance and we have a variety of communications between protocols and machines.

In this work we propose a new method for key distribution through publish/subscribe protocol.

Keywords-component; publish/subscribe, key distribution, MQTT, Internet of Things.

I. INTRODUCTION

Generally, the security of networks and information systems "IS" is indispensable to any entity or organization trying to protect their data and resources (hardware or software). In this perspective, several aspects must be controlled. First, the information availability in Information System (IS) can be ensured through redundancy mechanisms or back-up for example. The data integrity must be protected against intentional or accidental alteration or destruction. In addition, enhancement of the IS security also requires the establishment of authentication and authorization protocols (e.g. Identity Access Management or IAM [4]) to limit and control the resources access. Finally, in the case of sensitive Data (financial, industrial, chemical formulas ...), it should not be accessible to unauthorized person. Cryptography (symmetric or asymmetric) provides in this case an appropriate solution [5][6], although problems remain suspended and require more attention from the scientific community as the solution based on the sharing key [7].

The sharing of encryption keys has always been a challenge. The use of algorithms such as "Diffie-Hellman" with asymmetric cryptography used to establish a secret key (known as session key) between two entities. This is even more constraining if we need to share keys between a group of users. Several works were carried out in order to solve this problem; we classify them into three categories: "Centralized Group Key Distribution (CGKD)"[8], where the distribution is ensured by a central trusted entity. "Decentralized Group Key Management with Relaying (DGKM)"[8] where the set of users is divided into subgroups to which one distinctly assigns central distribution entities. Finally, "Contributory Distributed Group Key Agreement (CGKA)"[9] in this case the group was organized to establish a secret key without using a third party.

II. STATE OF THE ART

A. Publish/Subscribe Pattern:

The "Publish/subscribe" model is a concept based on the exchange of messages without prior knowledge of the different actors which are transmitters or receivers. This type of architecture avoids any coupling between actors, unlike traditional models that require exemplary knowledge of the IP or full DNS (Domain Name System) name destination addresses of individual, group or domain recipient. This decoupling model offers more dynamism, scalability and flexibility. Three components parts are essentials in a Publish/Subscribe model; the first part is the publisher that sends messages, the second part is a Publish/Subscribe server (called Broker) that handles and switch messages and the third part is a Subscriber that remains constantly and exclusively listening to the messages concerning him. This model can be implemented in two different ways. The Publisher tags its messages by a topic representing a subject or a specific class that the server loads and broadcast to all Subscribers who are subscribed on this topic. In the second type of implementation, the messages are routed to the Subscribers according to their content.

B. Problematic:

In the following, we consider that a group of users, named Trust Group (TG), wants to communicate securely over a Publish/Subscribe protocol.

In Publish/Subscribe architecture with several public brokers, all the users have the right to publish or subscribe to any Topic. This exchange of unsecured data is greatly threatening

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

the confidentiality of data exchanged. The first solutions is to limit the number of users listening on a Topic, several protocol implementations or provide an authentication system along with a communication encryption (e.g. SSL communication for tcp protocols). This approach becomes credible if the broker is under the management of a trusted organization. Otherwise we cannot be certain if the messages are routed only to users who belong to the Trust Group.

A second approach is to encrypt messages with an encryption key. Consequently, only users who have the decryption key can read the message content. However, this approach assumes that the set of TG have the encryption and decryption keys. In this case new users should receive the same key. Hence, this needs an implementation of a key distribution strategy for new users of TG.

In this work, we propose a new approach of communication and management of keys in a Publish/Subscribe architecture. This is based on a Public Key Infrastructure (PKI) [12].

III. Approach and proposed model:

The main objective of this work is to secure distribution of encryption-decryption keys in order to secure communication and messages exchanged between a set of users within a Publish/Subscribe protocol.

In this context, we propose to use two types of certificates that will be issued by a Certification Authority "CA". The first type is a certificate for TG's users, the second type will be dedicated to Topics, meaning that the messages exchanged in a given Topic will be encrypted by the public key of this Topic.

Our approach is based on a sending request system on a specific Topic named "Topic Demand" TD. All users must be subscribed to "TD". All users can respond to requests based on its type, and depending on a set of criteria and constraints. In fact we define three types of requests:

- The certificate request.
- The private key request.
- The revocation list request.

Several formal approaches to analyze security protocols are based on the Dolev-Yao[10] model. In this work, we use a formalism to describe the basic assumptions of this model.

In the rest of this work, we use the same formalism and notation to describe our Keys' distribution protocol.

Sending a message M on Topic T by a user U1 using a Broker B, and received by a user U2, can be modeled by the following:

$$U1 \rightarrow B: T, M$$
$$B \rightarrow U2: T, M$$

For simplification reasons, this will be expressed as (skipping the exchanges with the broker B):

$$U1 \rightarrow T\ U2: M$$

Each message sent by a member of the TG consists of two parts. The first part is a header describing the type of the message, the certificate and the revocation list versions. The second part contains the message body.

At the beginning, we will consider a set of a TG users aiming to communicate on a topic "T". Each one, noted UTGi, has a certificate CUTGi published by the CA and its own private key SKUTGi. We also assume that at least one TG user has the public key PKT and the private key SKT of the topic T.

We consider a new user UTG_N who wants to communicate in a secure manner through the topic T. The first step is to retrieve the certificate of the topic T, which contains the public key PKT. For this, UTG_N will send a request (containing the Response Topic RT) for topic T's certificate on the topic DT. The modeling of this is summarized in the following lines:

$$UTG_N \rightarrow^{DT} U_X: MT, TT, RT$$

UTG_N: New TG user.
DT: Demand Topic.
U_X: the User who handles the request.
MT: the message type, in this case: "the certificate request"
TT: Target Topic, in this case: the topic "T".
RT: Response Topic.

Algorithm 1: Certificate request

```
Input: MT, TT, DT output:
BEGIN
    RT ← generateRandomTopic ()
    subscribe (RT)
    M← {MT, TT, RT}
    publish (DT, M)
END
```

Figure 1: Certificate request algorithm

Any U_X having the Topic certificate TC can respond to the request above by including it in a response on RT. UTG_N will stop listening after he receives the first valid response:

$$U_X \rightarrow^{RT} UTG_N: MT, TT, TC$$

MT: the message type, in this case: "the certificate response"
TC: Topic Certificate of "T".

Algorithm 2: Certificate response

```
Input: {MT, TT, RT} output:
BEGIN
    RM← {MT, TT, RT}
    TT ← getTargetTopic (RM)
    RT ← getResponseTopic (RM)
    TC ← getTopicCertificate (TT)
    MT← getMTPublicKeyResponse ()
    M  ← {MT, TT, TC}
    publish (RT, M)
END
```

Figure 2: Certificate response algorithm

The second step is obviously to request the secret key SKT. For this, UTG_N will send a private key request for the topic "T" (containing the Response Topic RT along with his own certificate, both encrypted with PKT) on the topic DT.

$$UTG_N \rightarrow^{DT} U_X: MT, TT, \{CUTG_N, RT\}_{PKT}$$

MT: Message Type, in this case: "private key request".
$CUTG_N$: Certificate of the user UTG_N.
PKT: Public Key of the Topic "T".

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

Algorithm 3: Private Key request

Input: MT, TT, $CUTG_N$, DT output:
BEGIN
PKT← getPKT (TT)
RT ← generateRandomTopic ()
RT ← getResponseTopic (M)
EM← encrypt ({CUTGN, RT}, PKT)
subscribe (RT)
M← {MT, TT, EM}
publish (DT, M)
END

Figure 3: Private Key request algorithm

Only an U_X owning SKT can decrypt $\{CUTG_N, RT\}_{PKT}$. He checks the certificate $CUTG_N$ in order to send a private key response containing the SKT encrypted with the public key of UTG_N named $PKUTG_N$ on the topic RT.

$$U_X \rightarrow^{RT} UTG_N: MT, TT, \{CU_X, SKT\}_{PKUTGN}$$

MT: message type, in this case: "private key response".
SKT: secret key of "T".
CUX: the user UX certificate
PKUGCN: public key of UTG_N

Algorithm 4: Private Key response

Input :{ MT, TT, $\{CUTG_N, RT\}$ PKT}, CU_X output:
BEGIN
RM ← $\{MT, TT, \{CUTG_N, RT\}_{PKT}\}$
TT ← getTargetTopic (RM)
PKT← getPKT (TT)
DM ← decrypt ($\{CUTGN, RT\}_{PKT,}$ PKT)
$CUTG_N$← getUserCertificate (DM)
RT ← getResponseTopic (DM)
IS_UCOK← verifyUserCertificate ($CUTG_N$)
If IS_UCOK =true **then**
SKT← getPrivateKey (TT)
$PKUTG_N$← getUserPublicKey ($CUTG_N$)
ESM ←encrypt ($\{CU_X, SKT\}$, $PKUTG_N$)
M ← {MT, TT, ESM}
publish (RT, M)
End if
END

Figure 4: Private Key response algorithm

After receiving a private key response, UTG_N will decrypt $\{CU_X, SKT\}_{PKUTGN}$ (using $PSUTG_N$), check CU_X and stop listening to RT.

UTG_N can henceforth communicate with other users on the target topic T using the pair SKT and PKT:

$$UTG_N \rightarrow^{T} U_X: MT, TT, \{M\}_{PKT}$$

MT: the message type, in this case: "Encrypted Message"

A. Topic keys management:

Among the essential features of a "CA" is to update the revocation list (fraud, loss of keys, attack...). Indeed, to ensure that a revoked certificate is not used, we will introduce in all exchanged messages an additional field in the header representing the version of the revocation list.

The above list is simply a message signed by the "CA" containing the serial numbers of the revoked certificates along with the list's version which is updated after every modification.

To retrieve the last version, a user will send a revocation list request on the demand topic DT:

$$U_i \rightarrow^{DT} U_j: MT, RT$$

MT: message type, in this case: revocation list request

The answer to this request by the user U_j will be:

$$U_j \rightarrow^{RT} U_i: MT, RL$$

MT: message type: Response to revocation list request
RL: List of Revoked Certificates

This request is sent in two cases; the first is at the beginning, when a client lunches his program, the second case is the reception of a message with a version of the revoked list newer than the user's version. In the latter case, the user stops its activity (sending messages, handling requests...) in order to fetch the current revocation list. Subsequently, the user will update his certificates data base in compliance with the new revocation list.

B. handling of the hierarchical topics:

In most cases, topics are organized in a hierarchical structure, so that a user can subscribe for one or multiple levels in one time. In this case the user will not retrieve all the keys related to topics within a hierarchy. He will, on the contrary, wait until he receives a message on a giving topic, and run the process of key retrieving in case he does not have it.

C. System Setting:

For the proper functioning of the system, we will introduce two parameters:

The first parameter is a waiting time before restarting a new request. Basically, our approach is based on a request-response logic. However, the response may not arrive. Hence, the applicant should be able to resend a new one. For that, the protocol defines a period of validity for the requests so that the user ignores an out of date message.

For this reason, we add two additional fields in the header; the sending timestamp and the period of validity.

$$U_i \rightarrow^{DT} U_j: MT, RT, timestamp, PV$$

PV: period of validity

The second parameter is related to the processing control and the system constraints. To protect the users against DoS attacks, we use thresholds (e.g. energy level, number of simultaneous requests, bandwidth…) beyond which ~~we~~ the system stop processing requests.

Those two parameters depend on the type of the implementation. For example, in an "Internet of Things" environment the energy efficiency and processing capacity are primary constraints.

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

D. Modelisation :

To implement our approach, we choose the following scenario: we assume that a set of connected users want to receive metrics (e.g. temperature, humidity…) assessed by a Wireless Sensor Network (WSN) through their mobile devices (smart phones, tablets, laptops…) using a publish/subscribe broker. Our implementation is based on the MQTT [11] protocol.

MQTT is an open source M2M (Machine to Machine) protocol created by IBM and based on the *publish/subscribe* pattern, it is often used in an Internet of things "IoT" environment due to its low usage of bandwidth. It offers three quality service levels and a *last well* notification mechanism in case of a user disconnection. In November 2014, MQTT v3.1.1 has become a standard OASIS (Organization for the Advancement of Structured Information Standards). A new version of this protocol called MQTT - SN (MQTT for Sensor Networks) intended for a no IP protocols such as Zigbee is now available.

Our architecture (*Figure 5*) contains a Zigbee WSN connected to a Raspberry chip which publishes the measurements to a public MQTT broker.

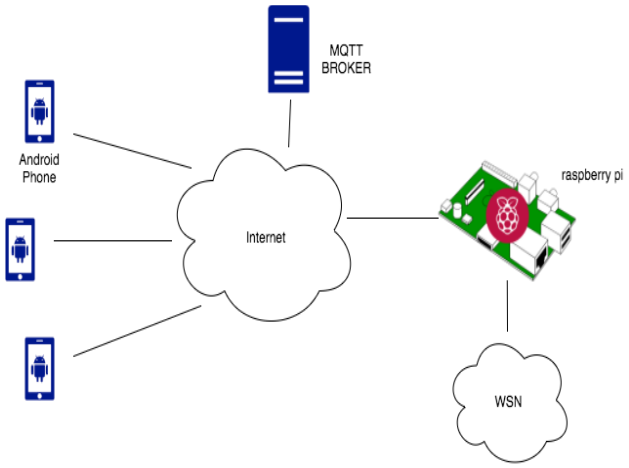


Figure 5: Global model of experimental plateform.

We used Java as a programming language due to its richness in functionality and its multiplatform aspect. We used the Eclipse PAHO Library that offers MQTT 3.1.1 implementation. We have developed our solution as a layer above the PAHO Library (*Figure 6*). This allows us to use the same code on multiple platforms (Android, Raspberry ...).

To secure all keys and certificates, we stored them in a Java "keystore" encrypted with a password. This can be easily replaced in a future work with a TPM (Trust Platform Module).

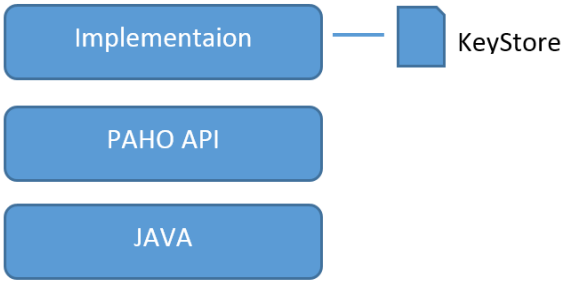


FIGURE 6: The architecture of a software application.

Our solution does not depend on the PAHO API. Indeed, we used a java interface named "PublishSubscribe" implemented with a class "PublishSubscribeImpl" (*Figure* 7). This will allow us to easily switch to another MQTT implementation if needed.

The Message class represents a sent message in our protocol, which contain a head (message type, target topic and the revocation list version) and a body.

For each message type we have a different class that implement the DetailMessage interface (SimpleMessage, CertifcateRequest…) *(Figure 8)*.

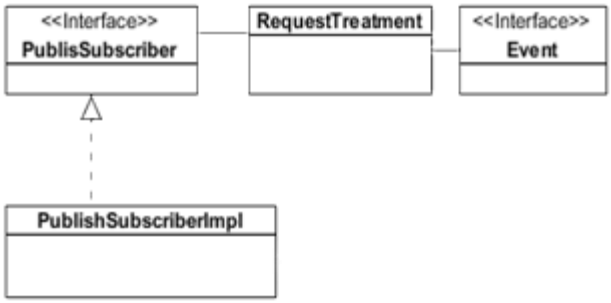


Figure 7: Class Diagram 1.

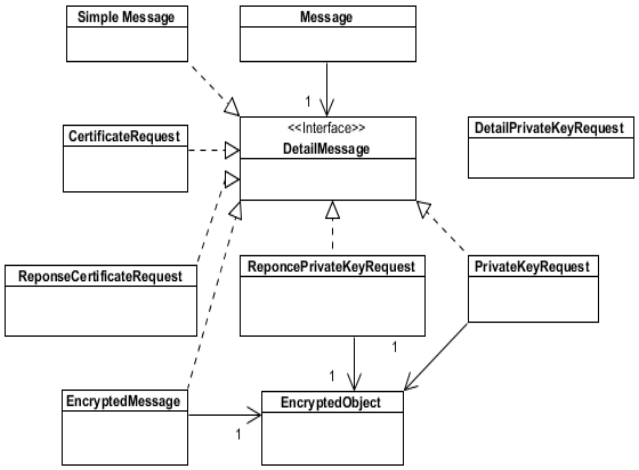


Figure 8: Class Diagram 2.

Any message type that uses an encrypted entity will instantiate the EncryptedObject class which contains a symmetric key "Ks" encrypted with a public asymmetric key and an object encrypted with "Ks".

IV. IMPLEMENTATION AND RESULTS :

We used the MEMSIC platform which is a Zigbee WSN~~, it's~~ composed of a based station, the "MIB520" (Figure 9), and a set of MICAZ sensors the (MDA100) (Figure 10). The said station is connected with a Raspberry PI board (Figure 11) through a USB port which publishes data to a MQTT server (Mosquitto).

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

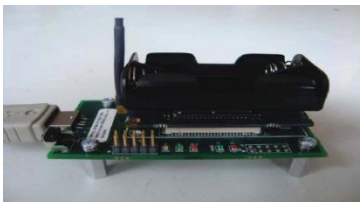


Figure 9: based station MIB520.

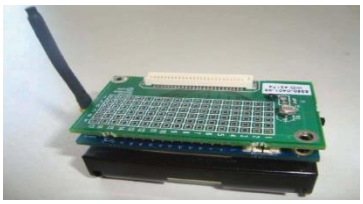


Figure 10 : MDA100 sensor.

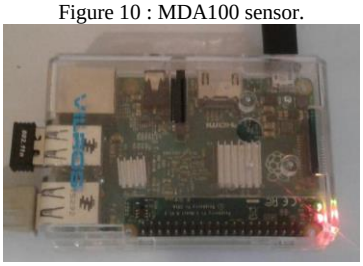


Figure 11 : Raspberry PI.

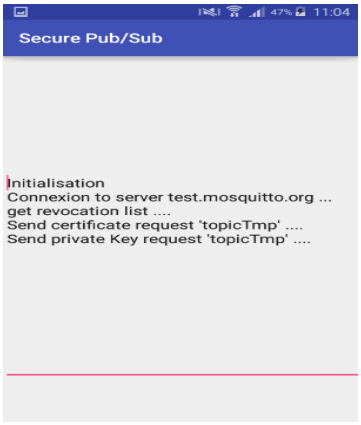


Figure 12: Initialization phase

The figure above shows the initialization phase through which we connect to a MQTT test server and lunch the process of key's retrieving.

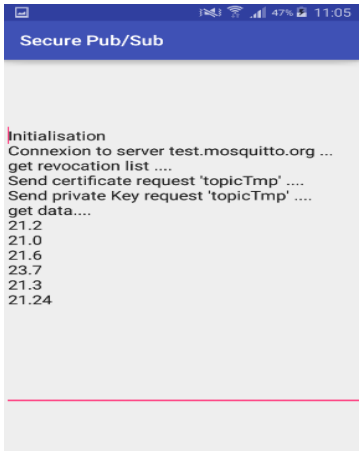


Figure 13: Getting Data

After obtaining the key pair, we start receiving data (*Figure 9*).

CONCLUSIONS AND FUTURE WORK

in fact, the Industry 4.0 revolution brings a new risk type for smart manufacturers and digital supply networks. In this industry, cybersecurity strategies should be secure.

However, our approach allows to be vigilante towards cyber attacks this through protecting the data exchange in no secure communication, thus we proposed a new approach to distribute encryption-decryption keys through a Publish/Subscribe protocol which is widely used in IoT environments. In the future work we will try to expand our approach to be multiprotocol with the same security level.

REFERENCES

[1] T. Jonathan Charity and H. Jian Hua, "Smart World of Internet of Things (IoT) and It's Security Concerns", 2016 IEEE International Conference on Internet of Things.

[2] M. Radovan*, B.Golub and Daimler AG, "Trends in IoT Security", 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics.

[3] Gartner Inc., Gartner Press Release - 6.4 Billion Connected, "Things" Will be in use in 2016, Stamford, Connecticut, USA : http://www.gartner.com/newsroom/id/3165317, 2015.

[4] E. Osmanoglu Identity and Access Management: Business Performance Through Connected Intelligence

[5] National Institute of Standards and Technology, "Specification for the Advanced Encryption Standard (AES)", Federal Information Processing Standards Publication 197, November 26, 2001.

[6] R.L. Rivest, A. Shamir, and L. Adleman , "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems ".

[7] C. Yuan Chen and H. Chieh Chao, "A survey of key distribution in wireless sensor networks", Security Comm. Networks (2011). Online Library

[8] P. Adusumilli, X. Zou, B. Ramamurthy, "DGKD: Distributed Group Key Distribution with Authentication Capability", 2005 IEEE Workshop on Information Assurance and SecurityUnited States Military Academy, West Point, NY

[9] Yu-Yi Chen ,Chuan-Chiang Huang, Jinn-Ke Jan" The Design of Secure Group Communication with Contributory Group Key Agreement Based on Mobile Ad Hoc Network" 2016 IEEE 2016 International Symposium on Computer, Consumer and Control

[10] D. Dolev and A.C. Yao. On the security of public-key protocols. IEEE Transactions on Information Theory, 29(8):198– 208, August 1983

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

[11] OASIS Message Queuing Telemetry Transport (MQTT) Version 3.1.1 http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/os/mqtt-v3.1.1-os.pdf

[12] C. Adams and S. Lloyd ,"Understanding PKI: Concepts, Standards, and Deployment Considerations"

AUTHORS PROFILE

Mektoubi Abdessamad currently a PhD student, University Hassan II /ENSEM Casablanca Morocco. Received an engineering degree in computer science in 2011 at ENSEM- Casablanca Morocco.

Research: Internet of things (Iot) security.

Olaf MALASS he is currently attached with NationalSchool of Arts/ and Crafts/ ParisTech in Metz/France asAssociate Professor in A3SI department.

Research: Automatic Signal Processing and ComputerEngineering.

Hicham BELHADAOUI currently working as a Professor Ability in UniversityHassan II /ESTC, Casablanca Morocco. Received his Phd degree at the National Polytechnic Institute of Lorraine/France.

Research: Security, Reliability, Automatic Signal Processingand Computer Engineering.

Mounir RIFI currently working as a Professor in University Hassan II /ESTC, Casablanca Morocco. Obtained his PhD Physical Sciences: ElectroMagnetic Compatibility, October 1996 (University Mohamed V of Rabat - Morocco) and PhD in Electronics, May 1987 (University of Lille - France) Director of the Research Laboratory: RITM (Networks, Computer, Telecom and Multimedia)

Research: Propagation of electromagnetic waves, ElectroMagnetic Compatibility, RFID, Microwave, Transmission Lines Theory, Antennas, Sensors, Networks.

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

DIFFERENTIAL EVOLUTION BASED SECURED ROUTING PROTOCOLS FOR VANETs

Er. Jayant Vats
Research Scholar Department of Computer Science & Engineering,
Shri venkateshwara, University,
Gujraula
jayantvasu@gmail.com

Dr. Gaurav Tejpal
Professor,
Shri venkateshwara, University,
Gujraula
gaurav_tejpal@rediffmail.co

Dr. Sonal Sharma
Assistant Professor, Department of computer Applications
Uttaranchal University
Dehradun, India
sonal_horizon@rediffmail.com

***Abstract*—** Vehicular Ad Hoc Networks (VANETs) are classified as a special application of mobile ad hoc networks (MANETs) which promise the new possibilities to improve traffic efficiency, road safety driving convenience. By providing the safety and non-safety applications and sharing the useful information through vehicle to vehicle (V2V) or vehicle to roadside (V2R) communications to avoid accidents and provide reliable information to travellers, such hot issues seeks much attention of researchers in this field. VANET and MANET having several common characteristics but VANET differ with applications, architecture, challenges and data dissemination. The survey of routing protocols in VANETs is important and necessary issue for smart ITS. The objective of this paper is to design an algorithm for the detection and correction of routing attacks made by obstructive nodes in VANETS and also drawn the comparison between various metrics like Cost, Average Packet loss, Throughput and Energy Consumed.

Keywords*— *VANETs (Vehicular Adhoc Networks), Attacks in VANETs, distance bounding technique and differential evolution

1. INTRODUCTION

Recently, it has been widely accepted by the academic society and industry that the cooperation between vehicles and road transportation systems can significantly improve driver's safety and road efficiency and reduce environmental impact. In light of this, the development of vehicular ad hoc networks (VANETs) has received more attention and research efforts [1]. VANETs are a special case of MANETs (Mobile Ad-hoc Networks). Their major feature is the high mobility of nodes. The immediate consequences are: topology changes and link disconnections. Managing this mobility is very important to ensure routing efficiency [2]. Increasing number of vehicles on the road has brought focus on improving road safety as well as in-vehicle entertainment [3]. VANETs play an important role for Intelligent Transport System (ITS) as well as road side units (RSU), providing an infrastructure to assist its services and applications. A particularly promising application is the data dissemination related to both the safety and commercial services to vehicles. That is, vehicles inside a geographical area of interest (AoI) attempt to obtain some common information, such as local weather, congestion status, or local business advertisement, from the RSU. A more important functionality for the RSU is to inform vehicles of the real-time traffic status and collision warnings such that the driver or the intelligent driving system can respond in time to avoid accidents. The primary requirement for the data dissemination in VANETs is to minimize the dissemination delay as much as possible. The dissemination delay refers to the duration from the start of data dissemination to the time when all the vehicles in the AOI successfully decode the entire data set. However, it is nontrivial to design low-latency data dissemination strategies for VANETs because the mobile channel [13, 14] is highly time variant, and the V2V channel modelling. These networks comprise of vehicles equipped with wireless communication devices and access points spread over streets and roads [4]. In general, the communication in VANET has been classified in two types: Vehicle-to-Vehicle (V2V) and Vehicle to- Infrastructure (V2I) [5]. The joint use of V2V and V2I supports many types of

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

ITS services, such as cooperative roads management and vehicles collision prevention [4, 5]. The architecture associated with VANET falls into three types [1]:

1. Inter-vehicle communication: It referred as vehicle to vehicle (V2V) communication as well as real ad hoc networking. Within class, the particular vehicles connect with one another without having infrastructure support. Any kind of useful data gathered through sensors on a vehicle as well as Communications by using vehicle could be forwarded to nearby vehicles.
2. Vehicle-to-roadside communication: It is referred as vehicle-to-infrastructure (V2I) communication. Within type, the vehicles could utilize gateways as well as cellular LAN accessibility things for attaching with the web as well as allow for vehicular purposes [6].
3. Inter-roadside communication: It is referred as hybrid vehicles-to-roadside communication (VRC). Vehicles may utilize infrastructure for transmitting collectively along with change data obtained through infrastructure or from other motor vehicles by the way of adhoc communication. In addition, vehicles may interact among infrastructure in single hop or multi-hop style based on their location during transferring or stationary. This kind of architecture contains V2V communication and offers higher flexibility within material revealing as well as raises network efficiency [1, 6].

VANETs are likely to provide support for cooperative driving applications, which would allow vehicles to navigate without driver intervention. The specific characteristics of VANETs favour the development of attractive and challenging services and applications, including road safety, traffic flow management, road status monitoring, environmental protection and mobile infotainment [11]. So the traffic safety and efficient warning message dissemination, where the main goal is to reduce the latency and to increase the accuracy of the information received by nearby vehicles when a dangerous situation occurs [12]. There are different types of ad hoc networks, the vehicle is moving with high speed value [2]. This network feature makes the topology very dynamic and increases the probability of communication links failure. Therefore, designing an efficient routing protocol for VANETs is a major challenge. In VANET, the routing protocols are classified into various categories [15]: position based routing protocols, topology based routing protocols and cluster based routing protocols. In position methods [16], each node maintains its geographical coordinates as well as its neighbour's positions using GPS service.

It doesn't share any routing information with neighbour nodes or keep any routing table. In order to take a decision, the data from GPS device is used. The pros of position based protocols are that route discovery phase which is not needed. Consequently, it is also appropriate for high speed node. However, this category has needs of position determining services. Topology based routing protocols are based on exchanging information state about the link in order to deliver the data packets from source node to destination [15]. It can be classified into classes. The first one is reactive protocols that are called also on demand routing approaches [16]. The route is discovered when a node seeks to send a data. The major highlight of this type is that updating routing table that is not needed. Nevertheless, flooding mechanism generates a large volume of overhead. The second category is proactive approaches which is called also table driven based schemes because information about all connected nodes is kept in routing tables. These tables are exchanged between neighbours' nodes. Once topology network is changed, each node updates its table. Since discovery route step is not required, proactive methods are judged as the most suitable category for real time applications with the lowest latency. However, many unused paths are generated which can take up an important part of the limited bandwidth. In Cluster-based routing protocols [15] the network is divided in different groups (clusters). Everyone has a single manager or cluster-head. It is responsible for managing its cluster members (intra cluster) and its neighbour's clusters (inter-cluster). While the intra-cluster communication is established using direct links, the inter-cluster is performed on the basis of cluster headers links. This kind of protocols is the most important to provide stability of communication link. However, creation of clusters as well as the selection of the cluster-head is a big issue in VANET. The performance of a cluster based approach is too related to the selection manner of the manager node. Several comparative studies [17] are demonstrated that topology based protocols are the most used category, particularly, the proactive methods which have the superiority of lower delay. However, these protocols are not suitable as they are in vehicular communication. The main problem with these protocols is that the control packets are flooded among every node to discover and keep a link path. Consequently, some of routes are never used. The IEEE 802.11 working group continues to actively develop 802.11p [6] for supporting Intelligent Transportation System (ITS) applications. The 802.11p standard will provide wireless devices with the ability to perform the short-duration exchanges necessary to communicate between a high-velocity vehicle and a stationary roadside unit. This mode of operation, called WAVE (wireless access in vehicle environments) will operate in a 5.9 GHz band and support the Dedicated Short Range Communications (DSRC) standard [7] sponsored by the U.S. Department of Transportation. These standards will support systems that communicate from vehicle-to roadside, vehicle-to-vehicle, or both. For supporting such wide range of applications, messages exchanged should be authenticated while at the same time the anonymity of the senders should be preserved. In vanets every vehicle performs life critical activities in light of the received messages from its neighbouring vehicles. VANET can't exclusively depend on the short lifetime testaments as a malicious vehicle can hurt different vehicles until its certificate lifetime terminates [8].

For detect the misbehaviour and malicious nodes in Vehicular Ad-hoc Networks the misbehaviour detection schemes can be broadly classified into following two types: Node-centric and Data-centric misbehaviour detection schemes [9].

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

1. Node Centric Misbehaviour Detection Schemes: Node-centric techniques need to distinguish among different nodes using authentication. Security credentials, digital signatures, etc are used to authenticate the node transferring the message. Such schemes emphasis on the nodes transmitting the messages rather than the data transferred [9].
2. Data Centric Misbehaviour Detection Schemes: Data-centric approach inspects the data transmitted among nodes to detect misbehaviours. It is primarily concerned with linking between messages than identities of the individual nodes [9 10]. The information disseminated by the nodes in the network is analyzed and compared with the information received by the other nodes, in order to verify the truth about the alert messages received.

2. ATTACKS IN VANETS

Security is an important issue for routing in VANETs, because many applications will effect life-or-death decisions and illicit tampering can have devastating consequences. Security is an important issue for routing in VANETs, because many applications will effect life-or-death decisions and illicit tampering can have devastating consequences. The characteristics of VANETs make the secure routing problem more challenging and novel than it is in other communication networks. Another challenge related to routing is efficient data dissemination and data sharing in VANETs.

In order to get better protection from attackers, it is essential to have the knowledge about the attacks in VANET against security requirements. These attacks are based on

(i) Identification and Authentication
(ii) Attack on privacy
(iii) Attack on availability
(iv) Routing attacks

3. DISTANCE BOUNDING TECHNIQUE

Secure distance-bounding protocols measure the time of flight to determine an upper bound on the distance between the prover and verifer. This measurement is typically performed during a challenge-response protocol—the main building block of the distance-bounding protocol. During each of the *m* fast-bit exchanges, the verifer measures the time between sending a challenge and receiving the corresponding response. Multiplying the time of flight with the communication medium's propagation speed (the speed of light for RF communication) provides an estimate of the distance between the prover and verifer. During authentication, the verifer wouldn't learn anything after having carried out the protocol. Most distance-bounding protocols want to preclude distance-fraud and relay attacks. (Some protocols focus on terrorist fraud attacks, but we don't discuss them here.) In a distance-fraud attack, a dishonest prover claims to be closer than he or she really is in relay attacks (also denoted by mafa fraud attacks) both the prover and verifier are honest, but there's a malicious intruder. This is a man-in-the middle attack, where the intruder *I* is modeled as a malicious prover *P* and verifer *V*. The malicious verifer *V* interacts with the honest prover *P*, and the malicious prover *P* interacts with the honest verifer *V*. The physical distances between the intruder and prover and intruder and verifier are small, so neither *P* nor *V* notices the attack.

Distance-bounding protocols can cryptographically enforce the concept of "proximity." They combine physical and cryptographic properties to let the user authenticate remotely and let the verifying party determine an upper bound on the distance between itself and the user (*prover*). Employing distance-bounding protocols avoids relay attacks, where an adversary close to the verifer impersonates an authorized user. Such attacks are an important threat in access control systems. However, simply enhancing a mutual authentication protocol with location information isn't sufficient. Physical access control mechanisms usually rely on a single security token.

4. DIFFERENTIAL EVOLUTION

DE Algorithm with this section describes the fundamental operator associated with DE which usually facilitates the easier explanation of the hybrid differential evolution algorithm later.
DE algorithm is an excellent evolutionary algorithm with regard to fixing or resolving mathematical optimization problems.
Let's Assume that min $f(X')$ is usually the type of the global optimization problem, *N'* is the dimension of the challenge, *NP* is actual population size of DE, X'_I=($x'_{I,1}, x'_{I,2,}, x'_{I,3}, \ldots\ldots x'_{I,N}$) is the *I*-individual of the population, *G* is the evolution generation. DE includes the three fundamental operators: mutation, crossover and selection on the basis of real coding.

A. Mutation

At generation G, this operator creates mutation vectors $M_{I,G}$based on the current population $\{X'_{I,G}|\ I=1,2,\ldots, NP\}$. Equation (1) is the classic mutation strategies.

$$M_{I,G} = X'_{r1,G} + F * (X'_{r2,G} - X'_{r3,G}) \quad (1)$$

In (1), the indices r1, r2 and r3, which are uniformly chosen from the set {1, 2... NP}\ {I}, are distinct integers. F is mutation factor which is fixed parameter.

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

B. Crossover

A crossover operator, which is based on Equation (2), creates the trial vector $T_{I,G}=T_{1,I,G}, T_{2,I,G}, T_{3,I,G}, \ldots \ldots T_{N,I,G}$ after mutation.

$$T_{J,I,G} = X' \begin{cases} M_{J,I,G}, \text{if rand}(0,1) \leq CR \text{ or } J = J_{rand} \\ x'_{J,I,G}, \text{otherwise} \end{cases} \quad (2)$$

In (2), r and (a,b) is uniform random number on the interval [a,b], J_{rand} is an random integer chosen from 1 to N and new for each I and the crossover probability CR is a fixed parameter.

C. Selection

the selection operator, which is based on (3), selects the better one between the parent vector $X'_{I,G}$ and the trial vector $T_{I,G}$. The selected vector is used as a parent vector in the next generation.

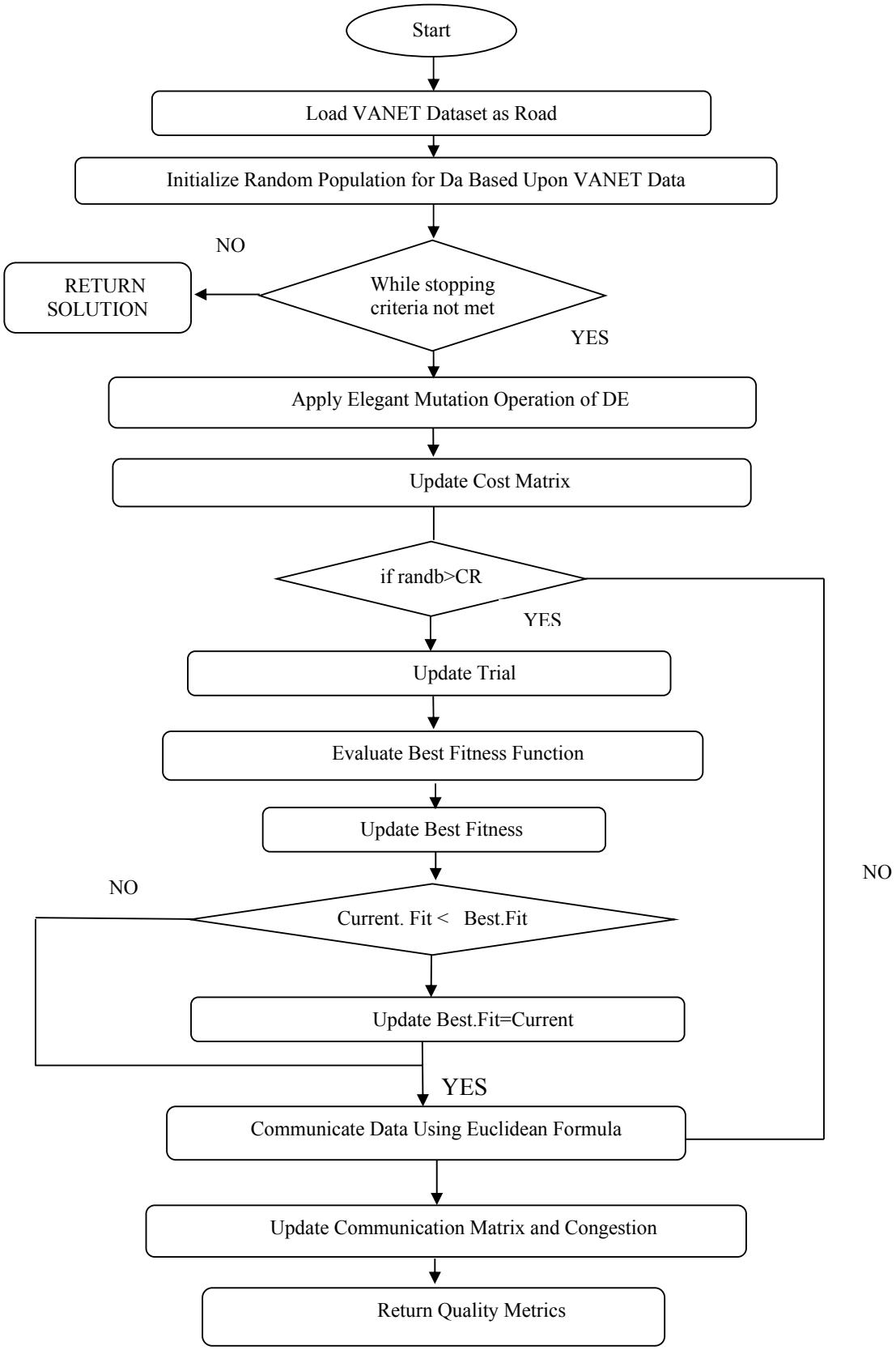
$$X'_{I,G+1} = \begin{cases} T_{I,G}, \text{if } f(T_{I,G}) \leq f(x'_{I,G}) \\ x'_{I,G}, \text{otherwise} \end{cases} \quad (3)$$

5. RELATED WORK

Fan Li and Yu Wang[1] Vehicular Ad Hoc Network (VANET) is an emerging new technology integrating ad hoc network, wireless LAN (WLAN) and cellular technology to achieve intelligent inter-vehicle communications and improve road traffic safety and efficiency. Irshad Ahmed et al. [2] proposed five different classes of attacks and every class is expected to provide better perspective for the VANET security. The main contribution of this paper is the proposed solution for classification and identification of different attacks in VANET. Ravi Kant Sahu et al. [3] proposed for DOS based attacks which use the redundancy elimination mechanism consists of rate decreasing algorithm and state transition mechanism as its components. This solution basically adds a level of security to its already existing solutions of using various alternative options like channel-switching, frequency-hopping, communication technology switching and multiple-radio transceivers to counter affect the DOS attacks. Swapnil G. et al. [4] Vehicular ad hoc networks (VANETs) have attracted a lot of attention over the last few years. VANETs are being used to improve road safety and enable a wide variety of value-added services. In this an author done a survey of existing approaches to solve the problems associated with vehicular networks. Senthil Ganesh N. et al. [5] VANETs would indeed turn out to be the networking platform that would support the future vehicular applications. Author also analyze the various security threats and the existing solutions to overcome the threat factors and show that there are active research efforts towards making VANETs a reality in the near future. Sumit A. Khandelwal et al. [6] Vehicular ad-hoc networking is an emerging technology for future on-the-road communications. Due to the virtue of vehicle-to-vehicle and vehicle-to-infrastructure communications, vehicular ad hoc networks (VANETs) are expected to enable a plethora of communication-based automotive applications including diverse in-vehicle infotainment applications and road safety services. Even though vehicles are organized mostly in an ad hoc manner in the network topology, directly applying the existing communication approaches designed for traditional mobile ad hoc networks to large-scale VANETs with fast-moving vehicles can be ineffective and inefficient. Gul N. Khan et.al [7] aims to fill these gaps by proposing a protocol that provides individual-owner-privacy, based on simple XOR and 128-bit pseudo-random number generators (PRNG), operations that are easily implemented on low-cost RFID tags while meeting the necessary security requirements thus making it a viable option for large scale implementations. Md. Humayun Kabir et al. [8] Vehicular Ad Hoc Network (VANET) which includes the characteristics of high node flexibility and fast topology changes. VANET is an energetic area of research, standardization, and development because it has tremendous potential to improve vehicle and road safety, traffic efficiency and convenience as well as comfort to both drivers and passengers. Kamal Deep et al. [13] discussed CR technologies for vehicular networks aimed at improving vehicular communication efficiency. CR for vehicular networks has the potential of becoming a killer CR application in the future due to a huge consumer market for vehicular communications. Sermpezis et al. [14] introduced a new routing technique designed exclusively for VANETs and present some initial performance results. The algorithm was named Junction-based Multipath Source Routing (JMSR). Its main characteristics comprise the multiple routes towards the destination, the junction-centric logic and the adoption of source routing mechanisms. Shikha Agrawal et al. [16] proposed algorithm DMN-Detection of Malicious Nodes in VANETs improves DMV Algorithm in terms of effective selection of verifiers for detection of malicious nodes and hence improves the network performance.

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

6. PROPOSED METHODOLOGY



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

7. RESULTS AND DISCUSSIONS

7.1 Experimental set up

To be able to implement the proposed algorithm, design and implementation have been done. Table 4.1 has shown various different constants and variables essential for simulating the work. These parameters are generally standard values utilized as standard for VANETS.

Table 1: Experimental setup

Parameter	Value
Level_of_agg	1:5
Speed_of_vehicle	10:10:50
D	50:50:150
N	100
min1	20
max1	80
oint1	8
oint2	12
simu_time	10
Total nodes	10

7.2 Results and Evaluation

1. Cost: A measure that needs to be paid or perhaps given up to receive something. In operation, expense generally is a economical valuation of work, product, resources, time and utility bills absorbed, risks incurred, plus ability forgone in production plus delivery of the excellent or perhaps service.

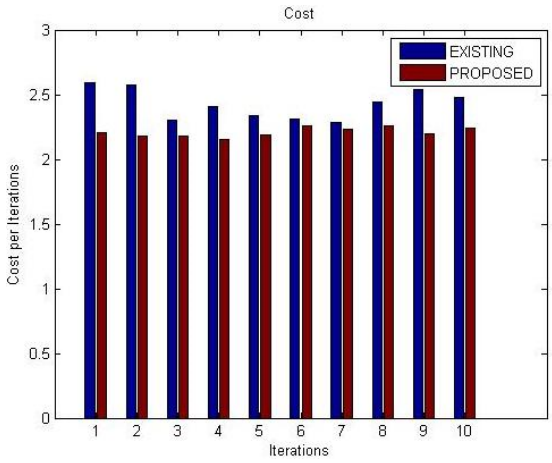


Fig 1: Represents the cost

Fig 1: Demonstrates the evaluation of cost among the pre-existing and our proposed technique in which the X-coordinate represents the Iterations and the Y-coordinate represents the Cost per Iteration. In our scenario the proposed technique cost is reasonably lower than the pre-existing technique.

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

2. Packet Lost Rate: Package damage occurs when number packages of internet data over a laptop fail to get to their particular destination. Package damage or loss is usually due to the system congestion. Package damage can be calculated as a percentage of package damage rates as compared to the package sent.

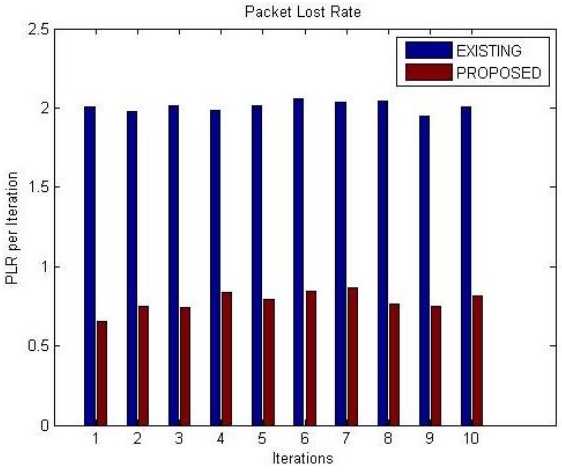


Fig 2: Represent the packet loss rate

Fig 2 demonstrates the analysis of Packet Lost Rate among the pre-existing and our proposed technique. In this figure, red line represents the proposed technique and blue line represents the previous one. In our case the proposed Packet Lost Rate are reasonably lower than the existing one.

3. Throughput: Throughput is actually the maximum charge connected with manufacturing or the maximum charge when something might be processed. Network throughput is actually the speed associated with the successive message supply on the communication channel.

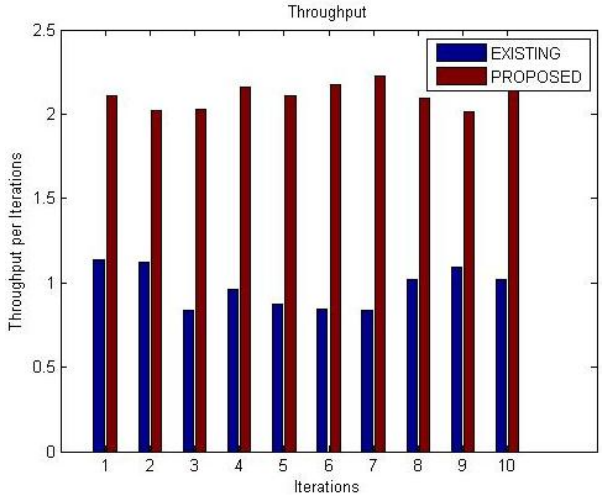


Fig 3: Represent the throughput

Fig 3 demonstrates the comparison of Throughput among the pre-existing and the proposed method where X-coordinate represents the Iterations and Y- coordinate values represents the Throughput per Iteration. In our scenario the proposed technique's Throughput reasonably higher than the pre-existing technique.

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

4. Energy consumed**:** Energy consumption is the amount of energy or power used.

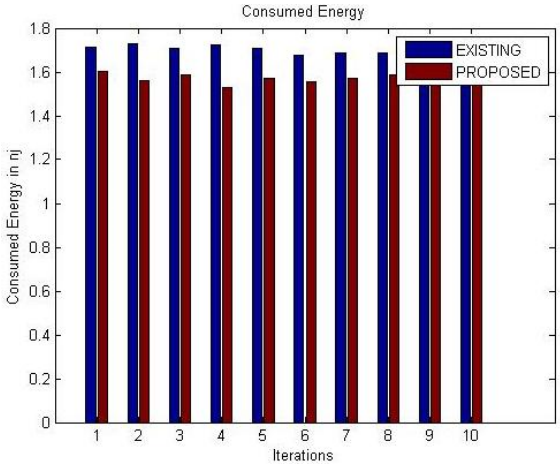


Fig 4: Consumed Energy

Fig. 4 demonstrates the comparison of Consumed Energy among the pre-existing and the proposed technique. In this figure the red colored lines represents the proposed technique and blue colored lines represents the pre-existing technique. In our scenario the proposed Consumed Energy is reasonably lower than existing one.

8. CONCLUSION

This paper represents the overview of recent issues and applications are discussed and emphasizing the importance of safety and non-safety applications in VANETs, current challenges and issues are elaborated to enhance improvements in technology and overcome these pitfalls from VANETs. The comparison has been drawn between the various metrics like Cost, Packet Loss Ratio, Throughput and Consumed Energy using differential equation algorithm (i.e. detection and correction of attacks in VANETs made by obstructive nodes). Result shown by above algorithm is better than existing results.

Security in VANETs as well as data dissemination and sender's privacy safety applications should be enhanced in future as they are the hot issues in recent times.

REFERENCES

[1] Fan Li and Yu Wang, University of North Carolina at Charlotte, "Routing in Vehicular Ad Hoc Networks: A Survey" IEEE Vehicular Technology Magazine, 2007 pp 12-22

[2] Irshad Ahmed Sumra,Iftikhar Ahmad, HalabiHasbullah, "Classes of Attacks in VANET", IEEE 2011

[3] AdilMudasirMalla, Ravi Kant Sahu, "Security Attacks with an Effective Solution for DOS Attacks in VANET", International Journal of Computer Applications, Volume 66– No.22, March 2013, pp 45-49

[4] Swapnil G. Deshpande , "Classification of Security attack in Vehicular Adhoc network: A survey", International journal of emerging trends and technology in computer science, Volume 2, Issue 2, March – April 2013, pp 371-377

[5] Senthil Ganesh N., Ranjani S., "Security Threats on Vehicular Ad Hoc Networks (VANET): A Review Paper", National Conference on Recent Trends in Computer Science and Technology (NCRTCST)-2013, Volume 4, Issue (6) NCRTCST-2013, pp 196-200

[6] Sumit A. Khandelwal, Ashwini B Abhale, "Topology base Routing Attacks in Vehicular Ad hoc Network – Survey", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, November 2013, pp 1352-1356

[7] VinhHoa LA, Ana CAVALLI, "Security Attacks And Solutions In Vehicular Ad Hoc Networks: A Survey", International Journal on AdHoc Networking Systems (IJANS) Vol. 4, No. 2, April 2014, pp 1-20

[8]. Kazemi, Babak, Masoumeh Ahmadi, and Siamak Talebi. "Optimum and reliable routing in VANETS: An opposition based ant colony algorithm scheme." In Connected Vehicles and Expo (ICCVE), International Conference on, pp. 926-930. IEEE 2013.

[9] Eiza, Mahmoud Hashem, Qiang Ni, Thomas Owens, and Geyong Min. "Investigation of routing reliability of vehicular ad hoc networks." *EURASIP journal on wireless communications and networking* 2013, no. 1 (2013): 179.

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

[10] Ayaida, Marwane, Mohtadi Barhoumi, Hacène Fouchal, Yacine Ghamri-Doudane, and Lissan Afilal. "Joint routing and location-based service in VANETs." *Journal of Parallel and Distributed Computing* 74, no. 2 (2014): 2077-2087. (ayida2014)
[11] Singh, Kamal Deep, Priyanka Rawat, and Jean-Marie Bonnin. "Cognitive radio for vehicular ad hoc networks (CR-VANETs): approaches and challenges." *EURASIP journal on wireless communications and networking* 2014, no. 1 (2014): 49.(singh 2014).
[12] Sermpezis, Pavlos, Georgios Koltsidas, and Fotini-Niovi Pavlidou. "Investigating a junction-based multipath source routing algorithm for VANETs." *IEEE Communications letters* 17, no. 3 (2013): 600-603. (macedio 2017)
[13] Macedo, Ricardo, Robson Melo, Aldri Santos, and Michele Nogueira. "Experimental performance comparison of single-path and multipath routing in VANETs." In *Global Information Infrastructure and Networking Symposium (GIIS), 2014*, pp. 1-6. IEEE, 2014. (macedio 2017)
[14] Cunha, F., L. Villas, A. Boukerche, G. Maia, A. Viana, R. A. F. Mini, and A. A. F. Loureiro. "Data communication in VANETs: survey, applications and challenges." *Ad Hoc Netw. doi* 10 (2016).
[15] Singh, Amandeep, and Sandeep Kad. "A review on the various security techniques for VANETs." *Procedia Computer Science* 78 (2016): 284-290. '(70 pdf)
[16] Khan, Uzma, Shikha Agrawal, and Sanjay Silakari. "Detection of malicious nodes (dmn) in vehicular ad-hoc networks." *Procedia Computer Science* 46 (2015): 965-972.
[17] Kadam, Megha, and Suresh Limkar. "Performance Investigation of DMV (Detecting Malicious Vehicle) and D&PMV (Detection and Prevention of Misbehave/Malicious Vehicles): Future Road Map." In *FICTA*, pp. 379-387. 2013.
[18] Wahab, Omar Abdel, Hadi Otrok, and Azzam Mourad. "A cooperative watchdog model based on Dempster–Shafer for detecting misbehaving vehicles." *Computer Communications* 41 (2014): 43-54.
[19] Ghandour, Ali J., Marco Di Felice, Hassan Artail, and Luciano Bononi. "Dissemination of safety messages in IEEE 802.11 p/WAVE vehicular network: Analytical study and protocol enhancements." *Pervasive and mobile computing* 11 (2014): 3-18.
[20] Sanguesa, Julio A., Manuel Fogue, Piedad Garrido, Francisco J. Martinez, Juan-Carlos Cano, Carlos T. Calafate, and Pietro Manzoni. "RTAD: A real-time adaptive dissemination system for VANETs." *Computer Communications* 60 (2015): 53-70.
[21] Cheng, Xiang, Cheng-Xiang Wang, Bo Ai, and Hadi Aggoune. "Envelope level crossing rate and average fade duration of nonisotropic vehicle-to-vehicle Ricean fading channels." *IEEE Transactions on Intelligent Transportation Systems* 15, no. 1 (2014): 62-72.
[22] Shen, Xia, Xiang Cheng, Liuqing Yang, Rongqing Zhang, and Bingli Jiao. "Data dissemination in VANETs: A scheduling approach." *IEEE Transactions on Intelligent Transportation Systems* 15, no. 5 (2014): 2213-2223.
[23] Chauhan, Saurabh, and Shashi Bhushan Tyagi. "Performance Evaluation Of Reactive Routing Protocols In VANET"." *International Journal of Innovations and Advancement in Computer Science* 3, no. 9 (2014): 189-193.
[24] Harrabi, Samira, Ines Ben Jaffar, and Khaled Ghedira. "Novel Optimized Routing Scheme for VANETs." *Procedia Computer Science* 98 (2016): 32-39.
[25] Dhankhar, Shilpi, and Shilpy Agrawal. "VANETs: a survey on routing protocols and issues." *International Journal of Innovative Research in Science, Engineering and Technology* 3, no. 6 (2014): 13427-13435.

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

Classification of Arabic texts using four classifiers

Ahmed H. Aliwy
Collage of CS and Mathematic
University of Kufa, Iraq
ahmedh.almajidy@uokufa.edu.iq,
ahmed_7425@yahoo.com

Kadhim S. Aljanabi
Collage of CS and Mathematic
University of Kufa, Iraq
kadhim.aljanabi@uokufa.edu.iq

Zena Abd Al_Retha AboAltaheen
Collage of CS and Mathematic
University of Kufa, Iraq
zena.retha@yahoo.com
zenaretha87@gmail.com

***Abstract*—Text Classification is one of the main parts of Text Mining (TM) field. Arabic text classification has its own difficulties and limits result from the nature of Arabic language. In this paper, four classifiers including a suggested algorithm, are applied to Arabic corpus. The used classifiers run after preprocessing of Arabic texts (tokenization, stemming, and stop word removal) and calculating the weights of each feature as tf-idf weights. The four used classifiers are Naïve Bayes, K-Nearest Neighbor, Logistic Regression, and suggested algorithm using the same data sets that applied for the first time and the same experimental settings. The used corpus consist of 16757 Arabic documents belongs to five different categories classified manually. The experiment shows that the suggested method give best results from the other classifiers.**

Keywords: Arabic corpus, Text Classification, Arabic Text Mining, Arabic Text Classification.

I. INTRODUCTION *(HEADING 1)*

Text Classification, also known as text categorization, topic spotting, document categorization or document classification, is one of the important tasks of text mining. It is a process of assigning and labeling documents to set of predefined categories based on their contents [1]. For example, the classes can be Literature, Arts, Economy, Sport, etc.. Text Classification is very useful task in many applications of Natural Language Processing (NLP) as Information Retrieval (IR).

For Arabic documents, the classification is very complicated results from the nature of Arabic morphology that creates ambiguity in the text [2]. There are many algorithms were used for Arabic document classification as Decision Trees [3, 4], k-nearest neighbor [5], Naïve Bayes [6], Maximum Entropy [7, 8] and others which give good accuracy.

In this work, three of well-known classifiers are compared with very simple suggested classifier. This comparison is done according to the performance of each classifier using the same datasets that applied for the first time.. the used classifiers are: k-Nearest Neighbor (KNN), Logistic Regression (LR), Naive Bayes (NB) and suggested algorithm. the used dataset consist of 16757 files collected from Alsabah1 Newspaper which is the formal Iraq newspaper.

1 See section 5 for more details

II. RELATED WORK

There are many works were done for comparing classifiers on Arabic texts. All of these works used a private dataset which make the comparison among them very difficult.

Duwairi et al. [9] compared the performance of three different classifiers Nave Bayes, k-nearest-neighbor (KNN) and distance based classifier on the Arabic language. The used corpus is a collection of 1000 documents classified into 10 categories. Each category contains 100 documents. the corpus was preprocessed by applying stopwords removal and stemming where the training was 50% and testing was 50%. She showed that the performance of Nave Bayes classifier outperformed the other two classifiers.

Danso et al. [10] worked a comparative study of four classifiers: Nave Bayes, Support Vector Machines and Decision Trees for verbal autopsy text. This study covers feature value representation, text classification and the effect of feature reduction. The experiment involves a total of 6407 Verbal Autopsy documents, consisting of two levels of groupings: the higher level has 5 categories and the fine grained level consists of 16 categories. The SVM algorithm was found to be the best performing algorithm.

Faidi et al. [11] compared Arabic NLP tools for hadith classification using three classification techniques: decision trees (DT), Nave Bayes algorithm (NB) and Support Vector Machines algorithm (SVM). Thes classifiers were implemented on WEKA toolkit that analyzed with a well-known Arabic tools for stemming and n-grams calculations. They used TF-IDF weighting for the term and the cross validation to evaluate the result of the classifiers. They show that Khojas stemmer outperformed the other tools as preprocessing and that the SVM classifier achieves the highest accuracy followed by the Naive Bayes classifier, and decisions trees classifier respectively.

Alsaleem [12] compared results of two classifiers on Arabic text collections of Saudi Newspapers (SNP) . the used classifier were Support Vector Machine (SVM) algorithm and Nave Bayesian (NB) algorithm. The results against different Arabic text categorization data sets reveal that SVM algorithm outperforms the NB with regards to all measures.

Our work is different than the others by using a suggested classifier and comparing it with the well-known classifiers but it similar to the other works by using private data set which is

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

the first in its type from Iraqi media. It will be freely available for the researchers soon.

III. COMPLEXITY OF ARABIC LANGUAGE IN TEXT CLASSIFICATION

Arabic language has many levels of difficulties in text mining results from its concatenative nature and richness vocabulary language. These difficulties rise in preprocessing stage as tokenization, stemming and stop word removal but it also occur in feature selection.

Tokenization of Arabic text is nontrivial task because of the concatenative nature of Arabic words. For example the word "وفي" that mean "Loyalty" or "and in".

In case of stop words removal, there is also problem in Arabic language because many stop words have number of different meaning where some of these useful in classification problems. Also, Arabic is complex for stemming due to the complexity of the morphological structure of the Arabic word.

The complexity, for Arabic text classification, may be occurs in feature selection resulting from the fact that Arabic language is the most richness language (many names for one thing). For example, Arabic language has 300 name for sword. These words should be unified with one feature not for different features; therefore, a dictionary of synonyms may a good solution.

IV. THE USED CLASSIFIER

There are many Classification techniques that applied to text. Like other tasks of NLP, text classification need to preprocessing stage (tokenization, stop-word removal, normalization and stemming). Almost all classifiers used labeled text for traing and evaluating the performance of the classifier by classifying another labeled text for testing.

Our work focused on comparing three well-known classifiers **Naïve Bayes, K-Nearest Neighbor and Logistic Regression with the suggested work.**

A. Naïve Bayes

Nave Bayes classifiers are simple probabilistic classifiers based on Bayes' theorem with strong independence assumptions between the features for this reason it is called naive. It is, also, the most commonly used generative classifier [13]. In general, most researchers employ NB method by applying Bayes rule:

$$c_{MAP} = \underset{c_j \in C}{\operatorname{argmax}}_c \hat{P}(c_j) \prod_i \hat{P}(x_i \mid c_j) \quad (1)$$

Where:

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}} \quad (2)$$

$$\hat{P}(x_i \mid c_j) = \frac{count(x_i, c_j)}{\sum_{x \in V} count(x, c_j)} \quad (3)$$

$p(c_j)$: the probability of class c_j among set of classes C.,
$p(x_i|c_j)$: probability that term x_i occurs in class c_j which maybe zero in training data, so the Laplace smoothing is chosen to estimate it [14].

There are two classes of models that commonly used for naive Bayes classification that based on the distribution of the words in the document. The major difference between these two models is the assumption in terms of taking (or not taking) word frequencies into account, **Multinomial Model (MNB)**, the document represented with bag of words, we compute frequencies of terms in it[15]. We can calculate the probability of occurrence of each feature i in a category y as $p(x_i|y)$ [16]. Another class is **Multivariate Bernoulli Model (BNB)**, In which the presence or absence of words in a text document can be used as features to represent a document. Thus, the word features in the text are assumed to be binary[15].

B. K-Nearest Neighbor

The k-nearest neighbor algorithm (k-NN) is a statistical learning Algorithm that is based on a distance or similarity function for pairs of observations, such as the Euclidean distance or Cosine similarity measures between k training data and the tested document. the degree of similarity between documents and k training data depending on value of k [17]. The best choice of k depends upon the data; generally, larger values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct [18]. If k is greater than one, then the object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor [18].

C. Logistic Regression

Also known as Log-linear Models (linear if you take log), also known as a conditional exponential classifier or logistic regression classifier. Multinomial logistic regression is known by a variety of other names, including polytomous LR [5, 15]] multiclass LR, softmax regression, multinomial logit, maximum entropy (MaxEnt) classifier, conditional maximum entropy model.

Multinomial logistic regression modeling is a general and intuitive way for estimating a probability from data and it has been successfully applied in various natural language processing tasks [19].

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

In text classification, Multinomial logistic regression is a model which assigns a class c of each word w based on its document d in the training data D. Conditional distributed p(c|d) is computed as follows:

$$p(c|d) = 1/z(d)exp(\sum_i \alpha_i f_i(d,c)) \quad (4)$$

Where Z(d) is a normalization function which is computed as:

$$z(d) = \sum_c exp(\sum_i \alpha_i f_i(d,c)) \quad (5)$$

α is weight of term; f is feature.

D. Suggested Work

In any supervised classification problems, we should have, Set of classes C =[1], Set of labeled examples (documents) and A classifier which map any input example to a one class (or set of classes) from C.

Almost all the known supervised classifier use weights for each term in the examples for predicting the most relevant class to the input document.

In this section we suggest very simple method for text classification. This method can be summarized by selecting the highest weight of the term among the classes. Then we used only these values for prediction the best class for any input example. The used weight in this work is Tf-Idf.

Formally:

Let we have set of classes C ={c_1,c_2 … c_n}.

Suppose that:

W_{imax}: is the maximum weight of the term i among all these classes.

C_{ij}: is a binary value which indicate the class j contain the maximum weight of the term i

$$c_{ij} = \begin{cases} 1, & if\ w_{imax} \in c_j \\ 0, & otherwise \end{cases} \quad (6)$$

Then the best class can be estimated by:

x

$$c_{best} = \underset{c_j \in C}{\operatorname{argmax}} \sum_{i=1}^{m} c_{ij} * w_{imax} \quad (7)$$

Where:

m=set of words in test document.

This is can be implemented simply where each feature will be related to one class not more. In our method, feature reduction is not matter for the same mentioned reason because each class has different subset of all features set and then feature reduction inclusively used.

V. Data Set

Dataset set used in our paper collected from AlSabah newspaper[2]. It is used for first time in machine learning. The corpus files were classified manually and it consists of 16757 Arabic documents belong to five different categories (see table [1]). All the used classifiers, in this study, are supervised learning; therefore the corpus used for learning and evaluating the performance. A corpus is divided into two parts training and test data.

Dataset used for training is 15080 documents (90%) and 1677 documents is used for testing (10%). This corpus is the first one of its type from Iraqi media. It will be freely available for the researchers in NLP and specially text classification.

TABLE I. Data Set

Category	#Total documents	#training documents	# testing documents
Literature and Arts	2175	1957	218
Family and community	1017	915	102
Economy	3411	3070	341
Sport	8546	7691	855
Science and Technology	1608	1447	161
Total	**16757**	**15080**	**1677**

VI. Implementation and Result

We apply four text classification algorithms on the new Arabic Data Set . All these classifier are done after many stages from preprocessing (Tokenization, Stop-Words Removal, Stemming, Stop-Words Removal second time to reduce text file). the weight of each term is calculated using tf-idf:

$$Weight = tf*idf \quad (8)$$

The result of implement KNN, NB, LR, and suggested work is 78.903%, 80.334%, 76.937%, 81.585% respectively. Our suggested work achieved the highest percentage of the algorithms applied to the corpus.

2 http://www.alsabaah.iq/

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

VII. Conclusion And Future Work

As we can see, the work focus on: (i) suggesting simple algorithm for classification, (ii) constructing a new Arabic corpus taken from Iraqi media which can be used by researcher in the future, (iii) testing the corpus with three well-known classifier as Naïve Bayes, K-Nearest Neighbor and Logistic Regression, (iv) applying the suggested algorithm on the same corpus.

By comparing the results, our simple algorithm gives a higher accuracy but the most gain is its simplicity and lease time for classification. From the explanation of the algorithm, each feature is relative to one class not more result in each class has its own features. In other words, each class has different subset of all features set and then feature reduction inclusively used. This algorithm was tested with a new Arabic corpus which was used, also, with three well-known classifiers for comparing the results.

The suggested algorithm can't be used for multi-label classification problems without modification.

In the future, we try for combining these classifiers in master-slave technique[20, 21].

References

[1] Abu-Errub, A., Arabic Text Classification Algorithm using TFIDF and Chi Square Measurements. International Journal of Computer Applications, 2014. 93(6).

[2] Al-Badarenah, A., et al., Classifying Arabic text using KNN classifier. 2016.

[3] Bahassine, S., M. Kissi, and A. Madani. New stemming for Arabic text classification using feature selection and decision trees. in Proceedings of the 5th International Conference on Arabic Language Processing. 2014.

[4] Witschel, H.F. Using decision trees and text mining techniques for extending taxonomies. in Learning and Extending Lexical Ontologies by using Machine Learning Methods, Workshop at ICML-05. 2005.

[5] Alhutaish, R. and N. Omar, Arabic text classification using k-nearest neighbour algorithm. Int. Arab J. Inf. Technol.(IAJIT), 2015. 12: p. 190-195.

[6] Sharma, N., CLASSIFICATION USING NAÏVE BAYES-A SURVEY.

[7] El-Halees, A.M., Arabic text classification using maximum entropy. IUG Journal of Natural Studies, 2015. 15(1).

[8] Zhu, S., et al. Multi-labelled classification using maximum entropy method. in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. 2005. ACM.

[9] Duwairi, R.M., Arabic text categorization. Int. Arab J. Inf. Technol., 2007. 4(2): p. 125-132.

[10] Danso, S., E. Atwell, and O. Johnson, A comparative study of machine learning methods for verbal autopsy text classification. arXiv preprint arXiv:1402.4380, 2014.

[11] Faidi, K., et al. Comparing Arabic NLP tools for Hadith classification. in Proceedings of the 2nd International Conference on Islamic Applications in Computer Science and Technologies (IMAN'14). 2014.

[12] Alsaleem, S., Automated Arabic Text Categorization Using SVM and NB. Int. Arab J. e-Technol., 2011. 2(2): p. 124-128.

[13] Nisha Mariam Daniel, a.K.K., Survey on Text Classification Methods. International Journal of Advanced Research in Computer Science and Software Engineering, 2016. 6: p. 585--588.

[14] Shen, D., et al. Web-page classification through summarization. in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. 2004. ACM.

[15] Aggarwal, C.C. and C. Zhai, A survey of text classification algorithms. Mining text data, 2012: p. 163-222.

[16] Usman, M., et al., Urdu Text Classification using Majority Voting. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, 2016. 7(8): p. 265-273.

[17] Khan, A., et al., A review of machine learning algorithms for text-documents classification. Journal of advances in information technology, 2010. 1(1): p. 4-20.

[18] Saad, M.K., The impact of text preprocessing and term weighting on arabic text classification. Gaza: Computer Engineering, the Islamic University, 2010.

[19] Jain, A. and J. Mandowara, Text Classification by Combining Text Classifiers to Improve the Efficiency of Classification. International Journal of Computer Application (2250-1797), 2016. 6(2).

[20] Aliwy, A.H., Arabic morphosyntactic raw text part of speech tagging system. 2013.

[21] Aliwy, A.H. Combining POS taggers in master-slaves technique for highly inflected languages as Arabic. in Cognitive Computing and Information Processing (CCIP), 2015 International Conference on. 2015. IEEE

AUTHORS PROFILE

Dr. Ahmed H. Aliwy prof in Collage of CS and Mathematic, University of Kufa, Iraq, Text Mining.

Dr. Kadhim S. Aljanabi prof in Collage of CS and Mathematic, University of Kufa, Iraq, Data Mining.

Zena AbdAlRetha AboAltaheen MSC student in , University of Kufa, Iraq.

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

Conducting Security Metrics for Object-Oriented Class Design

Dujan B. Taha
Dept. of Software Engineering
College of Computer Sc. & Math, University of Mosul.
Mosul, Iraq.

Osamah S. Mohammed
Dept. of Software Engineering
College of Computer Sc. & Math, University of Mosul.
Mosul, Iraq.

***Abstract*—Security issues often neglected until coding step in software development process, and changing in this step leads to maximize time and cost consuming depending on the size of the project. Applying security on design phase can fix vulnerabilities of the software earlier in the project and minimize the time and cost of the software by identifying security flaws earlier in the software life cycle. This work concerns with discussing security metrics for object oriented class design, and implementing these metrics from Enterprise Architect class diagram using a proposed CASE tool.**

Keywords-Software Engineering; Security metrics; Class design; SDLC; Design phase

I. INTRODUCTION

The increasing number of attacks has forced the organizations to integrate the characteristic of security during development process rather than considering during post development phase. It is estimated that 90 percent of reported security incidents result from exploits against defects in the design or code of software [1]. Software engineering on the other hand often take a long time to process and to complete, and any change in the software may lead to changes on early phases of the process which may take a lot of time and increases the costs.

Assessing security earlier in the software life cycle will minimize the time and cost of the software by identifying security flaws at design phase of software life cycle.

Willoughby said "You must think about security, reliability, availability, dependability at the beginning, in the design, architecture, test and coding phases, all through the software life cycle" [2]. Security must be designed and built into a system from the ground up. According to the CERT Coordination Center (CERT/CC) of the SEI, more than 90% of reported security incidents are the result of exploits against defects in the design or development flaws. Metrics can help to detect and analyze the software functionality and correct them during the software development process[3].

In this work, we discuss the security metrics for object-oriented class that is applied on design phase of the software life cycle. Identifying security flaws at this phase could help reducing them before software coding begins. If security is applied in design phase, it could minimize the vulnerabilities that can appear in coding or after releasing the software.

A proposed CASE tool has been used to conduct the security metrics from a class diagram designed using Enterprise Architect.

II. BACKGROUND

Security metrics are being widely used to measure the security level and try to fix any vulnerabilities, but most of these metrics were used at source coding which is considered late in the SDLC [4]. Chowdhury, Chan, Zulkernine[5], proposed a set of code-level security metrics which can assess a security level of a source code segment, and it can provide guidelines to improve the security level of that source code. Wang [1] proposed a new approach to define software security metrics based on vulnerabilities included in the software systems and their impacts on software quality. Kumar and Alagarsamy[6] proposed a set of security metrics to identify the most vulnerable parts of a Java source code at the level of methods. Gandhi et. Al [7] proposed a CASE tool that help the developer or the designer to calculate specific class security metrics from a class diagram designed in Rational Rose, and the tool was written in Java. Agrawal and Khan [8] proposed a new security metrics development framework which can in theory assess the security metrics of a software earlier in the software development life cycle, its security metrics collection consist of various type of metrics which may require number of software build or versions for it to be calculated. Smriti Jain and Maya Ingle[9] suggested a set of metrics based on security issues of software development from requirement phase to documentation phase. Saarela [10] measured software security from design phase, he reviewed the security metrics for assurance. His study reveals that these security metrics proposed by Alshammari were succeeded in discovering the current status of software security metrics and establishing the level of current understanding of secure software.

Security metrics was defined based on the quality properties specified by Bansiya[11]. The Metrics has a scale to fit with the range 0 to 1, with lower value indicating that the program's design is more secure. To apply these metrics to a given class, the class should be annotated with UMLsec and SPARK's annotations as shown in **Error! Reference source not found.**.

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

Enterprise Architect (EA) is a UML CASE tool used for analysis and design of software. It's used to cover all aspects of software including business, systems modeling and design, [12]. Designing automated security metrics tool is relatively easy, because EA can export a diagram as an (.XML) file which can be easily interpreted for use of automated processing on that diagram. The XML file contains all the details needed to elicit and calculate security metrics using XML file parser.

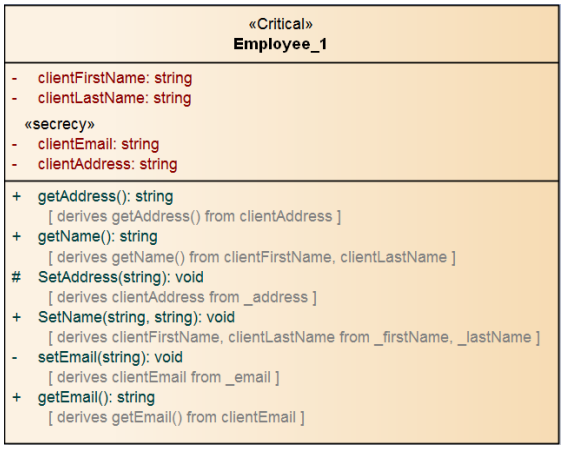


Figure 1, Class Design Structure.

III. SECURIY METRICS

A metric is a value that used to study some aspect of a project. Sometimes a metric is the same as an attribute. After calculating the metrics, study them to see if any of them are good indicators of the project's future. Two thing can be done with these indicators; first, it can be used to predict the future of the current project; the second thing can be done with indicators is making strategy improvements for future projects[13].

Some terminologies associated with security metrics are defined as follows [14][15] and shown in Figure II:

- Classified Attribute: An attribute which is defined in UMLsec as 'secrecy'.
- Instance Attribute: An attribute whose value is stored separately by each instance of a class.
- Class Attribute: An attribute whose value is shared by all instances of that class.
- Classified Method: A method which reads or writes at least one classified attribute.
- Mutator: A method that sets the value of an attribute.
- Accessor: A method that reads the value of an attribute.
- Classified Mutator: A method that sets the value of at least one classified attribute.
- Classified Accessor: A method that reads the value of at least one classified attribute.

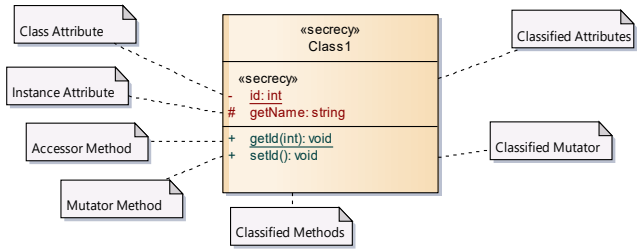


Figure II, Metric Terminology

Security metrics aim to measure the security metrics from two perspective of information flow: least privilege, and reducing attack surface.

The principle of least privilege states that an object should be given only those privileges that it needs in order to complete its task [11]. The main advantage of this principle is to minimize the interaction between privileged programs. In class design, this means the method do the least possible actions is the most secure. In this case, the class whose methods interact with the least possible classified attributes would be a secure design with respect to this principle.

The principle of reduce attack surface aims to limit the access to secret data. In class design, a class should have the least possible accessible methods each with the least number of parameters which can affect classified attributes needed for necessary tasks.

Security metrics are divided into two groups: One concerned with measuring accessibility level of attributes and methods; and another concerned with measuring the interaction level of methods attributes [14].

3.1 Security Accessibility Metrics

These metrics aim to measure the accessibility level of attributes and methods in a class from an access modifier perspective. These metrics are as follow:

A. Classified Instance Data Accessibility (CIDA)

These metric measures the direct accessibility of classified instance attributes of a particular class. It is defined as "The ratio of the number of classified instance public attributes to the number of classified attributes in a class". Higher values indicate higher accessibility to these classified attributes and hence a larger 'attack surface'.

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

Consider a set of classified attributes in a class C as CA = {ca1, ..., can} and its classified instance public attributes CIPA = {cipa1, ..., cipan} such that CIPA ⊆ CA.

$$CIDA(C) = \frac{|CIPA|}{|CA|} \quad (1)$$

B. Classified Class Data Accessibility (CCDA)

This metric measures the direct accessibility of classified class attributes of a particular class. It is defined as follows: "The ratio of the number of classified class public attributes to the number of classified attributes in a class". The result shows the ratio of classified class attributes which are directly accessible from outside its class. Higher values mean that confidential data of that class has a higher chance of being exposed to unauthorized parties. The lower the value of this metric enforce the security principle of reducing the attack surface.

Consider a set of classified attributes in a class C as CA = $\{ca_1, ..., ca_n\}$ and its classified class public attributes CCPA = $\{ccpa_1, ..., ccpa_n\}$ such that CCPA ⊆ CA.

$$CCDA(C) = \frac{|CCPA|}{|CA|} \quad (2)$$

C. Classified Operation Accessibility (COA)

This metric measuresthe ratio of the accessibility of public classified methods of a particular class. It is defined as: "The ratio of the number of classified public methods to the number of classified methods in a class". It aims to protect the internal operations of a class which interact with classified attributes from direct access. Lower values of this metric would reduce potential information flow of classified data which could be caused by calling public methods. This metric measures the potential attack surface size exposed by classified methods.

Consider a set of classified methods in a class C as CM = $\{cm_1, ..., cm_n\}$ and its classified class public methods CCPM = $\{ccpm_1, ..., ccpm_n\}$ such that CCPM ⊆ CM.

$$CCDM(C) = \frac{|CCPM|}{|CM|} \quad (3)$$

3.2 Security Interaction Metrics

These metrics measure the impact of class interaction between methods and attributes in respect of security principles. These metrics are as follow.

A. Classified Mutator Attribute Interactions (CMAI)

This metric measures the interactions of mutators with classified attributes in a class. It is defined as: "The ratio of the number of mutators which may interact with classified attributes to the possible maximum number of mutators which could interact with classified attributes". Higher interaction means stronger cohesion between mutators and classified attributes within a given class. With regard to the security principles, a lower value allows fewer privileges over confidential data and therefore adheres to the least privilege principle.

Consider a set of mutator methods in class C as MM_i, i ∈{1, ..., mm} and the classified attributes CA_j, j ∈ {1, ..., ca}. Let's $\alpha(CA_j)$ be the number of mutator methods which may access classified attribute CA_j. Then, the CMAI for class C can be calculated as follow:

$$CMAI(C) = \frac{\sum_{j=1}^{ca} \alpha(CAj)}{|MM| \times |CA|} \quad (4)$$

B. Classified Accessor Attribute Interactions (CAAI)

This metric measures the interactions of accessors with classified attributes in a class. It is defined as: "The ratio of the number of accessors which may interact with classified attributes to the possible maximum number of accessors which could have access to classified attributes". Higher interaction means stronger cohesion between accessors and classified attributes within a given class. Weak cohesion indicates fewer privileges are given to accessors over classified attributes.

Consider a set of accessor methods in class C as AM_i, i ∈ {1, ..., am} and the classified attributes CA_j, j ∈ {1, ..., ca}. Let's $\beta(CA_j)$ be the number of accessor methods which may access classified attribute CA_j. Then, the CAAI for class C can be calculated as follow:

$$CAAI(C) = \frac{\sum_{j=1}^{ca} \beta(CAj)}{|AM| \times |CA|} \quad (5)$$

C. Classified Attributes Interaction Weight (CAIW)

This metric is defined to measure the interactions with classified attributes by all methods of a given class. We define this metric as: "The ratio of the number of all methods which may interact with classified attributes to the total number of all methods which could have access to all attributes". The higher the value of this metric for a given class the more privileges are given to this class' methods over classified attributes, and therefore the less that class adheres to the security principle of least privilege.

Consider a set of attributes in class C as Ai, i ∈ {1, ..., a} and a set of classified attributes CA_j, j ∈ {1, ..., ca}. Let's $\gamma(CA_j)$ be the number of methods which access classified attribute CA_j. Let $\delta(Ai)$ be the number of methods which may access attributes A_j. Then, CAIW can be computed as:

$$CAIW(C) = \frac{\sum_{j=1}^{ca} \gamma(CAj)}{\sum_{i=1}^{a} \delta(Ai)} \quad (6)$$

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

D. Classified Methods Weight (CMW)

This metric is defined to measure the weight of methods in a class which potentially interact with any classified attributes in a particular class. It is defined as: "The ratio of the number of classified methods to the total number of methods in a given class". Higher values of this metric indicate more classified operations are offered by the given class. This leads to a higher chance of information flow of classified data by calling the class's methods and violations of the security principle of reducing the attack surface.

Consider a set of methods in class C as $M = \{m_1, \ldots, m_n\}$ and the classified methods as $CM = \{cm_1, \ldots, cm_n\}$ such that $CM \subseteq M$. CMW can be calculated as:

$$CMW(C) = \frac{|CM|}{|M|} \quad (7)$$

IV. Proposed CASE Tool

The proposed tool is a webapp CASE tool used to calculate specific-class and multi-class security metrics, which can be accessed with any device which contains an internet connection and a web browser. Users can register and access the functionalities of this tool, such as: uploading an XML file exported from EA to process and save the results into the database; and viewing the results of the diagrams. This tool can process a class diagram exported from EA as an XML file, and the class diagram has to be annotated with UMLsec and SPARK's annotations. The architecture of this tool is shown in **Error! Reference source not found.**. New users can register and login as a regular user, the user can view the last uploaded diagrams metrics which consist of design security metrics and specific class metrics. The tool will help the user to calculate, view, and easily compare these classes to select the most secure class. The use of this tool can reduce the total cost of a project by reducing the number of vulnerabilities discovered in the design phase, which reduces a large number of potential security vulnerabilities and flaws that may be discovered late in the project.[4].

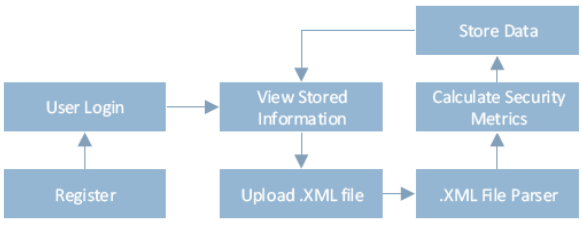


Figure 3, Proposed tool Architecture.

V. Case Study

The following case study illustrate security metrics for a single class using the proposed webapp CASE tool to elicit the design data from XMI file exported from Enterprise Architect diagram. The diagram must include UMLsec and SPARK's annotations in each class to help the application calculate security metrics.

This case study will consist of different designs of a single class which has a confidential information. The class diagram shows a different designs of a *User* class. This *User* class holds information of a user in the system. Its attributes consist of *userFullName* which holds the user full name, *userEmail* holds the user personal email address, *userPassword* holds the user own password that are used to enter the system, and *userTelephone* holds the user personal telephone number. The class's methods are used to access or mutate its attributes. User email, password, and telephone number are meant to be secret. Each classified class are labeled "Critical" in its stereotype field, all confidential attributes are labeled "secrecy" in its stereotype field and all method's behavioral are written in its behavior field.

As shown below, a few designs of the class *User* have been made to study the result of it in respect to security. **Figure** (a) shows the class User_1 which has all of its attributes and their methods set as public '+'.

Figure (b) shows the class User_2 has its attributes accessibility changed to private '-' unlike User_1 which is public.

Figure (c) shows User_3 class simillar to User_2 but it's methods are declaired private and added two extra methods to access and mutate the class attributes through its methods.

Figure (d) shows User_4 class, similar to User_3 but it has only two methods to access and mutate the attributes in the class and the rest of the methods was deleted to make the access/mutate point to the class through one method only.

User_5 class is shown in **Figure** (e), it has only one confidential attribute which has its own mutator and accessor method. The rest of the attributes has their accessor and mutator method.

VI. Results and Conclusion

Table 1 shows results after applying security metrics. The lower value of each metric are considered more secure.

Figure shows results on a radio chart to make it easier to compare between different classes.

User_1 class is determined as most insecure class regarding classified instance data accessibility metric (CIDA) since it's classified attributes are all declared as public, the other classes has the same level of security because they have their classified attributes declared as private. Regarding classified class data attributes metric (CCDA), all classes have the same results since they don't have any classified class data attributes.

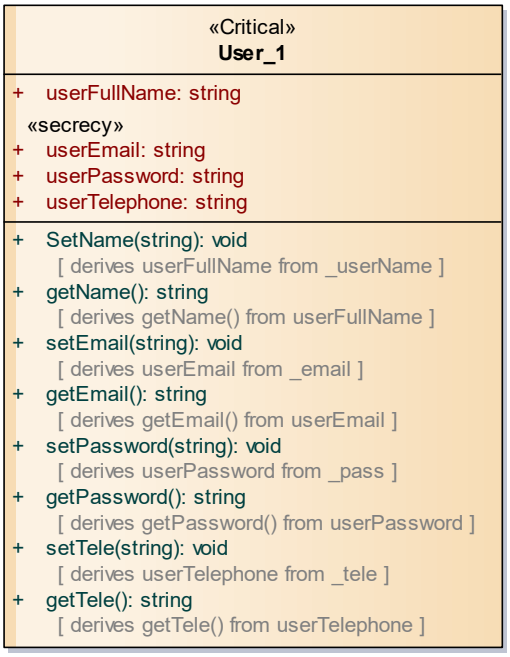
Class User_3 is considered the most secure regarding to classified operation accessibility (COA) since it has only two out of ten classified methods declared as public, the rest are

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

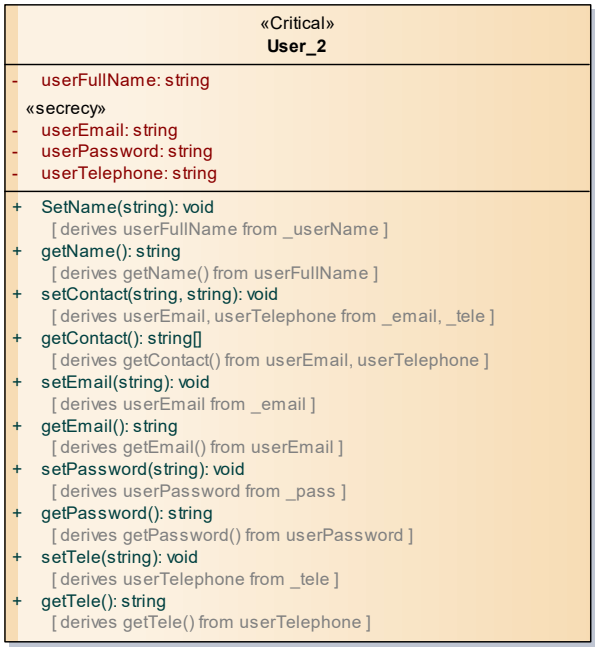
declared as private. Other class designs are considered insecure because all classified methods are declared as public.

In terms of classified mutator attribute interactions (CMAI), class User_1 has the most secure design since it has less connections between classified mutator methods and classified attributes and the classified mutator methods interact with less number of classified attributes. Sam results regarding classified accessor attribute interactions (CAAI), User_1 is the most secure design because it has the least interaction of classified attributes with classified accessor methods.

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

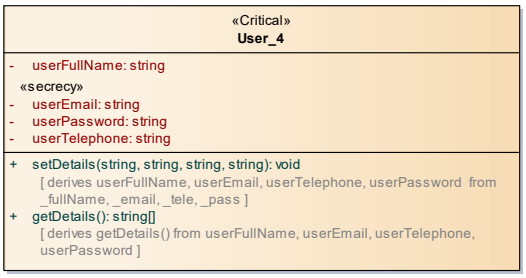


(a) User 1 Class.

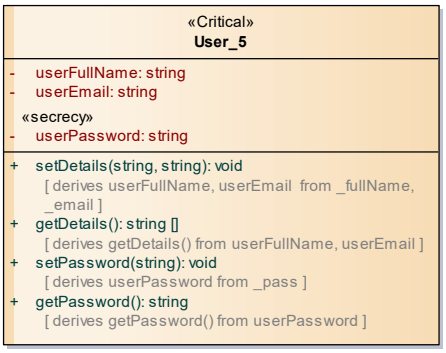


(b) User 2 Class.

(c) User 3 Class.



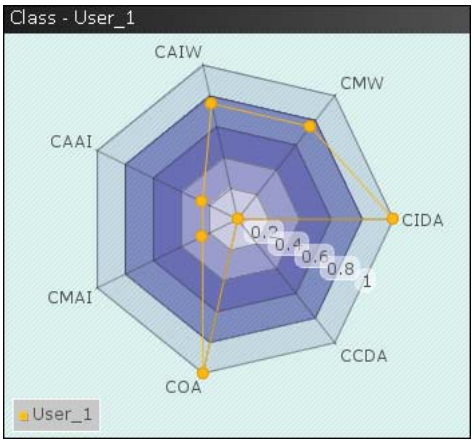
(d) User 4 Class.



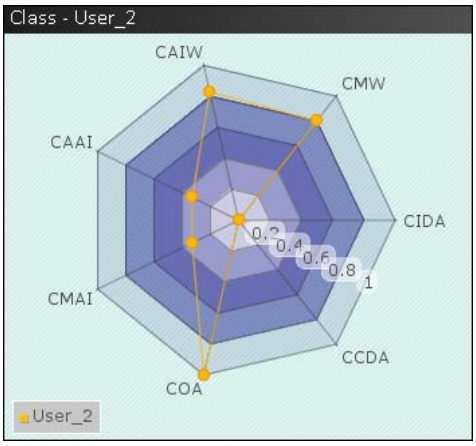
(e) User 5 Class.

Figure 4, User Classes Samples

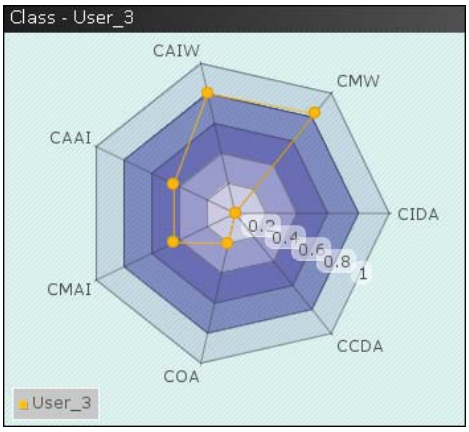
https://sites.google.com/site/ijcsis/
ISSN 1947-5500



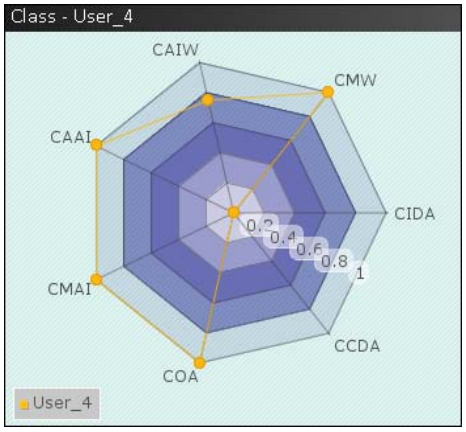
(a) User 1 Metric Results.



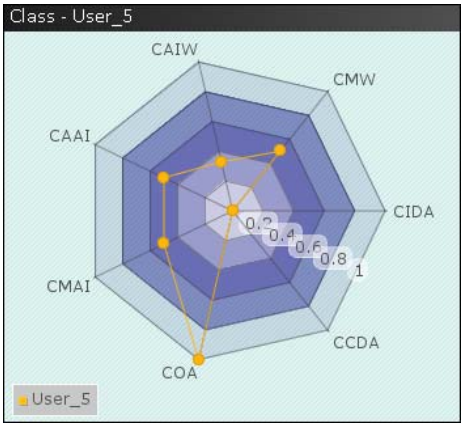
(b) User 2 Metric Results



(c) User 3 Metric Results.



(d) User 4 Metric Results.



(e) User 5 Metric Results.

Figure 5, Class Security Metrics Radio Chart Results.

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

Table 1, Class Security Metrics Results

Class Name	CIDA	CCDA	COA	CMAI	CAAI	CAIW	CMW
User_1	1	0	1	0.25	0.25	0.75	0.75
User_2	0	0	1	0.33	0.33	0.83	0.8
User_3	0	0	0.2	0.44	0.44	0.8	0.83
User_4	0	0	1	1	1	0.75	1
User_5	0	0	1	0.5	0.5	0.33	0.5

User_5 class design is considered the most secure design regarding classified attribute interaction weight (CAIW) since its classified attributes has less interaction with its methods. User_5 class design is considered a secure design regarding classified methods weight (CMW) since the number of classified methods is less than the total number of methods in this class.

After calculating the results of these metrics, we can see that User_5 class is the most secure class compared to other classes since it has the most overall low value results. User_2 and User_4 can be considered as insecure designs since they have high value results compared to other classes.

VII. References

[1] J. A. Wang, H. Wang, M. Guo, and M. Xia, "Security metrics for software systems," *ACM Southeast Reg. Conf.*, p. 1, 2009.

[2] M. Willoughby, "Q&A: Quality software means more secure software," 2005. [Online]. Available: http://www.computerworld.com/article/2563708/security0/q-a--quality-software-means-more-secure-software.html.

[3] N. R. Mead and G. Mcgraw, "A portal for software security," *IEEE Secur. Priv.*, vol. 3, no. 4, pp. 75–79, 2005.

[4] O. S. Mohammed and D. B. Taha, "Conducting multi-class security metrics from Enterprise Architect class diagram," *Int. J. Comput. Sci. Inf. Secur.*, vol. 14, no. 4, pp. 56–61, 2016.

[5] I. Chowdhury, B. Chan, and M. Zulkernine, "Security metrics for source code structures," *Proc. fourth Int. Work. Softw. Eng. Secur. Syst.*, no. October, pp. 57–64, 2008.

[6] S. R. K. T, "A Method Level Security Metrics Suite for Java Programs," *Int. J. Comput. Technol. Appl.*, vol. 3, no. 6, pp. 1991–1996, 2012.

[7] S. H. Gandhi, D. R. Anekar, M. A. Shaikh, and A. A. Salunkhe, "Security Metric for Object Oriented Class Design- Result Analysis," *Int. J. Innov.*

[8] a. Agrawal, "Software Security Metric Development Framework (An Early Stage Approach)," *Am. J. Softw. Eng. Appl.*, vol. 2, no. 6, p. 150, 2013.

[9] S. Jain and M. Ingle, "Security Metrics and Software Development Progression," *J. Eng. Res. Appl.*, vol. 4, no. 5, pp. 161–167, 2014.

[10] M. Saarela, "Measuring software security from the design of software Department of IT," 2016.

[11] J. Bansiya and C. G. Davis, "A Hierarchical Model for Object-Oriented Design Quality Assessment," *IEEE Trans. Softw. Eng.*, vol. 28, no. 1, pp. 4–17, 2002.

[12] G. Spark, D. O 'bryan, S. Mcneilly, N. Capey, J. Redfern, B. Maxwell, V. Kumar, H. Britten, and S. Meagher, "Enterprise Architect User Guide," p. 2888, 2012.

[13] R. Stephens, *Beginning Software Engineering*, vol. 1. Wrox Press Ltd., 2015.

[14] B. Alshammari, C. Fidgeand, and D. Corney, "Security metrics for object-oriented class designs," *Proc. - Int. Conf. Qual. Softw.*, pp. 11–20, 2009.

[15] B. Alshammari, C. Fidge, and D. Corney, "Security metrics for object-oriented designs," *Proc. Aust. Softw. Eng. Conf. ASWEC*, pp. 55–64, 2010.

AUTHORS PROFILE

Dr. Dujan B. Taha (Assistant Prof.) is currently a lecturer at Mosul University, College of Computer Science and Mathematics / Software Engineering Department. She received B.Sc. degree in Computer Science / University of Mosul in 1991, M.Sc. degree / University of Mosul in 1996 and Ph.D. degree / University of Mosul in 2005. Her research interests are in information and network security, Software Engineering, Image processing and pattern recognition.

Osama S. Mohammed is currently an M.Sc. student in Software Engineering Department / Collage of Computer Science and Mathematics / University of Mosul.

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

Integration of Principal Component Analysis and Support Vector Regression for Financial Time Series Forecasting

Utpala Nanda Chowdhury[1]*, Md. Abu Rayhan[1,2], Sanjoy Kumar Chakravarty[1], Md. Tanvir Hossain[1]
[1]Dept. of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh
[2]ICB Capital Management Limited, Dhaka, Bangladesh
*unchowdhury@gmail.com

***Abstract*–Financial time series forecasting has received tremendous interest by both the individual and institutional investors and hence by the researchers. But the high noise and complexity residing in the financial data makes this job extremely challenging. Over the years many researchers have used support vector regression (SVR) quite successfully to conquer this challenge. As the latent high noise in the data impairs the performance, reducing the noise could be effective while constructing the forecasting model. To accomplish this task, integration of principal component analysis (PCA) and SVR is proposed in this research work. In the first step, a set of technical indicators are calculated from the daily transaction data of the target stock and then PCA is applied to these values aiming to extract the principle components. After filtering the principal components, a model is finally constructed to forecast the future price of the target stocks. The performance of the proposed approach is evaluated with 16 years' daily transactional data of three leading stocks from different sectors listed in Dhaka Stock Exchange (DSE), Bangladesh. Empirical results show that the proposed model enhances the performance of the prediction model and also the short-term prediction gains more accuracy than long-term prediction.**

Keywords: Financial time series forecasting; Support vector regression; Principal component analysis; Dhaka Stock Exchange

I. INTRODUCTION

Along with the money market, capital market is also a vital component of the financial sector of any country in the world. The direct influence on people's life of this sector is increasing day by day. Some are investing their savings in this market for profits while some are directly related to this field. The investors are very eager to have a closer insight about the underlying patterns of this market. In fact, most of them now currently depend on Intelligent Trading Systems for predicting stock prices based on various conditions. Accuracy of these prediction systems is essential to make better investment decisions with minimum risk factors. That's why, the financial time series forecasting has gained extreme attention from both the individual and institutional investors. But, this field is characterized by data intensity, noise, non-stationary, unstructured nature, high degree of uncertainty, and hidden relationships [1]. Capital market trend depends on many factors including political events, general economic conditions, news related to the stocks and traders' expectations. Moreover, according to academic investigations, movements in market prices are not random. Rather, they behave in a highly non-linear, dynamic manner [2]. Therefore, predicting stock market price is a quite challenging task.

Technical analysis is a popular approach to study the capital market patterns and movement. The results of technical analysis may be a short or long-term forecast based on recurring patterns; however, this approach assumes that stock prices move in trends, and that the information which affects prices enters the market over a finite period of time, not instantaneously [3]. Technical indicators used in this analysis are calculated from the historical trading data. Researchers use various machine learning and artificial intelligent approaches to analyze these technical indicators to predict future trends or prices. The traditional statistical models include Box Jenkins ARIMA [4]. Continuous research has introduced plentiful approaches including Artificial Neural Networks (ANN), genetic algorithm, rough set (RS) theory, fuzzy logic and others [5-6]. The successful application of Support Vector Regression (SVR) in various time series problems has encouraged its adaptation in financial time series forecasting [7]. But the latent noise of financial time series data often leads to over-fitting or under-fitting and hence impairs the performance of the forecasting system. Over the years, several methods were proposed to negate the influence of such noisy data by detecting and removing those before applying prediction model. Lu et al. has proposed the use of independent component analysis (ICA) (both linear and non-linear) with SVR to elevate the forecasting accuracy [8-9]. In both approaches, at first the ICA was used to extract the most influential components from the technical indicators and then were fed to SVR for a better prediction. Cao et al. in [10] has shown that another method called principal component analysis (PCA) can improve the performance of support vector machine (SVM) in time series forecasting. Grigoryan proposed using PCA as a preprocessing tool in financial time series forecasting with ANN [11].

In this study, we proposed the integration of PCA in the financial time series forecasting model using SVR. Considering the fact that, technical analysis plays a vital role in the forecasting, it has been conducted to calculate technical indicators as the input features. Then PCA is used to extract the influential components from input features which are then filtered to transform the high-dimensional input into low-dimension features. The SVR then finally use the filtered low-

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

dimensional input variables to construct the forecasting model and predict future prices.

The reminder of this paper is organized into 6 sections. Section 2 provides a brief overview of the methodologies used in this study which includes PCA and SRV. Section 3 introduces the proposed method. Section 4 describes the research data. Section 5 reports the experimental results obtained from the study. Finally section 6 contains the concluding remarks.

II. METHODOLOGY

A. Principal *Component* Analysis (PCA)

Principal component analysis (PCA), invented by Karl Pearson [12], is a well-known statistical procedure for feature extraction. It finds smaller number of uncorrelated components from high dimensional original inputs by calculating the eigenvectors of the covariance matrix. Given a set of m dimensional input vectors $\boldsymbol{x}_i=(x_i(1),x_i(2),.....x_i(m))^T$ where $i=1,2,\ldots\ldots n$. PCA is a transformation of $\boldsymbol{x}_i$ into a new vector $\boldsymbol{y}_i$ by:

$$\boldsymbol{y}_i = U^T\boldsymbol{x}_i \qquad (1)$$

Where U is the $m \times m$ orthogonal matrix whose jth column u_j is the jth eigenvector of the sample covariance matrix $C = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i{}^T$. In other words, PCA solves the eigenvalue problem of equation (2).

$$\lambda_j u_j = Cu_j,\ j = 1, 2, \ldots m \qquad (2)$$

where λ_j is one of the eigenvalues of C. u_j is the corresponding eigenvector. Based on the estimated u_j, the components of $\boldsymbol{y}_i$ are then calculated as the orthogonal transformation of $\boldsymbol{x}_i$:

$$\boldsymbol{y}_i(j) = u_j{}^T\boldsymbol{x}_i,\quad j = 1, 2, \ldots m \qquad (3)$$

The new components are called principal components. By using only the first several eigenvectors sorted in descending order of the eigenvalues, the number of principal components in $\boldsymbol{y}_i$ can be reduced [13].

B. Support Vector Regression (SVR)

The SVR extends the basic principles of Vapnik's support vector machines (SVM) for classification [14] by setting a margin of tolerance ϵ in approximation and up until the threshold ε, 0 error is considered. Given a training set $(\boldsymbol{x}_i,y_i)$, $i=1,2, \ldots n$, where the $\boldsymbol{x}_i \in R^m$ is the m-dimensional input vector and $y_i \in R$ is the response variable. SVR generates the linear regression function in the form:

$$f(\boldsymbol{x},\boldsymbol{w})=\boldsymbol{w}^T\boldsymbol{x} + b \qquad (4)$$

Vapnik's linear ε-Insensitivity loss (error) function is:

$$|y - f(\boldsymbol{x},\boldsymbol{w})|_\varepsilon = \begin{cases} 0, & if\ |y-f(\boldsymbol{x},\boldsymbol{w})|\le\varepsilon \\ |y-f(\boldsymbol{x_i},\boldsymbol{w})|-\varepsilon, & otherwise \end{cases} \qquad (5)$$

Based on this, linear regression $f(\boldsymbol{x},\boldsymbol{w})$ is estimated by simultaneously minimizing $\|\boldsymbol{w}\|^2$ and the sum of the linear ε-Insensitivity losses as shown in equation (7). The constant c controls a trade-off between an approximation error and the weight vector norm $\|\boldsymbol{w}\|$, is a design parameter chosen by the user.

$$R = \frac{1}{2}||w||^2 + c\left(\sum_{i=1}^{n}|y - f(x,w)|_\varepsilon\right) \qquad (6)$$

Minimizing the risk R is equivalent to minimizing the following risk under the constraints mentioned in equations (8) – (10).

$$R = \frac{1}{2}||w||^2 + c\sum_{i=1}^{n}(\xi + \xi^*) \qquad (7)$$

$$(\boldsymbol{w}^T\boldsymbol{x}_i + b) - y_i \le \varepsilon + \xi_i \qquad (8)$$

$$y_i - (\boldsymbol{w}^T\boldsymbol{x}_i + b) \le \varepsilon + \xi_i^* \qquad (9)$$

$$\xi_i, \xi_i^* \ge 0, i = 1, 2, \ldots m \qquad (10)$$

Here, ξ_i and ξ_i^* are slack variables, one for exceeding the target value by more than ε and other for being more than ε below the target. As used in SVM, the above constrained optimization problem is solved using Lagrangian theory and the Karush-Kuhn-Tucker conditions to obtain the desired weight vector of the regression function.

SVR maps the input vectors $x_i \ R^m$ into a high dimensional feature space $\phi(\boldsymbol{x}_i) \in H$. A kernel function $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ performs the mapping ϕ(.). The most popular kernel function that is used in this study is Radial Basis Function (RBS) as shown in equation (11).

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = exp\ (-\gamma\ ||\boldsymbol{x}_i-\boldsymbol{x}_j||^2) \qquad (11)$$

where γ is the constant of the kernel function. The RBF kernel function parameter γ and regularization constant C are the design parameters of SVR.

III. PROPOSED PCA-SVR FORECASTING MODEL

In this study, a prediction model named PCA-SVR is implemented which integrates PCA with SVR to forecast financial time series. At first technical analysis is conducted on the training dataset and 29 technical indicators are calculated. Some important technical indicators and their formulas are shown in Table 1. All values of these constructed features are scaled into the range of [-1, 1] to eliminate the biasness towards larger value attributes. Then PCA is applied to the normalized data to extract the principal components (PCs) containing the most influential information. These PCs are filter according to the corresponding variance and thus the irrelevant features are discarded. Finally the selected PCs are used to construct the SVR prediction model. As mentioned earlier, the RBF kernel function is incorporated in this study. But the performance of SVR is highly influenced by the selection of the parameters: γ and C. A very popular method to select the best values of these parameters is the grid search approach with cross-validation [15]. This is a straightforward method of trying geometric sequences for the best (C, γ) value pair. The (C, γ) value pair generating the minimum mean absolute percentage error (MAPE) is considered the best values for the parameters. The complete grid search is a time-consuming task. That's why a coarse grid is used at first which identifies a better region on the grid. Then a finer grid search is conducted on that region.

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

TABLE I. IMPORTANT TECHNICAL INDICATORS AND THEIR FORMULAS.

ID	Feature	Description	Calculation Formula
1	10-day SMA	Simple 10-day moving average	$(\frac{1}{n})\sum_{i=t-n+1}^{t} C_i$, where C_t is the closing price.
2	10-day EMA	Exponential 10-day moving average	$EMA(n)_{t-1}+\propto\times(C_t - EMA(n)_{t-1})$, where α is a smoothing factor and $\propto=\frac{2}{n+1}$
3	10-day WMA	Weighted 10-day moving average	$\frac{(n)C_t + (n-1)C_{t-1} + \cdots + C_{t-n+1}}{n+(n-1)+\cdots+1}$
4	A/D Oscillator	Accumulation/ distribution oscillator. It is a momentum indicator that associates changes in price.	$\frac{H_t - C_{t-1}}{H_t - L_t}$, where Lt is the low price and This the high price at time t.
5	MACD	Moving Average Convergence/ Divergence.	$MACD(n)_{t-1} + \frac{2}{n+1} \times (DIFF_t - MACD(n)_{t-1})$, where $DIFF_t = EMA(12)_t - EMA(26)_t$
6	Stochastic K%	Stochastic %K. It compares where a security's price closed relative to its price range over a given period.	$\frac{C_t - LL_{t-(n-1)}}{HH_{t-(n-1)} - LL_{t-(n-1)}} \times 100$, where LLt and HHt mean lowest low and highest high in the last t days, respectively
7	Stochastic D%	Stochastic %D. Moving average of %K.	$\frac{\sum_{i=0}^{n-1} K\%_{t-i}}{10}$
8	Momentum (close price)	It measures the amount that a security's price has changed over a given time span.	$C_t - C_{t-9}$
9	Larry William's R%	Larry William's R%. It is a momentum indicator that measures overbought/ oversold levels.	$\frac{H_n - C_t}{H_n - L_n} \times 100$
10	Relative Strength Index (RSI)	Relative Strength Index. It is a price following an oscillator that ranges from 0 to 100.	$100 - \frac{100}{1+(\sum_{i=0}^{n-1} Up_{t-i}/n)/(\sum_{i=0}^{n-1} Dw_{t-i}/n)}$, where Upt means upward-price-change and Dwt means downward-price-change at time t.
12	Close price ROC	Price rate-of-change. It displays the difference between the current price and the price of n days ago.	$\frac{C_t}{C_{t-n}} \times 100$
13	CCI	Commodity Channel Index. It measures the variation of a security's price.	$\frac{(M_t - SM_t)}{(0.015 D_t)}$, where $M_t = (H_t + L_t + C_t)/3$, $SM_t = \frac{\sum_{i=1}^{n} M_{t-i+1}}{n}$ and $D_t = \frac{\sum_{i=1}^{n} \lvert M_{t-i+1} - SM_t \rvert}{n}$
14	Disparity 5	5-day disparity. It means the distance of current two moving averages of a stock's price.	$\frac{C_t}{MA_5} \times 100$
15	Disparity 10	10-day disparity. It means the distance of current two moving averages of a stock's price.	$\frac{C_t}{MA_{10}} \times 100$
16	OSCP	Price oscillator. It displays the difference between two moving average of a stock's price.	$\frac{MA_5 - MA_{10}}{MA_5}$

To evaluate the performance of the proposed model Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), relative Root Mean Squared Error (rRMSE) and Mean Squared Error (MSE) are used. Formulas of these evaluation measures are shown in equations (12) – (15).

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \frac{|A_t - F_t|}{|A_t|} \times 100 \quad (12)$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n} \frac{|A_t - F_t|}{|A_t|} \quad (13)$$

$$rRMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \left(\frac{A_t - F_t}{A_t}\right)^2} \quad (14)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (A_t - F_t)^2 \quad (15)$$

where A_t is the actual value and F_t is the predicted value. To calculate all these measures, a 5-fold cross validation method is used. These are the measures of deviation between actual and predicted prices. The prediction model should produce lower values for all four measures.

IV. RESEARCH DATA

To conduct the study and evaluate the performance of the proposed approach, the 16 years' historical data of daily transaction for the time period from January 2000 to December 2015 are collected from Dhaka Stock Exchange (www.dsebd.org). This data covers more than 3600 trading days and each data comprises five attributes: open price, high price, low price, close price and trade volume. We have considered three companies from three different sectors: *Square Pharmaceuticals Limited, AB Bank Limited* and *ACI Limited* as these are the most prominent stocks in DSE. The daily closing prices of these companies are shown in Fig. 1, Fig. 2 and Fig. 3 respectively. The first one is a leading company in pharmaceuticals sector, the second leads the banking sector and the last one belongs to the chemical sector.

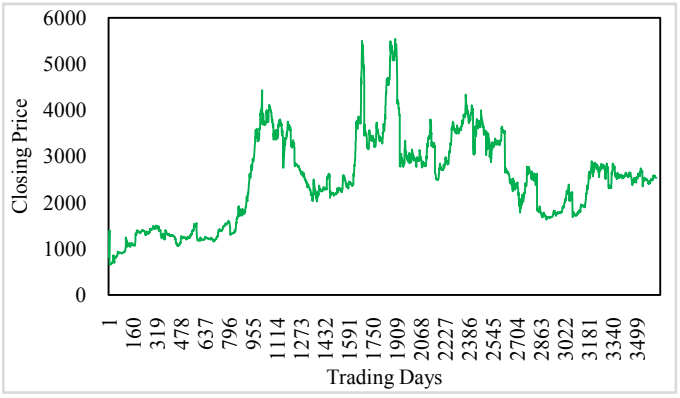


Figure 1. Closing prices of *Square Pharmaceuticals Limited*.

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

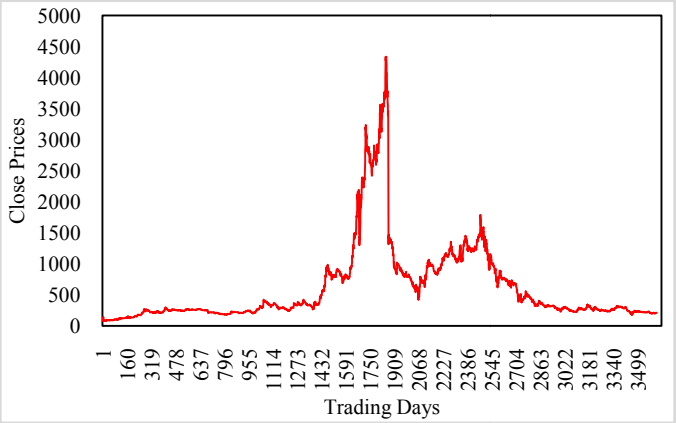


Figure 2. Closing prices of *AB Bank Limited.*

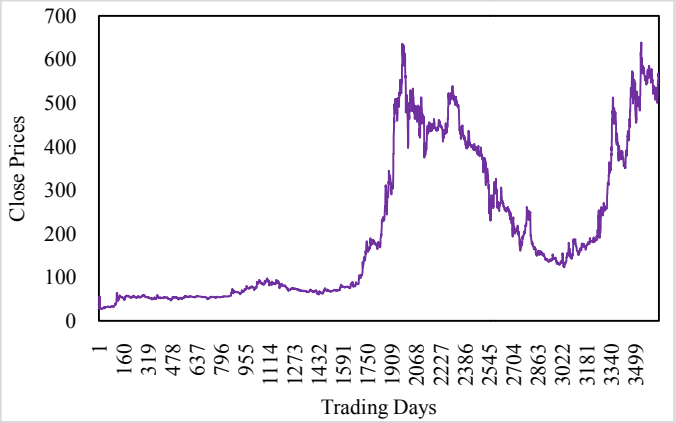


Figure 3. Closing prices of *ACI Limited.*

V. EXPERIMENTAL RESULTS

The principal component analysis on the original data shows that the first 10 components contribute over 98% cumulative covariance providing the most information. Fig. 4 shows the cumulative covariance contribution of principal components for *AB Bank Limited* and the same results are obtained for other two companies (not shown here). Hence the first 10 PCs are selected to form the input set for the SVR.

In this study, the radial basis function (RBF) is used as the kernel function of SVR. To find the best C and γ value pair we have considered e^{-5} to e^{10} for both parameters as our research space. For the data of *Square Pharmaceuticals Limited*, the coarse grid discovered the best (C, γ) as (e^{9}, e^{3}) with the 5-fold cross validation MAPE 2.23%. Then a finer grid search on the neighborhood of (e^{9}, e^{3}) produced a better cross-validation MAPE of 1.66% at (e^{9}, $e^{2.8}$). After the best (C, γ) is found, the whole training set is trained again to generate the final SVR model. The best value pairs for C and γ for every prediction task where minimum prediction error is exhibited by the grid search approach are shown in Table 2.

TABLE II. GRID SEARCH RESULTS FOR RBF KERNEL PARAMETERS.

Parameter	Square Pharmaceuticals Ltd.	AB Bank Ltd.	ACI Ltd.
C	$e^{8.0}$	$e^{5.0}$	$e^{6.8}$
γ	$e^{2.8}$	$e^{2.0}$	$e^{2.4}$

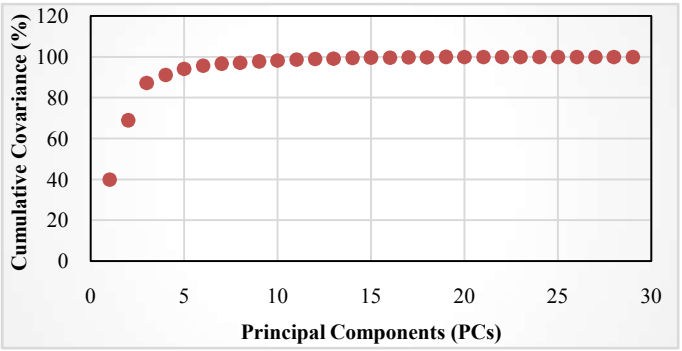


Figure 4. Cumulative covariance of PCs for *AB Bank Ltd.*

The prediction performance of the PCA-SVR model is compared with single SVR in terms of MAPE (%), MAE, rRMSE and MSE for 1, 5, 10, 15 and 30 days in advance targeting both short-term and long-term prediction. Table 3, Table 4 and Table 5 illustrate these comparative results for *Square Pharmaceuticals Limited, AB Bank Limited* and *ACI Limited* respectively.

TABLE III. COMPARATIVE PERFORMANCE OF SVR AND PCA-SVR WITH VARIOUS DAYS AHEAD FOR *SQUARE PHARMACEUTICALS LIMITED.*

Days Ahead	MAPE(%)		MAE		rRMSE		MSE	
	SVR	*PCA-SVR*	*SVR*	*PCA-SVR*	*SVR*	*PCA-SVR*	*SVR*	*PCA-SVR*
1	3.44	**2.86**	0.034	**0.028**	0.147	**0.098**	12827.48	**8908.66**
5	4.38	**3.86**	0.043	**0.039**	0.189	**0.151**	21385.53	**19054.59**
10	4.94	**4.73**	0.049	**0.047**	0.202	**0.166**	**29924.40**	30571.79
15	5.41	**5.33**	0.054	**0.053**	0.224	**0.200**	**36581.20**	39920.35
30	**6.26**	6.59	**0.063**	0.065	0.177	**0.156**	**52894**.77	63314.41

TABLE IV. COMPARATIVE PERFORMANCE OF SVR AND PCA-SVR WITH VARIOUS DAYS AHEAD FOR *AB BANK LIMITED.*

Days Ahead	MAPE(%)		MAE		rRMSE		MSE	
	SVR	*PCA-SVR*	*SVR*	*PCA-SVR*	*SVR*	*PCA-SVR*	*SVR*	*PCA-SVR*
1	11.99	**8.785**	0.119	**0.088**	0.663	**0.225**	20370.52	**9876.19**
5	12.88	**9.601**	0.128	**0.096**	0.787	**0.363**	21808.29	**12146.76**
10	13.15	**10.29**	0.131	**0.103**	0.815	**0.450**	17613.94	**10565.01**
15	13.97	**10.96**	0.139	**0.109**	0.942	**0.585**	22345.86	**14962.03**
30	14.71	**12.66**	0.147	**0.126**	0.947	**0.662**	33810.28	**29645.13**

TABLE V. COMPARATIVE PERFORMANCE OF SVR AND PCA-SVR WITH VARIOUS DAYS AHEAD FOR *ACI LIMITED.*

Days Ahead	MAPE(%)		MAE		rRMSE		MSE	
	SVR	*PCA-SVR*	*SVR*	*PCA-SVR*	*SVR*	*PCA-SVR*	*SVR*	*PCA-SVR*
1	13.01	**9.98**	0.130	**0.099**	0.314	**0.296**	502.22	**251.61**
5	13.76	**10.60**	0.137	**0.106**	0.488	**0.231**	616.08	**410.96**
10	13.98	**10.85**	0.139	**0.108**	0.638	**0.315**	700.67	**470.36**
15	14.02	**11.14**	0.140	**0.111**	0.720	**0.366**	671.94	**475.37**
30	14.16	**11.35**	0.141	**0.113**	0.869	**0.548**	711.05	**534.22**

It can be observed that the PCA-SVR has produced less MAPE(%), MAE, rRMSE and MSE than single SVR for all the cases for *AB Bank Limited* and *ACI Limited*. PCA-SVR exhibits moderate improvements over SVR for *Square*

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

Pharmaceuticals Limited. These results demonstrate the effectiveness of our proposed model. It is also observed that as the predictions are made for more number of days in advance, the performance of individual model and the improvement of proposed model decreases. This is obvious for any prediction system.

Fig. 5 shows the actual price and the prices predicted by SVR and PCA-SVR models for 1 day ahead of AB Bank Limited. Fig. 6 shows the result for the same prediction task but for 5 days in advance. The visual representations also validate the effectiveness of our proposed model. The graphical representations for other prediction tasks (not shown here) also verify the worthiness of the proposed approach.

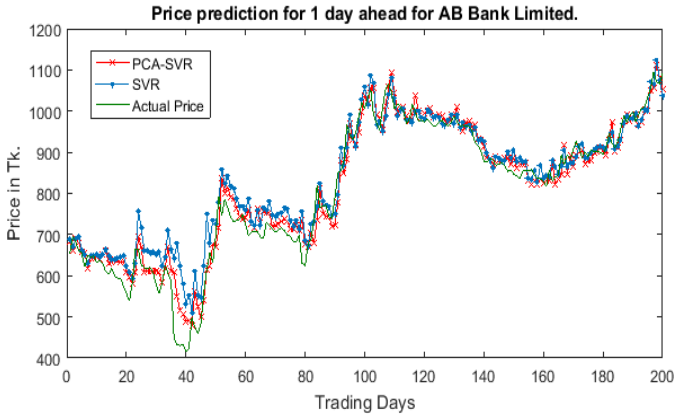


Figure 5. Prediction performance comparison of SVR and PCA-SVR for 1 day ahead for *AB Bank Limited*.

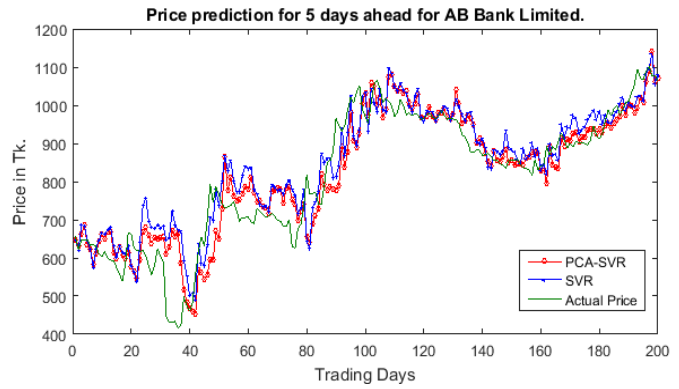


Figure 6. Prediction *performance* comparison of SVR and PCA-SVR for 5 days ahead for *AB Bank Limited.*

VI. CONCLUSION

This paper has proposed a price forecasting model integrating PCA and SVR for financial time series. This PCA-SVR model first uses the PCA to extract the most influential components from the input features in order to overcome the over-fitting or under-fitting challenge caused by the noisy nature of financial time series data. The filtered principal components are then used in SVR with RBF kernel function. The grid search for the best kernel parameters is conducted to improve SVR's performance. The experiments have evaluated 16 years' data for three commencing stocks from Dhaka Stock Exchange, Bangladesh. The performance of proposed model is compared with single SVR for various time durations with prediction error. Experiment results show that the proposed PCA-SVR model outperforms the single SVR model by generating less predictive error. The empirical results can conclude that the PCA can successfully unfold the influential information from the original data and uplift the forecasting accuracy of SVR. Future research may integrate Kernel PCA and other signal processing techniques like wavelet transformation with SVR. In this study, only the price related historical data is used to predict future prices. But, it is well known that various other aspects like general economic conditions, government policies, company performance, investor's interest etc. In future, these aspects can also be incorporated as input features for prediction which may buttress the accurate prediction.

REFERENCES

[1] Abu-Mostafa YS, Atiya AF. Introduction to financial forecasting. Applied Intelligence 1996; p.205–13.

[2] Huang, Wei, YoshiteruNakamori, and Shou-Yang Wang. "Forecasting stock market movement direction with support vector machine."*Computers& Operations Research* 32.10 (2005): 2513-2522.

[3] P. J. Kaufman, Trading Systems and Methods, John Wiley & Sons, 1998.

[4] Box, George EP, et al. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[5] Kim, Kyoung-jae, and Ingoo Han. "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index." *Expert systems with Applications* 19.2 (2000): 125-132.

[6] Yao, JingTao, and Joseph P. Herbert. "Financial time-series analysis with rough sets." *Applied Soft Computing* 9.3 (2009): 1000-1007.

[7] Trafalis, Theodore B., and HuseyinInce. "Support Vector Machine for Regression and Applications to Financial Forecasting." *IJCNN (6)*. 2000.

[8] Lu, Chi-Jie, Tian-Shyug Lee, and Chih-Chou Chiu. "Financial time series forecasting using independent component analysis and support vector regression." *Decision Support Systems* 47.2 (2009): 115-125.

[9] Kao, Ling-Jing, et al. "Integration of nonlinear independent component analysis and support vector regression for stock price forecasting." *Neurocomputing* 99 (2013): 534-542.

[10] Cao, L. J., et al. "A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine." *Neurocomputing* 55.1 (2003): 321-336.

[11] GRIGORYAN, Hakob. "Stock Market Prediction using Artificial Neural Networks. Case Study of TAL1T, Nasdaq OMX Baltic Stock." *Database Systems Journal BOARD*: 14.

[12] Pearson, Karl. "LIII. On lines and planes of closest fit to systems of points in space." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901): 559-572.

[13] Jolliffe, Ian. Principal component analysis. John Wiley & Sons, Ltd, 2002.

[14] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.

[15] Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A practical guide to support vector classification." (2003): 1-16.

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

P2P CACHE RESOLUTION SYSTEM FOR MANET

Dr. K.V.Prasad[1], Dr.G.Sanjay Gandhi[2]

[1]Professor, Dept of CSE, TKR College of Engineering and Techlogy, Meerpet,Hyderabad,
prasad_kz@yahoo.co.in

[2].Profssor, Dept of CSE, Visweswaraiah College of Engineering and Technology,Hyderabad.
sanjaygandhi.g@gmail.com

Abstract--**In this paper we explore the issue of store determination in a portable shared specially appointed system. In our vision reserve determination ought to fulfill the accompanying prerequisites: (i) it ought to bring about low message overhead and (ii) the data ought to be recovered with least postponement. In this paper, we demonstrate that these objectives can be accomplished by part the one bounce neighbors into two sets in view of the transmission run. The proposed approach lessens the quantity of messages overflowed into the system to discover the asked for information. This plan is completely circulated and comes requiring little to no effort as far as store overhead. The test comes about gives a promising outcome in view of the measurements of studies.**

Index Terms **– Mobile Ad Hoc Network, cooperative caching, caching strategies, cache invalidation**

I. INTRODUCTION

Late remote correspondence development and the nonstop change of portable terminal execution have empowered the utilization of these advances in various fields and to consider new applications. There are two classes of remote versatile systems: systems with framework utilizing access focuses to interface with an extensive variety of system and systems without framework which don't presuppose the nearness of a wired foundation. An impromptu system is a remote self configurable system comprised of versatile free moving terminals interconnected by remote associations. Impromptu systems give a method for conveying which can be rapidly and effectively conveyed. This makes the specially appointed systems a decent decision for applications in which different common what's more, military fields.

The objective of a PC system is to offer a route for clients with various terminals to impart with a specific end goal to share information and get to administrations gave by various servers. In portable specially appointed systems, gadgets by and large have constrained vitality saves and handling abilities. Data transfer capacity is additionally a rare asset, restricted by the way of the remote medium. In a datamanagement perspective, these limitations acquaint a few issues that need with be tended to. Information exchanges must be lessened and systems must be conveyed to go up against the incessant detachments and low transmission capacity requirements. Hence it is a testing assignment to exhibit the information productively by diminishing the postponement or holding up time to the end client.

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

Information storing is broadly utilized as a part of different areas to enhance information get to proficiency, by lessening the inertness experienced by the end clients. In remote portable system, holding as often as possible got to information things in a versatile hub's nearby stockpiling can diminish arrange movement, reaction time and server stack. Storing in specially appointed systems is viable in light of the fact that a couple of assets are asked for frequently by numerous clients, or more than once by a particular client, which is known as the region of reference. To have the full advantages of reserving, the neighbor hubs can coordinate and serve each other's misses, in this manner additionally decreasing the remote activity. This procedure is called agreeable storing. Since the versatile hubs can make utilization of the information put away in another hub's reserve the successful store size is expanded.

The usage of a decent helpful reserving strategy basically includes four noteworthy outline contemplations: reserve situation and determination, store affirmation control, reserve substitution and reserve consistency support [1]. The store affirmation control module chooses regardless of whether a got information thing is cacheable or not, and a reserve determination module chooses how to bring the information from the neighboring hubs when there is a nearby miss. Since the system hub has constrained memory, it can reserve different information things subject to its memory limit. The target of reserve affirmation control module is to store more unmistakable information things in the given reserve space. This decreases the information activity over the system since more number of solicitations can be served from the neighboring hubs. The reserve consistency module is to keep up store consistency, ie, guaranteeing that every hub reserving the information thing knows about the information refresh at the source. Because of the constrained reserve space, substitution approaches choose which things ought to be expelled from the current store to clear a path for new ones. In this paper we concentrate on reserve position and determination.

Towards the objective of enhancing the execution in helpful reserve, we propose a novel store determination conspire which utilizes a split table way to deal with resolve the reserved information ask for which in turn diminishes idleness and message overhead. All the more absolutely, our calculation parts the neighboring hubs into two sets in light of correspondence range. The cell phones keep the data about the neighboring hubs in two tables, best level closest hub store table and next level closest hub store table. The seeking time and the demand messages overflowed into the system is lessened by this strategy. Besides, because of straightforwardness.

CACHE RESOLUTION

Cache decision in ad hoc networks constitutes an difficulty that several researchers are that specialize in.There are a few interesting research in this difficulty. Two primary mechanisms used to discover cache in ad hoc networks are broadcasting and cluster based method. The first technique is based on request/response message exchanges wherein a consumer interested in discovering a cache pronounces a discovery request containing information on the carrier requested and the node presenting the statistics sends a respond. In the second one method, a node referred to as cluster head is used to temporarily sign up the statistics about the cached statistics inside the neighboring nodes. A node searching for facts will first look in to the cluster head node to find the statistics and then directly ship a request to the corresponding node. Although first method is the less difficult model, it relays on flooding to broadcast the data request. Flooding increases network competition and overhead

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

when the community density is high. In comparison, cluster based approach has been quality owed with the capabilities of reduced overhead by way of having a coordinator node which manages the cache resolution. The neighboring node which possesses the records may be effortlessly observed out checking the research desk maintained with the aid of the cluster head. The disadvantage of this technique is that institution renovation is tough because of the mobility of nodes. The manage node may additionally get disconnected which causes excessive overhead [4].The variety of entries inside the appearance up table will increase while the community density is high. To keep the accurate popularity of the network, these tables need to be regularly up to date. This involves facts alternate between the nodes which in turn will increase the site visitors overload in a dense community.

To evade these drawbacks we designed a cache decision protocol the various mobile nodes. Our approach is primarily based on a disbursed cache discovery algorithm and reduced cache table entries to locate the place of the requested information. An gain of the usage of this method is that it reduces network visitors as flooding is not used. This technique also avoids the downside of institution maintenance with the aid of having a distributed method. To reduce the number of entries in the look up table we break up the tables in two based totally on transmission variety.

RELATED WORK

The essential project in cooperative cache is to immediately discover a cache containing the desired statistics. To facilitate statistics discovery in cooperative cache, a few schemes [2][3][5][6][7] based totally on broadcasting and cluster primarily based technique [9][12] is proposed in literature. These approaches fluctuate in how the data request is resolved to find the vicinity of the data saved inside the neighboring nodes.

In broadcast based technique the cell nodes broadcast the request to discover a node with the desired information. On the opposite hand in cluster primarily based method, a few cell nodes are decided on as information coordinators which address the venture of finding the location of the desired information. We can also discover a few discovery mechanism [8] that doesn't comes below this category.Below, we describe a few consultant cache resolution techniques for cooperative caching from the one-of-a-kind companies referred to above. Aggregate caching scheme proposed in [5] tries to growth the information accessibility in an Internet based mobile community. A broadcast based statistics seek set of rules called simple seek is used to find the desired statistics object. Whenever a cellular nodes wishes a few statistics the request is broadcasted to its adjacent nodes .Upon receiving the published request, the adjoining nodes replies to the request if it has already cached the records, in any other case the request is forwarded to its buddies until it's miles acknowledged by way of an access point or some different nodes which have the requested information. Flooding is the technique used for broadcasting. This algorithm units a hop limit for the request packet to lessen the site visitors within the community. A caching method which uses cluster primarily based method may be visible in [9], wherein a coordinator node keeps the cluster cache country records of different nodes within its cluster area. If there may be a local cache leave out, the coordinator node will find whether the information object is cached in other clients inside its home cluster.

Another method for information discover other than the referred to schemes may be seen in [4] and [10].In [4] a distributed organization based totally technique in which

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

every node keeps a set desk and self table. Whenever a information request comes in it first checks in its self table, if statistics isn't observed checking is achieved inside the institution desk which stores the records identity and the node which posses the data. Here the mobile nodes interchange information about where the information is gift in the community. In [10], a cache resolution technique primarily based on adaptive flooding broadcast is used for looking statistics within the network. According to this scheme a cell node uses three schemes; adaptive flooding, profile-primarily based decision and avenue side resolution. In adaptive flooding, a node makes use of constrained flooding to look for objects in the community. In profile-based totally resolution, a node makes use of the beyond records of acquired requests. In street facet decision, forwarding nodes caching the requested object, respond to the requests in preference to forwarding them to the far off records source.

PROPOSED SCHEME

The goal of our cache decision scheme is to reduce the overall access price through reducing searching time and facts traffic. Most of the present day cooperative caching approach makes use of the wholelist of one hop acquaintances to find the cached information location. This will location useless load on each node due to the fact they need to go looking all the entries within the desk or vast cast a request to all of its buddies to discover the required facts. The complexity of records discovery will increase with the number of nodes within the network. In this paper we bear in mind a cache resolution approach which makes use of a break up table technique. In the given community each node Ni continues a table ncti1 , whose entries represent a fixed of pinnacle degree acquaintances nei1 ,which are positioned inner a wireless coverage of WR1.The second desk ncti2 represents a set of next level pals nei2 ,which can be positioned in a wi-fi range of WR2 . The desk entries in every node of an advert hoc community are maintained as follows. The nodes coming below the wi-fi communique range WR1 and WR2 are taken and the distance among the present day function and the node position is calculated using the Euclidian distance [5].When a node is farther faraway from the given node Ni the Euclidian distance can be greater. The nodes are organized on the basis of distance, signifying the primary access to be closest one .The correctness of the entries inside the table is ensured by periodically updating the neighbor node facts present inside the tables. The distance of every node from node Ni is decided by the use of the Euclidian distance. For cache discovery, the pinnacle degree friends are first of all checked, and if we are not able to retrieve the facts, request packet is forwarded to the nodes inside the variety WR2. By this the redundancies in flooding is decreased. The communique fee is also reduced via averting a couple of transmissions. Cache Admission and Replacement In cooperative caching nodes percentage the cache contents of neighboring nodes to utilize the overall benefit of caching. The to be had cache replacement mechanisms for advert hoc network can be categorised in to coordinated and uncoordinated relying on how alternative decision is made [15]. In uncoordinated scheme the replacement choice is made with the aid of person nodes. Coordinated cache alternative considers the information found in neighboring nodes for substitute.

In order to improve the content variety inside the cooperative cache, our scheme does now not cache any facts coming from the neighboring nodes. This will increase the supply of information for the person, as extra statistics objects are cached and additionally avoids additional request to the server. Previous research [16] have shown that the requests for smaller gadgets are greater in

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

comparison to larger objects, so the opportunity of attaining high hit rate is expanded if we store extra wide variety of small gadgets. In our proposed model we set a threshold price for the dimensions of the statistics item admitted to cache.

The threshold value is about as 50% of the full cache size. Any item larger than this threshold isn't added in to cache. The cache alternative is based totally on a key primarily based algorithm, which takes in to account the inter arrival time of new references, size and consistency for web page replacement.

SIMULATION SET-UP AND METHODOLOGY

We have advanced a simulation model in JAVA. In our simulation, nodes are randomly located in an area of 800X800 m2. Each node is recognized through a node id and a host name. The records server is implemented as a hard and fast node inside the simulation vicinity. The data server carries all of the facts items asked via the cellular nodes. The size of each information item is uniformly allotted among smin and smax. The nodes in the community circulate randomly based totally on a random course. Simulation begins with each node having empty caches and, then, iteratively caches the statistics object in every node's cache and next request for records are served from the cache. The nodes that generate information request are decided on randomly and uniformly. The time interval among consecutive queries generated from every node/customer follows an exponential distribution with suggest of 10sec. Each mobile node generates a unmarried circulation of examine handiest queries. After a question is sent out, the consumer does not generate new question until the pending query is served. The information access pattern follows a Zipf distribution [14] with a skewness parameter as 0.8

EXPERIMENTS AND RESULTS

We in comparison the performance of our approach with broadcasting for distinct node densities. It indicates the specific performance assessment of two schemes, cooperative caching with splittable method (ccs) and cooperative caching with broadcasting (ccb), as a function of various node densities. Fig 2 suggests the message overhead for each schemes below different node densities. The parent display that ccs outperforms ccb in any respect node densities. As the node density will increase, the distinction become extra tremendous which implies that ccs can gain from large node density. Fig 3 depicts the assessment of cache hit ratio for distinct node densities. From the parent we will see that the cache hit ratio for ccs and ccb carry out pretty intently. The relative overall performance of cache hit ratio remains particularly solid for better community densities.

VI. CONCLUSONS AND FUTURE WORK

In this paper we addressed the problem of cache resolution approach for peer to see cooperative caching in ad hoc networks. The objective of our hassle changed into to decrease the quantity of messages flooded in to the network, which in turn reduces the conversation value and bandwidth utilization. We designed a facts discovery procedure primarily based on break up table method. We evaluated the performance our algorithm with broadcasting approach through full-size simulations. Experimental results display that the proposed cache decision algorithm can drastically lessen the message overhead while as compared to broadcasting.

REFERENCES

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

[1] SridharIyer.(2000).MobileAdHocNetworks. *CIT*. p1- 106

[2] Sunho Lim Wang-Chien Lee Guohong Cao Chita R.Das.(2004).Performance Comparison Of Cache Invalidation Strategies For Internet-based Mobile Ad Hoc Network. AIEEE.

[3] N. Sabiyath Fatima,Dr. P. Sheik Abdul Khader. (2011). A Hybrid Cache Invalidation Technique for Data Consistency in MANET. International Journal of Computer Applications. 16 (5), p40-44.

[4] LI JIA. (2011). A Mobile Ad-hoc Network Data Cache Invalidation Method. Elsevier. p150-154.

[5] Yu Du. (2005). Improving On-Demand Data Access Efficiency In Manets With Cooperative Caching. Arizona state university. p1-150.

[6] Hugo Miranda Simone Leggio Lu´ıs Rodrigues Kimmo Raatikainen. (2005). A Stateless Neighbour-Aware Cooperative Caching Protocol for Ad-Hoc Networks. European Science Foundation. p1-28.

[7] Yaozhou Ma,Abbas Jamalipour. (2010). A Cooperative Cache-Based Content Delivery Framework for Intermittently Connected Mobile Ad Hoc Networks. IEEE. 9 (1), p366-373.

[8] Kuppusamy, P. and B. Kalaavathi. (2012). Cluster Based Data Consistency for Cooperative Caching over Partitionable Mobile Adhoc Network. ISSN. 9 (8), p1307-1315.

[9] Yu Huang, Beihong Jin, Jiannong Cao, Guangzhong Sun, Yulin Feng. (n.d). A Selective PushAlgorithm for Cooperative Cache Consistency Maintenance over MANETs. IEEE. p1-11.

[10] F.J. Gonzalez-Cañete and E. Casilari. (2011). Impact of the Mobility Model on CooperativeCaching Scheme for Mobile Ad Hoc Networks.inTech. p266-285.

[11] Roberto Beraldi and Roberto Baldoni. (2003). A Caching Scheme for Routing in Mobile Ad Hoc Networks and Its Application to ZRP. Published by theIEEE Computer Society. 52 (8), p1-12.

[12] R.Dhivya, V.Kavitha. (2014). Secured Client Cache Sustain for maintaining Consistency in Manets. International Journal of Research in Engineering and Technology. 3 , p1-6.

[13] Chih-Feng Chad, Ying-Hong Wang', Jenhui Ched, and Yi-Chein Lin'. (2005). A Cache Sharing Interface for Data Access in Mobile Ad Hoc Networks. IEEE. p78-82.

[14] Song Guo and Oliver Yang. (2005). Effects of Backup Routes and Cache Timeout Mechanism on Reliable Source Routing in Mobile Ad-hoc Networks. IEEE. p361-365.

[15] G. F. MariasI, K. Papapanagiotou†, P. Georgiadis‡. (2005). Caching Alternatives for a MANET-Oriented OCSP Scheme. IEEE. p1-9.

[16] Hassan Artail, Haidar Safa, and Samuel Pierre. (2005). Database Caching in MANETs Based on Separation of Queries and Responses.IEEE. p1-8.

[17] Ying-Hong Wang1, Jenhui Chen2, Chih-Feng Chao1*, and Chien-Min Lee1. (2005). A Transparent Cache-based Mechanism for Mobile Ad Hoc Networks. IEEE. p1-6.

[18] Shin-Jer Yang, Shih-Chun Chu. (2006). Performance Analysis of DSR Using Reclaim- Based Caching Policy on the MANET. IEEE. p1-6.

[19] Babar S. Kawish1, Baber Aslam2, Shoab A. Khan3. (2006). Reducing the Overhead Cost in Fixed & Low Mobility AODV Based MANETs.International Multiconference. p1-10.

[20] Babar S. Kawish1, Baber Aslam2, Shoab A. Khan3. (2006). Reducing the Overhead Cost in Fixed & Low Mobility AODV Based MANETs.International Multiconference. p1-10.

[21] Jiannong Cao,Yang Zhang and Guohong Cao,Li Xie. (2007). Data Consistency for Cooperative Caching in Mobile Environments. IEEE. p1-7.

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

Roty_Shift: a Proposed Method for Generating Secret Keys

Qusay Mohammed Jafar
Department of Computer Communication Engineering – AL-Rafidain University College
Iraq-Baghdad
qusay_mj@coalrafidain.edu.iq

Abstract:-

In this research a proposed algorithm to generate secret key**s** was accomplished, the proposed key generation algorithm will be called (Roty_Shift algorithm) and it will generate a series (list) of subkeys may be used for data encryption. Roty_Shift needs two secret seeds as secret keys, seed1 will be generated by true random number generator (TRNG) while seed1and seed2 will be used in pseudorandom number generator (PRNG), also this research utilizes the mechanism of the key derivation function (KDF) to generate the subkeys, on other hand Roty_Shift can be considered as PRNG. The proposed algorithm consists of five levels, and according to these five levels the subkeys will be generated and the size of each one is (N*N), each one of these subkeys is possible to use it in any encryption strategy of block cipher type. The proposed algorithm (Roty_Shift) was tested using different key sizes and obtained good results. In this research, the proposed algorithm proved the concept of diffusion on the key itself. Also according to some tests with 256 bits key and 5000 iteration there is no repetition in the generated subkeys, but in small key size a problem of repetition will be encountered.

Keywords:-

Secret keys, Seed, Subkey**s** generation, TRNG, PRNG, KDF, Rotation, Shift, Five Levels, Diffusion, and Encryption.

1- Introduction

Encryption is used to protect data that belong to a level of sensitivity, there are two main parameters in the encryption: the encryption algorithm and the secret key [1]. This research focused on the secret key(s) generation. There are two strategies of encryption: Symmetric and Asymmetric, in the symmetric encryption there is one key for encryption and decryption, whereas in the asymmetric encryption there are two keys; the first for encryption (public key) and the second for decryption (private key) [2].

The proposed secret key generation algorithm in this research work is suitable for symmetric encryption. Also this research solves the problem of selecting the keys for the encryption process, and the aim is for generating a series of secret keys without applying the encryption process itself.

2- Key Generation

The procedure of the key management must to be in secure form; because of without security in the key management the benefits of any level of encryption will be not useful. In firm institutions there is a life cycle for the key, the key lifecycle is as follow: key

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

generation, key establishment, key storage, key usage, key distortion, key change, and back to key generation and so on [3]. "Key generation is the process of generating keys for cryptography"[1]. This research focused on the key generation to generate a series of subkeys of size *n* bits.

To increase the complexity and difficulty of the cryptanalysis the algorithm of the subkey generation must be complex [4].

There are three types of random key generation (*True Random Number Generators* (TRNG), *Pseudo Random Number Generators* (PRNG), and *Cryptographically Secure Pseudorandom Number Generators* (CSPRNG)) [5], and each one of these three types needs a seed to complete its function. TRNG needs a nondeterministic source to generate the random numbers, whereas PRNG is a deterministic algorithm to produce sequences of random numbers, if the algorithm is excellent, the sequence of the random numbers will pass the tests of good randomness [4].

According to our works in this research the Roty_Shift will needs two primary secret keys (KEY1, KEY2), to make KEY1 as true random number generator (TRNG) we need a strategy to select KEY1 randomly, so that; the seed generation (seed1) of the KEY1 is as follow: suppose the key size is 256 bit (32 bytes), it is possible to go out to the high way and monitor the street at 7:13 AM and capture the cars plate numbers for 64 consecutive cars one after the other and apply the following procedure:-

For car=0 to 63
KEY1[car]=CarPlateNo[car] MOD 16
Next car

The loop in the above will create an array of 64 hexadecimal digits and then rearrange the KEY1 as a two dimension array of binary data. This procedure with assuming the key size is 256 bits. This mechanism will give a randomness of type TRNG to generate the secret key (seed1). The proposed procedure in the above is possible to make it as automated by using some equipment (camera with image processing) with secure timing.

KEY1 in the proposed method will be called (TKEY), while KEY2 will be called (InitiKEY).

PRNG needs a seed to generate sequences of random number [5]. It is possible to follow the manner in [5] with some modification to generate random numbers (secret keys) as shown below:-

z= input("enter how many number of subkeys you like to generate")
i = 0
S_i= seed1, (and seed2)
loop
S_{i+1} = f (S_i)
i=i+1
Until ((i=z) or (end of the plaintext))

In this research and as shown in the procedure above, we mixed seed1 and seed2 to complete the task of the PRNG. *f* is a function (algorithm) its input is the initiate seeds and the output of the previous step to produce a new secret key linearly. In this research we applied the PRNG concept to generate a series of secret keys, whereas the proposed algorithm (Roty_Shift) represents the (*f*), and according to this tacit concept of the TRNG and PRNG this research applied the TRNG and PRNG to generate a series of random secret keys.

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

3- Key Derivation Function (KDF)

Key distribution in one of its concepts deal with the mechanism of exchanging the key between the sender and the receiver in a manner that not make any threat on the generated keys [6]. There are two procedures to establishment the keys between parties (key transport and key agreement), key transport used for generating the key and distribute this key to other parties, whereas key agreement used for generating the secret key in a manner of joint [5] where each party have its parameter.

Key Derivation Function needs two parameters the first is not secret parameter (NSP), it is shared between the two parties, the second is secret parameter as a secret key (SKey), it is keep secret between the two parties (sender and receiver), these two parameters will be as input to the function to generate the secret key (Ksec), and then it will be used for encryption and decryption if the encryption is in mode of symmetric [5].

In this research we try to utilize the concept of the KDF to generate the secret keys but in manner differ in its essence; instead of making NSP not secret we make it secret with name KEY1 and it is shared between the sender and the receiver, where the sender will send this key frequently to the receiver using a secure channel, KEY1 needs to communicate between sender and receiver each time when the coordinator decide to update KEY1. KEY2 will be secret also but it is not need to exchange between sender and receiver each time. There are two secret keys and this will give more security to the key management and key generation, for example if we assume KEY1 will be updated every week and KEY2 will be updated every year using a very secure channel for keys transportation. In this research the function that will exploit these two keys is the proposed key generation method (Roty_shift). Figure (1) shows an overview of the proposed key generation method.

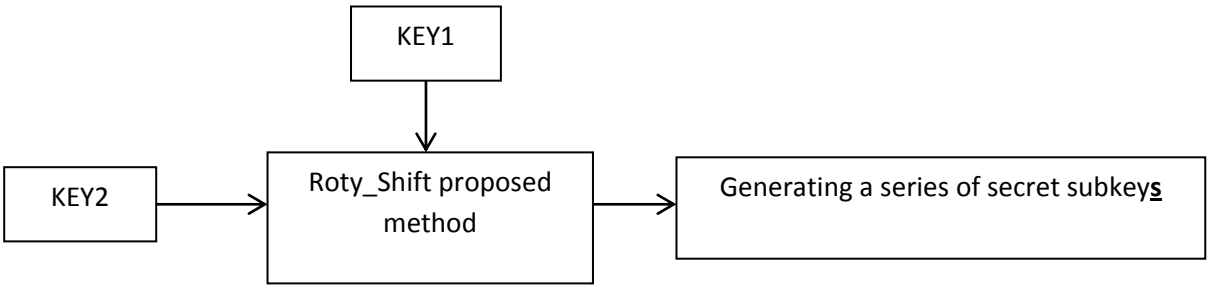


Figure (1) First view of the proposed keys generation method

4- Previous works:-

The researchers in [7] proposed a novel method for image encryption using dynamic secret key generation, where they are used the sunflower to create a random seed with the piecewise chaotic map methods and then applied the encryption algorithm.

The researchers in [8] proposed a new method for generating secret keys using the memristor device, whereas the master key is the initial profile of the memristor, and the session keys will be generated using the master key and other parameters that can be extracted from the memristor itself.

The researchers in [9] proposed a mechanism for key generation using the genetic algorithm, they used the random function to generate the initial population of 128 bits, and then by a fitness function to select individuals of maximum fitness for next processing. According to this it must to select the best two individuals and then apply the

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

one point crossover to produce the random number, and then a children and again fitness function and so on, and then a mutation to get the key.

The researcher in [10] gave an overview about the random number generator (methods, applications, theoretical concepts, practical implementation, studying three mechanisms of random number generating, comparisons, probabilities, and statistics).

5- The proposed method (Roty_Shift)

The aim of the proposed method is to design and implement a system to generate a series of secret keys in attempt to apply the concept of the PRNG with two secret seeds, the first seed will be called as TKEY, this seed will be selected randomly according to TRNG mechanism, and it will be selected as shown above in section 2, whereas the second seed will be called as InitiKEY to be as an initial secret key, TKEY and InitiKEY will be two arrays of size (N*N) of binary values, the keys in the proposed algorithm will need to be of two dimensions array because there is a level of rotation process as shown in the next sections. The seed (InitiKEY) will be selected according to agreement between the two parties (sender and receiver), of course the TKEY and InitiKEY will be secret, and by these two seeds (keys) it is possible to apply the KDF, and the function itself will consider the proposed algorithm (Roty_Shift algorithm).

After detecting (selecting) the two secret seeds (TKEY, InitiKEY), these two seeds will be the inputs to the Roty_Shift algorithm.

TKEY aims to increase the secrecy of the generated keys; TKEY will be XORed with the primary generated key ($PKEY_i$) to produce a new $GenKEY_i$, i=0,1,2…M, M is the number of the generated subkeys or the number of the blocks in the plain text, the block size of the plain text is equal to the array size of the InitiKey and TKEY. The XOR operation will be done after the InitiKEY process by the Roty_Shift algorithm to produce the GenKEY and the XOR operation will be applied between TKEY and $PKEY_i$ in each iteration of the Roty_Shift. TKEY itself will be modified in each iteration of the Roty_Shift algorithm by shifting process to produce a new linear TKEY. Roty_Shift will be repeated many times (1..M) to generates a list of GenKEYs (subkeys) ready for encryption operation.

To generate a list of GenKEYs, in each iteration of the Roty_Shift algorithm there are five levels (stages) as shown in the following:-

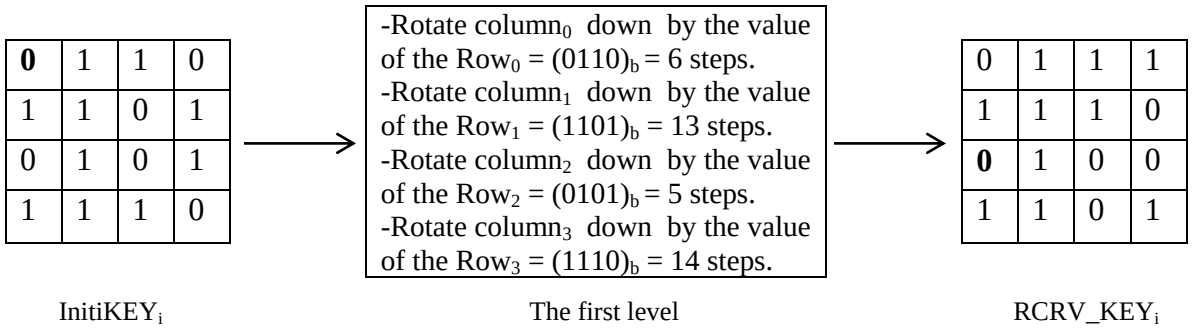
The first level (Rotate Columns by Rows Values (RCRV level)):-

As shown in the above the TKEY and InitiKEY will be arrays of two dimensions of size (N*N) of binary values, in this level each column of the InitiKEY it will be rotated down by the decimal value of the bits in each row after grouping them (row bits) according to the row number that is equivalent to the column number (t=0, Rotate $Column_t$ by the value of the row_t, t=t+1), figure (2) in below shows an example of the RCRV with assuming the secret key size (InitiKEY) is 4*4 (16 bits), the keys size (the two dimension array) in the proposed method is optional and in form of square array. There is a direct correlation between the key size and execution time of the program. In the Roty_Shift program we tested three types (sizes) of the key (4*4 (16 bits), 6*6 (36 bits) and 16*16 (256 bits)). The output of the first level is called $RCRV\_KEY_i$, and it will be the input to the next level (XX level).

In some methods of key generation the generated key called as subkey, so that the GenKEY is the subkey that will be ready to be used in encryption.

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

Note:- After finishing the iteration (i) of the Roty_Shift, the InitiKEY will be $GenKEY_i$ and the Roty_Shift will take $GenKEY_i$ as a new input (new InitiKEY) to generate $GenKEY_{i+1}$, and so on.



Figure(2): An example to explain the first level (RCRV)

The second level (XOR-XNOR (XX level))

The output of the first level will be the input to the second level, in this level the variable $RCRV\_KEY_i$ will be modified to produce XX_KEY_i through a process of applying (XOR) or (XNOR) operations by detecting the decimal value of each row in RCRV_KEY. Let no_1 contains the value of the Row_r, and no_2 contains the value of the Row_{r+1}, r=0...(Nr-1), Nr is the number of rows in RCRV_KEY, and then it must to examine the value of no_1, if it is even (no_1 MOD 2=0) the operation (XOR) will be implemented between no_1 and no_2, otherwise (no_1 MOD 2=1) the operation (XNOR) will be implemented between no_1 and no_2. The result of the logical operation will be assigned in Row_{r+1}, and so on for the next row, whereas the last row (Row_{Nr-1}) will be processed with Row_0 and the result will be assigned in Row_0. In briefly; (XX) level is responsible for the task of:- The bits of each row will be (XORed or XNORed) with the next row after detecting the decimal values of the Row_r. After finishing the (XX) level the XX_KEY_i will be ready for the third level. Figure (3) shows an example of the (XX level). The output of this level will be an input to the third level (ROTAT level).

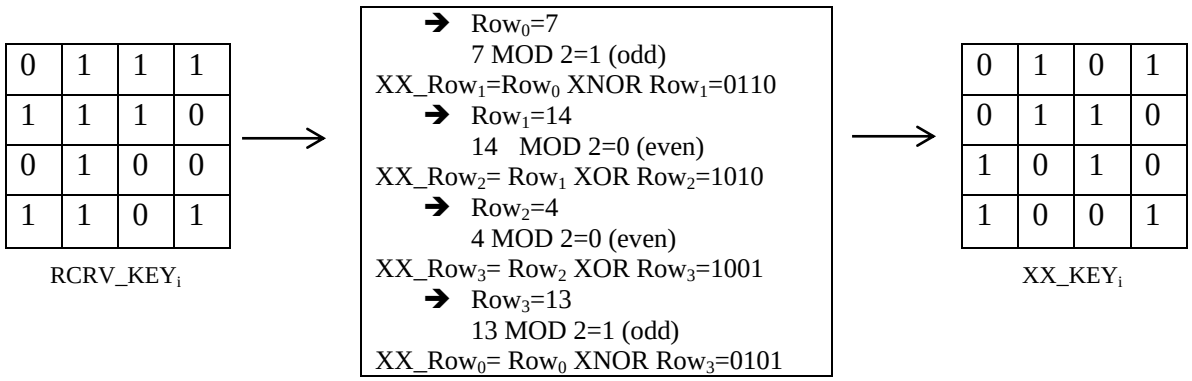


Figure (3): An example to explain the second level (XX level)

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

The third level (Rotate level)

In this level the rotation operation will be accomplished to generate ($RotateKEY_i$), the output of the second level will be processed by a rotation operation according to the summation of the binary values of each row with the next rows to get a decimal value (Sum), and then the Sum will be modulated by four to obtain a value ranging between (0..3), 0 to rotate the XX_KEY_i by 0 degree, 1 to rotate the XX_KEY_i by 90 degrees to the right, 2 to rotate the XX_KEY_i by 180 degrees to the right, 3 to rotate the XX_KEY_i by 270 to the right. In this level the XX_KEY_i will be treated as an image to make the idea a closer. After applying the Rotate level of course the values of the ($RotateKEY_i$) keeps its values as binary. Figure (4) shows an example of the (Rotate) level.

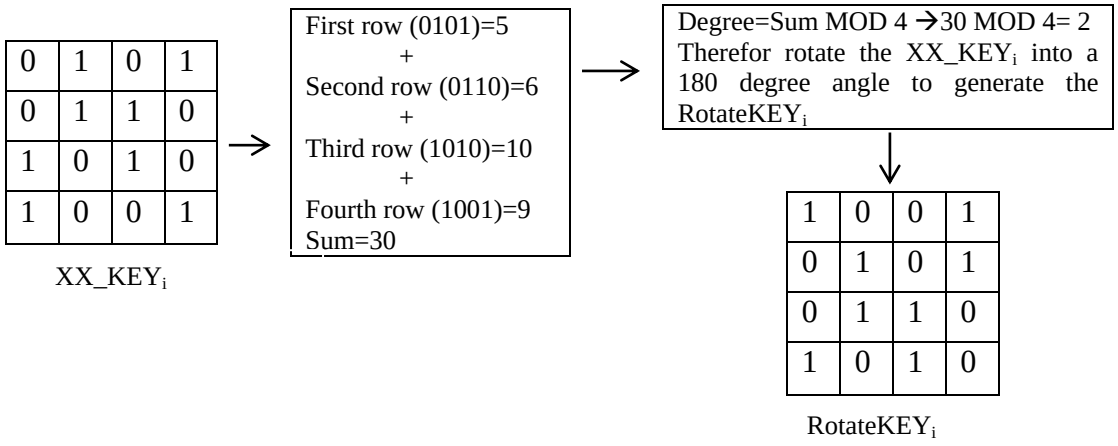


Figure (4) An example to explain the third level (Rotate level)

The fourth level (One Bit Mutation (OBM level))

After testing the proposed algorithm (Roty_Shift) a problem of some repetitions in the generated keys was encountered, and this led us to propose this level to perfection the job of the Rroty_Shift, so that it is possible to consider this level as an intermediate level to increase the entropy and the confusion after applying any encryption method.

OBM make one bit of the ($RotateKEY_i$) to be complemented ($0\rightarrow1$)($1\rightarrow0$) in each iteration of the (Roty_Shift). *Note*:- in each iteration of the Roty_Shift proposed algorithm there is one secret generated key (secret sukey), Roty_Shift will be looped to generate a list of M subkeys according to the number of blocks in the plain text and there is a conforming between the block size and the secret keys size.

The output of the OBM level is (OBM_KEY_i). Figure (5) shows an example of the OBM level for (OBM_R=0, OBM_C=0) (the first iteration only (i=0)):-

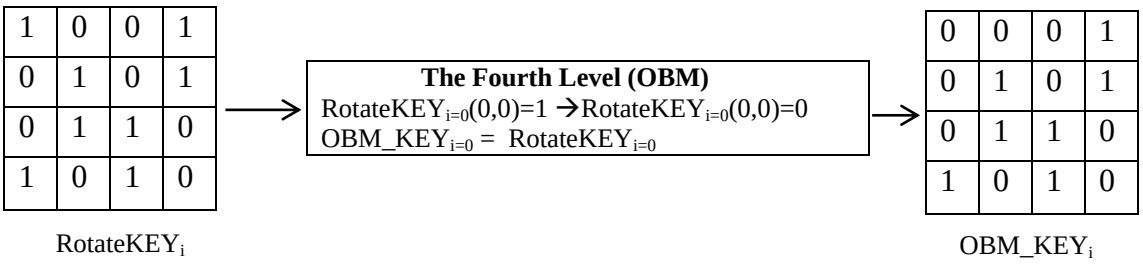


Figure (5) An example to explain the fourth level (OBM level)

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

In the next iteration (i=1) this level will modify (make a mutation) on the $RotateKEY_{i=1}(1,0)$ and so on for the next iteration (i=2) on the $RotateKEY_{i=2}(2,0)$, and so on. The mechanism of this level will be explained in more details in the section 5 (the algorithm)

The fifth level (XORing between OBM_KEY_i and TKEY with shift (XS level))

In this level the role of the TKEY will be started to make its action on the Roty_Shift proposed algorithm. TKEY itself is generated by the TRNG. X refers to XOR logical operation, while S refers to shift operation on the TKEY by one location to the left in each iteration of the Roty_Shift proposed algorithm. This level will complete the KDF by corresponding the TKEY as a first key, while $InitiKEY_i$ after four levels to produce OBM_KEY_i as a second key. Figure (6) shows an example of the XS level.

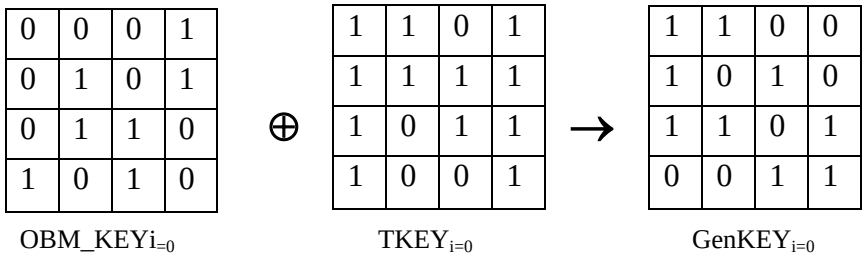


Figure (6) An example to explain the fifth level (XS level)

After applying this level the output will be $GenKEY_i$ that will be saved in a file for preparing a list of secret generated subkeys, or it is used directly by the encryption process. Before finishing the iteration i=0 a shift operation on the $TKEY_{i=0}$ will be applied to produce $TKEY_{i=1}$ that it will be used by the iteration i=1, and so on for each iteration as shown in figure (7).

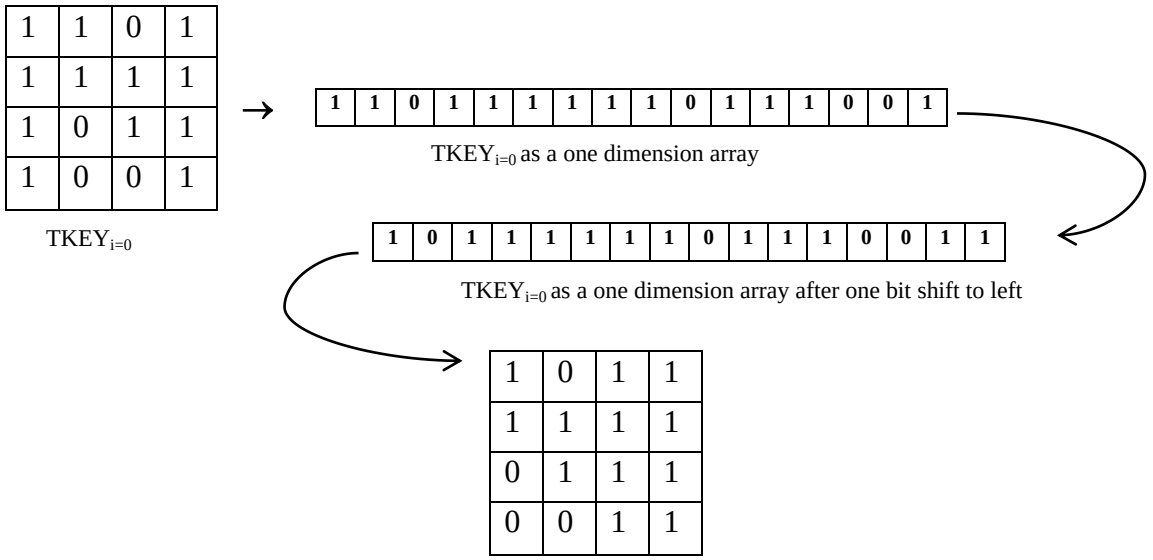


Figure (7) the shift operation on the TKEY

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

In some cases the shift operation that was accomplished as shown in figure (7) is called rotation but we used shift expression to distinguish between the operation that was applied with rotate level and the operation that was applied on TKEY in the fifth level.

In the second iteration of the Roty_Shift *i* will be 1, in the third iteration *i* will be 2, and so on. Figure (8) summarize the Roty_Shift proposed method.

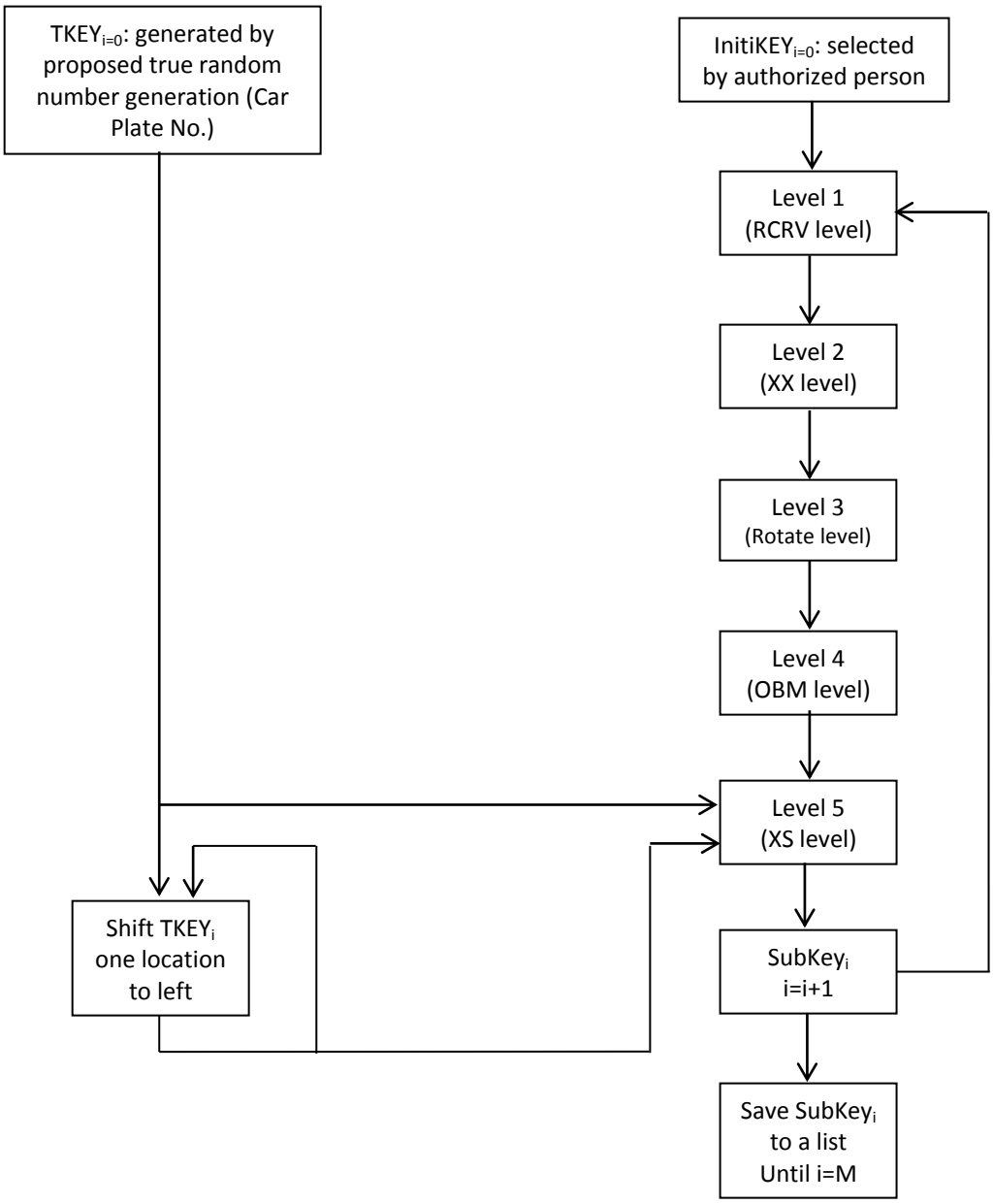


Figure (8) Block diagram of the Roty_Shift proposed method.

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

6- Roty_Shift Algorithm

Steps in below shows the proposed algorithm for secret key**s** generation (Roty_Shift algorithm):-

INPUT:- InitiKEY seed array of size (N*N) bits, TKEY seed array of size (N*N) bits, and M, where M is the number of blocks in the plain text or input directly by the user.

OUTPUT:- Series of generated secret keys ($GenKEY_i$), i=0,1,2,…M

1- Start

2- Initializations:- $InitiKEY_0$ = InitiKEY seed; *OBM_R=0; OBM_C=0;k=0;*

3- Loop i=0 to M

4- The first level (RCRV level):- its input is the two dimensions array ($InitiKEY_i$), its output is the two dimensions array ($RCRV\_KEY_i$).

4-1 Grouping the binary bits of each row of the $InitiKEY_i$ and convert them to decimal value (R_value), where R_value between $(0...2^N-1)$, and then apply the rotation process on each equivalent column by R_value as follow:-

Rotate the bits of the $column_c$ by (R_value), where (R_value) is the value of the Row_r , c=r, r=0…N-1.

4-2 Save the result of the step 4-1 in a variable of two dimension array ($RCRV\_KEY_i$)

5- The second level (XX level):- its input is the two dimensions array ($RCRV\_KEY_i$), its output is the two dimensions array (XX_KEY_i).

5-1 Loop Rn=0 to N-1 Rn is the row number in $RCRV\_KEY_i$

IF Rn != N-1 then

No1= decimal value of the binary bits of the row_{Rn}

No2= decimal value of the binary bits of the row_{Rn+1}

IF No1 MOD 2=0 then

XX_Row= No1 XOR No2

ELSE

XX_Row= No1 XNOR No2

END IF

The row_{Rn+1} of the XX_KEY_i=XX_Row (each bit in one cell)

ELSE IF

No1= decimal value of the binary bits of the row_0

No2= decimal value of the binary bits of the row_{N-1}

IF No1 MOD 2=0 then

XX_Row= No1 XOR No2

ELSE

XX_Row= No1 XNOR No2

END IF

The row_0 of the XX_KEY_i=XX_ROW (each bit in one cell)

END IF

5-2 Next Rn

6- The third level (Rotate level):- its input is the two dimensions array (XX_KEY_i), its output is the two dimensions array ($RotateKEY_i$)

6-1 sum=0

6-2 Loop For r=0 to N-1 (N is the number of rows in XX_KEY_i, while r will represent the row number in XX_KEY_i)

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

Convert the group of bits in row_r of the XX_KEY_i to decimal value (RowVal)
sum=sum+ RowVal
Next r

6-3 Make the rotation process on XX_KEY_i as follow:-

IF sum MOD 4=0 then
Rotate XX_KEY_i clockwise by 0 degree
$RotateKEY_i$= XX_KEY_i
Else IF sum MOD 4=1 then
Rotate XX_KEY_i clockwise by 90 degree
$RotateKEY_i$= XX_KEY_i
Else IF sum MOD 4=2 then
Rotate XX_KEY_i clockwise by 180 degree
$RotateKEY_i$= XX_KEY_i
Else IF sum MOD 4=3 then
Rotate XX_KEY_i clockwise by 270 degree
$RotateKEY_i$= XX_KEY_i
END IF

7- The fourth level (OBM):- its input is the two dimensions array ($RotateKEY_i$), its output is the two dimensions array (OBM_KEY_i):-

IF $RotateKEY_i$ [OBM_R,OBM_C]=”0” THEN
$RotateKEY_i$ [OBM_R,OBM_C]=”1”;
ELSE
$RotateKEY_i$ [OBM_R,OBM_C]=”0”;
END IF
OBM_C= OBM_C+1;
IF OBM_C= N THEN
OBM_C=0;
OBM_R= OBM_R+1;
IF OBM_R=N THEN
OBM_R=0;
END IF
END IF
OBM_KEY_i= $RotateKEY_i$;

8- The fifth level (XS level):- its input is the two dimensions array (OBM_KEY_i), its output is the two dimensions array ($GenKEY_i$):-

8-1 Apply the XOR operation between OBM_KEY_i and TKEY bit by bit to generate the secret key ($GenKEY_i$).

8-2 Apply the shift operation on the TKEY to generate the new TKEY.

To explain this level follows the following:-

For a = 0 To N-1
For b = 0 To N-1
$GenKEY_i$[a, b] = (OBM_KEY_i[a, b]) XOR (TKEY[a, b])
Shiftarray[k] = TKEY[a, b]
k = k + 1
Next b
Next a

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

```
        k = 0
    To shift TKEY one location to the left:-
    temp = shiftarray[0]
    For d = 0 To ((N*N)-2)
       Shiftarray[d] = shiftarray[d + 1]
    Next d
    shiftarray((N*N)-1) = temp
    For a = 0 To N-1
       For b = 0 To N-1
          TKEY[a, b] = shiftarray[k]
          k = k + 1
       Next b
    Next a
    k = 0
```

9- Save $GenKEY_i$ (the new generated subkey) in a file as string of hexadecimal values for using later or apply any encryption method directly with this generated secret subkey.
10- Go to step 3 (Next i).
11- End algorithm

7-The Experimental Results:-

In below will show some results of applying the Roty_Shift algorithm:-
The proposed algorithm accepts any key size, and for testing let the key size is 256 bits (16*16), and after applying this proposed algorithm there is a list of M secret subkeys. Table 1 shows a list of 40 secret generated subkeys without any repetition, and each one of these subkeys is considered as a PRNG.

Let InitiKEY and TKEY as follow:-
InitiKEY:- $(E78102DA3C917CBCEA370AEF4390CAF62782A1DFC65290ABCE4D7143A8327D4A)_h$
TKEY:- $(857369A3AC754457F495BCA9A6AE1DDBFDD3EEABA6AA55539CD3EEABA6AA5553)_h$

Table 1: Generating 40 secret subkeys of size 256 bits (64 digits in hexadecimal)

E78102DA3C917CBCEA370AEF4390CAF62782A1DFC65290ABCE4D7143A8327D4A
B165EDDB4EFF694C403975C8B98FFFECD647109D30D7DCE98B4E5AB91B4162E2
F05BE7E7CBC136387F5B92F64732EF40DF826318149D3CCCC0DD099F487817FA
C5FA23D7372CE2668E26D0F033142AA5538BDB18C8F44B718F8125CF80083941
F278169EBBB2AFD9AED253651542B154221CEF448D2B6CC8402BAD45F489ED0F
0D34F589C1E6E2DE75E963448BEE9A550480EE1AE53B000FD8DE87DA67521572
B821178362B53ED1F821C23D16CDF393FC121B86E1A583240E54AD43FA8F9838
7825B9DEE8868ED2CCE4FAC70894DB340F3D382C0B994D7C40250145C61E2B85
D491DFC9CC7166EC3A7BD8A770FBC662FA733636A3117CCC4145FEBCED5B8DF4
63A62E53027DE52DDC72275019AC4AA22967264BC524200E48A3374855E4A18D
48320EC6D1E51F35466E5DFC6EDEADEDA2B6260115CF8783FCD321932D38227F
D93A17F1E3BD6FDD9430F61FE1E347B57BCF6B0834506D329FB77BF7EBCD733D
2726EDC8A54165442330BD9BCBA88341BD1967EDFBBF279A48299BC507A54216
6E323746A4F95B7CD659C69041DD0DCFA1DD244A8A78A871A2CE2766E4E5BD4D
05724C9DD6C0D189AC420208F8C1399CF2E0CD97C656E1433B8CD4315BA0E369
F0C3B20134AF10AE0E44C5FC88DD4C0F801B6A41A37C31BBBA1B540FEDCEF408
184E32C9F3B6D977B0280FFB90E870BE1D24732C1F92F01F4B916ADFE58EBB8C
AAA7005F9C09C30CCC7FABF4D9D1426E5356F6890FE3F7DD7747DCB10167A5B0
6C96C270B15AD15433AC923D86845E4289C09DE9385B3EF8F34D62C833FD011B
B93F7E83C04F0C5E931163E19E2D32E80B96BE8C1B513C35C11C2651B05C3663

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

1A81B6B4BFCE9340942852E327B72897653E23C1637A4011C2B1BBEBD23F1271
7B07505D79B30BA3EFB463894156261119C7F85519C61DF88C55445AE9398199
938001340B4FA5BF38060A3A76D4AA02C81352EB1D865ECA659EB8C993A08188
5237DB172508E45A8538D3559C2E9CE3A48CD1E2D3EBD606C668659A76C36786
F20BE5303AF65962D7EA8CB6FC8113EA185D8CFC55862F94D99B15635B74A9DA
0F9380A26C5BE5A27593022E613966681662F117ABAB0FD733AB3F8AB52A0F5C
52C7AC26253AF2DBBBE9D096F078AECAEA82AE22D14FE2C587FC826C34731609
EEA15CC4B39763AAADAC3D9BA267D8406278DAFB61C13B6194328B636185C5A0
BB152862077F9A76B558F80C3119A84C4CA281C60FF8B16464DBE9B6966B7528
4911936A9164F3FBEDA6BCAB1ACA5B1954FBC7E4A3637D03D67ED1221CB45D54
742AB69E0F369AE3C57FFDD1B2188988878063020DE7585F4A652F27B8D6C8B4
2ADE87E7DC7C4ABF1B7FC15F171C22395870D97EB34EF8EAA008AFD02E3619EB
4C8E0742863311238776907D823BB3D5D2776197C710737F3DA6B497D41C9988
3FA4D80FD0C1333A7E54F615F57A4302BF53E6646F0369B45F58FF775F4BE26A
317C3A0A79104C3D0D3D3AFFE0E5F73EBE57C65FFF395352ABF8B6ACEFAB3AD1
3EA4318BB783ECAD75C1FC3F1B9DD81C368EC84EC99D24508AD11EEE3EC9287F
63F9286E023E595C80071FB8CAEB709FB180005B85D3D17874823109B28E15F5
55E295A19630A0662FB297C062EA5FE4E7851EAE183EA1318EC00D1CB8CBB5BF
EA664779045EA6287201A68C03E1B10E9E3D9FC675D9F69C3D87FD3B0FDE4753
D39D5A8499146AFB1ED46881DA2AAD503775DF924A836044E3978A9CC2285B98

Now we try to change one bit only in the TKEY (the fifth bit 0→1) to see how changing one bit on generating the 40 subkeys.

InitiKEY:- $E78102DA3C917CBCEA370AEF4390CAF62782A1DFC65290ABCE4D7143A8327D4A)_h$

TKEY:- $(8\mathbf{D}7369A3AC754457F495BCA9A6AE1DDBFDD3EEABA6AA55539CD3EEABA6AA5553)_h$

As shown in table 2 the effect of changing begins from the fourth subkey to make the seventh subkey and forth are totally different from the first try (table 1), and this prove the diffusion concept.

"Diffusion means that if we change a character of the plaintext, then several characters of the cipher text should change" [11], and depending on this definition if we change one bit of the InitiKEY or TKEY a new series of the subkeys will be generated.

Table 2: Generating 40 secret subkeys of size 256 bits (64 digits in hexadecimal) with modifying one bit only on TKEY.

E78102DA3C917CBCEA370AEF4390CAF62782A1DFC65290ABCE4D7143A8327D4A
B965EDDB4EFF694C403975C8B98FFFECD647109D30D7DCE98B4E5AB91B4162E2
E05BE7E7CBC136387F5B92F64732EF40DF826318149D3CCCC0DD099F487817FA
E5FA23D7372CE1668E26D0F0331C2AAD538BDB18C8F44B718F4125CF80083941
B278169E8BB6AFDDAED2536571C21154235CEE448D2B6CC8402BAD49F489ED0F
0D3535BB49F6097E75E57B448BE6CA5D0488DE1AE5310009DD0EEFCA6B501573
934BE38D3103F2A4C720723F8478D1B3542EB0F4B9B01BD76061608B8AA0CEED
D2A97C9B20EDF4AB9F5A058E09F73553565FA8823D6228B81E7FB75524BC9AD2
84574E786D7D50247441906BE6E1137B9B902C7BB81E70B49228CE3860D2EEFA
0A91E883770D9365EFAFC68EA2AB13DE20A6C0A3F2B6F3C59AC919E41E87CD13
99D8DD2D871B36505ED10CB698AC88FE1A069254FB45529C7A45FEFAFAF2F5E4
076FECF69F33DCF2E807348A7288CE7F896B97C90D08B103EB7BAAC88B135966
17A08C23243A0FDE36D04390F5DA25BCA654AFBE0BD3D83B41782544D022235A
3A820E1D73DF0CEBBD76865CA03BC4D80D5783F6497BBCAD6B25C38DBE7930E7
DF8FB0C0D20F42AF3219978FC1DE0D88F8C00AB74C63623391EE558AF8098CE3
A5D68B587E08323E7F107D2ABDE5FEFDF7075104FBCA57323F1FF30F5C754E7E
93BA7FBA82CF5444354F6B945378058F1BF7F34B40D27C98C6EA311164C30543
E344AE58ADD0EA272211883776F7211C7B23D3CAEDEC9D4CE4D49A8C24F9DB49
174E215795CDEADD74A5CB292E9FF1F9E84D8FE8C4D9761D02940BEE173B8AC5

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

5C153F11AEFD39AB9755F99A54820CE47AFBDADAE4A398225885B9F05FA0083C
E96A6F00D6BB178BE102FDB09D4EB28F910A06A495245E973F2A9E41A123D898
1FF49EF569E26F6727EAC56DA0439DFAC9B8551B64CCC389D6DED391D0B2A497
312A4ED6AA55733D703BAD412C3A34C5F237DEB4E750247DF0D3CA314F893AEF
2033AD482C938A498AE4F46FC3D5029790D80C30530B85F89E1E7C094CAF47C8
F761C786FF256206E37A9B2CDF7C52125C4EDE4445F43C1A3EC38E01CDF8D4D0
C3B9D3707A897E606796157BF8C9E3C63952B393F1460304E06B5CA82CCB7F93
A71A7C8BD378DE161E57AFFAA0F7971708BE0B77117F53771EB0B218D9421E14
8A7599B320F6E47EDB6C50E056FB626EF8213DC48767900A1FD20DA8CF076E86
01309F3AAE046AE1C1D63D7AA13BE7B0184145AE3D7A8BF71B740C9A1472F77D
81EED0018E44875764B36F2700ACBF2F67D2DFA7EAAC9592335575DBCAF72247
39FD69F1FDC86C39D53E0E6E23BF9E894E12EA90D829AA52410E106A8F2DA306
60515A2D7B11A86B01E3E00860CCF8B43E036287D1DE31B2EB4A9935FF8C68B9
9346D3C15CACF566394F70882F55CA9E3693210E4C065F0C5F862DCC11378A73
BD4934AC4604B87E04EC3F016C44C4D798FEBCDAC797ACEE4D8ADC1F527D552B
704047315CB66E4FE7CFA3F9233DF2D9DB831B87B3A6EC09E5BBF30923158653
5A2E8986EE381519482FC3C61415663008E0CF9446705BE3F749C375AED57959
C20562A24D094A276CA4F920FFF29BCD3C251C1419775EC763A78F03D399B898
13D65B5F083CF7CDD43A2ADAB7474E5E4D80CB6F6BD131C78019B3AAD34EDB11
00D31CD845E85E3686364D2D5BCBD5371A7A10F5E1C2A1F9A835995C5165CA0D
278D0E20C488E9A96FB35B5C85095DA3DD12A57B39576EC06953F325670253B7

Now we try to change one bit only in the InitiKEY (the bit 31th) to see how changing one bit on generating 40 subkeys.

InitiKEY:- $(E78102D\mathbf{\underline{8}}3C917CBCEA370AEF4390CAF62782A1DFC65290ABCE4D7143A8327D4A)_h$

TKEY:- $(857369A3AC754457F495BCA9A6AE1DDBFDD3EEABA6AA55539CD3EEABA6AA5553)_h$

As shown in table 3 the effect of changing begins from the second subkey to make the generated subkeys totally different from the first try (table 1).

Table 3: Generating 40 secret subkeys of size 256 bits (64 digits in hexadecimal) with modifying one bit only on InitiKEY.

E78102D83C917CBCEA370AEF4390CAF62782A1DFC65290ABCE4D7143A8327D4A
889DBE1CE458FDBFA90402C0CAD1340D11966A53203B207E0465BFECB88B7D7C
BF8AB4353C801A04B306968267B7B9DBE5682972E2D2DD7DA513479C7F689EE4
B1FCB85E02326B6C065D885078E20203736E9014DDC57EF1916E66A9B0AE4E1C
A8DB4D806510B83F603123B337CE2BC8BC20D5AAF0DD0A731745D20A934C06C6
BD46BCF779CA9A91A4262E62B2490856C35209C021E878AD8B34FBF247CBB6BB
BE89E9A8DB0A06A7F6595ED97904FFD2F10E68C2D783D244997FDF481D9E9465
123BEAE76DBA061EF3A73318EF0BAFC10615E2A3D6CC865DD27B84B282BD431F
D77D77D9896D3F3C340DDB76EFDD34174571A230179AB3720101D2BC9425442B
0965ECAF700B98CBF57489D45C253A04D8D81DE3D7A8FB8DB9FFD65FA417EC3E
0F1BD24F8A9C555097A18912C70598B3B5292B94AEF1BC043A827F4BBC03B69D
6E60D57C722900C102898D9B0E7DF3B2BEB63008A66A263256F110F5D06E35B1
271C798D35FEF8D707E416400D95A34715C194B5BDE06CB2FFBB4B822711A11D
A8C8E1231AD0F1C6FE3AD1032FEBC73172FB4C1B1F9FCE62FF9AB31BC28FE22E
92C9200548971C7178E41C861862F669F11EFF2CB4B3256F8A928CD1DB00EF51
6C65CE89B8E2DF04B55E25C8EB6F4F729EACF59EBF68B9623FEB663D3CF19131
49ADADF96958291166C3DFF6E0A3B36699E7EB1AA78C97748D9C7D84A9777C06
62C9ECD0059474E9ADFB76A1A6B89D3544A22466554E1E53E0A9A5293050D3A3
B77F2C0F6BE89011420B26DF1283C7DB98754DC2576C7B7E81CECF928D8B96C0
790C60215727F309801CF3F6016F3F296ED130FDF7C7F604702BC4CABAC1D436
F6C9B599BDBC5BB89593441120C75F010CD02D212C129D91DDA2F5AB66FF0046
30B68B799BDD7958B147FB9DCE148C1142848B731E6A1177F465454E68E224ED
5EBE4CDED8D639A3C148B92E76923C74C72E24F6E54A0C51FDA341014413FD3B
5988C8C40B6C3DDF2AFB1FBF9747EA4F8790B71248CA7DFBA7F123EDFD0A9A56

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

D15279FEA02D93039362C3187C07B6FEC529CAA3C1E7E414EE48BEFB938098B4
E082E285E3593E7CAA48003083DDEA09727F766537E9C423C8719F78D16C07AB
BD526A319212938ED1F06356C188C87C6AA1610E25E4437D103025377E58BBE7
89836FB5F8ECC858E6B0BA3A1E58E27FB3503EF5242E7BBF7BE655B32F480146
DB5F84FD77BA818676094931949BF00AA3C3FC78677AA3A99B012AB96F5FA72E
67C9B0CFE2587F57220DD029CBA4F4F8E7183DEE7AD890B1B2013885C3CEB6C5
2DB4417950A4469C3772F87F8F0EA86F3E2757885C4D5812E45C86DCDF39312E
C86D7AABA177C44B81AFA0E947EF48D89A35EB689370D50E71E309B4BD108E21
E2D9C91601E4E8E8AFD453DE55D0A6A45CDFB67C02D3366BEE395F76FE6F73FD
80DBDC13F9746AD3E30949A6895B90BF96FF407D96EE3B0328C312E0A76A1C97
EB5D48FB47E1594A95C3BE389AB3707ACAB90A1E1C1FF44AC55219D160E53E8A
6F8C2522FABC6E5D6C7591D4613E9CAED7958D11AA2191C56596625CEBE1BCF5
771D3931460F18D69AF71015D1371BA3C6DB0241D1ED32F9ACAAB1CB1BAA20D2
D24CA30F96E6EE4F5DC6BDF54379D3654B4A718E58EBD6F9DF648F8B97289259
140351EFA6721BD994542ACC8ABD6BAF995E7B638BF804FA323B897CE6B7E12C
075F4A0F0CFC44067423C4F0588746EB443451659370E37F1D086E92160A0732

Now we try to change one bit only in the InitiKEY (the bit 199th) to see how changing one bit on generating 40 subkeys.

InitiKEY:- $(E78102DA3C917CBCEA370AEF4390CAF62782A1DFC65290ABC\mathbf{\underline{F}}4D7143A8327D4A)_h$

TKEY:- $(857369A3AC754457F495BCA9A6AE1DDBFDD3EEABA6AA55539CD3EEABA6AA5553)_h$

As shown in table 4 the effect of changing begins from the sixth subkey to make the sixth generated subkey and forth are totally different from the first try (table 1).

Table 4: Generating 40 secret keys of size 256 bits (64 digits in hexadecimal) with modifying one bit only on InitiKEY.

E78102DA3C917CBCEA370AEF4390CAF62782A1DFC65290ABCF4D7143A8327D4A
B165EDDB4FFF684C403975C8B98FFFECD647109D30D7DCE98B4E5AB91B4162E2
F05BE7E7CBC136387F5B92F64732EF40DF806318149D3CCCC0DD099F487817FA
C5FA23D7372CE2668E26D0F033142AA5898BDB18C8F44B718F8125CF80083941
F274969EBBB26FD96ED253652542B154221CEF448D2B6CC8402AAD44F489DD0F
777010F6210508E27DFDC1278BCD8028CAD08E31537C8019C48E40DE59F5372C
0F0269DF00313364035905330DC7ACE4A0E3036599594DFF63E28C0540F1DCD5
F0095893B904D0FCED3CFD43A2E8926CB5632A7E3728266C098F4BDC61C4772A
9A859EA434A817649DA19C5D1EE2FB87A3E4BC492C2FDF2370CFC5A31BD8A586
81DCF934EBB102507C661F608089A503CBDE4CEA9D313803264324DD730EFFCA
4F90350AD81FDA575C78A955974DE1A23FD15D8F60E2FFD7AA75DE011EE5E79F
A03876207575C7F69AE8132A9966ABADC9EBAE0724F7964E1BA2489E604B33A3
FF611F6E73047F6D2F20ED2B68BFC682BE861524D2A6BFA6BC77D9ADE2EBA00D
582DF17DB994C862A67AAECC07BD3795F180F041DAA0393EE08B32A315A75BF7
F2B69B6525DAB3C482317C3AB548B001C404B7AAB1B98632F7653E651DDC9157
80F821C33D9A89C6ACE97DB7DC0D06B8F04C2357A17004FF6C86D89C6E07A806
671B721BF6E92877ED5464312F54B518DA907CF004434EA15B1C707D2E8C1172
B6D01411B5494145E7FBA7D63B6D972021FDCBCE4BD3D609E506DE25734F4B88
86B8CB0B8DFC6170DE31C2D485C28EBA65F9E329BB567F42074687F42D6CF54C
8EB28DA8CF6DDB59FD75BB560C1FFB3B949E947722CAB8BF4A3ED6D2EB7629D9
F9561834EA777AFB34A91812BFC57B9EF5F46DD8EC28C1141F262641D34AF9B6
8AD2A4BB0421AD78E21DBF513C2012288D172B05408B842C6E98302522EE403E
88378259BA23A9A80BB3E7B4BFF01BE5CCF1246A443F7B042DA86E462560EC50
BB84D7926B79625A727F3A8B04B25ED2C5EFBCD6A06EC4E3B608CBE817F22A11
D0506AF87A37E3E75B553D1D4484C78B6BA70BBF61980807C946E6A0F348D834
51B818F993153C4612DAB84A21DA198CEDF54E0AB9F4ED33F4332D76EF335B26
5812328DC54686D9350836097ECD6F048BE68EC44DAF2F056D37EBD418A932A9
B19195EBB157D5034E6F1B723B5A8F9372C45E4F5AAAE0A041580821D517495A
55551A2FA33E1B7B43A0955264B6ED15CE09726A3D47160A545B2B91E426F75F

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

999C9ED567BD9CDE691CCEFE15275C40265A7A2A2D39582192882A4E6104EC5F
74A5A433BC4EF8DF2DDFA89E91CC280F8A9796F1A7BD5B5D36A72CEAAD457D0A
95ABB0565B866F30F2B322F8BF915D52A5A43C4F428F737861AB703323DAB716
64138D39D94C4FB896B3925CE40BE9DEAB54C9E88347B0FD44DECA6FCA4D209C
97D668C0C87C31819A5FEC7E0871ED323F6A10FC395C6197B21BC26E6C47AC7F
6CDC32225A01103FCDB683D7E4FD95627F15C859BBB58A59ABC15E6B3BBABF6B
60124D54C62AAE191ED285F15C2F8AEFC380DB9B161364822797CB63C5F6C505
4A5244116226BF7112F99DC978E1456C6345A4DD565424B4290D9BF65EFDD289
7D08F9FCC4904DE545750B499F9137F8281EA07BA255C34E94448699960BA067
3D4DE39866D0C569DA910C8FA88CA0C735ABAF0C159A378C6F428C7FC8FB13B8
2F00F411BF4D9E09EE49A874681E93A9B2D0E8858DD5812CD7BED56A33CCEF76

If we change the bit 40 in the InitiKEY this will not make any changing on the list of the subkeys if we compare them with table (1), but if we change the bit 41 the changing will begin from the subkey number 6, also if we change the bit 42 the changing will begin from the subkey number 6.

If we change the bits 3 and 20 simultaneous on the InitiKey this will not make any changing on the series of the subkeys if they are compared with table 1, this test decrease the chance of diffusion.

If we change the bits 17 or 18 on the TKEY the action of changing on the series of the subkeys will begin from the subkey number 4 and forth if they are compared with table 1.

Any way according to the tests in the above with changing only one bit on TKEY or InitiKEY we conclude the subkeys will be different from the subkey number 6 and forth, so that the Roty_Shift proved the concept of the diffusion. Also there is no repetition so that there are M unique secret subkeys using Roty_Shift with any TKEY and InitiKEY, with some tests the system generated 5000 subkeys of size 256 bits and all of them were unique. The role of TKEY is more effect to give a fully diffusion than InitiKEY, so that Roty_Shift with InitiKEY is partially diffusion, while with TKEY is fully diffusion.

8- Conclusions:-

After preparing and testing Roty_Shift as a proposed algorithm is possible to conclude the following:-

1- Roty_Shift is good proposed method for generating a series of subkeys using two seeds; so that, it is possible to consider it as a KDF where the first seed is the first secret key, while the second seed is the second secret key.
2- This research applied the TRNG using the car plate number to give the first seed.
3- Roty_Shift algorithm applied the PRNG principle using the two seeds of n bits, the first seed selected randomly by the TRNG concept, while the second seed selected by the authorized person.
4- Roty_Shift proposed algorithm mixed the TRNG and PRNG to generate a series of subkeys.
5- Multi levels in Roty_Shift algorithm to increase the complexity of the subkeys generation process to create strong subkeys.
6- This research depends on the concept of rotation with other process to get a series of secret subkeys.
7- There are M unique subkeys generated by the Roty_Shift, this research tested with 16, 36, and 256 bits key size, and we concluded; if we maximize the key size then the probability of uniqueness will be greater (series of M unique subkeys).

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

8- Roty_Shift method it near to be as fully diffusion (for the generated subkeys).
9-In diffusion the TKEY (seed1) is more effect than InitiKEY (seed2).
10-This research gives a chance for the researchers for more testing and may modify and apply it with any encryption method; especially this research generates all the required secret subkeys for the encryption process.
11-Just by applying XOR operation between one block of the plaintext with one subkey an encryption method will be proposed, and it is good to make some confusion by systematic unordering the generated list of the subkeys.
12- The main problem of the proposed method is the runtime, it is take long time to generate a list of M unique subkeys, where as it is take 517 seconds to generate 2000 unique subkeys of size 256 bits using Microsoft visual studio ultimate (VB) with Celeron CPU of 2.0 GHz, 4GB RAM, and windows 7 operating system.

References:-

[1] E. Barker, A. Roginsky, "Recommendation for Cryptographic Key Generation", NIST Special Publication 800-133, December 2012.
[2] A. Menezes, P. Oorschot and S. Vanstone, "Handbook of Applied Cryptography", CRC press, 1997.
[3] K. M. MARTIN, "Everyday Cryptography: Fundamental Principles and Applications", OXFORD university press, 2012.
[4] W. Stallings, "Cryptography and Network Security", 4th Edition, Prentice Hall, 2006.
[5] C. Paar, J. Pelzl, "Understanding Cryptography", Springer-Verlag Berlin Heidelberg, 2010.
[6] D. E. Denning ,"Cryptography and Data Security", Addison-Wesley, 1982.
[7] L. M. Jawad, G. Sulong, " A Novel Dynamic Secret Key Generation for an Efficient Image Encryption Algorithm", Modern Applied Science; Vol. 9, No. 13; 2015.
[8] H. Abunahla, D. Shehada, C. Y. Yeun, B. Mohammad, and M. Jaoude, "Novel Secret Key Generation Techniques Using Memristor Devices" API advances, volume 6, issue 2, 2016.
[9] A. Soni1, S. Agrawal, "Key Generation Using Genetic Algorithm for Image Encryption", IJCSMC, Vol. 2, Issue. 6, June 2013.
[10] P. R. Gurubilli, "Random Number Generation and its Better Technique", A thesis submitted in partial fulfillment of the requirements for the award of degree of Master of Engineering in Computer Science and Engineering, 2010.
[11] W. Trappe, L. C. Washington, "Introduction to Cryptography with its Coding Theory", Second Edition, Person Prentice Hall, 2006.

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

« Ambulance diversions reducing and dispatching theory for rescue operations »

[(1)] Y. BOUHALLAF, [(1)] L. RADOUI, [(2)] O. MALASS, [(1)] H. BELHADAOUI, [(1)] M. RIFI

[(1)] Laboratory of Network, Computing, Multimedia & Communication, EST de Casablanca, University of Hassan II, BP. 8012, Morocco.

[(2)] Ecole Nationale Supérieure des Arts et Métiers, 4 rue Augustin Fresnel, 57078 METZ Cedex 3, France.

Abstract— World population keeps growing up and injuries related death statistics is increasing. Optimizing healthcare logistic processes became then a vital need to lead patient cares to higher performances. Moreover, Worldwide healthcare systems are facing the challenge of the sophistical facilities rising costs as well as patients' requirement of high-quality care at lower cost. In the other hand, undetected behaviors of citizens and environmental constraints are influencing the quality of deployment which amplifying the response time threshold. In the present paper, we regulate vehicles capacities to optimize patients picking for each incident nature. We proposed also a dynamic vehicle relocation and routing using a decision making processes. We are considering for each decision to take, the aspect of the variable emergency constraints influence to satisfy different scenarios of daily life.

Keywords: *Ambulance services, Dynamic dispatching, Route changing, Petri Net Processes Modeling.*

I. Introduction

In order to arrange transport insurance for citizen requests needs a smart strategy to keep main transport services smooth against environment parameters disrupting. Moreover, rescue services, such as ambulance transportation services, represent a substantial system since it is the most intervening one. It is characterized by additional parameters related to emergency conditions and patient health itself. Indeed, people are affected or killed when rescue operations are slowed by bottlenecks, public works or natural disasters. Customers must be delivered to emergency units within the shortest response time [1][2]. A response time represents time interval between call arriving at emergency platforms and the arrival time of deployed ambulance to patient locations. Also to minimize the so-called response time, deployment should take into consideration human behaviors. such as inappropriate driving, routing/walking through emergency lanes and hostage situation to satisfy the coverage level [3]. This coverage level is a notion of commonly used metric parameter defined as the proportion of patients who are responded to within a given distance/time threshold [4]. The survivability rate has been reported by several works in the literature review [5][6], and this, by working on the balance between the minimum response time and/or the maximum coverage. With the ultimate goal to save lives, efficient operations depend on resource allocation decisions, vehicles' position, and demands zones. Allocation and ambulance dispatching assign appropriate vehicle toward the call point; the most common dispatching rule used in practice is to send the closest unit available [7][8] to lead customers to the closest hospital. This policy is rational since the objective is to minimize the response time. However, this theory requires servers with equal technical capabilities to treat several incidents risk level [9], as well as an unlimited hosting capacity of each emergency unit, which is not the case in reality. Patients transportation and hosting capacities are technically limited. Indeed, several works such as [10] mark the impact of ambulance diversion on response time. However, when the hospital is overcrowded, this situation requests of incoming units to seek for available hospitals. Vehicles are rerouted then so their paths become farther and their responses time become longer too. A number of root causes have been proposed to explain Emergency Department(ED) overcrowding and ambulance diversions. However, most of discussions are limited either to qualitative commentary and surveys [11] [12] [13] [14] or to single hospital empirical studies [15] [16] [17] [18]. The work [19] proposes a simple queuing network model to describe the patient flow between EDs and the inpatient department. It aims to derive two separate sets of measures for inpatient occupancy and ED size. Also, this work uses these sets of measures to form hypotheses and test them by estimating a sample selection model using data on a cross section of hospitals in California. The work [20] sets up a cooperative strategy to reduce ambulance diversions where emergency services must coordinate dispatching between them. Thence, EDs operating independently will change to a completely centralized system. On the other hand, literature works treat also transport capacities management, such as the work [21] which proposes capacity management for transport within isolation facilities, which are reserved for patients with Highly Infectious Disease (HID) that impose safety measures and desires rapid relocation of patients. The work is representing a capacity regulation using different ambulances availability and technical specifications, and reflects different preparedness levels among sixteen European nations. Hence, regulations for technical specifications and operational procedures should be harmonized in order to promote both patient and health-care staff safety.

The frequency of disasters, whether the natural or the human-made, has increased to an unprecedented level in the last decade [22]. Disasters, as well as terrorist attacks, are characterized by long-term Socio-economic, psychological and physical hazard impacts. For that reason, an enormous amount of works and theses focus on transportation systems performance. They represent quantitative and qualitative measures. Some of them propose conceptual frameworks and performance metrics to set up strategic decisions that improve

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

preparedness and reduce the duration of recovery. Other works provide models responding to disasters (e.g., evacuation planning, resource allocation). The work [23] represents a detailed overview of literature review papers related to transportation system performance for disaster events. It provides a synthesis and classification of recent topic works based on host criteria, methodology and approaches.
Optimizing patient flow improves the quality of care through the use of limited resources [24] [25]; especially those routed through complicated or damaged infrastructures. It treats ambulance diversion effect which contributes to an improving response time. It aims also to manage capacities needed to response incidents with different risk levels. It also regulates a dispatching policy base on the relation between the available resources and demand level. And so, we aimed by the previous work [26] to minimize response time by reducing ambulance diversion. And by the present, to optimize patient transport performances. It is invoking a dynamic dispatching policy which is based on the capacity variation along the day. The aim of this work is to make suit decisions to improve vehicles performance responding to daily life scenarios and to serve disaster events where impacts are heavy and critical.

II. PROPOSED APPROACH

The present paper advances a decision aid tool modeling which represents an executive stage of the strategy we proposed in work [26] to define each patient destination. It is reserved to execute rescue operations, after being sure of avoiding ambulance diversion. To manage capacities, it starts by calculating transportation needs process, which defines the number of ambulances needed for each discipline to treat per scenario. It tries next to relocate, deploy and then route all ambulances till rescue operations end. This study takes into consideration the presence of psychological and social cases that can tackle services intervention, amplify ambulance responses and increase vehicles busy rate.

A. Treatment of Transport needs (**Sub process**: **TraitementTypeAmbulanceNecessaire**)

This sub-process is designed to calculate the transportation needs. This requires defining the nature of each incident, which represents the number of the patients to carry. It also requires a knowing of the category of ambulances to deploy according to each medical discipline imposed. The transport capacity "CT" result then according to each incident size and risk level. The approach is proposed in the interest of making the aid tool more dynamic and prepared to resolve the daily scenarios as well as catastrophic cases. Indeed, disaster scenarios represent incidents with a large number of injured persons, who need to be carried to one or different emergency centers. Figure 1 explains in details the methodology to balance capacity treatment.

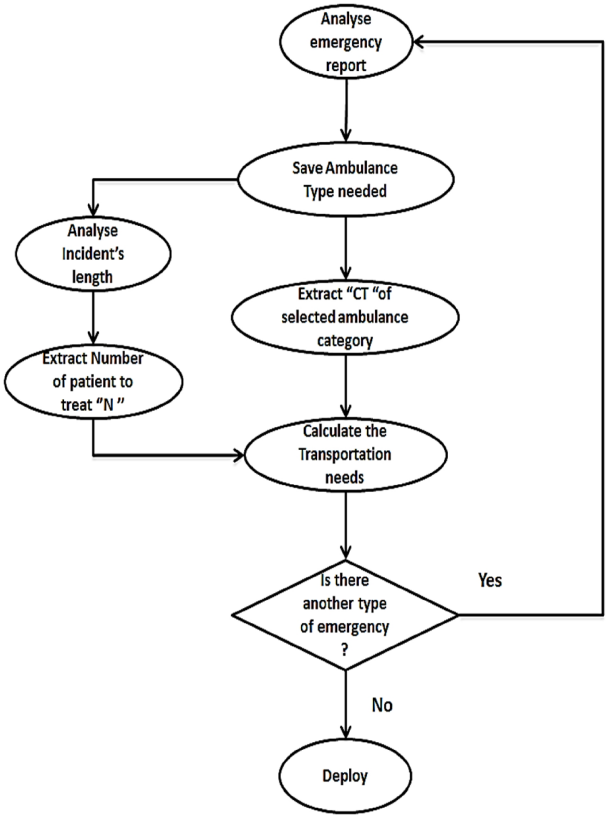


Fig.1 – Methodology of calculating transport needs per scenario

The methodology of defining transport needs is modeled by the Petri Network in figure 2.

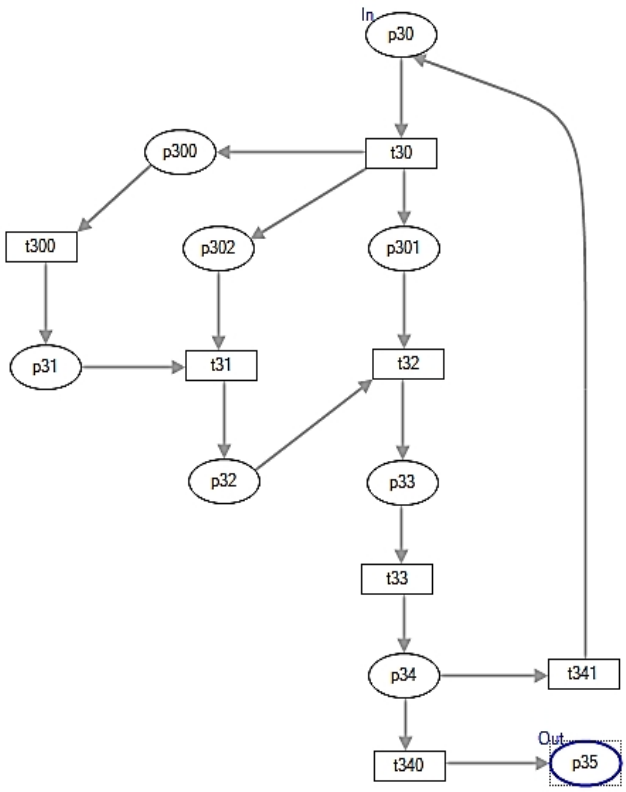


Fig.2 – Sub-process of calculating transport capacity requested

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

Table.1 – Descriptive transportation needs sub process

Place/ transition	Designation
P30/T30	Load report analysis/report loaded
P300/T300	Extract incident size / Incident size saved
P301	Extract ambulance category needed / Ambulance category defined
P302	Extract transport capacity of selected category
P31	Define the number of patient to transport
T31	Save data extracted
P32/ T32	Calculate number of ambulances needed / Number calculated
P33/T33	Save result (Amount of Ambulance to deploy + Category)
P34	Is there another category needed?
T340/T341	No, there is only one discipline to treat /Yes, emergency represent multiple disciplines to treat
P35	Dispatch via next sub-process

*B. Multi Ambulance relocation (**Sub-process: Localisation2Ambu**)*

The sub-process of relocation that we proposed in work [26] aims to check, complete diagnostics, and give the first aid by sending the closest vehicle to incident places. This deployment has to valorize decisions to take by collecting key information that patients could not communicate. Whereas this sub-process represents an advanced version of ambulance relocation. It is reserved for rescue operations. It provides multiple relocations of ambulances to satisfy the pre-calculated number of vehicles requested to treat every incident. It improves also vehicles availability by prioritizing deployment of ambulances on free roam before sending ambulances of nearest base stations. Figure 3 illustrates the sub process modeled by the Petri Net.

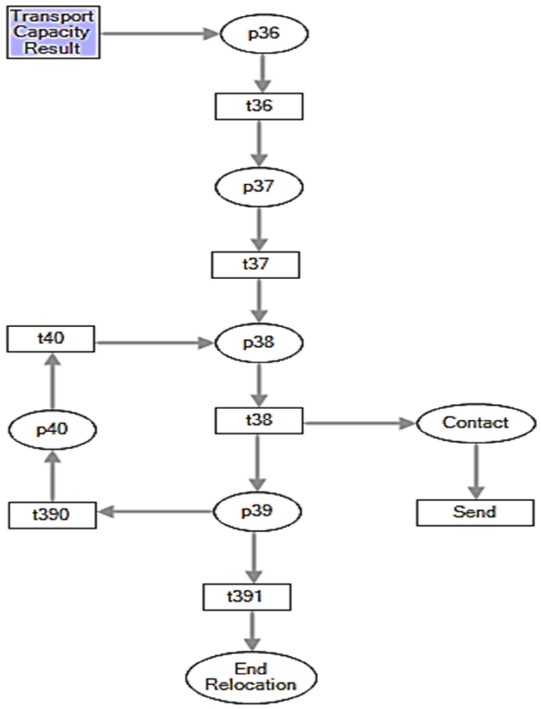


Fig.3 –Multi Ambulance relocation sub-process

Table.2 –Multi Ambulance relocation sub process

Place /Transition	Designation
Transport Capacity Result	Result of calculated need treated by the previous sub-process
P36/T36	Load Ambulances Coordinates
P37/T37	Localize ambulance type calculated / ambulances located
P38/T38	Select the closest one
P39	All Transport needs satisfied?
T390/T391	No, Transport need not satisfied / Yes, Transport need is satisfied
P40/t40	Update ambulances location data / Data updated
End relocation	End of sub-process "Multi Ambulance Relocation"
Contact/ Send	Contact and deploy ambulance selected

*C. Ambulance Post-carriage (**Sub-process: PostAcheminement**)*

The sub-process post-carriage provides flexible trips and offers the same advantages of sub-process pre-carriage detailed in work [26], in which ambulance routes throughout dynamic paths. Pre/Post-carriage use traffic flow monitoring to anticipate bottlenecks presence and create new routes. Indeed, the pre-carriage is used to lead vehicles to incident places (incident places coordinates used) whereas the post-carriage orients ambulances to hospitals destinations (hospitals coordinates used). The post-carriage is modeled by the Petri Net in figure 4.

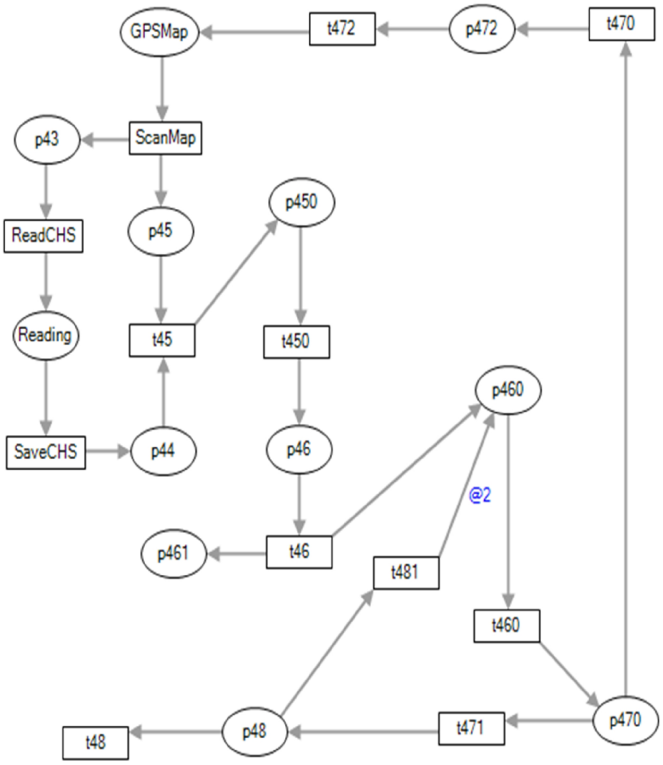


Fig.4 – Ambulance Post-carriage sub-process

Table.3 – Ambulance Post-carriage sub-process

Place/transition	Designation
GPSMap/ ScanMap	Load ambulances and hospitals coordinates/data loaded

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

P45	Read dispatched ambulance coordinates
P43	Load Hospital coordinates data
P44	Save Hospital coordinates data
T45	Data needed is uploaded
P450/T450	Calculate all possible ways/ways calculated
P46/T46	Choose optimum way (shortest trip time estimated)
P461	Inform driver to switch to the newest path
P460/T460	Load traffic data
P470	Bottleneck detected?
T470/T471	Yes/No
P472/T472	Calculate New Path
P48	Did ambulance arrive?
T48/T481	Yes, End of sub process / No

The end of Post-carriage sub-process marks the end of rescue operations. Ambulances return to their initial stations marked as free roam vehicle. The analysis takes into consideration, the presence of social problems (e.g., hostages) to bypass the access to patients if needed. Figure 5 represents the main model of decision-making tool. It unit approaches proposed in previous and present work, to establish interactions between processes.

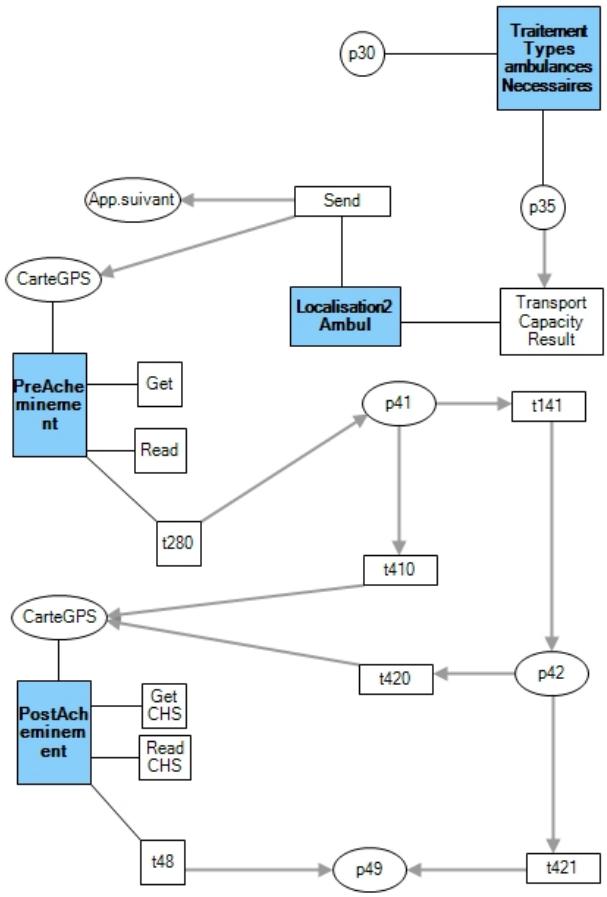


Fig.5 – Model of the proposed approach for this study

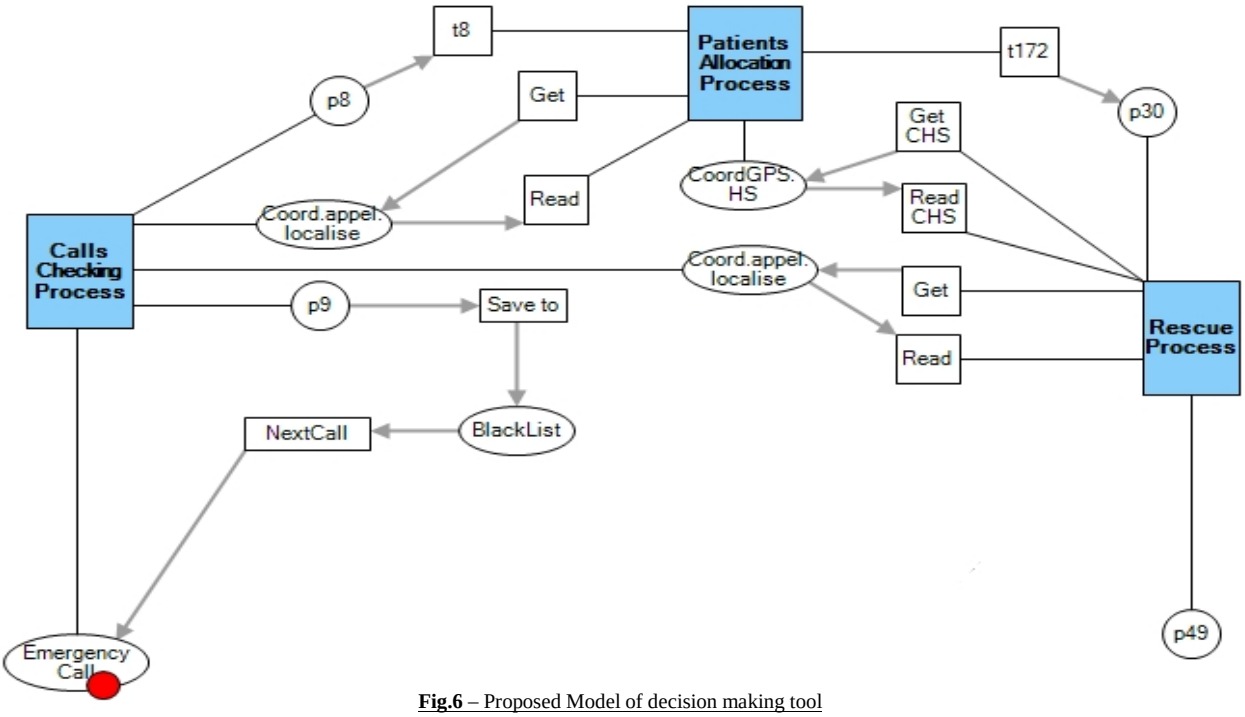


Fig.6 – Proposed Model of decision making tool

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

Table.4– Additional details for decision-making tool

Place /Transition	Designation
P41	Prepare patient's transportation
T410/T141	Transport prepared / No, There is a social problem
P42	Asks for Police department assistance
T420/T421	Transport allowed/ Transport not allowed
P49	Free vehicles on free roam mode

This study proposes a model for decision support tool. As represented, it is a decisional system based on several constraints cited by researchers. We include new constraints for improvement to decrease the response time. Social problems are slowing down or blocking rescue actions (robbery, hostage). It requires the presence of the police department interventions or any defense department which may or not allow the continuity of saving operation. Thus, the interaction of the three main processes is done using the General model presented in figure 6. Simulation results, illustrated in the next section, reflect the system performances.

III. SIMULATION

The work [26] shows how decisions are avoiding false deployment via the checking call approach. The sub process proposed in the same work "TraitementHopitalDestination" allocates the number needed for hosting hospitals to avoid ambulance diversions. Figure 7 shows the total time consumed by this sub-process to treat different incident types. Thus the maximum time needed to treat emergencies goes up to 1400 sec/100 ambulances.

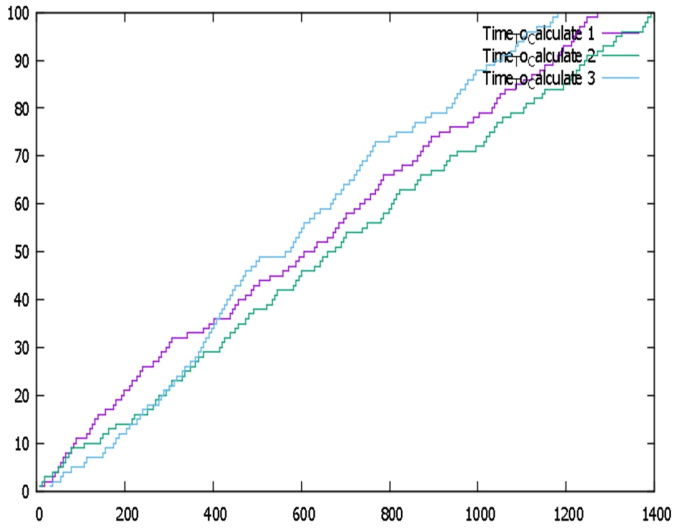


Fig.7 –Total time needed for hospitals needs calculating

Transportation needs in vehicles

In this part, the simulation shows that incidents may contain a unique type of emergency (only one discipline needed to treat patients) or they may be composed of several types of emergencies. According to the results in figure 8, incidents may require up to three different disciplines in one emergency case.

Statistics				
Name	**Avrg**	**StD**	**Min**	**Max**
Urgence_Mixtes				
count_iid	1.800000	0.918937	1	3
max_iid	1.800000	0.918937	1	3
min_iid	1.000000	0.000000	1	1
avrg_iid	1.351754	0.409820	1.000000	1.894737

Fig.8 – Average number of disciplines required peer emergency

Ambulance by using rate

Basing on the illustration of figure 8, the present study proposes the use of three vehicle types. Each vehicle has a proper technical capability to cope against discipline varieties. Figure 9 presents the rate of use for each ambulance category. The proposed ambulances type A is reserved for daily use, supposed to carry one patient. They are equipped for the most extreme cases (defined as code red by authors in literature review). It should represent 54% of the fleet of transport according to detected needs. The category B is reserved for incidents with medium urgency. It can carry up to 4 patients and finally the category C for a low emergency degree with a transport capacity of 5 patients.

	Ambulance A	Ambulance B	Ambulance C
Transport Needs (%)	54,00%	27,00%	19,00%
Transport Capacity(patient)	1	4	5

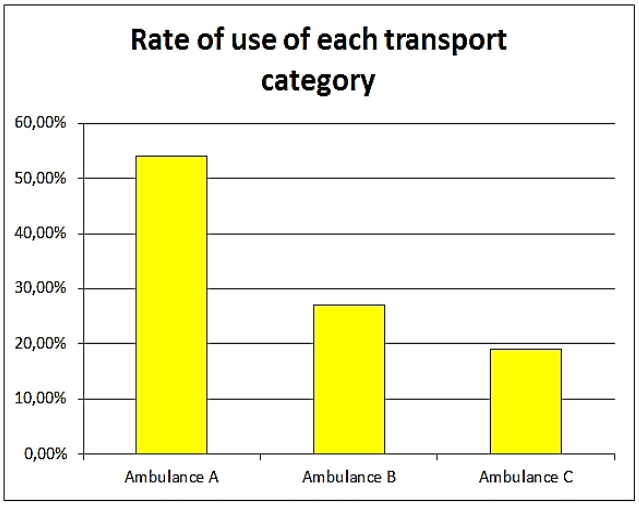


Fig.9 – Rate of use of proposed ambulances categories

Multiple relocation

Figure 10 is the result of the transport needs treatment. In extreme cases, the need in vehicles (CT) can reach up to 16 ambulances/emergency. Also, the result (fig11) shows the number of busy ambulances during deployments. It varies

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

between 26,7% and 50% depending on the different periods of the day.

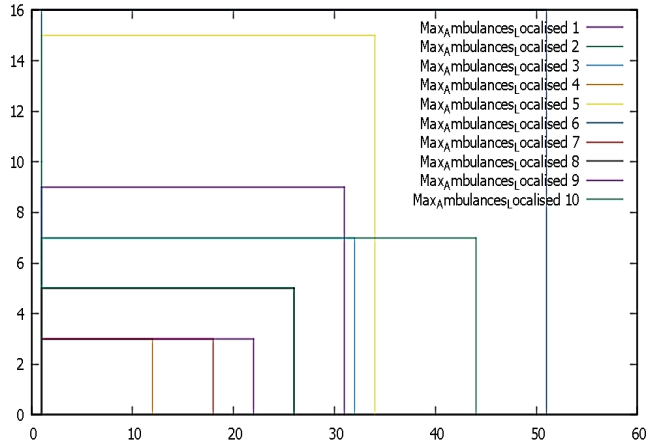


Fig.10- Ambulances location to satisfy the transport needs

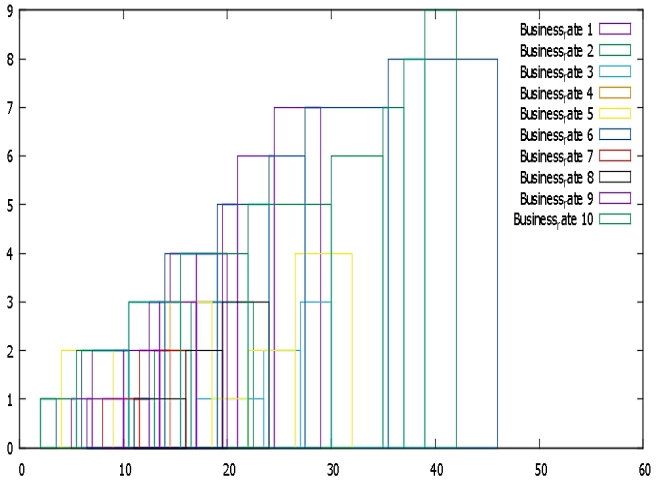


Fig.11 – Rate of busy ambulances for emergency deploying

IV. CONCLUSION

The present article is a proposed study for a decision-making system; it aims to improve the response time through an optimization model. The proposed sub-processes suggest solutions against the influence of the environment on the smooth functioning of ambulance services. It exposes results from analysis of inserted decisions toward regulation objectives. It aims to cope with several scenarios of daily life and disasters representing large damages. On the other hand, we combined between different objectives proposed by the previous research. So far, we replaced the verbal verification process by more measurable and more vivid one. It adapts also to the bundling and unbundling of the disciplines within the emergency centers of different areas architectures. It responds properly to required capabilities treatment with an efficient use of emergency vehicles and carries patients through the shortest paths. As perspective, the study will integrate the tool to make it usable by emergency platforms to serve the patient in urban and rural areas.

V. REFERENCES

[1] Nogueira, L. C., Pinto, L. R., & Silva, P. M. S. (2016). Reducing Emergency Medical Service response time via the reallocation of ambulance bases. Health care management science, 19(1), 31-42.
[2] Nogueira, L. C., Pinto, L. R., & Silva, P. M. S. (2016). Reducing Emergency Medical Service response time via the reallocation of ambulance bases. Health care management science, 19(1), 31-42.
[3] Degel, D., Wiesche, L., Rachuba, S., & Werners, B. (2015). Time-dependent ambulance allocation considering data-driven empirically required coverage. Health care management science, 18(4), 444-458.
[4] Echoka, E., Kombe, Y., Dubourg, D., Makokha, A., Evjen-Olsen, B., Mwangi, M., ... & Mutisya, R. (2013). Existence and functionality of emergency obstetric care services at district level in Kenya: theoretical coverage versus reality. BMC health services research, 13(1), 113.
[5] Sanchez-Mangas, R., García-Ferrer, A., de Juan, A., & Arroyo, A. M. (2010). The probability of death in road traffic accidents. how important is a quick medical response? Accident Analysis and Prevention, 42, 1048–1056.
[6] McLay, L. A., & Mayorga, M. E. (2011). Evaluating the impact of performance goals on dispatching decisions in emergency medical service. IIE Transactions on Healthcare Systems Engineering, 1, 185–196.
[7] Dean SF (2008). Why the closest ambulance cannot be dispatched in an urban emergency medical services system. Prehospital Disaster Medicine 23: 161–165.
[8] Hayes J, Moore A, Benwell G and Wong B (2004). Ambulance dispatch complexity and dispatcher decision strategies: Implications for interface design. Lect Notes Comput Sc 3101: 589–593.
[9] El Sayed, M. J. (2011). Measuring quality in emergency medical services: a review of clinical performance indicators. Emergency medicine international, 2012.
[10] Deo, S., & Gurvich, I. (2011). Centralized vs. decentralized ambulance diversion: A network perspective. Management Science, 57(7), 1300-1319.
[11] Derlet RW, Richards JR (2000) Overcrowding in the nation's emergency departments: Complex causes and disturbing effects. Ann. Emergency Medicine 35(1):63–68.
[12] The Lewin Group (2002) Emergency Department Overload: A Growing Crisis—The Results of the AHA Survey of Emergency Department (ED) and Hospital Capacity (American Hospital Association, Falls Church, VA).
[13] GAO (2003) Hospital Emergency Departments: Crowded Conditions Vary Among Hospitals and Communities (U.S. General Accounting Office Washington DC).
[14] Burt CW, McCaig LF (2006) Staffing, Capacity, and Ambulance Diversion in Emergency Departments, United States 2003–2004 (US Department of Health and Human

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

Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Atlanta).
[15] Schull MJ, Lazier K, Vermeulen M, Mawhinney S, Morrison LJ (2003a) Emergency department contributors to ambulance diversion: A quantitative analysis. Ann. Emergency Medicine 41(4):467–476.
[16] Han JH, Zhou C, France DJ, Zhong S, Jones I, Storrow AB, Aronsky D (2007) The effect of emergency department expansion on emergency department overcrowding. Acad. Emergency Medicine 14(4):338–343.
[17] McConnell KJ, Richards CF, Daya M, Bernell SH, Weathers CC, Lowe RA (2005) Effect of increased ICU capacity on emergency department length of stay and ambulance diversion. Ann. Emergency Medicine 45(5):471–478.
[18] Forster AJ, Stiell I, Wells G, Lee AJ, Van Walraven C (2003) The effect of hospital occupancy on emergency department length of stay and patient disposition. Acad. Emergency Medicine 10(2):127 133.
[19] Allon, G., Deo, S., & Lin, W. (2013). The impact of size and occupancy of the hospital on the extent of ambulance diversion: Theory and evidence. Operations Research, 61(3), 544-562.
[20] Hagtvedt, R., Ferguson, M., Griffin, P., Jones, G. T., & Keskinocak, P. (2009, December). Cooperative strategies to reduce ambulance diversion. In Winter simulation conference (pp. 1861-1874). Winter Simulation Conference.
[21] Schilling, S., Maltezou, H. C., Fusco, F. M., De Iaco, G., Brodt, H. R., Bannister, B., ... & Ippolito, G. (2015). Transportation capacity for patients with highly infectious diseases in Europe: A survey in 16 nations. Clinical Microbiology and Infection.
[22] Guha-Sapir, D., Vos, F., Below, R., and Ponserre, S. (2011). "Annual disaster statistical review 2011: The numbers and trends, Centre for Research on the Epidemiology of Disasters (CRED), Institute of Health and Society (IRSS), Université Catholique de Louvain, Brussels, Belgium.
[23] Faturechi, R., & Miller-Hooks, E. (2014). Measuring the performance of transportation infrastructure systems in disasters: A comprehensive review. Journal of infrastructure systems, 21(1), 04014025.
[24] Kriegel, J., Jehle, F., Dieck, M., & Tuttle-Weidinger, L. (2015). Optimizing patient flow in Austrian hospitals–Improvement of patient-centered care by coordinating hospital-wide patient trails. International Journal of Healthcare Management, 8(2), 89-99.
[25] Kriegel, J., Jehle, F., Moser, H., & Tuttle-Weidinger, L. (2016). Patient logistics management of patient flows in hospitals: A comparison of Bavarian and Austrian hospitals. International Journal of Healthcare Management, 9(4), 257-268.
[26] Bouhallaf. Y, Malassé. O, Rabbah. N, Belhadaoui. H, Rifi. M,.(2017) Decision Support Tool for Traffic Management. International Journal of Computer Science and Information Security. (pp. 222-230).

AUTHORS PROFILES

Youssef BOUHALLAF is currently a PhD student, Research laboratory RITM (Networks, Computer, Telecom and Multimedia) in High School of Technology /ENSEM University Hassan II Casablanca Morocco. Obtained Master degree in Industrial Logistics.
Research: Urban and health care transportation, traffic management, e-logistic processing.

Lamiae RADOUI is currently a Ph.D. student, Research laboratory RITM (Networks, Computer, Telecom and Multimedia) in the ESTC - University of Casablanca Hassan II researching on intelligent transport systems which manages the road traffic with safety, comfort and saving time.
Research : Smart city & intelligent transport, smart traffic management.

Olaf MALASS is currently attached with National School of Arts/ and Crafts/ ParisTech in Metz/France as Associate Professor in A3SI department.
Research: Automatic Signal Processing and Computer Engineering.

Hicham BELHADAOUI is currently working as a Professor Ability in University Hassan II /ESTC, Casablanca Morocco. Received his PhD degree at the National Polytechnic Institute of Lorraine/France.
Research: Security, Reliability, Automatic Signal Processing and Computer Engineering.

Mounir RIFI is currently working as a Professor in University Hassan II /ESTC, Casablanca Morocco. Obtained his PhD Physical Sciences: Electromagnetic Compatibility, October 1996 (University Mohamed V of Rabat - Morocco) and PhD in Electronics, May 1987 (University of Lille - France) Director of the Research Laboratory: RITM (Networks, Computer, Telecom and Multimedia)
Research: Propagation of electromagnetic waves, ElectroMagnetic Compatibility, RFID, Microwave, Transmission Lines Theory, Antennas, Sensors, Networks.

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

The application of predictive analytics in healthcare sector

FATIMETOU ZAHRA MOHAMED MAHMOUD
Faculty of ICT, International Islamic University Malaysia (IIUM), Kuala Lumpur, Malaysia
E-mail: fatimetou1991@hotmail.com

ABSTRACT

In fact, predictive analytics have been used in many and different sectors and industries such as manufacturing, education, market and in healthcare. As a matter of facts, the predictive analytics are considered as an opportunity for the healthcare sector to be able to extract valuable information from data and predicting the future. Moreover, this opportunity can transform the healthcare to not only predictive but to a preventive sector by the early detection of risks and the ability to take better decisions and saving more people lives. this paper, study the current application of predictive analytics in healthcare, why the predictive analytics have been used, what are the approaches followed by the previous research and the points of focus, what are the tools used in predictive analytics application, what are the algorithms and models used for the predictive analytics application, what are the main benefits and challenges of the predictive analytics application in healthcare? Moreover, this research suggests the necessity to have and developing a framework that guide the application of predictive analytics in addition to have a new holistic process and methodology to get better results and overcoming challenges and low revenues.

Keywords: **healthcare, predictive analytics, models accuracy, tools, algorithms, integration, automation**

1- INTRODUCTION

Nowadays, organizations can predict the future by benefiting from the evolution of business intelligence systems especially the power of the combination of predictive analytics and big data. Thus, predictive analytic refer to set of techniques used to predict the future based on historical or static data (Bas Geerdink. 2013). While (Nishchol Mishra, & Dr.Sanjay Silakari.2012) see predictive analytics as an advanced branch of data engineering which utilize the data mining techniques to recommend best decisions based on the prediction of the future events. Furthermore, its capable to deal with constant and discontinuous changes. Moreover, it utilizes analytical techniques, statistical models and algorithms to allow finding meaningful information's and future trends in the data (Bas Geerdink. 2013; Nishchol Mishra, & Dr.Sanjay Silakari.2012) . In fact, the base of predictive analytics are the models used which are developed by the predictive analytics techniques analytical and statistical, those models are continuously optimized, modified and trained depending on the users alerts and environment (Bas Geerdink. 2013; Nishchol Mishra, & Dr.Sanjay Silakari.2012). Furthermore, those models can take different forms depending on the data used (Nishchol Mishra, & Dr.Sanjay Silakari.2012). Another (James Ogunleye. 2014) define Predictive analytics as the set of expertise, skills, and technology to extract, analyze and converting data to clear and meaningful information that help in better decision making and is one of the predictive analytics advantage its ability to transform the decision making from intuition based to

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

facts based decision. In addition, the fast and significant increase of big data was one of the rationale to promote the necessity of predictive analytics and appreciate it by organizations. Thus, organizations use predictive analytics to be able to make strategic business decisions based on facts, patterns and accurate future trends predicted with these systems with lower costs. For instance, predictive analytics can be used in retailers such as supermarket which use predictive analytics for analyzing the sales data the historical and current data in order to extract patterns in customer behavior and utilize it to forecast what is the products that the customers will probably buy. Moreover, predictive analytics, is used in banks, healthcare, and insurance industry (James Ogunleye. 2014; James Ogunleye. 2015).

Healthcare predictive systems are analytic systems which aim to minimize the future medical cost and help to provide in hospital a high level of healthcare and preventive healthcare due to the early detection of risks and possibility to take better actions and decisions. In fact, those predictions are based on the historical patients' data including detailed information about the patient, his medical history and diagnoses (Mohammad Ahmad Alkhatib, Amir Talaei-Khoei, & Amir Hossein Ghapanchi. 2015).

In fact, predictive analytics have many benefits by its application in different industries such as detecting patterns in data and discovering opportunities, testing variables (Mohammad Ahmad Alkhatib. 2015; Hoda Moghimi, Stephen Vaughan, Steven McConche, & Nilmini Wickramasinghe.2016; Mohammad Ahmad Alkhatib, Amir Talaei-Khoei, & Amir Hossein Ghapanchi. 2015), increasing the organizations profit (Mohammad Ahmad Alkhatib. 2015; Hoda Moghimi, Stephen Vaughan, Steven McConche, & Nilmini Wickramasinghe.2016), predicting risks and prevent it (Mohammad Ahmad Alkhatib. 2015; Hoda Moghimi, Stephen Vaughan, Steven McConche, & Nilmini Wickramasinghe.2016; Chandra J, Dr Nachamai.M, & Dr Anitha S Pillai.2015;Mohammad Ahmad Alkhatib, Amir Talaei-Khoei, & Amir Hossein Ghapanchi. 2015), augment the organizations revenue by enhancing the key metrics and making strategic corrections in the organization (Prasada Babu, & S.Hanumanth Sastry.2014; Meryem Ouahilal, Mohammed El Mohajir, Mohamed chahhou, & Badr Eddine El Mohajir.2016).
Moreover, by its use in healthcare predictive analytics have shown a lot of benefits such as minimizing the healthcare costs (Mohammad Ahmad Alkhatib, Amir Talaei-Khoei, & Amir Hossein Ghapanchi. 2015; Raid Lafta, Ji Zhang, Xiaohui Tao, Yan Li, & Vincent S. Tseng.2015; Muhammad Kamran Lodhi, Rashid Ansari, Yingwei Yao, Gail M. Keenan, Diana J. Wilkie, & Ashfaq A. Khokhar.2015), providing better quality of healthcare services (Muhammad Kamran Lodhi, Rashid Ansari, Yingwei Yao, Gail M. Keenan, Diana J. Wilkie, & Ashfaq A. Khokhar.2015), early detection of risks and transform the healthcare to a preventive healthcare which help significantly in saving people lives(Osama T. Ali, Ali Bou Nassif, & Luiz Fernando Capretz.2013 ;Mohammad Ahmad Alkhatib, Amir Talaei-Khoei, & Amir Hossein Ghapanchi. 2015), improve the decision making in hospitals and better resource allocation (Aly Megahed, Guang-Jie Ren, & Michael Firth.2015; Samet Ayhan, Johnathan Pesce, Paul Comitz, David Sweet, Steve Bliesner, & Gary Gerberick. 2013; N. Ayyanathan, & A. Kannammal.2015; Nawal N. Alotaibi, & Sreela Sasi.2015; GINA GUILLAUME-JOSEPH, & JAMES S. WASEK.2015; Raid Lafta, Ji Zhang, Xiaohui Tao, Yan Li, & Vincent S. Tseng.2015).

Although, as any other systems predictive analytics face many challenges and issues such as the level of accuracy of the models results can be less or lower than the expected due to some issues in the predictive analytics process and models used (Samir E AbdelRahman, Mingyuan Zhang,

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

Bruce E Bray, & Kensaku Kawamoto.2014; Ravi S. Behra, Pranitha Pulumati, Ankur Agarwal, & Ritesh Jain.2014; Alexey V. Krikunov, Ekaterina V. Bolgova, Evgeniy Krotov, Tesfamariam M. Abuhay, Alexey N. Yakovlev, & Sergey V. Kovalchuk. 2016; Noura Al Nuaimi. 2014), issues in data such as the lack of input data or insufficient data to make the analysis (Samir E AbdelRahman, Mingyuan Zhang, Bruce E Bray, & Kensaku Kawamoto.2014; Eman AbuKhousa, & Piers Campbell.2012; Riad Alharbey.2016; George Mathew, & Zoran Obradovic.2012; Johannes Kunze von Bischhoffshausen, Markus Paatsch, Melanie Reuter, Gerhard Satzger, & Hansjoerg Fromm.2015), problems in data integrity such as the incomplete data (Eman AbuKhousa, & Piers Campbell.2012; Johannes Kunze von Bischhoffshausen, Markus Paatsch, Melanie Reuter, Gerhard Satzger, & Hansjoerg Fromm.2015; Yang Xie, G¨unter Schreier, David C.W. Chang, Sandra Neubauer, Stephen J. Redmond, & Nigel H. Lovell.2014), issue of data sparsity (Isak Karlsson, & Henrik Bostrom. 2014), the use of predictive analytics is decentralized and fragmented in the organizations and hospitals which make its use time consuming and represent a barrier in the ability of taking decisions faster (Alexey V. Krikunov, Ekaterina V. Bolgova, Evgeniy Krotov, Tesfamariam M. Abuhay, Alexey N. Yakovlev, & Sergey V. Kovalchuk. 2016; Michael Goul, Sule Balkan, & Dan Dolk2015), also making the right choice of model variables is crucial to have successful results from the application of predictive analytics(Prasada Babu, & S.Hanumanth Sastry.2014), making the models run faster and improving their scalability (Nishchol Mishra, & Dr.Sanjay Silakari.2012; Michael Goul, Sule Balkan, & Dan Dolk2015; Rui Henriques, & Cl´audia Antunes.2014), wrong application of predictive analytics by their use for tactical purposes only (Prasada Babu, & S.Hanumanth Sastry.2014), in addition some organizations consider predictive analytics expensive and the return from investing on it is not as expected and did not give the desired results (James Ogunleye. 2015; Alexey V. Krikunov, Ekaterina V. Bolgova, Evgeniy Krotov, Tesfamariam M. Abuhay, Alexey N. Yakovlev, & Sergey V. Kovalchuk. 2016), and transforming the information get from the predictive analytics systems into business value (James Ogunleye. 2014).

As can be seen from the issues and challenges mentioned that predictive analytics in many cases are unable to attain the purpose of its use due to wrong practice and issues in the process of its application which lead to the creation of problems in models quality, and low revenue from predictive analytics.

2- LITERATURE REVIEW

2-1- Introduction

Healthcare predictive systems are analytic systems which aim to minimize the future medical cost and help to provide in hospital a high level of healthcare and preventive healthcare due to the early detection of risks and possibility to take better actions and decisions. In fact, those predictions are based on the historical patients' data including detailed information about the patient, his medical history and diagnoses. For instance, in one of hospitals in Texas due to predictive analytics the hospital could save more than half million dollars by implementing a predictive system that predict any complications of heart failure patients to prevent it. Thus, the predictive analytics use models which suggest algorithms to help in the medical treatment and disease detection; or it can be used beside the electronic health records systems (Mohammad Ahmad Alkhatib, Amir Talaei-Khoei, & Amir Hossein Ghapanchi. 2015).

2-2- Tools and Techniques

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

As a matter of fact, many tools and techniques are used to support the predictive analytics in the analysis of healthcare data such as Hadoop Distributed File System (HDFS) it help in portioning the huge amounts of healthcare data to small cluster and disseminate it to other systems and its able to eliminate the redundancy in data a similar system to the HDFS is Casandra File System (CFS); MapReduce system is able to fragment tasks into subtasks and combining its outputs, it also allow to perform operational calculations effectively in a huge number of data and it save time due to parallelism technique which enable to run multiple task in the same time; JAQL which is a functional query language it can support and function with the MapReduce ; Hive it also able to address a huge number of data but with slow process and it can does all the excel functions which does not require the real time performance; Presto engine is characterized by its ability to analyze a great amount of data in a very fast and rapid time of performance; Vertica can also analyze huge amount of data rapidly but with less cost and known by its scalability; Complex Event Processing(CEP) it allow to link events in data to the real time; Mahout which help in producing application that assist Hadoop systems for Healthcare analytics (Mohammad Ahmad Alkhatib, Amir Talaei-Khoei, & Amir Hossein Ghapanchi. 2015).

In fact, the predictive analytics systems and algorithms have been used to solve many healthcare problems. For instance, in order to predict the level of gravity and risk of certain disease in patients some algorithms are used such as Logistic Regression, Genetic and decision tree algorithms based on medical data such as the patients' symptoms and laboratory measures. Moreover, to predict the early stages of certain diseases in patients some algorithms are used such as decision tree, Bayesian classifiers, back propagation neural network and logistic regression; while the detailed and clear medical patients' data help in resulting with better prediction accuracy and decision making (Raid Lafta, Ji Zhang, Xiaohui Tao, Yan Li, & Vincent S. Tseng.2015).

2-3- predictive analytics use to assist in health care improvement

The predictive analytics have been widely used by researcher to solve the readmission problems. Thus, to enhance the quality of care provided to patients with less cost there have been a strategy to emerge the hospitals readmissions as a quality indicator to prevent it. Thus, the CMS (Centers for Medicare and Medicaid Services) started a new law since October 2012 to penalize hospitals by making them paying 1% for readmissions excess especially for patient who suffer from pneumonia, heart failure, and acute myocardial infarction. Then the punishment has been increase to attain 2% in 2013 and 3% in 2014 and extended to be applied for all causes of unplanned readmission in 30 days of discharge including the surgeries. Thus, to set the benchmarks matching algorithms will be used based on administrative data (Elizabeth M Hechenbleikner, Martin A Makary, Daniel V Samarov, Jennifer L Bennett, Susan L Gearhart, Jonathan E Efron, & Elizabeth C Wick.2013).

In this context, a study proposes a big data solution for the prediction of 30- day readmission risk of the patients of congestive heart failure (CHF). To attain this, firstly helpful factors were extracted from patients' datasets by using Hadoop, Hive which is an open source data warehousing solution, and Casandra which is also an open source big data tool for the distributed data management and it makes the extraction of the data faster and it can process a large number of data distributed across various servers with no single point of failure because of replication , then predictive models have been developed to predict the risk of readmission by utilizing the

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

mentioned datasets. The algorithms used in the prediction are multiple classification algorithms especially logistic regression and the random forest algorithm due to its ability to perform with all the types of predictor variables and the analytical platform used is Mahout which has as feature that the increase of input data does not cause an augmentation of the time required during the training phase. The experimental analysis results show a high level of accuracy and a 32scalability for large datasets. This research can be extended to predict other clinical risks (Kiyana Zolfaghar, Naren Meadem, Ankur, & Brian Muckian.2013).

Another research in the same context but with different approach, it focuses on the administrative data to predict the potential readmissions in hospitals. The method used is the tree-based classification based on a hybrid approach which allow to directly emerge the patients' history of readmission and changes over time of risk factors. The data used in this research is from Michigan State patients from 2011 to 2012 Veterans Health administration data for patients who were hospitalized due to pneumonia, acute myocardial infarction, or chronic obstructive pulmonary, and heart failure disease. Thus the main goal of this research is firstly to create and prove an algorithm to predict all types of readmissions by using administrative data, and secondly to make the model able to make real-time predictions of readmissions risks of more than 80% of accuracy. The lack in this study is the use of homogeneous data which lack of diversity, and secondly the data was only from administrative data which make the predictions inaccurate. Moreover, the accuracy of the results would be better with the involvement of other algorithms to get better results (Issac Shams, Saeede Ajorlou, & Kai Yang.2014). Additionally, in one of the studies the researchers show that administrative data is not sufficient to efficiently identify and differentiate between the preventable and non-preventable readmissions (Elizabeth M Hechenbleikner, Martin A Makary, Daniel V Samarov, Jennifer L Bennett, Susan L Gearhart, Jonathan E Efron, & Elizabeth C Wick.2013).

Also, to resolve this problem a research goal was to create a predictive model to predict the congestive heart failure (CHF) readmission risk by utilizing the operational data which are dynamic and change over time and incomplete. Thus, the method followed to attain this goal pass by 3 steps starting by the data pre-processing divided into deduction and validation data; then the model development by integrating the statistical analyses to uncover the relations among the groups of data, the normalization methods, classifiers, feature selection, and individualization methods; and the analysis of the risk factor. The data used in this research is about 2787 hospitalization at one the healthcare in a university from 2003 to 2013. The results of the research showed an 86.8% accuracy. The lack in this study was the small number of data and lack of diversity where its conducted in a small health system which will decrease the validity and accuracy of work results, secondly the model need to be applied to other diseases and domains to evaluate its scalability. Moreover, the information provided by the system can be integrated with the electronic health records systems and the model process can be automated which will make the repetition of the process automatic (Samir E AbdelRahman, Mingyuan Zhang, Bruce E Bray, & Kensaku Kawamoto.2014).

Moreover, the predictive analytics have been used in health care to be able to prevent high risk from many chronic diseases which will help in saving patient lives and providing better quality of care. For instance, one of the research utilized the predictive analytics systems to predict the types of diabetes diffuse, its complications and identifying the possible treatments by utilizing the

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

Hadoop/Map Reduce environment. In fact, Hadoop is an open-source distributed data processing platform from Apache. Hadoop can have two functions as a data analytics tool and a data organizer tool with an ability to process huge amount of data due to its two components which are the Map/Reduce and the Distributed File System. Thus, the Map/Reduce is based on programming models to handle huge datasets by their division to small blocks of functions. It utilizes the distributed algorithms on a set of computers in a cluster to handle these huge data sets. While, the Hadoop Distributed File System (HDFS) to ensure accuracy will copy the data blocks which remain on other computers in the data center and it will manage the transfer of data to different part of the distributed system. Many prediction models have been developed to predict diabetes. Thus, those models were based on a collection of algorithms such as the classification, regression, neural network and genetic. Moreover, the predictive analytics systems pass by various phases including the collection of the data, data warehousing, making the predictive analysis based on defined models and algorithms and then the process of the analyzed reports. Indeed, this system after analysis shown an ability to deliver and efficient way to cure and care diabetes patients by its early detection and it will reduce costs especially for people in Indian rural areas (Dr Saravana kumar, Eswari T, Sampath P, & Lavanya S.2015). However, the use of Hadoop in this model as the only analysis tool can limit its capabilities, because Hadoop does not give the query functionality, in addition it considered slower in comparison with other tools and database management systems.

This research aim is to develop a predictive analytics system to predict the wellness and chronic conditions especially for diabetes patients to enhance the decision making in the hospital. To attain the aim, the researchers used 2 classifiers algorithms to make the predictions which are the multilayer Perceptron and Bayes Network and the models were developed by utilizing WEKA tool analyze and classify the results. Then those models accuracy have been tested by two techniques root mean squared error and Area under ROC. And the Data used for the development of the model is from National CDC-NHANES in USA, this data is based on nutritional and health situation of USA individuals. The results show that for the wellness prediction the accuracy was 55.36% for both models, while it was 89.68% for the occurrence of diabetes prediction. More work is needed in this research to improve the accuracy results and creating models for other chronic conditions (Ravi S. Behra, Pranitha Pulumati, Ankur Agarwal, & Ritesh Jain.2014).

Actually, some diseases their early detection is the way to be able to cure them, thus a research was done in the early detection of Liver disease. The research aim was to develop intelligent medical decision support systems by utilizing specific algorithms such as decision trees J48, artificial neural network, ZeroR, Naïve Bayes and VFI to classify the diseases and comparing their effectiveness and accuracy rate. Thus, the correctness in data classification will lead to improvements in the predictive and descriptive models accuracy by reducing the computing time needed to construct model. To perform the analysis WEKA tool is utilized which is a combination of machine learning algorithms for data mining functions. WEKA execute algorithms for data pre-processing, classification, clustering, feature reduction, association rules, and regression. And it contains visualization tools. Moreover, for the classifier selection, 6 classifiers utilized for prediction classification are selected and their performance is analyzed based on liver disease data sets as shown in the table 1(Tapas Ranjan Baitharu, & Subhendu Kumar Pani.2016)

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

After the analysis, the results show that Multilayer Perceptron algorithm is the most accurate and gives the best classification result among other classifier while Naïve Bayes gives the less accuracy (Tapas Ranjan Baitharu, & Subhendu Kumar Pani.2016).

Another research for another disease suggests the idea of the development of a complex model of clinical episode to support decision in treatment of Acute Coronary Syndrome (ACS) which will be based on data driven approach and the enhancement of its predictive capability by merging various models such as outcome prediction, simple classifier and states-based prediction in a complex data-driven model. This combination comes as a result to the low accuracy and unsatisfactory prediction results when running the models separately. Thus, the proposed model for the possible advance prediction it utilizes some models output as inputs, and when the data is unavailable for prediction it takes into consideration the incoming data which is about different medical procedures or events that are hard to predict at the admission time of patients into the hospital. The research focus is on the information of constant episodes which are a combination of data about different medical procedures and events (example: clinical test results, surgery, therapy, etc..), data about the doctor medical checkup (example: height, blood pressure, weight, etc…), and general information about the patient (example: gender, age, allergies, etc..). This model can be used also to predict risk and unwanted events such as clinical death, and to predict any future development of the clinical episode as shown in the figure below (Alexey V. Krikunov, Ekaterina V. Bolgova, Evgeniy Krotov, Tesfamariam M. Abuhay, Alexey N. Yakovlev, & Sergey V. Kovalchuk. 2016).

While, another study review 5 models built from composite and single techniques to support the prediction and diagnosis of heart disease. Each of those systems gives an automatic pattern identification and try to disclose the relationships between various parameters and heart disease symptoms. The table 2 show a description of the predictive analytics systems used, the techniques and their advantages and limitations (Eman AbuKhousa, & Piers Campbell.2012).

In fact, the composite models goal is to improve and raise the accuracy of the overall classification by minimizing the difference of estimation errors and try to avoid the biased decisions. Moreover, one of the main challenges to be able to create efficient predictive models is the lack of input data. Thus, the size of data is very small and characterized by many incorrect and missing values which make the construction and training of predictive models limited by those barriers (Eman AbuKhousa, & Piers Campbell.2012).

Another research for same disease in which an intelligent system recommender was developed to make predictions and assessment of the disease risk in short term for the patients of heart failure in the tele-health environment. The goal of this system is to improve the decision making and minimizing the cost and time for patients. To achieve this a recommendation algorithm was used based on time series data analysis. This algorithm helps in making the decision for the needed medical measurement such as the test of the heart rate for patient. The results of system prediction accuracy varying from 75% to 100% among various patients and it shown an ability to minimize the patients' workload from their medical test of 10%. Moreover, the system need to be improved to reducing more the workload of patients and more tests. and experiments on larger number of patients might be made to ensure of the accuracy and ability of the system (Raid Lafta, Ji Zhang, Xiaohui Tao, Yan Li, & Vincent S. Tseng.2015).

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

Furthermore, a model was developed to predict the COPD aggravation risks before it happens to prevent it by using the multi-layer neural network algorithm and the data were trained by using backpropagation algorithm. The data used in this research are simple data and it incorporate 10 attributes classified in 3 domains the mental, functional and symptoms stets. Thus, this research allows to discover what are symptoms of COPD patients that can be controlled in patients' homes, and how the COPD aggravation can be predicted, and finally discovering what are the signs that need the intervention and their prioritization. The process of this algorithm is as follow; it started by identifying the datasets which are five in this research while 4 are for test and 1 for training. Then a normalization of the output data this is because the neural network backpropagation algorithms process with numerical data thus the other types such as categorical must be transformed to numerical by the data normalization. After that, the backpropagation algorithm has been applied to train the data. The results of the research demonstrate a positive relation among the computed and the desired outputs. The limitation of this research is the use of a limited number of data which decrease the accuracy and correctness of results (Riad Alharbey.2016).

In one of the researches three predictive algorithms were discussed, those algorithms allow the prediction of patients' health and the management of the patient diseases. The algorithms are:
-Collaborative and Recommendation Engine(CARE): this methodology opts to predict and analyses diseases the collaborative filtering concept. Furthermore, it includes methods such as clustering, Inverse Frequency, Vector similarity. The feature of this algorithm that it addresses large datasets.
2. Collaborative Health Care System Model (COHESY): this methodology is based on collaborative filtering mechanism to recommend and analyze practices which can enhance the persons' health. This algorithm includes 4 phases the categorization of events, the chosen of users from diverse datasets, calculation of activities utility, and creation of recommendations.
3. Hierarchical Association Rule Mining Model (HARM): it's based on associative mining concept to predict the future patients' symptoms from the past and current symptoms (Chaitanya Kaul, Ashmin Kaul, & Saurav Verma.2015).

One of the approaches adapted was the use of multiple models to get better results such as in this research suggests a multiple predictive model to predict the physiological status of patients. The reason behind this is that with the variety of types of clinical data there is no perfect prediction algorithm to make predictions for all the disease conditions and cases. Thus, in this research 4 various Neural Network algorithms were utilized for the prediction engine in short term and a comparison of their accuracy was made. Those algorithms are the probabilistic Neural Networks which is a convenient algorithm for the small training data amounts and its data classification is non-parametric and its less sensitive to the unusual data, then the Generalized Feed Forward Multi-layer MLP with Levenberg-Marquardt (LM) it works better for huge amount of training data and its convenient for many applications, and support vector machine(SVM) which is known by its high accuracy. Moreover, this study takes into consideration some ensemble schemes and compared their performance, those schemes are the vote-based schema, accuracy based ensemble which search to discover a measured sum of algorithms that minimize error, diversity based schema, optimization based schema its characterized by its ability to take into consideration the

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

accuracy and the diversity based schemes to enhance the accuracy. This approach has been tested by employing clinical data of 1073 patients comprising 255 patients with Deep Vein Pulmonary Embolism. The results show that the mutli-model approach especially when combined with multiple schema it can enhance the prediction accuracy not only for the cases study in his research but it can be applied to other medical cases (Peter K Ghavami, & Kailash C. Kapur. 2013).

In fact, other researches focused on the management of length of stay and hospitalization days of patients in the hospital. Thus, to manage the hospital resources efficiently and be able to plan for preventive interventions for the patients' acute conditions, the researchers present a model for Predicting Hospital Length of Stay (PHLOS). Thus, to build a model with better accuracy than previous models that were constructed by utilizing statistical methods based on correlation and regression analysis. The model in this research utilize a multi-tiered data mining approach which use clustering to make the training sets to train various classification algorithms. The research start by identifying the similar sets of claims in hospital by utilizing clustering which is created by employing the K-means clustering; then those sets are used to predict the length of stay and a method was applied to class diverse classifiers for various levels of clustering those classifiers are such as SVM (support vector machine) , Bayesian Network, JRIP, J48; after that identification of length of stay classes the next method is to recognize which patients require fast and early interventions or normal interference to prevent any complication that may lead to length of stay. Thus, the result shows after the comparison of various classifiers performance that clustering a portent to compose the training set gives superior prediction results than those based on non-clustering training sets. Moreover, SVM, J48, JRIP, and Bayesian Network demonstrate a better performance than other classifiers. The weakness in the study is that the findings of the research were validated with only one expert in the emergency medicine (Ali Azari, Vandana P. Janeja, & Alex Mohseni. 2012).

In the same context, a research, the aim was to predict the number of hospitalization days by utilizing the data of health insurance claims. This, research helps to provide better quality of care, lower the costs and good allocation of hospital resources. To achieve this, a bagged regression decision tree algorithm was employed and implemented by using one of MATALAB functions called "tree Bagger"; with 300000 insurance claims data during 3 years, while the predictions were made to know the hospitalization days in the third year based on medical data of the 2 first years. Moreover, after the training dataset, the model was tested by utilizing an unseen testing dataset. The results of this research show an accuracy level of 84.3%. In fact, this research would be more accurate if the hospitalization days were categorized depending on the patient disease and medical conditions in addition to having more detailed information about patient medical history will lead to higher accuracy in the prediction especially with the incompleteness and low data quality and missing values in the insurance especially the clinical data such as the codes of diagnoses. Moreover, the employment of a combination of various models (Yang Xie, G¨unter Schreier, David C.W. Chang, Sandra Neubauer, Stephen J. Redmond, & Nigel H. Lovell.2014).

In another research in healthcare predictions the aim was to develop a predictive system to predict the need to transfer a stroke-in-patients to the intensive care unit. Thus, the ability to take this decision based on prediction will allow to reduce costs and better allocation of hospitals resource and in the other hand it will save patients' lives by making earlier interventions. Thus, to achieve the research goal various model have been used, tested and compared such as artificial

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

neural network(ANN), logistic regression (LR), support vector machine(SVM), and decision tree (DT). In previous researches, the ANN algorithm were considered as the more accurate and sensitive classification algorithm in comparison with LR and DT. This is being tested in a research to predict the survival patients of cardiac surgery. The datasets used in this research include 1415 observations with 6 variables comprising the heart and respiratory rate, oxygen saturation, blood pressure, temperature and the last variable is to transfer or not the patients to the intensive care unit. Thus, the comparison of these 4 algorithms by using the mentioned dataset show that SVM, DT, and LR have an accuracy of 0.96 while ANN has 0.94 of accuracy. In fact, this research can get better results with larger and diverse amount of data and an ensemble approach can be used such as Bagging and genetic algorithms (Nawal N. Alotaibi, & Sreela Sasi.2015).

Moreover, the Intensive Care Units (ICU) datasets include challenges such as the patient and variables heterogeneity and time synchrony to overcome this a postsurgical decision support system was developed with analytical tools comprising categorization of data and its pre-treatment, feature extraction and selection, and predictive modeling to predict the mortality rates in Intensive Care Units. The model used in this study is the multilayer neural network, and to minimize the bias in the neural network algorithm a cross validation, K-fold and random sampling have been used. The experimental results of this research were positive and show an important possibility to employ data-driven analytics to enhance the quality of services in healthcare (Yun Chen, & Hui Yang.2014).

In addition, in another context of solving healthcare challenges a research suggested an analytic framework based on data gathered from the remote health monitoring systems (RHMS) and physiological data. Moreover, when taking into consideration the variation of patient population, the use of single statistic predictive model will be inadequate to predict the adverse events. For this, the proposed technique is a multiple prediction modelling based on a collection of prediction models with high accuracy the focus of this study is to predict the readmission of patients with heart failure. Thus, after the data collection and preprocessing, the features are extracted and feature selection algorithms were applied to choose the most revealing features, then the incomplete and incorrect data were deleted. Furthermore, clustering algorithms were applied utilizing a collection of the extracted features to divide datasets into 8 clusters by employing K-means clustering algorithm. Moreover, in this method the researchers do not use a single classification model for the whole data but they have construct predictor models for the classification of data of each cluster individually. The framework was tested by using 600 records for heart failure patients the classifier used was the random forest and 10-fold cross validation technique. The results show a higher predictive accuracy and improvement in performance than the single predictive models (Mohammad Pourhomayoun, Nabil Alshurafa, Bobak Mortazavi, Hassan Ghasemzadeh, Konstantinos Sideris, Bahman Sadeghi, Michael Ong, Lorraine Evangelista, Patrick Romano, Andrew Auerbach, Asher Kimchi, & Majid Sarrafzadeh.2014).

In addition, some of the predictive models were developed based on the electronic health records data. For instance, one of the research aim was to develop a predictive model that will make the process of health data more simple and fast. This study started by an investigation in the statistical frameworks and patient features. The platform developed is a parallel predictive modeling (PARAMO) that firstly build a dependency graph of functions from specifications of predictive modeling pipelines, secondly topologically in the organization of the graph scheduling

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

the tasks and finally executing in parallel those tasks. This platform has been implemented by utilizing Map-reduce to allow the independent tasks to work in parallel in a cluster computing environment. Then an assessment of the platform performance has been done by utilizing 3 different EHR systems databases and creation of conformable predictive modeling pipeline structure comprising one patient data set, a settled number of input features such as labs, diagnoses, procedures, medications, and symptoms with a mean accumulation function for the labs and count accumulation function for others. The algorithms used were four of the classification algorithms which are the Random Forest, Naïve Bayes, K-Nearest Neighbor, and Logistic Regression. Therefore, the results have shown an important improvement of speed of research workflow and reutilization of health information compared to standard approaches while PARAMO can construct 800 various models on a 300,000-patient data set in 3 h in parallel compared to 9 days if running sequentially. Although, the weakness of this research is that it focuses only on the scalability of PARAMO platform and it doesn't take into consideration the accuracy of the predictive models (Kenney Ng, Amol Ghoting, Steven R. Steinhubl, Walter F. Stewart, Bradley Malin, & Jimeng Sun.2014).

In fact, the development of predictive models based on EHR data have improve its success during its application on several targets comprising cancer, heart failure, bipolar disorder, physiology status, life expectancy, chronic obstructive pulmonary disease, and kidney disease. The table below (table3) show examples of the application of predictive modeling on EHR data (Kenney Ng, Amol Ghoting, Steven R. Steinhubl, Walter F. Stewart, Bradley Malin, & Jimeng Sun.2014).

Examples	Time scale	Value of predictive model	Who benefits
Risk of chronic progressive disease	12-36 months	Slowing progression, preventing onset	Patient and payer
Risk of disease progression	12-60 months	Slowing progression, preventing rapid decline	Patient and payer
Optimizing choice of interventions and treatments	variable	Improve chance for more optimal outcome	Patient and others
Time to inpatient discharge	days	Improve discharge preparation, reduce readmission	Hospital, payer, and patient
Risk of 30-day readmission	Days to weeks	Reduce risk of readmission	Hospital, payer, and patient
Identifying future costly patients	12-36 months	Prevention and case management to reduce cost of care	Payer, and patient

Table 3: examples of the application of predictive modeling on EHR data

Kenney Ng, Amol Ghoting, Steven R, Walter F. Stewart, Bradley Malin, Jimeng Sun (2014)

In another research using HER data, the research addresses the issue of data sparsity when predictive models are built by utilizing electronic health records data to predict the adverse drug effects (ADE). Previously, the random forest algorithm was used to address this type of data with sparsity but the issue is that the researches prove its bias toward the majority class and that may

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

lead to low performance and less predictive accuracy. In fact, the random forest algorithm can be described as a technique to integrate multiple classification trees in a forest where each tree is constructed utilizing a collection of random subspace and bagging methods. Thus, this research proposes approaches to solve the mentioned issues; those approaches are tested by employing 14 adverse drug effects data and 3 performance metrics which are the AUC to categorize the true positive from the false positive, F1-score to examine the trade-off among recall and precision, and Brier score for the predicted probabilities accuracy. The result show that the selection of convenient approach to process the data sparsity is dependent on the task or metric performance. Thus, a sampling approach is suggested if the task is to allocate an ADE to a patient record accurately; the selection of approach does not have great value if the task is to classify patient depending on the risk of some ADE; and baseline approach is suggested when the task is to allocate probabilities for some ADE accurately (Isak Karlsson, & Henrik Bostrom. 2014).

In addition, a research aim was building predictive models to define the factors affecting the death apprehension. Thus, the modeling techniques employed to develop a fine-grained model which assist to predict the result at the end of each shift and coarse-grained models which assist to predict the result at end of hospitalization. This research experiments were based on various algorithms comprising the support vector machine (SVM), K- neural network, and Naïve- Bayes. Thus, the outcome of these tests on big data from the HER nursing system show a high accuracy of the built models which can contribute to minimize the healthcare costs and improving the quality of care (Muhammad Kamran Lodhi, Rashid Ansari, Yingwei Yao, Gail M. Keenan, Diana J. Wilkie, & Ashfaq A. Khokhar.2015).

In fact, one of the fears of using predictive analytics is the privacy issue of patient medical data. Thus, a research was done to show how predictive analysis can be done privately by the application of Homomorphic encryption which is a tool that allow to conserve the privacy by encrypting data and operating on it without having need to decrypt it or having a key as the traditional encryption methods. To demonstrate this, the researchers run a prediction service in the cloud by using Microsoft's Windows Azure to know the prospect to suffer from cardiovascular disease based on some measures and encrypted health data as input. Although, the homomorphic encryption scheme is not able to securely and correctly performing random computation, thus the scheme is able just to compute tasks with a limited complexity and in a prefixed level. Also, what makes the implementation of leveled homomorphic encryption in real-world application difficult is the necessity to choose its right parameters. Thus, to overcome these limitations the researchers proposed to select parameters automatically to determine the secure parameters for the practical homomorphic encryption scheme to ensure the security and accuracy of the results when assessing tasks utilized in predictive analysis such as Cox proportional hazard regression and logistic regression (Joppe W. Bos, Kristin Lauter, & Michael Naehrig.2014).

Moreover, one of the researches highlight the problem of having insufficient or few number of records in hospital which will make the hospital unable to create accurate local models and making good decisions. Thus, the researchers suggest an algorithm DIDT to construct a distributed model that can be utilized collaboratively which can employ only the statistics from diverse hospitals databases to overcome the privacy issues that makes hospitals worry and refuse to reveal their patients' data. Moreover, the researchers specified 9 hospitals in the Nationwide Inpatient Sample (NIS) 2009 data, and then for each of those hospitals where the number of diabetes patients is less than 100 a local model was built, after that a comparison was done between the local models

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

and the distributed model using the DIDT algorithm (Distributed Id3- based Decision Tree). The DIDT algorithm output a decision tree like the one generated on an equivalent centralized data accumulation and it has an integrated mechanism to investigate the distributed databases utilizing logical constructs based on identified attributes of interest. The results show an improvement of the accuracy of 9.9% of the distributed model than the local models (George Mathew, & Zoran Obradovic.2012).

In a different context, the researchers focus on the issue of the lack of predictive models that can predict the health conditions based on varied time sequences acquired from the health records repositories. To resolve this issue and lack they proposed a predictive model that can acquire information from an expressive temporal structure and a varied time sequence. Thus, the model can capture cross attribute and temporal dependencies. The researchers to accomplish their goal started by proposing a data mapping which will merge the diverse attributes in a single temporal structure. Then, 2 predictive models are used for the dependencies capture. Those models are P2MID depend on the classification rules to possess the integrated profiles, and M2ID model depending on the customized hidden markov models to capture the changing length and sparseness degree of health records of patients. The results of the experiments show a high accuracy of the suggested model to predict the health conditions for instance as the need for surgeries (Rui Henriques, & Cl´audia Antunes.2014).

In addition, a research focus was on creating model to predict the risk modulation of patients' expenditures. The strategy followed called the "divide and conquer" to operate large and diverse amount of data to get better accuracy results. The platform used in this work is MapReduce in order to perform machine learning algorithms such as random Forest regression on the data. This platform is a distributed computing platform that give a scalable, portable and distributed file system which handle a huge data amounts on a group of machine with high combination bandwidth across the groups. Moreover, in previous researches, the algorithm used for the risk adjustment was Linear regression, but in this research the algorithm Random Forest is employed because it can be appropriately used for the healthcare data with high dimensions and complex relationships which will result of an improvement of predictive model accuracy. Furthermore, this algorithm is robust toward noisy data and outliers, and it's appropriate for the parallelism concept. The steps of this model are 3, it starts by designing n boots sample from the main data of patients; then growing a regression tree for each boot and at each node arbitrarily sample m risk factors and select the better part between those variables; last step is to forecasting new sample by combining the prediction of the n trees. The results of the research show the efficiency to use this algorithm and the distributed computing platform. In fact, this study can be extended by merging more factors such as the insurance, disease history, and income level (Lin Li, Saeed Bagheri, Helena Goote, Asif Hasan, & Gregg Hazard.2013).

3- LITERATURE FINDINGS

How the predictive analytics have help to improve the healthcare sector (benefits):

- Minimizing the costs
- Better quality of healthcare
- Preventive healthcare
- Early detection of risks

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

- Better decision making
- Better resource allocation and management
- Saving people lives

Purpose of use:

- Types of heart diseases
- Diabetes
- Liver disease
- Acute coronary syndrome
- COPD
- General physiological status
- Mortality rates in ICU
- Stroke in transfer to ICU
- Hospitalization days
- Risk of readmission

Algorithms used for data analysis and data training:

- Multiple classification algorithms:
    - Random forest
    - Logistic regression
    - Tree based classification
    - Multilayer perceptron
    - Bayes Network
    - Decision tree: J48, DT Id3, C.4.5
    - Artificial neural network
    - Zero R
    - Naïve Bayes
- Support vector machine
- Neuro Fuzzo
- Genetic algorithms: 10-fold cross validation
- Time series algorithm
- Neural network:
    - Probabilistic NN
    - Levemborg-Marquardt
- Back propagation
- Clustering algorithms: K-means
- Feature selection algorithms
- Class; K-nearest neighbor
- Hidden Markov model

Tools used in the predictive analytics:

- HADOOP : HDFS; MAP REDUCE
- Hive
- Casandra
- Mahout

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

- WEKA
- MATLAB; Tree bagger function
- Microsoft windows azure
- IBM SPSS

Challenges of predictive analytics application in healthcare:

- In some cases, the predictive analytics are considered expensive and the return from it is not as expected.
- The lack of data input and availability.
- The issue of data integrity (incomplete, missing values)
- The decentralization or the fragmented use of predictive analytics and data sparsity
- The accuracy and speed of results in some cases is not as expected which mean that there is need to enhance the quality of models used
- Wrong choice of models variables lead to problems in the models results and accuracy

Lessons learned from previous studies:

- ✓ Using the multi model instead of single model this can significantly improve the accuracy of predictions
- ✓ Using the parallelism methodology can improve the speed to get results and save time
- ✓ Distributed data warehouses and the possibility to share statistics between hospitals while keeping the privacy in order the solve the problem of data availability
- ✓ The integration of predictive analytic systems with the other hospital systems for instance its integration with the electronic health records (HER) will help in improving the data accuracy, saving time and help in getting better work process. Furthermore, the predictive analytic systems integration with the enterprise resource planning systems can also help in reducing time and enhancing the decision making.

4- CONCLUSION

Indeed, he previous studies, have focus in research to develop models for the predictive analytics for different purposes and by using various algorithm and tools. Although, we can see a lack of a guidance framework, process and methodology to implement and apply the predictive analytics which can be based on the integration, automation and models accuracy to enhance the use of predictive analytic systems in hospitals and get the desired results from its application.

5- REFERENCES

Alexey V. Krikunov, Ekaterina V. Bolgova, Evgeniy Krotov, Tesfamariam M. Abuhay, Alexey N. Yakovlev, & Sergey V. Kovalchuk. 2016.“Complex data-driven predictive modeling in personalized clinical decision support for Acute Coronary Syndrome episodes”. Volume 80. Pages 518–529. The International Conference on Computational Science.

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

Ali Azari, Vandana P. Janeja, & Alex Mohseni. 2012.“Predicting Hospital Length of Stay (PHLOS) : A Multi-Tiered Data Mining Approach”. IEEE 12th International Conference on Data Mining Workshops.

Aly Megahed, Guang-Jie Ren, & Michael Firth.2015. “Modeling Business Insights into Predictive Analytics for the Outcome of IT Service Contracts”. IEEE International Conference on Services Computing

Bas Geerdink. 2013. “A Reference Architecture for Big Data Solutions Introducing a model to perform predictive analytics using big data technology”. The 8th International Conference for Internet Technology and Secured Transactions (ICITST).

Chaitanya Kaul, Ashmin Kaul, & Saurav Verma.2015.” Comparitive Study on Healthcare Prediction systems using Big Data”.IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems.

Chandra J, Dr Nachamai.M, & Dr Anitha S Pillai.2015. “ An Integrated Approach to Predictive Analytics for Stock Indexes and Commodity Trading Using Computational Intelligence”. Vol. 5, No. 8, 142-148. *World of Computer Science and Information Technology Journal (WCSIT).*

Dr Saravana kumar, Eswari T, Sampath P, & Lavanya S.2015. “Predictive Methodology for Diabetic Data Analysis in Big Data”. Procedia Computer Science 50.

Elizabeth M Hechenbleikner, Martin A Makary, Daniel V Samarov, Jennifer L Bennett, Susan L Gearhart, Jonathan E Efron, & Elizabeth C Wick.2013. “Hospital Readmission by Method of Data Collection”. Elsevier.

Eman AbuKhousa, & Piers Campbell.2012.” Predictive Data Mining to Support Clinical Decisions: An Overview of Heart Disease Prediction Systems”. International Conference on Innovations in Information Technology.

George Mathew, & Zoran Obradovic.2012.” Distributed Privacy Preserving Decision System for Predicting Hospitalization Risk in Hospitals with Insufficient Data”. 11th International Conference on Machine Learning and Applications.

GINA GUILLAUME-JOSEPH, & JAMES S. WASEK.2015.” Improving Software Project Outcomes Through Predictive Analytics”. VOL. 43, NO. 3. IEEE ENGINEERING MANAGEMENT REVIEW.

Hoda Moghimi, Stephen Vaughan, Steven McConche, & Nilmini Wickramasinghe.2016.” How Do Business Analytics and Business Intelligence Contribute to Improving Care Efficiency?”. 49th Hawaii International Conference on System Sciences.

Isak Karlsson, & Henrik Bostrom. 2014.” Handling Sparsity with Random Forests when Predicting Adverse Drug Events from Electronic Health Records”.IEEE International Conference on Healthcare Informatics.

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

Issac Shams, Saeede Ajorlou, & Kai Yang.2014. “A predictive analytics approach to reducing 30-day avoidable readmissions among patients with heart failure, acute myocardial infarction, pneumonia, or COPD”. Health Care Manag Sci.

James Ogunleye. 2014. ” The Concepts of Predictive Analytics”. International Journal of Knowledge, Innovation and Entrepreneurship Volume 2 No. 2, pp. 82—90.

James Ogunleye. 2015. “Challenges in Operationalising Predictive Analytics”. Research Papers on Knowledge, Innovation and Enterprise, Volume 3, PP.69-79.

Johannes Kunze von Bischhoffshausen, Markus Paatsch, Melanie Reuter, Gerhard Satzger, & Hansjoerg Fromm.2015. “.An Information System for Sales Team Assignments Utilizing Predictive and Prescriptive Analytics”.IEEE 17th Conference on Business Informatics.

Joppe W. Bos, Kristin Lauter, & Michael Naehrig.2014.” Private predictive analysis on encrypted medical data”. Journal of Biomedical Informatics.

Kenney Ng, Amol Ghoting, Steven R. Steinhubl, Walter F. Stewart, Bradley Malin, & Jimeng Sun.2014. “PARAMO: A parallel predictive modeling platform for healthcare analytic research using electronic health records”. Journal of Biomedical Informatics.

Kiyana Zolfaghar, Naren Meadem, Ankur, & Brian Muckian.2013.” Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients”. IEEE International Conference on Big Data.
Lin Li, Saeed Bagheri, Helena Goote, Asif Hasan, & Gregg Hazard.2013.” Risk Adjustment of Patient Expenditures: A Big Data Analytics Approach”. IEEE International Conference on Big Data.

Meryem Ouahilal, Mohammed El Mohajir, Mohamed chahhou, & Badr Eddine El Mohajir.2016. “A Comparative Study of Predictive Algorithms for Business Analytics and Decision Support systems: Finance as a Case Study”. IEEE

Michael Goul, Sule Balkan, & Dan Dolk2015. Predictive Analytics-Driven Campaign Management Support Systems” .Hawaii International Conference on System Sciences.

Mohammad Ahmad Alkhatib, Amir Talaei-Khoei, & Amir Hossein Ghapanchi. 2015.” Analysis of Research in Healthcare Data Analytics”. Australasian Conference on Information Systems.

Mohammad Ahmad Alkhatib. 2015. “Analysis of Research in Healthcare Data Analytics”. Australasian Conference on Information Systems.

Mohammad Pourhomayoun, Nabil Alshurafa, Bobak Mortazavi, Hassan Ghasemzadeh, Konstantinos Sideris, Bahman Sadeghi, Michael Ong, Lorraine Evangelista, Patrick Romano, Andrew Auerbach, Asher Kimchi, & Majid Sarrafzadeh.2014.” Multiple Model Analytics for Adverse Event Prediction in Remote Health Monitoring Systems”. Health Innovations and Point-of-Care Technologies Conference.

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

Muhammad Kamran Lodhi, Rashid Ansari, Yingwei Yao, Gail M. Keenan, Diana J. Wilkie, & Ashfaq A. Khokhar.2015. “Predictive Modeling for Comfortable Death Outcome Using Electronic Health Records”.IEEE International Congress on Big Data.

N. Ayyanathan, & A. Kannammal.2015. “Combined forecasting and cognitive Decision Support System for Indian green coffee supply chain predictive analytics”.IEEE.

Nawal N. Alotaibi, & Sreela Sasi.2015.” Predictive Model for Transferring Stroke In-Patients to Intensive Care Unit”. Conference on Computing and Network Communications. IEEE.

Nishchol Mishra, & Dr.Sanjay Silakari.2012. “Predictive Analytics: A Survey, Trends, Applications, Oppurtunities & Challenges”. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3).

Noura Al Nuaimi. 2014. “Data Mining Approaches for Predicting Demand for Healthcare Services in Abu Dhabi”. IEEE

Osama T. Ali, Ali Bou Nassif, & Luiz Fernando Capretz.2013. “Business Intelligence Solutions in Healthcare A Case Study: Transforming OLTP system to BI Solution”. IEEE

Peter K Ghavami, & Kailash C. Kapur. 2013.” The Application of Multi-Model Ensemble Approach as a Prognostic Method to Predict Patient Health Status”. VOL. 33, CHEMIICAL ENGIINEERIING TRANSACTIIONS.

Prasada Babu, & S.Hanumanth Sastry.2014.” Big Data and Predictive Analytics in ERP Systems for Automating Decision Making Process”. IEEE

Raid Lafta, Ji Zhang, Xiaohui Tao, Yan Li, & Vincent S. Tseng.2015.” An Intelligent Recommender System based on Short-term Risk Prediction for Heart Disease Patients”. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

Raid Lafta, Ji Zhang, Xiaohui Tao, Yan Li, & Vincent S. Tseng.2015.” An Intelligent Recommender System based on Short-term Risk Prediction for Heart Disease Patients”. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

Ravi S. Behra, Pranitha Pulumati, Ankur Agarwal, & Ritesh Jain.2014.” Predictive Modeling for Wellness and Chronic Conditions”. IEEE 14th International Conference on Bioinformatics and Bioengineering.

Riad Alharbey.2016. “Predictive Analytics Dashboard for Monitoring Patients in Advanced Stages of COPD”. 49th Hawaii International Conference on System Sciences.

Rui Henriques, & Cl´audia Antunes.2014.” Learning Predictive Models from Integrated Healthcare Data: Extending Pattern-based and Generative Models to Capture Temporal and Cross-Attribute Dependencies”. 47th Hawaii International Conference on System Science.

Samet Ayhan, Johnathan Pesce, Paul Comitz, David Sweet, Steve Bliesner, & Gary Gerberick. 2013.“PREDICTIVE ANALYTICS WITH AVIATION BIG DATA”.IEEE.

Samir E AbdelRahman, Mingyuan Zhang, Bruce E Bray, & Kensaku Kawamoto.2014.” A three-step approach for the derivation and validation of high-performing predictive models using an operational dataset: congestive heart failure readmission case study”. BMC Medical Informatics and Decision Making.

Tapas Ranjan Baitharu, & Subhendu Kumar Pani.2016.” Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset”. Procedia Computer Science 85.

Yang Xie, G¨unter Schreier, David C.W. Chang, Sandra Neubauer, Stephen J. Redmond, & Nigel H. Lovell.2014. “Predicting Number of Hospitalization Days Based on Health Insurance Claims Data using Bagged Regression Trees”. IEEE.

Yun Chen, & Hui Yang.2014. “Heterogeneous Postsurgical Data Analytics for Predictive Modeling of Mortality Risks in Intensive Care Units”. IEEE.

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

Towards Education Delivery Through E-learning in an Environment of Scarcity of Teachers: A Case of A-level Science Education in Tanzania

[1]Mr. Zakaria Ezekiel Moshi, [2]Dr. Zaipuna O. Yonah, Member IEEE

[1]Master's Scholar, School of Computational and Communication Science and Engineering, Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania. Email: moshiz@nm-aist.ac.tz
[2]Senior Lecturer, School of Computational and Communication Science and Engineering, Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania.
Email: zaipuna.yonah@nm-aist.ac.tz

***Abstract*- E-learning platforms are transforming the way education delivery is done in a significant way. Key to this transformation is the broadband-enabled internet, now being promoted under broadband-for-all global policies, which is facilitating how students learn beyond the walls of the classrooms. As students learn at their own pace with e-learning, teachers serve and act as facilitators guiding and helping out where students find difficulties in certain concepts. This is transforming the way education is delivered and allows teachers to put more attention on students with difficulties (slow learners) while the more capable students (fast learners) can move on to new concepts at their own pace. There are many e-learning platforms in use today. This paper focuses on those platforms dealing with delivering secondary level education; specifically science education at the advanced-level (A-level) secondary schools. A review of existing e-learning platforms is presented with the aim of identifying requirements for and justification of developing an improved e-learning platform that supports interactivity, collaboration and motivational engagement of learners throughout the process of education delivery. The output of this analysis review shows that most of these platforms currently in use do not support a holistic way to engage students in an interactive and collaborative manner, which is known to motivate learning and at the same time develops within learners collaboration, creativity, communication and critical thinking (4Cs) mindset necessary for the 21st century learning. It is desirable to have a platform that supports interactive education delivery and student collaboration in learning by engaging students, teachers (facilitators) and parents (guardians).**

Keywords: E-learning, Platform, Education delivery, Science studies, Secondary schools.

I. INTRODUCTION

Globally, there is shortage of teachers for science subjects. In the case of sub-Saharan countries, there is a need to increase the number of teachers by 68% from 2.4 to 4 million in less than a decade [1]. The shortage of teachers is a result of many factors including poor working conditions and minimum wages, which have contributed to about 77.2% of teachers' terminations and turnover seeking for green pastures for the case of Tanzania secondary schools [2].

Another cause of shortage of teachers is increased student enrollment mostly due to government policy of free basic education, which has been attributed by programs such as the Primary Education Development Plan (PEDP) and Secondary Education Development Plan (SEDP). These programs have contributed to an increase in students-teacher ratio especially in the science subjects [3]. This shortage of science teachers in secondary schools is degrading education delivery, which in a way undermines scientific contribution in a country like Tanzania.

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

Furthermore, education delivery at advanced secondary level lacks skills delivery mechanism that prepares the student for todays' digital competitive world. However, e-learning blending with traditional learning at advanced secondary school level is known to have the potential to improve student's digital competency skills and support teachers in education delivery [4]. It is also known that, in traditional educational delivery, learning is more teacher centered with single path progression as student's act more as receivers of information. This tends to isolate the students rather than activating them towards a collaborative way of exchanging information [4].

Today the traditional education delivery is being replaced by broadband-enabled e-learning in which education delivery no longer focuses on producing information. Instead, education delivery focuses on guiding, facilitating the learner (student) to validate information, synthesize information, leverage information, communicate information, collaborate information and problem solve with information. Encouragingly, there are many e-learning platforms in use today most of which are too generic with varying content; hence fail to meet the learning requirements for a specific scholar group, i.e. secondary schools. Other platforms that target or support secondary schools education delivery fall short on the ability to effectively engage and motivate student learning.

As a case study, this paper provides a comprehensive analysis of the education delivery effectiveness of popular e-learning platforms in use to support/supplement provision of Tanzania secondary education. The analysis is intended to establish important features/criteria that need to be followed when developing a holistic e-learning platform. The scope of the presented analysis is limited to advanced-level government secondary schools, excluding vocational studies.

The paper is organized as follows: Part II shows analysis of student's performance in science and non-science studies. Part III discusses the proposed criteria for e-learning platforms; while Part IV shows the selection of study platforms. Part V presents the findings from the analysis on the selected platforms and Part VI concludes the paper.

II. WHY A DIFFERENTIAL PERFORMANCE IN SECONDARY SCHOOLS BETWEEN SCIENCE AND NON-SCIENCE LEARNERS?

Figure 1 shows the performance, in terms of graded learning outcomes, of advanced-level secondary school students for the period of 10 years. The data shows just an average performance. Tables 1, 2 and 3 show average aggregate division points for 30 government advanced-level secondary schools; while Tables 4, 5 and 6 show corresponding results for 30 non-government advanced-level secondary schools, all having both science and non-science subjects streams, and placed in a category of schools with more than 30 students per stream in the National Examination Council of Tanzania (NECTA) grading system. The 30 schools in each category of government and non-government represent 10 best performers, 10 average performers and the 10 least performers in the Advanced Certificate of Secondary Education (ACSEE) results for the year 2014 and 2016. In the NECTA grading system, grades points are assigned as follows: A (1), B (2), C (3), D (4), E (5), S (6), and F (7). A student at A-level sits for a combination of three subjects and the division scored is assigned based on total points for three subjects as follows: Division I (3 - 9); Division II (10 -12); Division III (13 - 17); Division IV (18 -19) and Division 0 (20 -21). Lower average aggregate points means better performance in terms of learning outcomes – corresponding to higher division; and higher aggregate points means not as good performance in terms of learning outcomes – corresponding to lower division. Table 7 shows overall average aggregate points for the year 2014 in which science students passed

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

at higher aggregate points of 10.75 (Division II) whereas non-science students passed at lower aggregate points of 8.87 (Division I). Again, for the year 2016 science students passed at higher aggregate points of 13.22 (Division III) whereas non-science students passed at lower aggregate points of 11.78 (Division II). The overall poor/lower performance of science students compared to the performance of non-science students is a result of many factors among which scarcity of teachers and shortage of educational materials in secondary schools especially in the science streams is more critical. Convincingly, these results justify the research work being done.

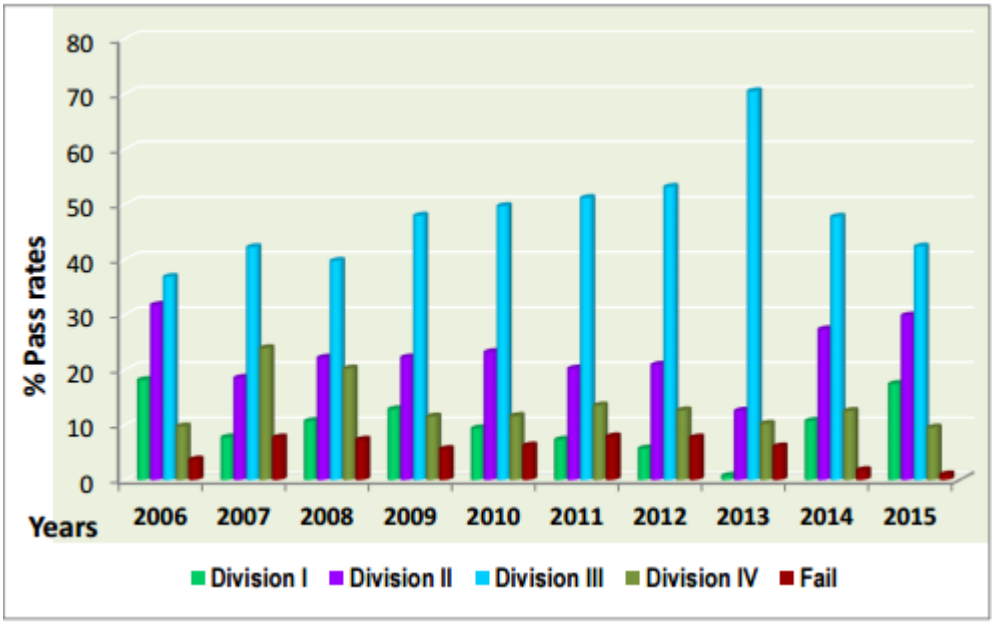


Figure 1: Secondary Education Pass Rates Summary of Form 6 Examination (ACSEE) in Government and Non-Government Schools, 2006-2015, BEST 2016 – Case of Tanzania [2].

TABLE 1: SECONDARY EDUCATION AGGREGATE DIVISION POINTS FOR TOP PERFORMING SCHOOLS 2014 -2016 NECTA – CASE OF TANZANIA [5].

2016			**2014**		
Schools	**Science**	**Non-science**	**Schools**	**Science**	**Non-science**
Tabora boys	8.31	10.46	Uwata	9.20	7.72
Kibaha	9.69	8.55	Kibaha	7.89	5.96
Mzumbe	9.24	10.15	Mzumbe	8.08	7.96
Ilboru	9.72	9.09	Msalato	8.58	7.00
Tabora girls	9.95	9.50	Kisimiri	8.36	6.00
Msalato	10.21	9.95	Ilboru	8.46	8.11
Runzewe	11.57	10.06	Tabora girls	8.81	8.20
Uwata	10.22	10.28	Tabora boys	8.44	8.75
Msolwa	11.23	10.54	Mpwapwa	9.84	7.98
Tlawi	12.55	10.21	Weruweru	11.56	7.86
Average	**10.27**	**9.88**	**Average**	**8.92**	**7.55**

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

TABLE 2: SECONDARY EDUCATION AGGREGATE DIVISION POINTS FOR AVERAGE PERFORMING SCHOOLS 2014 -2016 NECTA – CASE OF TANZANIA GOVERNMENT SCHOOLS [5].

2016			2014		
Schools	**Science**	**Non-science**	**Schools**	**Science**	**Non-science**
Loleza	13.35	11.43	Nganza	11.12	8.47
Nganza	13.79	10.83	Longido	10.83	8.53
Pamba	13.92	11.85	Mbeya	10.88	9.77
Kondoa girls	13.53	10.88	Kondoa girls	11.01	7.50
Longido	13.61	11.01	Ifunda girls	10.33	9.06
Mbeya	14.16	12.21	Lugalo	10.45	9.16
Kazima	13.94	11.99	Iringa girls	10.32	9.42
Bugene	14.88	11.37	Malagarasi	10.93	8.56
Tunduru	14.45	10.89	Korogwe girls	10.78	8.07
Ndanda	13.47	11.31	Machame girls	10.58	8.64
Average	**13.91**	**11.38**	**Average**	**10.72**	**8.72**

TABLE 3: SECONDARY EDUCATION AGGREGATE DIVISION POINTS FOR LOWER PERFORMING SCHOOLS 2014 -2016 NECTA – CASE OF TANZANIA GOVERNMENT SCHOOLS [5].

2016			2014		
Schools	**Science**	**Non-science**	**Schools**	**Science**	**Non-science**
Lufilyo	15.37	13.08	Tambaza	11.68	10.36
Same	14.02	11.53	Same	11.18	9.02
Minaki	14.11	13.21	Minaki	10.67	10.35
Pugu	14.09	13.50	Ndanda	11.18	10.47
Jangwani	15.91	12.80	Mahiwa	11.33	9.85
Songea boys	15.12	14.56	Kalangalala	10.78	9.65
Shinyanga	15.44	13.63	Shinyanga	11.10	9.90
Bagamoyo	15.52	13.42	Bagamoyo	11.31	10.17
Lumumba	15.25	15.20	Lumumba	10.72	11.08
Azania	16.06	14.10	Azania	10.72	9.62
Average	**15.09**	**13.50**	**Average**	**11.07**	**10.05**

TABLE 4: SECONDARY EDUCATION AGGREGATE DIVISION POINTS FOR TOP PERFORMING SCHOOLS 2014 -2016 NECTA – CASE OF TANZANIA NON-GOVERNMENT SCHOOLS [5].

2016			2014		
Schools	**Science**	**Non-science**	**Schools**	**Science**	**Non-science**
Feza boys	8.09	9.89	Feza boy's	7.34	7.23
Alliance girls	9.07	9.41	St.mary's mazinde	9.30	7.53
Feza girls	9.05	9.37	Marian girls	8.04	7.50
Marian boys	8.78	10.24	Feza girls	8.30	7.00
Marian girls	9.42	10.31	Kirinjiko islamic	10.33	8.14
St mary's mazinde	9.70	9.75	St.mary goreti	9.24	8.87
Pandahill	10.23	9.92	St.joseph cathedral	9.62	8.17
Kifungilo	9.76	10.60	Alfagems	10.80	8.46
Donbosco-didia	12.02	9.20	Al-muntazir	9.03	9.42
St.james	11.63	9.91	Barbro-johansson	10.71	8.10
Average	**9.78**	**9.86**	**Average**	**9.27**	**8.04**

TABLE 5: SECONDARY EDUCATION AGGREGATE DIVISION POINTS FOR AVERAGE PERFORMING SCHOOLS 2014 -2016 NECTA – CASE OF TANZANIA NON-GOVERNMENT SCHOOLS [5].

2016			2014		
Schools	**Science**	**Non-science**	**Schools**	**Science**	**Non-science**
St.maurus chemchemi	15.70	12.25	St.antony	10.34	9.34
Sanu seminary	13.45	12.18	Taqwa	11.69	9.60
Winning spirit	14.50	11.95	Airwing	10.88	9.30
Loreto girls	12.98	12.05	Thaqaafa	11.28	9.60
Simba wa yuda	12.92	10.67	St.christina girls	10.86	9.70
Thaqaafa	14.30	11.50	Alpha	10.66	9.09
Masama girls	13.73	11.65	Bendel memorial	10.44	8.80
Rosmini	13.50	11.68	Consolata seminary	10.38	8.57
Masjid qubah	12.90	10.90	Tusiime	10.89	8.98
Bigwa	15.06	11.00	Jitegemee	11.63	9.26
Average	**13.90**	**11.58**	**Average**	**10.90**	**9.22**

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

TABLE 6: SECONDARY EDUCATION AGGREGATE DIVISION POINTS FOR LOWER PERFORMING SCHOOLS 2014 -2016 NECTA – CASE OF TANZANIA NON-GOVERNMENT SCHOOLS [5].

2016			**2014**		
Schools	**Science**	**Non-science**	**Schools**	**Science**	**Non-science**
Ben bella	17.50	14.29	Ben bella	11.47	11.02
Green bird boy	16.95	14.21	Majengo	11.05	9.67
Tanzania adverntist	17.48	12.63	Lutheran junior	10.86	10.78
Al-ihsan girls	16.09	14.08	Edmund-rice-sinon	11.87	9.76
St.mathews	16.36	14.51	Lord baden powel	11.33	9.53
Etatha seminary	16.5	12.77	Mazizini	11.79	11.15
Imboru	15.51	12.37	Al-falaah muslim	11.73	9.43
Mlima mbeya	16.44	14.56	Makita	11.73	9.43
Anne marie	16.55	13.46	Iwalanje	11.58	8.42
Fidel castro	14.80	13.56	Taqwa	11.48	9.70
Average	**16.42**	**13.64**		**11.49**	**9.89**

TABLE 7: COMBINED 30 SCHOOLS AVERAGE FROM HIGHER, MEDIUM AND LOWER PERFORMANCE SCHOOLS.

	2014		2016	
	Science	Non-science	Science	Non-science
Government	10.95	8.70	13.09	11.87
Non-government	10.55	9.05	13.36	11.69
Overall average	**10.75**	**8.87**	**13.22**	**11.78**

As reported in [6], and confirmed by the foregoing performance analysis, most students prefer non-science subjects due to lack of science teachers and study materials, where schools have few teachers and sometimes these teachers have to borrow books from students to teach. The impact of such inadequacy and scarcity often translates into student's performance where non-science students show higher performance in terms of graded learning outcomes compared to science students as shown in Table 7.

III. PROPOSED CRITERIA FOR E-LEARNING PLATFORMS

In order to get a comprehensive analysis of the existing platforms, it was deemed necessary to categorize the features of e-learning platforms into two main criteria, namely: motivation to learn and support systems for learning. In this categorization, learning motivation has 3 sub criteria and support systems for learning has 5 sub criteria.

A. Motivation to Learn Criteria

For students to learn in a personalized learning environment such as e-learning platforms, the platforms need to motivate learners through interactivity, collaboration and engagement of learners to learners and learners to tutors. This main criterion is intended to determine how a platform motivates learners to love coming back to learn more [7]. Items in each feature are shown in Table 8 and the criteria are described as follows:

a) *Interactivity:* The platform should be interactive with contents that motivate learning, i.e. structured notes, animated/graphics and video tutorials [8].
b) *Collaboration:* The platform should have embodied collaboration tools that create a connected environment for the learners and tutors. Such tools include chat rooms, discussion forums, and E-mail [9].
c) *Engagements:* The platform should engage a learner into participation through assessment activities such as projects, assignments, interactive exercises and feedback.

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

B. Support Systems Criteria

In accordance with the P21's Framework for 21st Century learning, education provision does not stop only on the content delivery (key subjects) alone as a single output a student is supposed to possess. Students do need knowledge, skills, expertise and literacy to blend well in the new digital competitive environment of today. In Figure 2, The P21's framework for 21st century defines the outputs a student should possess (shown as rainbow) and the learning structures as inputs to support the student outputs (represented at bottom) [10].

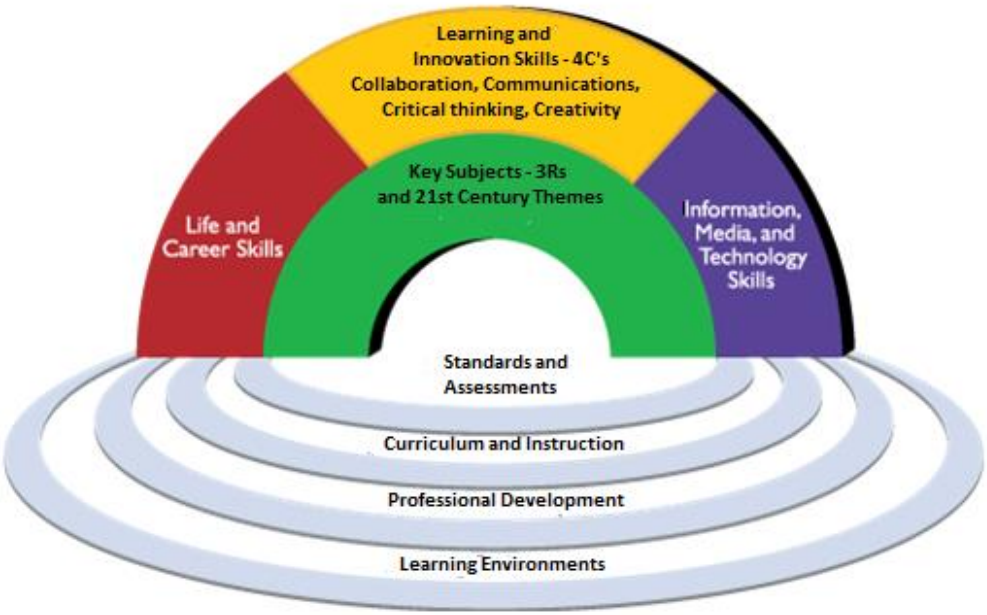


Figure 2: Partnership 21st Century learning Framework [9].

For an e-learning platform to comply with P21 Framework for 21st Century learning it should ensure that it is built on supporting systems that ensure appropriate, supportive and empowering standards and assessments, curriculum and instruction, professional development and learning environment, all are in place. And all these should be well defined to meet the desired outputs to the learner. These outputs are centered around the 4C's: Critical thinking, Communication, Collaboration and Creativity. The list of support systems is shown in Table 9 and described as follows:

a) *Standards:* The key question here is: does the e-learning platform emphasize on deep understanding of knowledge rather than shallow learning; engages students in problem solving, following national standards for a particular level and allows multiple measures of mastery?

b) *Assessments:* The learning platform should have a standard assessment testing, provision of useful feedback on student performance and also tracks on student portfolio of work, which elaborates student's mastery of concepts. The platform should be able to assess overall education system effectiveness [8, 11].

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

c) *Curriculum and instructions:* Does the platform follow a defined curriculum for the intended audience, enabling innovative learning methods that inspire creativity and critical thinking?

d) *Professional development*: The platform should support cultivation of teachers' ability to identify students' particular learning styles, intelligences, strengths and weaknesses. In addition, the platform should support continuous evaluation of students' 21st century skills development and encourage knowledge sharing among communities of professionals (practitioners) through online webinars, which inspire professional careers.

e) *Learning environments*: Create learning practices, human support and physical environments that support the teaching and learning of 21st century skill outcomes and enable students to learn in relevant, real world 21st century contexts (e.g., through project-based or other applied work).

TABLE 8: MOTIVATION TO LEARN CRITERIA AND THEIR FEATURES.

Criteria
1.**Interactivity**
1.1 The platform should have structured contents and easy to use by students with clear visibility in color and font.
1.2 The platform should have animation contents for more elaborations
1.3 The platform should have graphics/pictures contents for visual understanding
1.4 The platform should have live or on-demand videos tutorials of elaborations.
2.**Collaboration**
2.1 Platform support for instant chat that supports text, voice note, images or videos.
2.2 Platform support for discussion forums.
2.3 Real time information of other students participating in solving challenges or learning similar content.
2.4 Inter-schools challenges.
3.Engagements
3.1 Platform support for interactive exercises and self-assessment
3.2 Prompt feedback support
3.3 Platform support for projects and assignments related to the content learnt
3.4 Platform support for guardian/parent engagement with their students activities through reports of student status at all time.

TABLE 9: SUPPORT SYSTEMS CRITERIA AND THEIR FEATURES.

Criteria
4. **Standards**
4.1 Platform supports for deep knowledge understanding
4.2 Engages students in problem solving
4.3 Allows multiple measures of mastery
5. **Assessments**
5.1 Standard assessment testing
5.2 Feedback on student performance
5.3 Track student progress and work portfolio
5.4 Overall education system effectiveness assessment.
6. **Curriculum and Instruction**
6.1 Define curriculum for learners
6.2 Enable innovative learning methods that inspire creativity and critical thinking
7. **Professional development**
7.1 Ability to identify students' particular learning styles, intelligences, strengths and weaknesses
7.2 Encourage knowledge sharing among communities of practitioners through webinars that inspires professional careers
7.3 Support continuous evaluation of students' 21st century skills development
8. **Learning environment (s)**
8.1 Support learning in real world 21st Century context (e.g., through project-based or other applied work).

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

IV. METHODOLOGY

A. *Criterion for Selection of Study Platform*

After scanning through existing e-learning platforms that are currently used to support delivery of secondary education in Tanzania, the four e-learning platforms; hereby named as LMP1, LMP2, LMP3 and LMP4; were found suitable for analysis as per the research objectives of the reported study. The reasons for selecting thee platforms are briefly described in the following paragraphs:

a. *The LMP1*: The ***Shuledirect platform*** (www.shuledirect.co.tz) currently focuses on providing learning contents for O-level secondary schools. The platform has tried to identify the curriculum and syllabus for secondary schools in Tanzania from form I to IV and added materials in the form of text and graphics for student to read and learn. Currently the platforms have instructional materials for History, Civics, Geography, Physics, Chemistry, Biology, English, Mathematics, Kiswahili, Bookkeeping, Commerce and Life skills. Figure 3 shows the screen shot of the platform. Unfortunately, LMP1 is yet to include contents for advanced level education.

Figure 3: Screenshot of LMP1.

b. *The LMP2*: The ***mElimu platform*** (http://www.melimu.com) is an online e-learning environment where learners and tutors exchange information in the sense that tutors/teachers add learning contents on the platform and form classes for learners to benefit on the contents. Teachers get paid by the students/learners in exchange for the information. Figure 4 shows a screenshot of the platform. Since anyone qualified/authorized can upload content to the platform, it can be used for both O-level and A-level secondary education.

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

Figure 4: Screenshot of LMP2.

c. *The LMP3*: The ***Studi web platform*** (www.studi.co.tz) provides learning contents for secondary schools currently for form I and II that are rich in animations and video material easy for a student to understand. Currently, it has science subjects: Mathematics, Physics, Biology and Chemistry. Figure 5 shows the Screenshot of the platform. Unfortunately, LMP3 as well is yet to include contents for advanced level education.

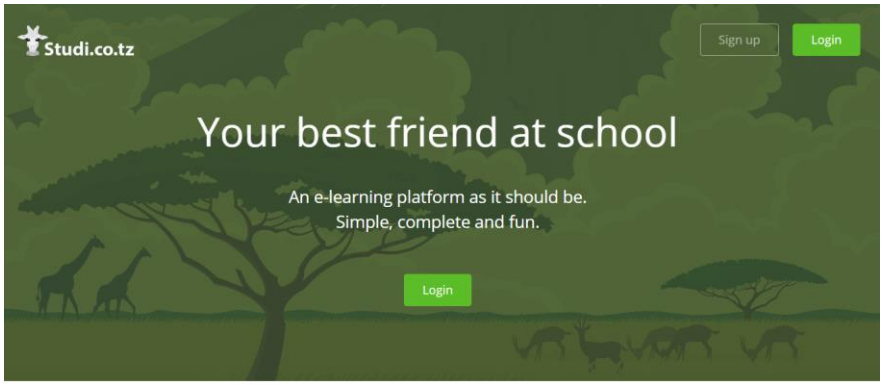


Figure 5: The Screenshot of LMP3.

d. *The LMP4*: The ***Thl 2.0 Revised edition platform*** (http://thlpc.com/) focuses on Tanzania secondary education for both O-level and A-level (from form I to form VI). The students can access learning notes and past papers. It is mostly accessible through mobile phones (smartphones) and has a software version for windows PC. The platform requires a student to pay subscription fees to access its contents. Currently the platform has notes for Biology, Physics, Chemistry, History, Commerce, Civics, Kiswahili, Bookkeeping, Geography, Basic Mathematics, Computer, English and Agriculture for O-level and Biology, History, Accountancy, Physics, Basic Applied Mathematics, Economics, General Studies, Advance Mathematics, English, Computer, Chemistry, Geography, Commerce, Kiswahili and Agriculture for A-level. Figure 6 shows a screenshot of the platform.

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

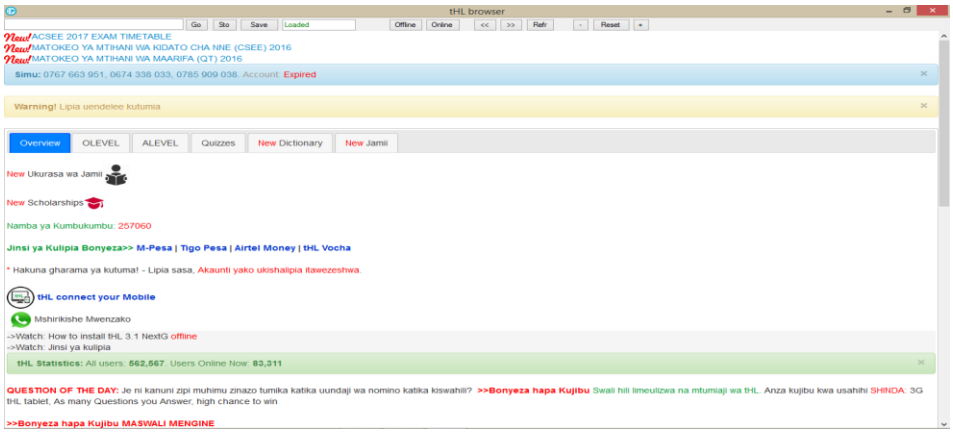


Figure 6: LMP4 screenshot.

B. *The Evaluation Process*

From the criteria given in Tables 8 and 9, the selected four e-learning platforms were analyzed and compared against each other to check whether they fit into any of the proposed criteria for an interactive motivational e-learning platform with 21st Century education delivery skills mindset. For each criterion, the associated features were checked against platforms to know how many are unsupported or not present/violated by the platforms.

V. FINDINGS AND DISCUSSION

From a total of 12 criteria in Table 8 and a total of 13 criteria in Table 9, a sum of 17, 7, 16 and 20 violations for e-learning platforms LMP1, LMP2, LMP3 and LMP4, respectively were observed. For the case of LMP1, 8 violations were related to learning motivation and 9 to support systems criteria. For the case of LMP2, 3 violations were related to motivation and 4 to support systems criteria. LMP3 had 6 violations related to motivation and 10 to support systems criteria. Finally, LMP4 had 6 violations related to motivation and 14 to support systems criteria as shown in Table 10.

TABLE 10: VIOLATIONS MATRIX FOR THE LEARNING MANAGEMENT PLATFORMS

Category	**Motivational criteria**			**Support systems criteria**					**Total Violations**
Criterion	1	2	3	4	5	6	7	8	
LMP1	3	3	2	2	2	1	3	1	17
LMP2	1	1	1	1	1	1	0	1	7
LMP3	0	4	2	3	2	1	3	1	16
LPM4	2	3	3	3	4	1	3	1	20

From the results of the initial analysis, it is seen that most existing selected platform exceeded over 50% of the total violations hence renders them incompetent in education delivery. Observable from Table 10, only the LMP2 has below 50% violations. It is a challenge to have a violation free platform but an acceptable optimum

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

level of below 20% creates an ideal learning environment for student especially advanced level secondary students who are less exposed to the digital environment. It is important that when designing a learning platform to consider motivational and support systems features as they are key to creating a friendly learning environment where students communicate and collaborate effectively and at the same time are part of the learning process.

VI. CONCLUSION

The paper has reported on an ongoing research on the identification of criteria to guide the design of a holistic learning platform for delivery of secondary education in an environment of scarcity of teachers and learning materials. The reported initial work involved the analysis for compliance of existing learning platforms to deliver into learners the 4C's of the 21st Century learning; Collaboration, Creativity, Communication and Critical Thinking.

These initial results reveal that education delivery has many challenges. Among them is the influx of students amidst few teachers especially in science subjects. Apart from few teachers, study resources have been a challenge for both students and teachers as they fall short on materials such as books for learning. As observed, there are many e-learning platforms that try to supplement education delivery but many lack features that engage the students towards today's 21st Century competitive world. Therefore, an enhanced featured e-learning platform with 21st Century learning mindset that engages the students in a more collaborative and interactive manner is required for the case of Tanzania secondary schools especially advanced level Secondary School studies where the scarcity of teachers and resources is more critical.

REFERENCES

[1] UNESCO Institute for Statistics. (2006). Teachers and educational quality: monitoring global needs for 2015 (Vol. 253). UNESCO Inst for Statistics.
[2] BEST (Basic Education Statistics), 2016 Tanzania.
[3] Wedgwood, R. (2007). Education and poverty reduction in Tanzania. International Journal of Educational Development, 27(4), 383-396.
[4] Tarkelson, E., Sinclair, J., Yook, S., & Egidio, R. (2011). An Analysis of e-Learning Impacts & Best Practices in Developing Countries With Reference to Secondary School Education in Tanzania.
[5] NECTA. (2012 - 2016). Advanced Certificate Secondary Education (ACSEE). Retrieved from www.necta.go.tz: http://www.necta.go.tz/acsee_results
[6] Ndalichako et al. (2014). Students' Subject Choice in Secondary Schools in Tanzania: A Matter of Students' Ability and Interests or Forced Circumstances? Open Journal of Social Sciences, 2(08), 49.
[7] Zaharias, P. (2008). Developing Usability Evaluation Methods for E-Learning Applications: From Functional Usability to Motivation to Learn. Journal of Human Computer Interaction, 25(1), 75–98.
[8] Reeves, T. C., Benson, L., Elliott, D., Grant, M., Holschuh, D., Kim, B., … Loh, S. (2002). Usability and instructional design heuristics for e-Learning evaluation. In World Conference on Educational Multimedia, Hypermedia and Telecommunications 2002 (pp. 1615–1621). Denver, Colorado: Chesapeake, VA: AACE. Retrieved from http://editlib.org/noaccess/10234
[9] Albion, P. R. (1999). Heuristic evaluation of educational multimedia: From theory to practice. In 16th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education, ASCILITE (pp. 9–15). Brisbane, Austrlia. Retrieved from http://www.ascilite.org.au/conferences/brisbane99/papers/albion.pdf
[10] Learning, T. P. (2015, 5). Framework For 21st Centurey Learning. Retrieved from http://www.p21.org: http://www.p21.org/our-work/p21-framework
[11] Alsumait, A. a., & Al-Osaimi, A. (2010). Usability heuristics evaluation for child e-learning applications. Journal of Software, 5(6), 654–661. doi:10.4304/jsw.5.6.654-661

Comparison Between PSO and HPSO In Image Steganography

Ziyad Tariq Mustafa Al-Ta'i
Department of Computer Science
University of Diyala - College of Science
Baghdad , Iraq
Ziyad1964tariq@gmail.com

Enaam Rabah Mohammad
Department of Computer Science
University of Diyala - College of Science
Diyala , Iraq
Anaamrabah83@gmail.com

***Abstract*— Efficient steganography techniques are needed for the security of digital information over the Internet and for secret data communication. Therefore, many techniques are proposed for steganography. One of these intelligent techniques is Particle Swarm Optimization (PSO) algorithm. Recently, many modifications are made to Standard PSO (SPSO) such as Human-Based Particle Swarm Optimization (HPSO). Therefore, this paper presents image steganography using HPSO in order to find best locations in image cover to hide text secret message. Then, a comparison is done between image steganography using PSO and using HPSO. Experimental results on six (256×256) cover images and different size of secret massages, prove that the performance of the proposed image steganography using HPSO has been improved in comparison with using SPSO.**

Keywords- Steganography , PSO, HPSO, LSB.

I. INTRODUCTION

In the growing linked modern world, one may wish to be able to protect not only secrecy of the communication but also privacy of the communicators [1]. Therefore, information hiding gets its way in modern communication. Information hiding is the process of hiding amount of data called secret massage into a cover media that may be audio , video or image in an imperceptible way to build a covert channel[2]. Information hiding system should satisfy two basic requirement the first requirement is usually referred to as transparency , the stego- medium should be similar to cover medium according to a suitable distortion measure. The second requirement is referred to as robustness, hiding message should survive the application of any data processing technique with a certain class to stego - medium [3].

Since the images are much better means of communications between human users, it is convenient to develop an image hiding systems, specifically an image steganographic systems[4].A number of steganography techniques [5] are available for embedding information in an image. These can be broadly classified as spatial domain techniques and transform domain techniques. In the spatial domain [6], the simplest technique is to embed the data in the Least Significant Bits (LSBs) of each pixe l in the cover image.

Swarm Intelligence is part of artificial intelligence. The idea of SI comes from systems found in nature, including ant colonies, bird flocking and animal herding that can be effectively applied to computationally intelligent system. Particle swarm optimization (PSO) algorithm can be used as powerful swarm intelligence search technique [7].

Recently HPSO is presented as a modified version of SPSO based on human behavior [8]. Therefore, this paper presents an image hiding scheme using HPSO in addition to a comparison with SPSO algorithm .

II. PSO AND HPSO

A.Particle Swarm Optimization(SPSO)

Particle swarm optimization, is a heuristic optimization technique based on the behavior of a colony or swarm of insects, such as ants, termites, bees, and wasps; a flock of birds; or a school of fish. The particle swarm optimization algorithm mimics the behavior of these social organisms[9]. The word particle denotes, for example, a bee in a colony or a bird in a flock. Each individual or particle in a swarm behaves in a distributed way using its own intelligence and the collective or group intelligence of the swarm. The Standard PSO (SPSO) model consists of a swarm of particles, which are initialized with a population of random candidate solutions. They move iteratively through the d -dimension problem space to search the new solutions, where the fitness ,f , can be calculated as the certain qualities measure. Each particle has a position represented by a position-vector xi, (i is the index of the particle) and a velocity represented by a velocity-vector vi [9]. Each particle has a memory to store its historically best solution (i.e., its best position ever attained in the search space so far, which is also called its experience). The standard version of the PSO algorithm is essentially described by the following two simple velocity and position update equations, shown in 1 and 2 respectively[10].

$$Vid(t+1)= wVid(t)+ c1R1(Pid(t) – Xid(t))+c2R2(Gbid(t) – Xid(t)) \quad (1)$$

$$X id (t+1) = Xid (t) + Vid (t+1) \quad (2)$$

Where:

1) Inertia weight (w): is linearly decreased from 0.9 to 0.4 to balance the global and local search abilities of particles in the search space, which is given by equation (3).

$$w= wmax –(wmax – wmin) /T \times t \quad (3)$$

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

Note that wmax is the initial weight , wmin is the final weight , t is the current iteration number and T is the maximum iteration number.

2) Vid represents the rate of the position change (velocity) of the ith particle in the dth dimension, and t denotes the iteration counter.

3) Xid represents the position of the ith particle in the dth dimension.

4) Pid represents the historically best position of the ith particle in the dth dimension (or the position giving the best ever fitness attained by the Xi).

5) Gbid represents the position of the swarm ‘s global best particle in the dth dimension (or, the position giving the global best fitness value attained by any particle among the entire swarm).

6) R1 and R2 are two n-dimensional vectors with random numbers uniformly selected in the range of [0.0, 1.0], which introduce useful randomness for the search strategy.

7) c1 and c2 are positive constant weighting parameters, also called the cognitive and social parameters, respectively, which control the relative importance of particle‘s private experience versus swarm‘s social experience (or, in other words, it controls the movement of each particle towards its individual versus global best position[6].

According to the aforementioned equations (1) and (2), the procedure for implementing PSO algorithm is given by the pseudo code No.(1).

Pseudo Code No.(1): PSO Algorithm; (adopted from[10])

Randomly generate an initial population with positions and velocities
Repeat
For i = 1 to population size do
if (f(Xi) < f(Pi) then Pi= Xi
G = argmax (f(Pbest (i)))
For j = 1 to D do
Velocity update with Eq.(1);
Position update with Eq.(2);
End//end for loop j;
End//end for loop I;
Until termination criterion is met.

B. Human – Based Particle Swarm Optimization

Hao Liu et al. proposed [8] a modified version of SPSO based on human behavior, which is called HPSO. This modification is proposed to improve the performance of SPSO. In SPSO, all particles only learn from the best particles Pbest and Gbest. Obviously, it is an ideal social condition. However, considering the human behavior there exist some people who have bad habits or behaviors around us, at the same time, will bring some effects on people around them. If we take warning from these bad habits or behaviors it is beneficial to us. Conversely, if we learn from these bad habits or behaviors, it is harmful to us. Therefore, we must give an objective and rational view on these bad habits or behavior. To simulate the human behavior, the global worst particle was introduced into the velocity equation of SPSO, and the learning coefficient r3 which obeys the standard normal distribution that is ($r3 \in N(0,1)$). Learning coefficient can balance the exploration and exploitation abilities by changing the flying direction of particles. When learning coefficient is positive, it is called impelled leaning coefficient, which is helpful to enhance the "flying " velocity of the particle , therefore it can enhance the exploration ability. When the learning coefficient is negative, it is called penalized learning coefficient, which can decrease the “flying” velocity of the particle. Therefore, learning coefficient is beneficial for improving the exploitation ability. If (r3=0)this represents that these bad habits or behaviors have no effect on the particle. At the same time, the acceleration coefficients c1 and c2 have been replaced with two random numbers, whose sum is equal to 1 in [0,1]; this strategy decreases the dependence on parameters of the solved problem[8] . Therefore the velocity equation has been changed as equation (4).

$$Vid(t+1)=wVid(t) + R1 (Pid(t) – Xid(t))+ R2 (Gbid(t) – Xid(t)) + R3 (Gwid(t) – Xid(t)) \quad (4)$$

$$Xid\ (t+1)= Xid\ (t)+ Vid\ (t+1) \quad (5)$$

Inertia weight w is linearly decreased from 0.9 to 0.4 to balance the global and local search abilities of particles in the search space, which is given by equation (3).

In HPSO, the global worst particle is introduced , which is the worst fitness in the entire population at each iteration. It is denoted as Gworst and defined as equation (6):

$$Gworst(t) = argmin\{f(Pbest1),f(Pbest2),\ldots,f(PbestN)\} \quad (6)$$

where f(.) represents the fitness value of the corresponding particle.

According to the aforementioned equations (4) and (5), the procedure for implementing HPSO algorithm is given by pseudo code No. (2).

Pseudo Code No.(2): HPSO Algorithm; (adopted from [8])
Randomly generate an initial population with positions and velocities initialize Pbest , Gbest and Gworst. Pbest = X; Gbest = argmax{f(Pbest1),f(Pbest2),….,f(PbestN)} Gworst = argmin{f(Pbest1), f(Pbest2), …., f(PbestN)} For t = 1 to T do For each Particle I = 1, 2, …., N Update velocity according to equation (4); Update position according to equation (5); End for. Update Pbest , Gbest , Gworst; End for. Return the best solution

Identify applicable sponsor/s here. *(sponsors)*

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

III. THE PROPOSED METHOD

A.Embedding Side

The proposed framework for embedding side is shown in Fig. (1).

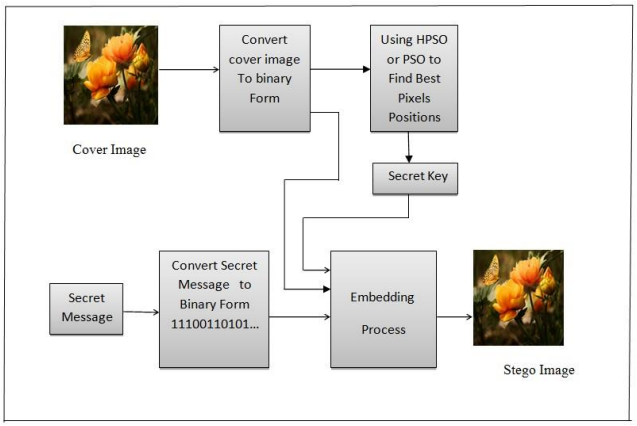


Figure.(1)The Proposed Method in The Embedding Side

The embedding technique finds optimum locations of pixels in the cover image, using PSO or HPSO. These output locations of the PSO or HPSO module are then used to embed the secret massage data in order to obtain the stego image. The steps of embedding secret message in a cover image are shown in algorithm (1).

Algorithm (1): The Proposed Embedding Side Algorithm.
Input: Cover image size (M*N); Secret message size (L).
Output: Stego image.
Step1: Convert Cover image and secret message to binary form.
Step2: Using PSO (Pseudo code(1)) or HPSO(Pseudo code(2)) as a search technique in order to obtain the best pixels' locations in the cover image. Where: No. of best locations= 8 * No. of secret characters.
Step3:Secret message bits are embedded in best locations (obtained by step 2) of the cover image using LSB technique.
Step4: Convert binary sequence into pixels.
Step5: Get the output of (step 4) as a stego image.

B.Extracting Secret Massage

The proposed framework for extracting side is shown in Fig. (2).

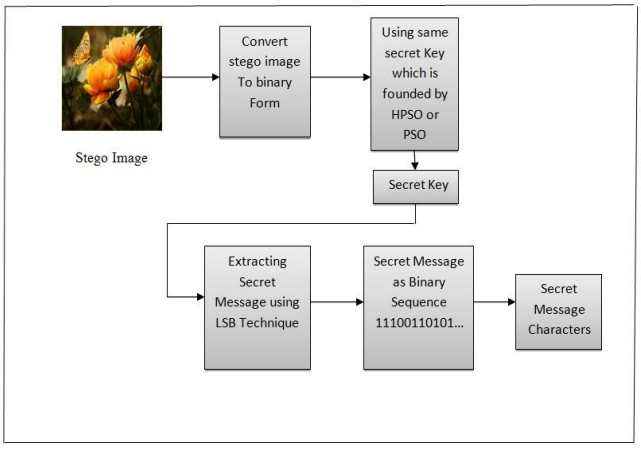


Figure. (2) The Proposed Method in Extracting Side

The steps of extracting the secret message from stego image are shown in algorithm (2).

Algorithm (2): The Proposed Extracting Side Algorithm.
Input: Stego image size (M*N).
Output: Secret message size (L).
Step1: Convert Stego image to binary form.
Step2: Using the secret keys that are founded by PSO (Pseudo code(1) or HPSO (Pseudo code(2)) in order to obtain the locations of hidden bits.
Step3:Secret message bits are extracted from locations (obtained by step 2) of the stego image using LSB technique.
Step4: Convert binary sequence into characters.
Step5: Get the output of (step 4) as a secret message.

IV. EXPERIMENTAL RESULTS

The test samples that are used as (256*256) cover images are shown in table (1).

TABLE(1) Cover Images

	Name Image	Image
1	Deer Image.bmp	

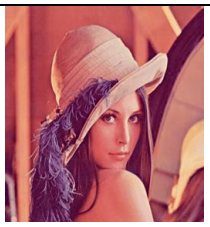
https://sites.google.com/site/ijcsis/
ISSN 1947-5500

2	Lena Image.bmp	
3	Flower Image.bmp	
4	Butterfly Image.bmp	
5	Nature View Image.bmp	
6	Baby Image.bmp	

Different objective quantitative measures are used for comparison between the cover images and stego images. These measures are: Peak Signal Noise Ratio (PSNR),Mean Squared Error(MSE),Number Of Pixel Change Rate Test (NPCR) and Unified Average Changing Intensity Test(UACI).The results of these measures on stego images for the proposed system (PSO and HPSO) with secret message of lengths: (240, 320, and 400), are shown in table (2).

The stego images for the proposed system with PSO and HPSO are shown in table (3).

TABLE (3): Stego Images for the Proposed System with PSO and HPSO

Cover Image Name		Stego Image with PSO	Stego Image with HPSO
1	Deer		
2	Lena		
3	Flowers		
4	Butterfly		
5	Nature View		
6	Baby		

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

TABLE(2) Compression Results

Cover image	Evolution Criteria	240- Secret Bit		320- Secret Bit		400- Secret Bit	
		PSO	***HPSO***	***PSO***	***HPSO***	***PSO***	***HPSO***
1.Deer	PSNR	75.12288995835	75.32650278483	73.81587025129	74.068438203440	73.052778361939	73.285302958276
	MSE	0.001998901367	0.001907348632	0.002700805664	0.0025482177734	0.0003219604492	0.0003051755781
	NPCR	0.19989013671875	0.1861572265625	0.27008056640625	0.24261474609375	0.32204345703125	0.299072265625
	UACI	0.000780820846	0.000745080596	0.001055002212	0.0009953975677	0.0012576580047	0.0011920928955
2.Lena	PSNR	75.43200460816	75.76481848008	73.89011043209	74.336606390823	73.073409967576	73.462590627880
	MSE	0.001861572265	0.001724243164	0.002655029296	0.0023956298828	0.0032043457031	0.0029296875
	NPCR	0.1861572265625	0.17852783203125	0.25787353515625	0.23040771484375	0.3173828125	0.30059814453125
	UACI	0.000722717666	0.000673532485	0.001037120819	0.0009357929229	0.0012516975402	0.0011444091791
3.Flowers							
	PSNR	75.54013330099	75.96136536004	74.36435693137	74.62242956734	73.114969565288	73.307072150819
	MSE	0.001815795894	0.001647942187	0.002380371093	0.0022430419921	0.003173828125	0.0030364990234
	NPCR	0.18768310546875	0.1800537109375	0.2532958984375	0.2471923828125	0.31280517578125	0.299072265625
	UACI	0.000709295272	0.000643730163	0.000929832458	0.0008761882781	0.0012397766113	0.0011861324310
4.Butterfly							
	PSNR	75.29189746374	76.04254426226	74.09452203451	74.392285933213	73.135899460346	73.285302958276
	MSE	0.001922607421	0.001617431640	0.002532958984	0.0023651123046	0.0031585693359	0.0030517578125
	NPCR	0.2044677734375	0.17001953125	0.26397705078125	0.2410888671875	0.32501220703125	0.29754638671875
	UACI	0.000751018524	0.000631809234	0.000989437103	0.0009238719940	0.0012338161468	0.0011920928955
5.Nature View							
	PSNR	75.3265027848	75.5767828418	74.09452203451	74.30903204537	73.073409967576	73.440029824838
	MSE	0.00190734863	0.00180053710	0.002532958984	0.0024108886718	0.0032043457031	0.0029449462890
	NPCR	0.19073486328125	0.177001953125	0.25482177734375	0.2466650340625	0.3173828125	0.296020507815
	UACI	0.00074505805	0.00070333480	0.000989437103	0.0009417533874	0.0012516975402	0.0011503696441
6.Baby							
	PSNR	75.80342268821	75.92133793550	74.281631671711	74.505833441984	73.30702150819	73.623885630886
	MSE	0.001708984375	0.001663208007	0.0024261475093	0.00230407714843	0.00303649902343	0.00282287597656
	NPCR	0.1922607421875	0.16479991875	0.24871826171875	0.2461923828125	0.3021240234375	0.286865234375
	UACI	0.000667572021	0.000649606280	0.0009477138519	0.00090003013610	0.00118613243103	0.00110268592834

The stego images for the proposed system (PSO and HPSO) with hiding locations are shown in table (4).

The comparison between the required time for the proposed system with PSO and HPSO and No. of iteration that each algorithm needed to obtain best location without repetition is shown in table (5).

CONCLUSION

The results of this work show that both PSO and HPSO algorithms are good search techniques. However, from steganography point of view, HPSO algorithm is better than PSO algorithm for these reasons:

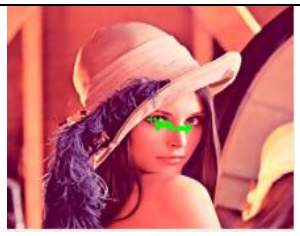
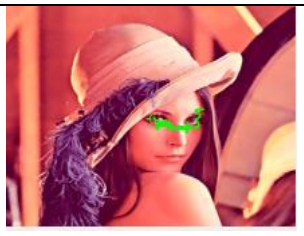
1) HPSO algorithm selects the best hiding positions as shown in tables (4).

2) Table (2) shows that the HPSO stegocovers have better quality metrics than PSO stegocovers.

3) Table (5) shows that the time required by HPSO is better than PSO in embedding and extracting.

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

TABLE (4): Stego Images for the Proposed System (PSO and HPSO) with hiding Locations

Cover image	Algorithm	240 secret- bit	320 secret- bit	400 secret- bit
Deer	PSO			
	HPSO			
Lena	PSO			
	HPSO			
Flowers	PSO			
	HPSO			

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

Butterfly	PSO			
	HPSO			
Natural View	PSO			
Baby	PSO			
	HPSO			

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

TABLE(5) Time Compression

Cover Image	Algorithm	NO. of Iteration	240 Secret -Bit	No. Of Iteration	320 Secret- Bit	NO. Of Iteration	400 Secret- Bit
Deer	PSO	677	00:00:05.626801	884	00:00:07.334865	1065	00:00:08.750805
	HPSO	563	00:00:04.952431	724	00:00:06.244141	959	00:00:08.281493
Lena	PSO	542	00:00:04.544128	692	00:00:07.322857	866	00:00:08.784935
	HPSO	704	00:00:06.364627	860	00:00:09.072650	1033	00:00:10.537757
Flowers	PSO	560	00:00:040955160	717	00:00:07.1593016	968	00:00:09.6984334
	HPSO	614	00:00:05.2074545	768	00:00:07.9859102	952	00:00:09.6214929
Butterfly	PSO	584	00:00:06.1000475	848	00:00:08.7724821	993	00:00:09.3298196
	HPSO	622	00:00:06.5964593	818	00:00:08.6668760	1043	00:00:08.5717672
Natural Veiw	PSO	389	00:00:03.888580	546	00:00:05.6628458	668	00:00:06.6605633
	HPSO	430	00:00:04.749482	568	00:00:06.2497119	715	00:00:07.6379359
Baby	PSO	501	00:00:04.273835	775	00:00:06.2607371	986	00:00:08.0710223
	HPSO	521	00:00:04.5865807	686	00:00:05.6707950	894	00:00:07.7527664

REFERENCES

[1] Ziyad Tariq Mustafa Al-Ta'i, " Development of Multilayer New Covert Audio Cryptographic Model ", International Journal of Machine Learning and Computing, Vol. 1, No. 2, June 2011.

[2] Musrrat Ali, Chang Wook Ahn, and Millie Pant, " Data Hiding Schemes: A survey", Embodying Intelligence in Multimedia Data Hiding, The authors; licensee Science Gate Publishing P.C. - CC BY-NC 4.0 International License DOI: 10.15579/gcsr., GCSR Vol. 5 ch.1, pp. 1-19, 2016.

[3] Pierre Moulin, and Joseph A. O'Sullivan, " Information-Theoretic Analysis of Information Hiding", IEEE Transactions on Information Theory, Vol.49, No. 3, March 2003.

[4] Ziyad Tariq Mustafa Al-Ta'i, " Simulation of New Covert Audio Cryptographic Model", 3rd International Conference on Machine Learning and Computing (ICMLC 2011), Singapore, 26-28 February 2011.

[5] Abbas Cheddad , Joan Condell, Kevin Curran, and Paul Mc Kevitt, " Digital image steganography: Survey and analysis of current methods", Elsevier Journals- Signal Processing, Contents lists available at ScienceDirect, journal homepage: www.elsevier.com/locate/sigpro, Signal Processing 90 (727–752), 2010.

[6] Chang CC, Lin MH, and Hu YCm, "A fast and secure image hiding scheme based on LSB substitution" , International Journal of Pattern Recognition and Artificial Intelligence, Volume 16, Issue 04, June 2002.

[7] Ziyad Tariq Mustafa Al-Ta'i and Omer Younis Abd Al-Hameed , "Comparison between PSO and Firefly Algorithms in Fingerprint Authentication", International Journal of Engineering and Innovative Technology (IJEIT) , Volume 3, Issue 1, July 2013.

[8] Hao Liu, Gang Xu, Gui-yan Ding, and Yu-bo Sun, " Human Behavior-Based Particle Swarm Optimization" , Hindawi Publishing Corporation , The Scientific World Journal, Volume 2014, Article ID 194706, 14 pages, 17 April 2014.

[9] Ajith Abraham , He Guo, and Hongbo Liu, "Swarm Intelligence : Foundations, Perspectives and applications", Studies in computational Intelligence (SCI) , Springer – verlag Berlin Heidelberg 26 ,3-25 (2006).

[10] Ahmed H. , Glasgow J. , " Swarm Intelligence Concepts , Models And Applications" Technical Report 2012-585, School of Computing Queen's University Kingston, Ontario, Canada K7L3N6, February 2012.

AUTHORS PROFILE

Ziyad Tariq Mustafa Al-Ta'i was born in Baghdad (1964). Received BS.C degree in Electrical Engeineering from university of Baghdad (1987) and MS.C degree in computer science from university of Technology- Baghdad (1995) , and PH.D in computer science from university of Technology- Baghdad(2002).

Enaam Rabah Mohammed was born in Diyala (1983). Received BS.C in computer science from university of Diyala(2006) and research student for a master's degree in computer science from the University of Diyala(2017).

Persuasive Cued Click Point Password *with OTP*

Anita Chaudhari,Payal Shahapurkar ,Asmit Patil
Department of Information Technology
St. John College of Engineering and Management, Palghar, India
anitac@sjcet.co.in

***Abstract*—Authentication plays a major role in Digital environment. In this environment we have different methods which generally use alphanumeric characters and special characters for password creation. These methods have some problems like hard to remember password because it has no meaning and easily breakable by third parties or attackers. To address these issues, many techniques for authentication are proposed from which graphical password method is best in terms of cost and usage. Basically, Graphical passwords use images for password creation and it has some demerits like hotspot and shoulder surfing problem. A persuasive cued click-point based method reduces hotspot problem. To prevent persuasive cued click-point based method from shoulder surfing we include one time password. For more user convenience we provide two login methods one which requires internet and other which does not require internet.**

Keywords*—*Persuasive Cued Click Points,International Mobile Equipment Identity ,One Time Password

I. INTRODUCTION

Various graphical password schemes have been proposed as alternatives to text-based passwords. Research and experience have shown that text-based passwords are fraught with both usability and security problems that make them less than desirable solutions. Psychology studies have revealed that the human brain is better at recognizing and recalling images than text. Graphical passwords are intended to capitalize on this human characteristic in hopes that by reducing the memory burden on users, coupled with a larger full password space offered by images, more secure passwords can be produced and users will not resort to unsafe practices in order to cope.

In this project, we propose a new click-based graphical password scheme called Persuasive Cued Click Points (PCCP) with OTP. A password consists of one click-point per image for a sequence of images. The next image displayed is based on the previous click-point so users receive immediate implicit feedback as to whether they are on the correct path when logging in. PCCP offers both improved usability and security and OTP prevents it from shoulder surfing.

II. PREVIOUS WORK

A new click-based graphical password scheme called Cued Click Points (CCP). It can be viewed as a combination of PassPoints Passfaces, and Story. A password consists of one clickpoint per image for a sequence of images. The next image displayed is based on the previous clickpoint so users receive immediate implicit feedback as to whether they are on the correct path when logging in. CCP offers both improved usability and security[2].It is very difficult for the user to remember the exact pixel point.PassPoints is a new and more secure graphical password system. This work proposed a password scheme in which the user is presented with a predetermined image on a visual display and required to select one or more predetermined positions on the displayed image in a particular order to indicate his or her authorization to access the resource. Beyond this, This system was developed early in the evaluation of graphical passwords, and in this, the user is given with an image. The click points on the image are used as the password for user authentication. The user has to remember the order and position of the click points. The click points are not stored as such, but as a hashed value. For correct validation, discretization square is used which is the tolerance area around the original click point. The user should click on the discretization area. Here, the system does not have any influence over the selection of the click points. The user is free to set the password which the user can easily remember. Since it is being very simple, it can easily be attacked. In PassPoints, passwords consist of a sequence of click-points on a given image. Users may select any pixels in the image as click-points for their password.Drawback is As it is very difficult to remember the random points, user chooses to select points on images that can be easily recognized in the image[1].

In this system an alternative gaze-based authentication scheme that supports users in selecting secure gaze-based graphical passwords. To tackle the problem of hotspots, our scheme uses a computational model of visual attention – also known as saliency maps – to mask out those areas of the image most likely to attract visual attention. We show that this approach significantly increases the security of gaze based cued-recall graphical passwords. The specific contributions of our research are 1) a shoulder surfing resistant gaze-based authentication scheme that allows the user to select a sequence of arbitrary points in an image, 2)the introduction of computational models of visual attention to increase the security of gaze-based cued-recall graphical passwords, and 3. a security evaluation of three different gaze based graphical passwords – PIN, picture without a saliency mask, and picture with a saliency mask – in a user study with 12 participants guessing passwords after watching close-up videos of the eye movements of other users.Users often create memorable passwords that are easy for attackers to guess, but strong system-assigned passwords are difficult for users to remember[3]. A password authentication system should encourage strong passwords while maintaining reputation. We propose that authentication schemes allow user choice while influencing users towards stronger passwords. In our system, the task of selecting weak passwords (which are easy for attackers to predict) is more tedious, discouraging users from

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

making such choices. In effect, this approach makes choosing a more secure password the path-of-least-blocking.

The password system provides security against unauthorized access but the evolution of different attacks made this system ineffective. To make complex text passwords was a solution but it was very difficult for users to remember so an alternative for this came graphic based passwords, which again had its own disadvantages like in passpoints it is very difficult to remember the random points, user chooses to select points on images that can be easily recognized in the image. Then in cued click points it is very difficult for the user to remember the exact pixel point. To overcome the disadvantages of the above system came Persuasive Cued Click Points method. This is more efficient and user friendly but the only attack possible on this method was shoulder surfing, so to avoid this disadvantage we are coming up with a new two way authentication system which would include Persuasive Cued Click Point technique along with an OTP.

III. IMPLEMENTATION

The existing system consists of text passwords and OTP which is sent to user via SMS or email. The problems of knowledge-based authentication, typically text-based passwords, are well known. Users often create memorable passwords that are easy for attackers to guess, but strong system-assigned passwords are difficult for users to remember. In the passpoints technique it is very difficult to remember the random points so user chooses to select points on images that can be easily recognized in the image. Whereas in cued click points it is very difficult for the user to remember the exact pixel point. For the existing OTP method email spoofing or man in the middle attack can occur. Hence user's security can be compromised. Also the text passwords are either predictable or if made complex using a combination of alphabets, symbols and numbers becomes difficult for the users to remember. Hence this approach is not user convenient and prone to attacks[5].

Our project uses the image based password as the 1st key of authentication and OTP as the 2nd key of authentication. Also it eliminates email spoofing, man in the middle attack and includes another level of security.
Objective of Our project is To make our password system more user convenient and easy to remember,To make it more secure,To allow users to login with the help of internet as well as without internet,To allow users to login who are not having a Android mobile phone,To increase the security level without making the user to take an effort to remember the complex text passwords,To educate people and promote the Digital India Movement.

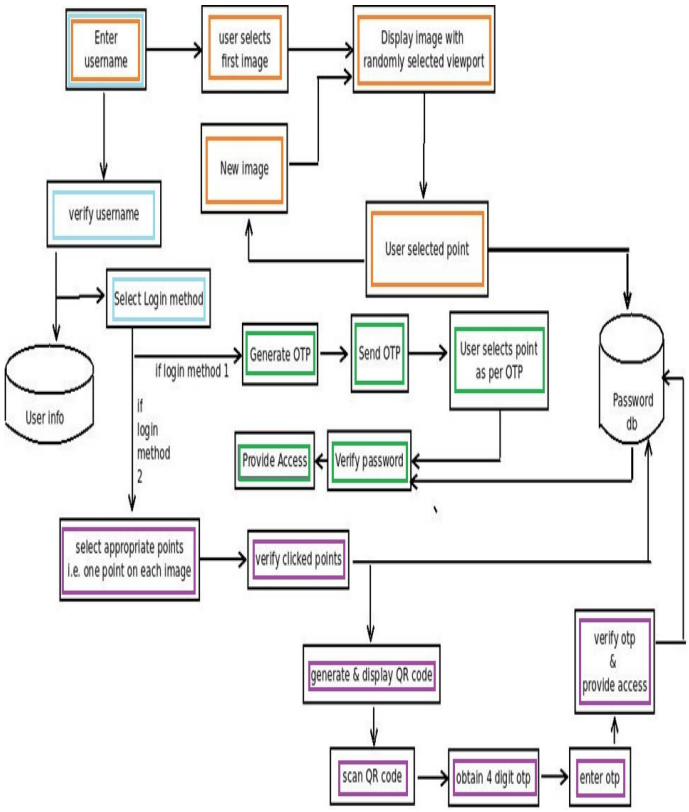


Fig 1: Architecture of System

IV.RESULTS

Following form gives the login option ,we can select one option from it,

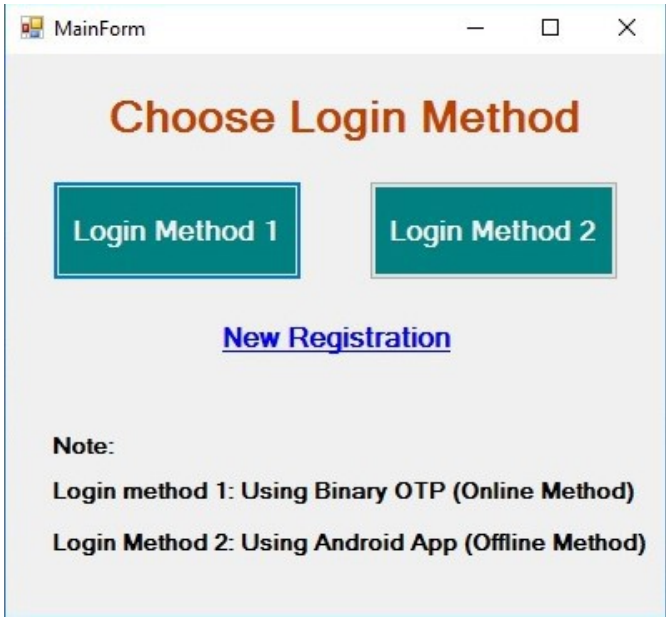


Fig 2.Login Method

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

The User will select the login method by which he wants to login or if he is a new user then he can select the new registration option.

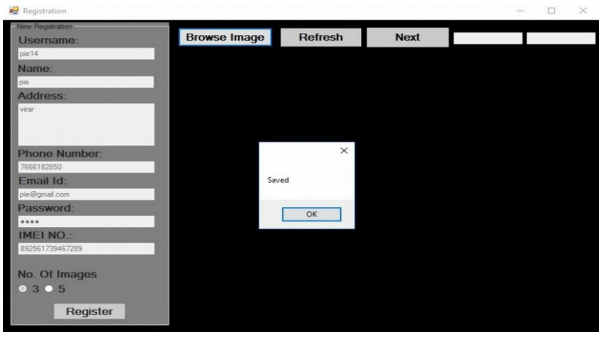


Fig3: Registration

The user needs to enter all his information in order to register himself.

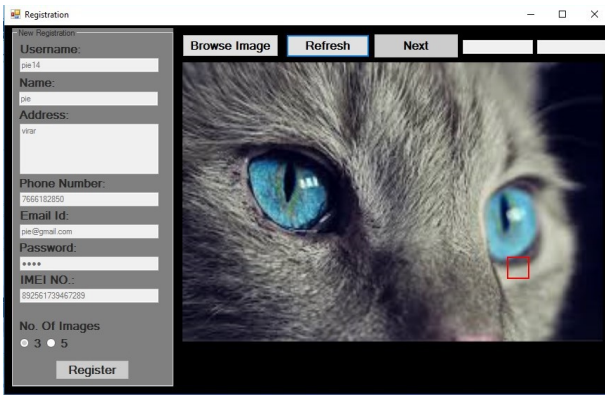


Fig4:select point on image

The user needs to browse an image and then he has to select a point in the randomly generated view port i.e. the red color square in the above fig 4 The user has to select the point on each image depending on the number of images selected by him at the time of registration. After all the images the password would be saved successfully.

If user chooses login method 1 then after entering the correct user name he will obtain a binary OTP on the registered email id.As per the obtained OTP user will have to select the same points which he had selected at the time of registration for the images when the bit in the OTP is 1 and select any other point except the point selected while registration for the image when the bit in the OTP is 0.

If user chooses login method 2 then after entering the correct user name he will obtain the 1st image.The user will have to select the same points which he had selected at the time of registration and after selecting all the points if the points selected are correct then he/she will be given a QR code which would be generated by the system.Now the user has to scan the QR code using our Android application. After scanning the QR code with the registered IMEI number device the user will obtain an OTP and will have to enter the obtained OTP.If user selects wrong point, random images will be generated and at the end he will get a message as Invalid.If user scans the QR code with another device then he will obtain message as INVALID USER.If user enters the wrong OTP, then he will get message as Access Denied.After following login method 1 or login method 2 accurately, user will be provided access to the system.

V.CONCLUSION

Thus we have successfully implemented a secure and more user friendly authentication system called Persuasive cued click point with OTP which contains graphical password which are easy for the users to remember as compared to text based passwords and difficult to attack for the attackers. Our system provides two login methods for the users so that users can login even when there is no internet or even if they don't have an android phone.In future we would like to improve many aspects of our project. We would like to include a forgot password and change password option in our system, so that user can change the password whenever he wants or if he forgets the password. We would also like to save the image points by encrypting them to increase security. For increasing the security at the second level of authentication we would like to include the encryption for generating QR code and decryption for obtaining OTP.

REFERENCES

[1] Sonia Chiasson, Elizabeth Stobert, Alain Forget, Robert Biddle, and Paul C. van Oorschot, "Persuasive Cued Click-Points: Design, Implementation, and Evaluation of a Knowledge-Based Authentication Mechanism," *IEEE transactions on dependable and secure computing,* vol. 9,no. 2, March/April 2012. .

[2] S. Chiasson, E. Stobert, A. Forget, R. Biddle, and P. van Oorschot, "Persuasive Cued Click-Points: Design, Implementation, and Evaluation of a Knowledge-Based Authentication Mechanism," *Technical Report TR-11-03, School of Computer Science, Carleton Univ.*, Feb. 2011.

[3] P.C. van Oorschot and J. Thorpe, "Exploiting Predictability in Click-Based Graphical Passwords," *J. Computer Security*, vol. 19, no. 4, pp. 669-702, 2011.

[4]Farnaz Towhidi, Maslin Masrom , "A Survey on Recognition-Based Graphical User Authentication Algorithms", *(IJCSIS) International Journal of Computer Science and Information Security,* Vol. 6, No. 2, 2009

[5]B. Rodrigues, A. Chaudhari and S. More, "Two factor verification using QR-code: A unique authentication system for Android smartphone users," *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, Noida, 2016, pp. 457-462. doi: 10.1109/IC3I.2016.7918008

Mrs. Anita Chaudhari completed her M.E **from** Mumbai University in 2013. Currently she is working as Assistant Professor in St. John College of Engineering and Management, Palghar, Mumbai University. She has published one national and seven international papers. Her research areas include network Security, and Data Mining.

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

The use of Technology in Discovering Money Laundry

Ramadan Mahmood Ramo
Assistant Lecturer
University of Mosul
Department of Management Information systems
rmrramo@yahoo.com

Dr. Khalil Ibrahim Alsaif
Professor
University of Mosul
Department of Computer Science
khalil_alsaif@hotmail.com

Abstract
After the emergence of globalization and the tremendous growth in money markets as well as the spreading of bank centers in many world countries, all that has the flexible role in the transportation of capitals amongst those countries. However, this process led to increase the money laundry crime (M.L. is an economic crime aims at giving legal legitimacy to illegally gained money in order to own, use, manage, keep, exchange, deposit, invest, transport, or manipulate it). Moreover, this crime is considered as one of the biggest challenges facing the financial institutions. Consequently, what makes this crime more dangerous is that the more efforts being done to fight money laundry processes, the more fight back and resistance will face these efforts. Money laundry processes witnessed the entering of new groups of professionals from high cultural levels for different specializations such as accountants, legists, information organizers, managers, and other specializations being employed by money laundry criminals to help them in this crime making these processes an integrated industry.

With the recently increased globalization, a technological revolution broke out in communications, information, sets, and equipment which are started to be used largely in all state institutions, especially financial and banking directorates.

This paper casts light on the phenomenon of money laundry and what is the technology which can be used in banks or even to use to discover money laundry and the role of banks in monitoring as well as to prevent this phenomenon which began to determine the financial and the state institutions.

I The Concept of Money Laundry

There is no unanimous agreement among the countries concerning the concept of money laundry which makes the attempts to fight these crimes harder especially at the international level, since some of these countries take the broad concept of money laundry as for the considerations of the financial revenues for all the criminal acts as ways to money laundry like trading, drugs smuggling, slavery, terrorism, bribe, political corruption, adultery, currency trading, embezzlement, arms trading, money counterfeiting, tax evasion, stealing, and other crimes and illegal acts. However, some other countries take narrow concept of money laundry in which these processes are limited only to the attempt of hiding the financial revenues to smuggle drugs [17].

II The Meaning of Money Laundry

Ghasala in Arabic (cleans) refers to the action of purging and purifying something by pouring water on it to remove its dirt. Furthermore, in Arabic alghusl is the noun of ghasala, alghusool is the cleaning water, and almaghsal is the place of cleaning [18][19].

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

III STAGES OF MONEY LAUNDERING

1. Placement : This is the movement of cash from its source. On occasion the source can be easily disguised or misrepresented. This is followed by placing it into circulation through financial institutions, casinos, shops, bureau de change and other businesses, both local and abroad. The process of placement can be carried out through many processes including[7] :

- Currency Smuggling
- Bank Complicity
- Currency Exchanges
- Securities Brokers
- Blending of Funds
- Asset Purchase[6]

2. Layering : The purpose of this stage is to make it more difficult to detect and uncover a laundering activity. It is meant to make the trailing of illegal proceeds difficult for the law enforcement agencies. The known methods are[9]:

- Cash converted into Monetary Instruments
- Material assets bought with cash then sold

3. Integration : This is the movement of previously laundered money into the economy mainly through the banking system and thus such monies appear to be normal business earnings. This is dissimilar to layering, for in the integration process detection and identification of laundered funds is provided through informants. The known methods used are[4]:

Property Dealing

- Front Companies and False Loans
- Foreign Bank Complicity
- False Import/Export Invoices

Figure (1) represents the stages of money laundering, which started with the first step and ended with the third step.

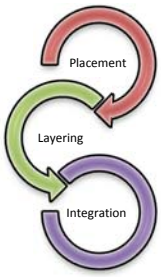


Figure (1) : Stages of Money Laundering

IV- Money laundering risks

1. Draining the national economy
2. Economic recession
3. Increasing balance of payments deficit and rising foreign indebtedness
4. The devaluation of the national currency exchange rate [5].
5. Rising interest rates on local currency
6. Inflation rates rise
7. Increase the tax burden
8. Fluctuation of stability in the stock market
9. The spread of the phenomenon of financial corruption[1].
10. Rising unemployment
11. High rates of economic crimes
12. The emergence of new criminal categories
13. The emergence of gangs specializing in money laundering.[13]

V- concept of banking technology

Modern technological methods have emerged as a quick means to process the funds which lost him the possibility of control over the source of such money laundering.

Also highlights the importance of advanced electronic means , which came as a result of the communications revolution and the evolution of its network through the comparison between traditional methods and modern methods in the stages of money laundering

Traditional methods rely on deposits in banks and smuggling across the border is safe to deposit operations while the use of modern means smart tickets, computers and through the Internet via the system protection and encryption to ensure the confidentiality of deposit operations.[20]

In terms of employment, in traditional ways are through remittances or through the means of payment other than cash such as checks tourist drafts drawn on banks abroad, but in the modern electronic means is up through a series of complex and rapid and successive operations that can be with them separated from their sources of illegal

Regarding the integration phase, they would be in a traditional means through fictitious transactions and invoices (bills) counterfeit and acts the role of gambling and brokerage , but in the electronic means is up through the purchase of physical assets and gambling by credit cards and by means of a personal computer without the mediation of banks , and in a manner that accuracy, speed and confidentiality so difficult with him the possibility tracked, and managed to rid the modern technological tools as follows[8].

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

VI Forms of banking technology

In light of the shift from the information age to the age of knowledge and wisdom and extensive use
For information technology and telecommunications, the banking and financial services industry by providing systems, applications and methods
New achieve maximum use of modern technology in the provision of banking services by high efficiency
And the positive impact on attracting customers, as many forms of technology emerged emerge as a Photo
Different uses of banking technology, which can be summarized in both directions who come.
Below are some forms of banking technology[11] :

- Plastic cards
- Electronic Cash
- Electronic Checks
- Home Banking
- Electronic Funds Transfer
- Internet Banking
- Interactive TV Banking
- Mobile Phone Banking

VII- The use of modern technologies in the discovery of money laundering

Technology is one of a number of components in the framework of effective compliance with the global anti-money laundering (AML). Using current technology tools, organizations can improve their ability to mitigate the risks of financial crime. There are many tools to combat money laundering has been the development of these tools with the passage of time poor in terms of where the financial services, data, and technology, and these tools are designed to help customers meet the challenges of compliance to fight their own complex money laundering.

There are at least four categories of technologies that may be useful in the analysis of wire transfers. These technologies can be classified by the task they are designed to accomplish:

- wire transfer screening to determine where to target further investigations.
- knowledge acquisition to construct new profiles for use during screening
- knowledge sharing to disseminate profiles of money laundering activities quickly, reliably, and in a useful form
- data transformation to produce data that can be easily screened and analyzed.[15]

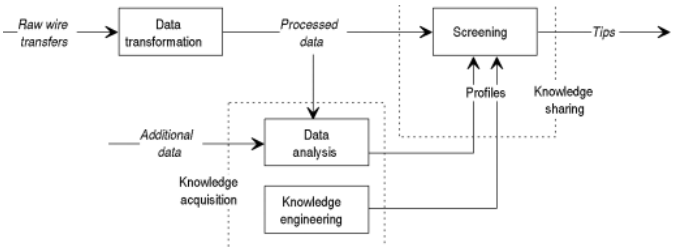


Figure 2: How Technologies Relate to Each Other

Technology is one of a number of components in an effective global anti-money laundering (AML) compliance framework. By using current technology tools, organizations can improve their ability to mitigate financial crime risk. Below is a set of proprietary AML automated tools and techniques that can help. [16] :

- ✓ **Computer Assisted Subject Examination and Investigation Tool (CASEit)** : A Web-based tool that facilitates AML compliance, AML transaction monitoring, trade surveillance, operational risk and anti-fraud case management

- ✓ **Customer Due Diligence Tool (CDD)** : Web-based tool that acts as the single data entry point and risk rating for all existing and new customer and account data in support of Know Your Customer (KYC) requirements. Additional customer and account information captured includes, ultimate beneficial owners, officers/directors (non-individuals and financial institutions only), power of attorney, co-signers, and other related parties

- ✓ **Name/entity matching** : Sophisticated matching and scoring tools and techniques that improve the searching of account and transaction information across systems, regions and business lines to create one view of the customer or to improve the name/entity screening (e.g. OFAC, PEP, etc) and matching processes (e.g. 314a, subpoenas, NSL, ad-hoc searches, etc)

- ✓ **Know your customer quick reference guide** : A user-friendly Web-based guide to anti-money laundering legislation and regulatory requirements for nearly 50 countries

- ✓ **Suspicious activity detection tuning** : Advanced methods and techniques that improve the

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

efficiency and effectiveness of transaction surveillance technology. We apply an empirical approach with an emphasis on statistical analysis of historical transaction data and alert output. By analyzing the population of data, institutions can identify trends and patterns and better determine which behaviors fall outside an acceptable range. Statistical analysis can be a first step in selecting appropriate rules and thresholds. Equally important is the reassessment of the monitored behaviors and thresholds over time. On-going analysis can be used to determine correlations and trends between productive and non-productive alerts allowing refinements that better target potentially suspicious activity, reducing overall review efforts.

- ✓ Fuzzy Computing Applications for Anti-Money Laundering : Fuzzy computing (FC) has made a great impact in capturing human domain knowledge and modeling non-linear mapping of input-output space. In this paper, we describe the design and implementation of FC systems for detection of money laundering behaviors in financial transactions and monitoring of distributed storage system load. Our objective is to demonstrate the power of FC for real-world applications which are characterized by imprecise, uncertain data, and incomplete domain knowledge. For both applications, we designed fuzzy rules based on experts' domain knowledge, depending on money laundering scenarios in transactions or the "health" of a distributed storage system. In addition, we developped a generic fuzzy inference engine and contributed to the open source community.[14]

- ✓ Genetic Clustering Algorithms for Detecting Money Laundering : Genetic Algorithms are one of the most applied class of algorithms for solving global/multi-modal optimization problems and have been extensively studied for solving NP-hard optimization problems. This work presents the development of genetic algorithms for detecting money-laundering by _nding the clusters in a graph, constructed using _nancial and customer data. The developed algorithm can be applied in other related areas as well. We present two algorithms based on Genetic Algorithms for (i) detecting all the clusters in a graph and (ii) detecting the cluster of any given node [15]

VIII The role of banks in money laundering discovery

The subject of anti-money laundering of the most important hot on international and regional issues, which explains the increased interest in him by a lot of countries that are keen on giving a true picture of the banking business in the state mode, through the issuance of a number of laws or to take several measures and procedures that emphasize its seriousness in combating money laundering through banks or related financial institutions .

Where the central bank plays a key role in the money laundering process, and it must be on the central bank and the obligation under the law of the following:

1. The balance of the funds coming from out of state or transferred to the outside through the size of financial institutions and reporting and movement and associated activate .
2. Balance control and no apparent unusual balance resulting from the movement of funds in the state does not coordinate with the economic reality .
3. monitor the activities of financial institutions in order to make sure that they are free to deal in transactions or money laundering .
4. Create a unit is doing the necessary investigations to uncover the ways and means to keep track of money laundering.
5. issuing bulletins and instructions for the issuance of the audit in the field of combating money laundering.

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

Reference

[1] Adam J. Szubin, “National Money Laundering Risk Assessment “ acting Under secretary, Terrorism and Financial Intelligenece,2015.

[2] Adha , Ibrahim (2003) The role of commercial banks on money laundering , Master Thesis , Department of Administrative Sciences , Naif Arab Academy for Security Sciences.

[3] Awad Allah , Safwat (2005 m) the economic impact of the operations of money laundering and the role of banks in the fight against these operations , Journal of Law , Vol. 29 , No. 2 , pp S13-138 .

[4] Dennis Cox, Handbook of Anti Money Laundering,2014.

[5] fatf National,” Money Laundering and Terrorist Financing Risk Assessment”, 2 rue André Pascal 75775 Paris Cedex 16, France , 2013.

[6] Financial Action Task Force (FATF) , www.fatf-gafi.org Centre for Tax Policy and Administration work on tax crimes and money laundering ,2009.

[7] Friedrich Schneider “Money Laundering and Financial Means of Organized Crime: Some Preliminary Empirical Findings”, Economics of Security Working Paper 26, Berlin: Economics of Security,2010.

[8] Dr.. Iftikhar Muhammad & Dr. Manhel Mustafa, “The role of banks to cope with financial fraud and money laundering operations “ Baghdad University / College of Administration and Economy & Iraq Central Bank,2008

[9] John madinger , money Laundering A guide for criminal investigators,2012.’

[10] Llu Alsed, Abhishek Awasthi2, “Genetic Clustering Algorithms for Detecting Money-Laundering” Department of Mathematics, Autonomous University of Barcelona, Spain alseda@mat.uab.cat 2 Department of Electrical Engineering and Computer Science , University of Applied Sciences Zittau/G orlitz, Germany {aawasthi,jlaessig}@hszg.de

[11] Dr.. Rafeai brahim al-Hamdani ,”The effect of using technology in the banking phenomenon of money laundering and international efforts to combat”, Mosul / Iraq University,2005

[12] Dr.. Rafeai brahim , Job environment and its impact on the profitability of banks and risk – study Analysis of a sample of banks , unpublished PhD thesis , Faculty of Management and Economics , University of Mosul , 2003.

[13] Dr. Saleh Saad , “Damage and the risk of money laundering” Arwas, Oman .2007

[14] Yu-To Chen, Johan Mathe,” Fuzzy Computing Applications for Anti-Money Laundering and Distributed Storage System Load Monitoring”. Google, Inc. 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA{ytchen, johmathe}@google.com

[15] Mahmood Shah, Steve Clarke , “E-Banking Management: Issues, Solutions, and Strategies” , Lancashire Business School, University of Central Lancashire, UK, University of Hull, UK,2009

[16] Ali Khayrallah, Nimish Radia, Jason Hickey, Jasvinder Singh, Vicky Xu, ” Technology & Banking “ Ericsson, Google Inc, Synopsys, VMware, 2015

[17] Fayrouz Abadi: Majd al-Din Muhammad ibn Ya`qub: The Surrounding Dictionary, Dar al-Hadith, Cairo

[18] Al-Rubaie: Zuhair Said: Money Laundering, the plague of the age and the mother of crimes, Al-Falah Library, Kuwait, 1, 1425 – 2005

[19] Tantawi, The Legislative Confrontation of Money Laundering in Egypt, The Arab Renaissance House, Cairo, 2003

[20] Punitha Sahaya Mary Francis,” Banking Technology “,Assistant Professor, MBA Dept., St. Xavier’s Catholic College of Engineering, Chunkaankadai, Nagercoil. punithaf2007

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

Improve the Offloading decision by Adaptive partitioning of task for Mobile Cloud Computing

Neha Goswami
Computer Science and Engineering Department, Chandigarh University
National Highway95, Mohali, 140413, Punjab, India;
nggoswamineha@gmail.com

Sugandha Sharma
Computer Science and Engineering Department, Chandigarh University
National Highway95, Mohali, 140413, Punjab, India;
Sugandha.ss1@gmail.com

Abstract- Adaptive Offloading in Mobile Cloud Computing by automatic partitioning approach of tasks is the idea to augment execution through migrating heavy computation from mobile devices to resourceful cloud servers and then receive the results from them via wireless networks. Offloading is an effective way to overcome the resources and functionalities constraints of the mobile devices since it can release them from intensive processing and increase performance of the mobile applications, in terms of response time. Offloading brings many potential benefits, such as energy saving, performance improvement, reliability improvement, ease for the software developers and better exploitation of contextual information. Parameters about method transitions, response times, cost and energy consumptions are dynamically re-estimated at runtime during application executions.

Keywords- ***Cloud Computing, Mobile Cloud Computing(MCC), Offloading, System model (cost model, energy model, weighted model)***

I. INTRODUCTION

With the computing technology advancements, the desktop computer usage expands and main frame is also expanded to a wider range of embedded applications and mobile applications which includes environmental sensing, surveillance, mobile phones and GPS navigation, etc. These several applications on systems run with resources that are limited. Fore.g, battery powered is the mobile phones. The environmental sensors comprises of smallersize physically, small storage amounts and slow processors. Wireless applications are used by most of these applications and low are their magnitude orders in comparison with wired networks. Meanwhile, programs that run on these systems have increasing complexity—for example, video processing on mobile phones and object recognition on mobile robots. An increasing gap is there among the complex programs demand and the limited resources availability.

A solution is offloading for augmentingthe capabilities of this mobile system with computational migration for more computers that are resourceful (i.e., servers). The client-server architecture traditionally utilized is different from this, in which always a thin client migrates to a server computation. There is a difference between the computation offloading and the migration model which utilizes in grid computing and multiprocessor system, where for load balancing migrating the process [1-6]. The key difference is that computation offloading differs mainly when it migrates programs to servers outside the computing environment of the immediate users; typically migrating the process to grid computing which occurs within computer in asimilar computing environment, that is, the grid. In similar principle for efforts is offloading like SETI@home [8], where for computation performance sending request to surrogates. The terms like “surrogate computing” and “cyber foraging” are also utilized for describing computation offloading.

Computation offloading[2-3, 6, 9] is growing topic on which a significant amount of research has been done which makes it feasible, making decisions of

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

offloading, and developing infrastructures of offloading. Most of the researches done over offloading prior to 2000are on making it feasible. Primarily, this is because of its limitations in wireless networks, such as lesser bandwidths. The focus shifted to developing algorithms utilized for making offloading decisions in early 2000s that is, deciding*whether* mobile users are benefited by offloading. The offloading direction is shifted with the improvements in network bandwidths, cloud computing infrastructures, and virtualization technology. More practical has become computation offloading with these developments.

On mobile systems, saving the energy and performance improvement may be done by utilizing computation offloading depending usuallyon various parameters like as the data exchanged amounts through various networks and network bandwidth. Various algorithms have been proposed for making the offloading decisions to save energy or improve performance [10, 13, 15-16, 18]. Usually decisions are made with the parameter analyzation including server speeds, bandwidths, server loads, available memory, and the data exchanged amount among mobile systems and servers. In the execution environment [11], application behavior variation of predicted parametric and partitioning programs [15] is included in the solutions.

Resourceful computers access is required in offloading through networkswireless or wired for short durations. *Virtualization* may be used by these servers for providing offloading services so as various programs and their data can be protected and isolated. Protections and isolations have motivated the researchers to developoffloading infrastructures at different granularities [16]. Performing offloading may be at the levels of tasks [9],methods [3], virtual machines [5], or applications [11]. .NET remoting, RPC (remote procedure call), and Java RMI are the various mechanisms which enables offloading at the object level and class level. At the virtual-machine level, techniques have been proposed for offloading enabling [8, 17]. The offloading and elastic resources allowed by cloud computing for multiple servers; which enables the computation offloading. Several solutions and infrastructures and solutions have been proposed to improve offloading: they deal with various issues forprivacy, users, security, mobility, etc.The goal is paper to improve the utilization of computing resources and reduce energy consumption.

II. LITERATURE REVIEW

For the mobile codes execution on the remote server suchas the wall-powered PCs or cloud, a technique has been employed in previous work, called offloading execution, which is the execution transferring act between two machines run time. By computational loads relieving, the labors technique for bring benefits of smartphones in terms of execution time and batteryin their proximityfrom the servers. In offloadingexecution, two maintasks are there which involves before the remote execution: state migration and code partitioning. In recent years, greater deal of research has been conducted there for finding or supportingof distributed systems optimal partitioning with mobile devices.The static partitioning schemes are proposed [10] by several researchers where the assigned jobin the system to each machine at compile timeis fixed. More doable ought to be the static partitioning if the resource configurations computationallysuch as memory capacity, processor speed, network characteristics, and energy consumption, remain fairly constant as process is launched. In mobile computing, however, the change in configurations is possible due to user mobility even in the process execution mid. Therefore, in other works mostly [4], [5], [8], [11] have been on the semi-dynamic or dynamic partitioning schemes development to execute offloading. In a partitioning schemedynamically, annotated is the code commonly withdirectivesdelimiting the code regions which can be delegated for remote executionto the server, if profitable? Which regions run on the server actuallyis decided during the execution of code when for execution the resource configurations areknown. Finally, once certain region at run time is selected, execution needs for current state tobe captured and server migration with the control command in which execution resumption is directed.In other approach [5], the entire state including the existing stack is included in entire state and migrated all heapreachableobjects for offloading the full process. In anotherapproach [4], the stack is not to be migrated as the functions set to run remotely will be newly invoked in the server. Clearly, these two

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

approaches are having trade-offs. Above all, the usuallythe state transferred amount, that is a major decisive factor for themobile networks over execution offloading efficacy and have latter of smaller size.

III. CLOUD COMPUTING

Cloud computing is an important model in the world of information technology. It is a type of computing which provides a shared pool of virtualized and managed computing resources to the customers on their demand over the internet and other available networks rather than having local servers or their own personal devices to handle the applications. The word cloud is used as an analogy for “the internet”, so the term cloud computing means “ a type of internet based computing” where different applications and services are provided to the organizations over the internet. Cloud computing is a network of large group of servers interconnected to each other i.e basically it a shared infrastructure which contains large pools of systems that are linked together. Due to the various features like elasticity, scalability and availability, cloud computing becomes more popular these days.

There are three types of cloud computing (i) public cloud (ii) private cloud and (iii) hybrid cloud.

Public cloud provides an infrastructure where the information is shared among different organizations. In this model, services providers provides resources like storage and applications to the general public. It’s a pay per usage model i.e the users only have to pay for what they used.

Private cloud provides an infrastructure that is dedicated to a particular organization i.e the resources and applications are not shared with the other organizations. Private cloud provides more security to the users than that of public cloud but also more expensive.

Hybrid cloud is the composition of public cloud and private cloud. In hybrid cloud, the important and critical files or applications are hosted in private cloud. While the applications with less security concerns are hosted on public cloud. Only the authorized organizations can access the applications from private cloud while the applications and resources from the public cloud are easily accessible by the users.

Three service models of cloud computing are (i) Saas (ii) PaaS (iii) IaaS.

Software as a service (SaaS) is a model that allows users to access the internet based software’s. Users don’t have to install, manage and upgrade the applications on their devices as SaaS providers will manage all this. Users need not to worry about their data as equipment failure will not results in data loss because data is secure in cloud.

Platform as a service (PaaS) is a cloud computing model that provides users a platform with tools where they can develop, manage and deliver their applications. Providers are responsible for managing security, backups and also for handling operating systems and server software’s.

Infrastructure as a service (IaaS) is a cloud computing model that allows users to access the resources, storage and servers. Instead of purchasing and maintaining own hardware devices the users can pay per usage for IaaS on demand.

Charactristics of cloud computing are(i)*On-demand self-service:* A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider. (ii)*Broad network access:* Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops and workstations). (iii)*Resource pooling:* The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state or datacenter). Examples of resources include storage, processing, memory and network bandwidth. (iv)*Rapid*

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

elasticity: Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time. (v)*Measured service:* Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth and active user accounts). Resource usage can be monitored, controlled and reported, providing transparency for the provider and consumer.

IV. MOBILE CLOUD COMPUTING

Mobile cloud computing (MCC) is termed as the combination of mobile computing, cloud computing and networks in order to provide various services to the mobile users like providing various computational resources, network operators etc. Mobile cloud computing is a model in which the mobile applications are built and hosted using cloud technology. It is different than that of mobile computing as the devices can run cloud based applications rather than that of remote applications. MCC provides a lot of opportunities for mobile industry and also allows the mobile users to utilize the various resources offered by cloud. Three approaches of MCC are : (i) providing access to cloud services (ii) allows mobile devices to work together as cloud service providers work. (iii) augmenting the execution of various mobile applications on the portable devices using cloud resources.

Characteristics of Mobile Cloud Computing are

(i)*Convenience of data sharing:*- A huge amount of data is stored in the backend servers which allows the users to access the data at any time without depending on the mobile device.

(ii)*Effective task Processing* :- As there is an interface in the cloud and mobile devices the users are able to see the output directly on the mobile device.

(iii)*Elimination of the regionality:*- Mobile cloud computing eliminates the regionality for accessing the data. The user can access the data at any time and at any place.

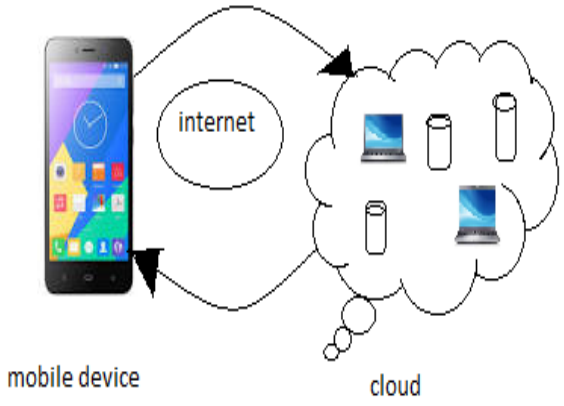


Figure 1:- Mobile Cloud Computing (MCC)

V. OFFLOADING IN SMARTPHONES

Smartphones has some constraints like limited energy, small processors and memory capacity. As smartphones becomes more popular in our life and with the advancement in the technology they uses more powerful processors, different operating systems like android, apple iOS , windows phone and larger memory provide sharper screens , multiple sensors but these together put a burden on battery's energy consumption. One approach for saving mobile energy is task offloading.

Mobile cloud computing enhance the computing capabilities of the smartphones by offloading the computation to the cloud infrastructure. In the cloud each mobile device has a clone cloud, which runs on virtual machine and executes the mobile applications.

Offloading a task to the cloud is a critical technique as in some cases the energy consumption of the smartphones increases. To make offloading beneficial, the energy cost of a particular task should be estimated by comparing it with the task that is executed locally on the mobile device. The energy consumed during the task offloading process is mainly due to the networking activities.

There are two types of offloading (i)Partial Offloading (ii)Complete Offloading.

Partial offloading:- The compute intensive applications are partially offloaded to the cloud in order to reduce the load of the smartphones. In this

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

offloading process the heavy computation is done on the virtual machines i.e clone cloud and the other part of the application is executed on the smartphone. After completing both the executions the smartphone collect all the results and provide them to the customer in the required format.

Complete Offloading :- In complete offloading the whole task is uploaded to the cloud in order to save the battery life where the execution of the task is performed on the cloud and then download the task results on the mobile device. The cloud acts as an exact replica of the mobile device having same data.

VI. SYSTEM MODEL

The cost for a task is a user-defined metric, which could be a task's response time, usage of CPU power, and consumption of battery energy by the task, etc. In this dissertation, we investigate modeling and optimality by considering three parameters: the energy consumption on the mobile device, the application response time and cost. We deploy three models- cost model, energy model and weighted model.

A. Cost Model: In this model, we calculate the memory gain of offloading. We also calculate the time of offloading for a particular tasks. After this calculation we perform offloading.

Algorithm- Cost Model

for(i=1 to k);
{
Calculate the memory gain of offloading
gain $\leftarrow \sum_{i=1}^{n}$ (mobile $_I$ – cloud $_i$)
end
$U_i \leftarrow$ max { gain$_{i \text{ to } i}$ $1 < i \leq k$ }
For (j=2 to k);
{
calculate the time of offloading of j^{th} task
$X_j = x_{j-1} – (\text{mobile}_{j-1} – \text{cloud}_{j-1})$ }
Offloading point

B. Energy Model : In this model, we calculate the energy gain of offloading. Energy is the total amount of work performed over a period of time. We also calculate the time of offloading of given tasks. After this calculation we perform offloading.

Algorithm- Energy Model

for(i=1 to k);
{
Calculate the energy gain of offloading
energy $\leftarrow \sum_{i=1}^{n}$ (mobile $_I$ – cloud $_i$)
end
$U_i \leftarrow$ max { gain$_{i \text{ to } i}$ $1 < i \leq k$ }
For (j=2 to k);
{
calculate the time of offloading of j^{th} task
$X_j = x_{j-1} – (\text{mobile}_{j-1} – \text{cloud}_{j-1})$ }
Offloading point

C. Weighted Model: In this model we have the memory gain on the cloud and mobile, which is not available in previous models. Then the EM algorithm is applied to partition the task optimally.

In E step-using the current best guess for the parameters of the data model, we construct an expression for the log-likelihood for all data, observed and unobserved, and, then, marginalize the expression with respect to the unobserved data.
In M step-given the expression resulting from the previous step, for the next guess we choose those values for the model parameters that maximize the expectation expression. These constitute our best new guess for the model parameters.

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

Algorithm – weighted model

for (i=1; i< k; i++)

{

Calculate the memory gain of offloading

gain $\leftarrow \sum_{i=1}^{n}$ (mobile $_i$ – cloud $_i$) – input $_i$ – return $_i$

end

$U_i \leftarrow$ max { gain$_{i \text{ to } i}$ $1 < i \leq k$ }

Integrating Point $\leftarrow$ EM (max (gain $_{i \text{ to } i}$ $1 < i \leq k$))

for (j=2; j< k; j++);

{

calculate the time of offloading of j^{th} task

$X_j = x_{j-1} – (\text{mobile}_{j-1} – \text{cloud}_{j-1}) + \text{input}_{j-1} – \text{input}_j$

}

Offloading point

VII. EXPERIMENTAL RESULTS

TASK	TIME (response)	TIME (energy)	TIME (weighted)
twenty five	6	2	3
fifty	85	7	1
hundred	174	1	1
one hundred fifty	159	1	0
two hundred	254	1	2
two hundred fifty	363	1	1
three hundred	567	1	9
three hundred fifty	753	2	1
four hundred	1106	2	1
four hundred fifty	1443	1	1
five hundred	1781	1	4

Table 1: Comparison table of Time parameter among Response, energy and weighted

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

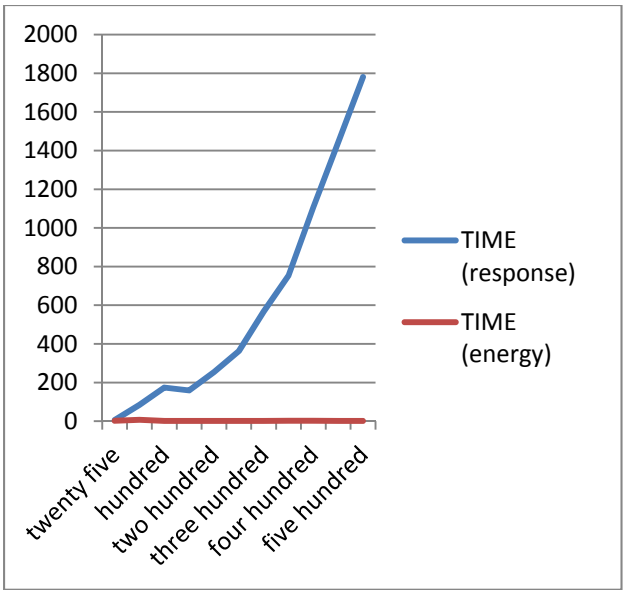


Figure 2: Comparison graph of time parameter between Response and Energy

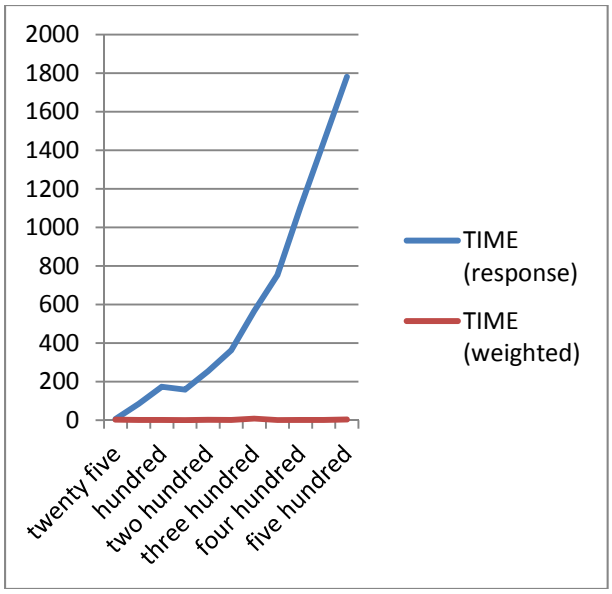


Figure 3: Comparison Graph of time parameter among Response and weighted

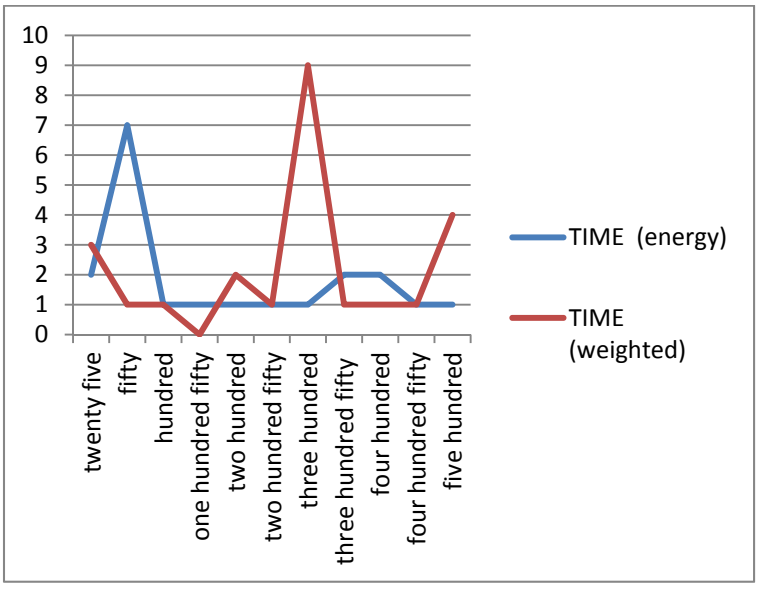


Figure 4: Comparison graph between energy and weighted (time)

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

TASK	MEMORY (weighted)	MEMORY (response)	MEMORY (energy)
twenty five	0.206481	0.29795	0.2061004
fifty	0.132324	0.62301	0.1318359
hundred	0.143814	0.98295	0.14331054
one hundred fifty	0.155487	3.92187	0.15498352
two hundred	0.166625	4.17924	0.16603088
two hundred fifty	0.1765747	4.40298	0.17607116
three hundred	0.1895751	12.10267	0.189071655
three hundred fifty	0.1997375	15.560867	0.19914245
four hundred	0.2097167	8.11734	0.20921325
four hundred fifty	0.22369384	16.32435	0.223190307
five hundred	0.233764648	21.154418	0.2332611

Table 2: Comparison table of Memory parameter among Response, energy and weighted

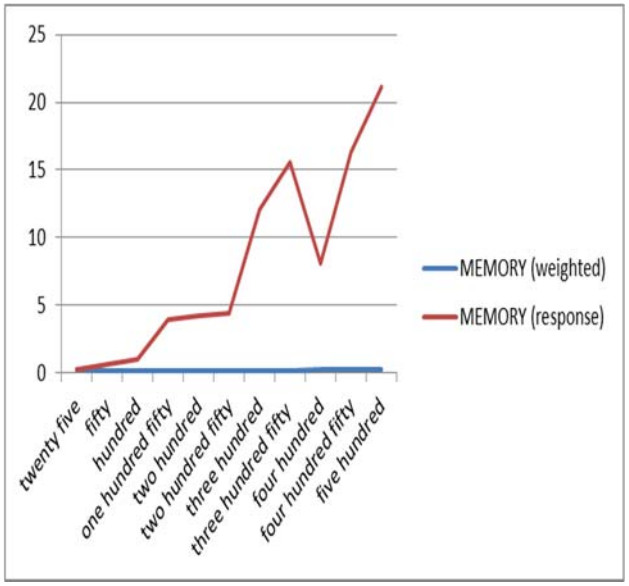


Figure 5: Comparison graph of Memory parameter among Response and weighted

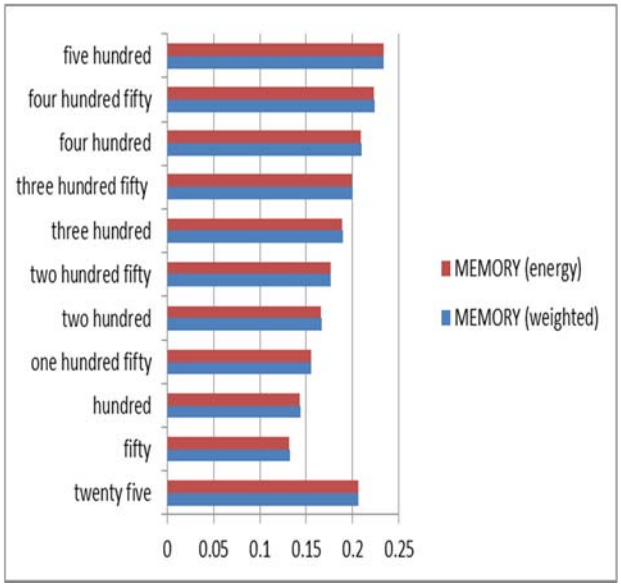


Figure 6: Comparison graph of Memory parameter among energy and weighted

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

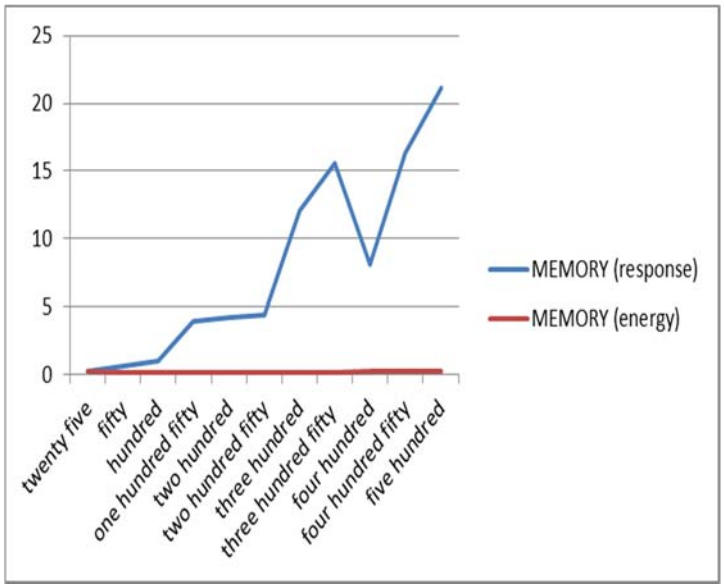


Figure 7: Comparison graph of Memory parameter among Response and energy

VIII. CONCLUSION

Several solutions have been proposed to enhance the CPU performance and to manage the disk and screen in an intelligent manner to reduce power consumption. However, these solutions require changes in the structure of mobile devices, or they require a new hardware that results in an increase of cost and may not be feasible for all mobile devices.

By offloading the computational intensive part of the application which requires less communication to the rest of the application, to the cloud, a mobile device can save significant amount of battery energy and provide more responsive user experience. In this thesis, we have presented a computation offloading scheme on virtual devices. Our partition algorithm finds optimal program partitioning for given program input data. Experimental results show that our computation offloading scheme can be significantly improve the performance and energy consumption on handheld devices.

IX. FUTURE SCOPE

Computation offloading technique is proposed with the objective to offload the large computations and complex processing from resource-limited devices (i.e., mobile devices) to resourceful machines (i.e., servers in clouds). This helps in reducing the application execution time and also power consumption.

More work is needed to be done for offloading field. Future work can be extended in the following fields: Here, we use EM algorithm for optimal partition the application running on Smartphone, in future we will use metaheuristic algorithm, since metaheuristics are one of the most practical approaches for solving hard optimization problems. This approach is interesting for very large problem instances. We will also take our operations to hardware for better understanding the calculate delay and memory usage.

REFERENCES

[1] Li Z,WangC,XuR,"Computation offloading to save energy on handheld devices: a partition scheme",InProceedings of the 2001 international conference onCompilers, architecture, and synthesis for embeddedsystems (CASES),238–246,2001

[2] ChengWang and ZhiyuanLi,"A computation offloading scheme on handheld devices", Journal of Parallel and Distributed Computing, 2003

[3] Karthik Kumar and Yung-Hsiang Lu,"Cloud computing for mobile users: Can offloading computation save energy?",Computer, vol. 43, no. 4, pp. 51–56,IEEE 2010

[4] S. Kosta, A. Aucinas, P. Hui, R. Mortier, and X. Zhang, "Thinkair: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading." in Proc. IEEE INFOCOM 2012

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

[5] J.Lei, F.Roy, FU.Xiaoming, *S.*Stefano, S. Zbigniew and T. Hannes,"Cloud-based Computation Offloading forMobile Devices: State of the Art, Challenges and Opportunities",Future Network and Mobile Summit Conference Proceedings, IIMC,2013

[6] SokolKosta. Alessandro Mei, JulindaStefa,"To offload or not to offload? The bandwidth and energy costs of mobile cloud computing", INFOCOM, Proceedings IEEE 2013

[7] Huber Flores, SatishNarayanaSrirama, RajkumarBuyya,"Computational Offloading or Data Binding? Bridging the Cloud Infrastructure to the Proximity of the Mobile User"IEEE,2014

[8] Roopali, Rajkumari,"Overview of Offloading in Smart Mobile Devices for Mobile Cloud Computing",Vol 5, Number 6, 2014

[9] Jaya Ashok Suradkar and R. D. Bharati, "Computation Offloading: Overview, Frameworks and Challenges" International Journal of Computer Applications Volume 134 – No.6, January 2016

[10] KarthikKumar·JibangLiu·Yung-Hsiang Lu· Bharat Bhargava,"A Survey of Computation Offloading for Mobile Systems", Springer Issue on Mobile Network Application,10 April, 2012.

[11] Monika Dudeja and KritikaSoni," Offloading Schemes in Mobile Cloud" International Journal of Computer Applications Volume 96– No.8, June 2014

[12] Xu Chen," Decentralized Computation Offloading Game For Mobile Cloud Computing", IEEE 2015

[13] Chung.J, Park.Y, Park.J, and Cho.H,"Adaptive Cloud Offloading of Augmented Reality Applications on Smart Devices for Minimum Energy Consumption",Ksii Transactions On Internet And Information Systems VOL. 9, NO. 8, 31 August 2015

[14] D. Murali, J. Vamsinath,"Cloud Computing For Mobile Users: Be Capable of Offloading Computation save Energy?", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 9, September 2014

[15] Ying-Dar Lin, Chu, E.T.-H., Yuan-Cheng Lai and Ting-Jun Huang, "Time-and-Energy-Aware Computation Offloading in Handheld Devices to Coprocessors and Clouds," in Systems Journal, IEEE , vol.9, no.2, pp.393-405, June 2015.

[16] R. Kemp, N. Palmer, and T. Kielmann, "Cuckoo: A computation offloading framework for smartphones," presented at the Proc. 2nd Int. Conf. Mobile Comput., Appl., Serv., Santa Clara, CA, USA, 2010.

[17] Kaushik. N, Kumar. J,"A Computation Offloading Framework to Optimize Energy Utilisation in Mobile Cloud Computing Environment", International Journal of Computer Applications & Information Technology Vol. 5, Issue II April May 2014

[18] S.Rathnapriya, S. Sumithra," Techniques to Minimize State Transfer Cost for Dynamic Execution Offloading In Mobile Cloud Computing", Int. Journal of Engineering Research and Applications, Vol. 5, Issue 5, (Part -5), pp.32-35, May 2015

AUTHORS PROFILE

Neha Goswami, Student of ME(maters of engineering in CSE) at Chandigarh University.

Er. Sugandha Sharma, Assistant Professor at Chandigarh university.

The Quantification of Human facial expression Using Fuzzy Logic.

Dileep M R [1] and Ajit Danti [2]
N E S Research Foundation,
Department of Computer Applications,
Jawaharlal Nehru National College of Engineering, Shimoga, Karnataka, India
[1] dileep.kurunimakki@gmail.com
[2] ajitdanti@yahoo.com

Abstract: Fuzzy logic is an interesting theory that allows the natural description, in linguistic terms, of problems that should be solved rather than in terms of relationships between precise numerical values. In this paper, an effective approach is proposed that quantifies the human facial expression using Mamdani implication based fuzzy logic system. The new technique involves in extracting mathematical data from the face and fed to a fuzzy rule-based system. Fuzzification and Defuzzification operation issues trapezoidal membership functions for both input and output. The distinct feature of a system is its simplicity and high accuracy. Experimental results on Image dataset indicate good performance of the proposed technique. Comparative analysis reveal that the proposed technique is uniqueness and robust with reference to other state of the art methods.

In this paper, a legitimate procedure proposed for quantification of human facial expression recognition from Facial features using Mamdani-type fuzzy system. It is Fuzzy Inference System (FIS), which is capable to set up an easy membership relation between the different dimensions of the happy expression. The FIS recognizes three levels of same happy expression namely No happy, Bit Smiley and Loud Laugh based on membership function modeled on different psychological studies and surveys.

Index Terms – Fuzzy Rule, Quantification of Expression, Membership Function

1. INTRODUCTION

Facial expression is one of the most important subjects in the field of biometric, which has wide range of applications such as Business, Managerial, Organizational, Cultural contexts, Telecommunication, Medical, Human Computer Interactions (HCI). The ideal human computer interaction system is the one that the computer is able to communicate and respond to the user actions, based on emotional state of human's face. In this way the user will be able to communicate with it more effectively. For this aim, automatic recognition of human's facial expressions has been very active research area in machine vision within the last several years. Facial expressions are generated by movement of face muscles that makes facial features such as stretching corner lips, raising eyebrows, opening eyes, etc.

The traditional approach to building any system controllers requires a prior model of the system. The quality of the model, that is, loss of precision from linearization and/or uncertainties in the system's parameters negatively influences the quality of the resulting control. At the same time, methods of soft computing such as fuzzy logic possess non-linear mapping capabilities, do not require an analytical model and can deal with uncertainties in the system's parameters. Although fuzzy logic deals with imprecise information, the information is processed in sound mathematical theory. Based on the nature of fuzzy human thinking, Zadeh, originated the "fuzzy logic" or "fuzzy set theory", in 1965. Fuzzy logic deals with the problems that have fuzziness or vagueness. In fuzzy set theory based on fuzzy logic a particular object has a degree of membership in a given set that may be anywhere in the range of 0 (completely not in the set) to 1

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

(completely in the set). For this reason fuzzy logic is often defined as multi-valued logic (0 to 1), compared to bi-valued Boolean logic.

Facial expression process usually extract facial expression parameters from a static face image. This process is called as "Quantification" of expression. These extracted features are then fed to a classifier system for facial expression quantification by defining the range of expression with their membership function. In this paper, a complete system for quantification of facial expression i,e different range of happy faces is proposed. The core of our system is a Mamdani-type Fuzzy Rule Based system which is used to quantify facial expression from facial features. The comparison of proposed methodology with other state of the art methods is carried out.

Aruna Chakraborty et al, 2009, introduced a method for emotion recognition from facial expressions and its control using fuzzy logic, where the expressions of the human face would be recognized by applying the fuzzy rule implementation. An effective application is invented by B. K. Bose, 1994, about the expert systems, fuzzy logic, and neural network application in power electronics and motion control. It was a multiple application system. Dae-Jin Kim and Zeungnam Bien, 2003, developed an algorithm of Fuzzy Neural Networks (FNN) - based approach for Personalized Facial Expression Recognition with Novel Feature Selection Method, where this method was the combination of Artificial Neural Network and Fuzzy Inference System. This method was used to recognize the different expressions of the human face. Dennis Gillette and Ping Zhang did an effective survey on Human-Computer Interaction And Management Information Systems: Applications. S. Dongcheng, J. Jieqing , 2010, introduced a method of facial expression recognition based on DWT-PCA/LDA, the multiple application approach. Esau N, et al, 2007, given an approach for Real-Time Facial Expression Recognition using a Fuzzy Emotion Model, that effectively recognizes the emotions of the human face. Francisco Herrera and Luis Magdalena, 1997, published a tutorial on genetic Fuzzy Systems gives a detailed description on biological features of human faces using fuzzy logic. P. S. Hiremath and Ajit Danti, 2005, developed a methodology on fuzzy-rule based method for human face detection, that effectively detects the human faces by applying fuzzy inference system. A. Jamshidnezhad, 2011, designed an approach that learns Fuzzy Model for Emotion Recognition that recognizes the emotions of the human face. B. Jaychandra, simulated Speed Sensorless Operation of Vector Controlled Induction Motor Using Neural Networks. Khanum A et al, 2009, contributed a research on Fuzzy case-based reasoning for facial expression recognition, that efficiently recognizes the human facial expressions based on Fuzzy rule. G. Klir and B. Yuan, 2010, published a study material on Fuzzy sets and Fuzzy Logic – Theory and Applications, that gives a detailed description on the applications and usage of the fuzzy rule. Kyoung- Man Lim et al, designed an algorithm for face recognition system using Fuzzy Logic and Artificial neural network, that effectively recognizes the face in a given still image. S. Y. Lee et al, 1996, introduced a method that recognizes the human front faces using knowledge based feature extraction and neuro-fuzzy algorithm, a multiple application approach, that uses geometrical facial features of the face along with Artificial neural networks plus fuzzy inference systems. Maedeh Rasoulzadeh, 2012, did a research on facial expression recognition using Fuzzy Inference System. Milki et al, 1991, designed an effective application for Vector control of induction motor with fuzzy P-I controller. Muid Mufti and Assia Khanum, 1991, proposed an approach on Fuzzy Rule-Based Facial Expression Recognition in a precise manner. Mufti M and Khanam A, 2006, solved an unique problem on Fuzzy Rule Based Facial Expression Recognition. M. Nasir Uddin, 2002, pursued a study on performances of Fuzzy-Logic-Based

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

Indirect Vector Control for Induction Motor Drive, an effective approach on fuzzy logic. Ralescu A and Hartani R, 1995, studied some issues in Fuzzy and Linguistic Modeling. D. H. Rao and S. S. Saraf, 2007, conducted a detailed study of Defuzzification Methods of Fuzzy Logic Controller for Speed Control of a DC Motor.

Several researchers contributed their work on emotions, emoticon recognition using facial expressions using fuzzy inference system and neural networks etc. namely M.S.Ratliff and E. Patterson, 2008; T. A. Runkler, 1996; M. Schmidt et al, 2010; N.Sebe et al, 2005; Starostenko O et al, 2010; TakKuen John Koo, 1996; T. Takagi and M. Sugeno, 1985; L. H. Tsoukalas and R. E. Uhrig, 1997; Ushida H et al, 1993; M. Usman Akram, et al, 2008; V. P. Vishwakarma et al, 2010; T. Xiang et al, 2008;

1.1 Fuzzy Inference System

Fuzzy inference is the process of formulating the mapping from a given input faces to an output facial expressions using fuzzy logic and membership function. In this paper, Mamdani's fuzzy inference method is used and it expects the output membership functions to be fuzzy sets. After the aggregation process and defuzzification, crisp decision is made on facial expression as shown in Figure-1.

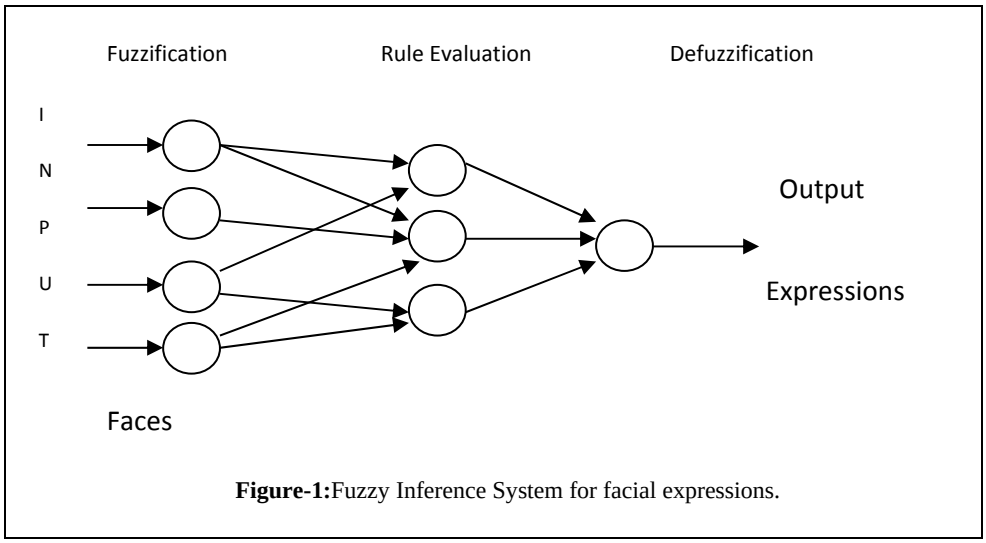


Figure-1: Fuzzy Inference System for facial expressions.

1.2 Face Database

The proposed methodology is experimented on database of the faces of people of different dimensions of happy expressions viz Normal, Bit smiley and Loud Laugh. There are 1000 facial images in this database, among them 700 images were used for training and remaining 300 images were used for testing purpose. Each image is normalized to a size of 64 × 64 dimensions for optimum computational cost. Sample facial expression images are shown in figure-2.

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

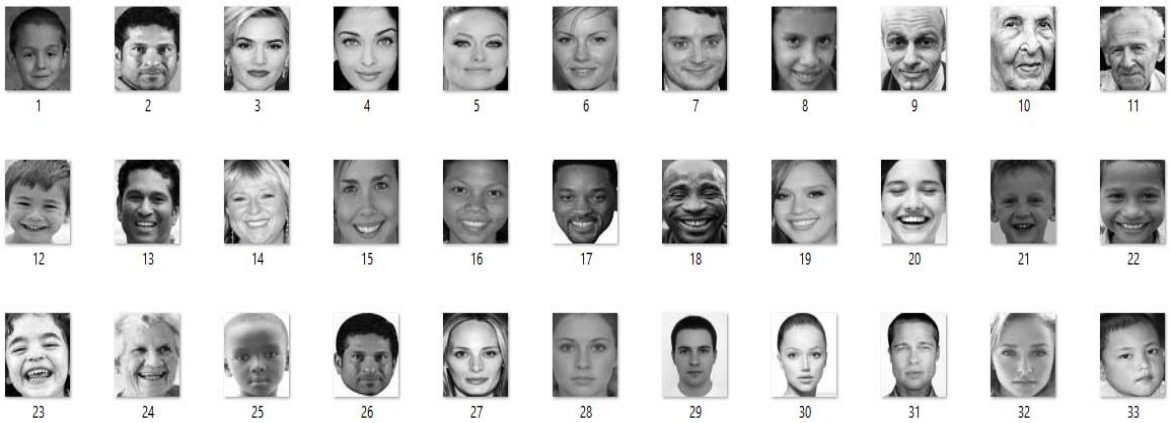


Figure-2: Samples of the facial images in the dataset

The rest of this paper is organized as follows. The Proposed Methodology is explained in Section 2. The Proposed Algorithm is described in section 3. In Section 4, Experimental results are presented and Conclusions are drawn in section 5.

2. PROPOSED METHODOLOGY

This paper proposes an effective method for quantification of human facial expressions from facial images. Block diagram of the proposed method is given in Figure-3.

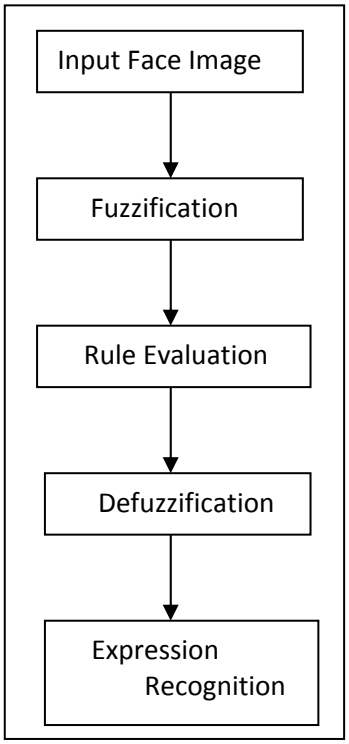


Figure-3 : Block diagram of the proposed Methodology

The Proposed algorithm has been implemented to classify input images into one of three happy expression using Mamdani-type Fuzzy Rule Based system In order to improve the

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

efficiency of the performance, mean vector is obtained for each image is used in the fuzzy inference system as shown infigure-1.

2.1Quantification of Facial Expressions

It is assumed that every human face is having the same geometrical configuration. Feature extraction method uses the mean of the image as input to the fuzzy system. This improves the efficiency of the performance. Geometrical features of the face are fed to mamdani-type fuzzy system for quantification using trapezoidal membership functions. This system is capable of quantifying 3 basic forms of happy expressions that are No Happy, Bit Smiley, Loud Laugh as shown in Figure-4 & 5.

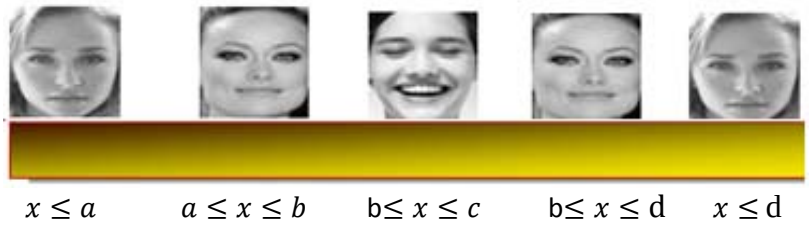


Figure-5 : Representation of Quantification of Happy Expression

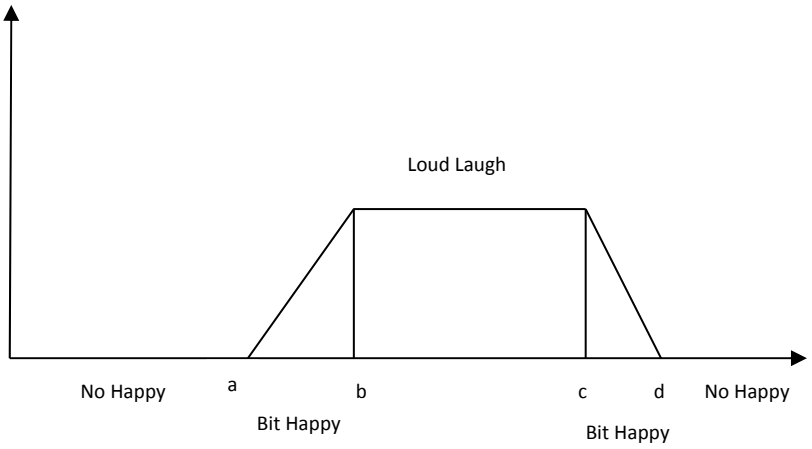


Figure-5: Representation of Trapezoidal Membership Function

The output of the fuzzy system is defuzzified to depict either the given face is in happy, or bit happy or loud laugh. This can be represented by the equation(1).

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

$$M(x,a,b,c,d) = \begin{cases} 0 & x \le a \quad No\ Happy \\ \dfrac{x-a}{x-b} & a \le x \le b \quad Bit\ Happy \\ 1 & b \le x \le c \quad Loud\ Laugh \\ \dfrac{x-d}{d-c} & c \le x \le d \quad Bit\ Happy \\ 0 & x \ge d \quad No\ Happy \end{cases} \quad \textbf{(1)}$$

Where a,b,c,d values are determined empirically for quantification of expression.
The experimental results of quantification of happy expressions are shown in the figure-6.

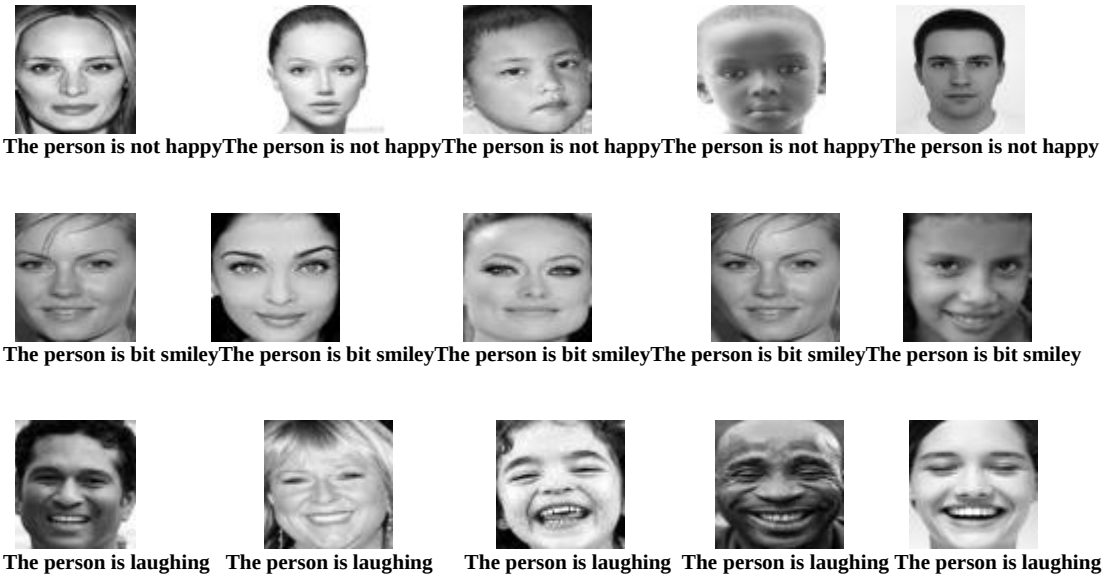


Figure-6 : Sample Experimental Results of happy expression

3. PROPOSED ALGORITHM

Proposed algorithm for expression quantification from the given image is as given below:

Input: Query face
Output: Quantification of happy expression into No happy, Bit happy or Loud laugh

Step 1: **To Train**: Input all n face images to the Fuzzy Inference System.
Step 2:Define the Trapezoidal membership function for different expressions and compute membership values for a,b,c,d as shown in Figure 5
Step 3: To Test: Compute membership value x for the query face using equation (1)
Step 4: Evaluate facial expression using equation (1).

4. EXPERIMENTAL RESULTS

In this research, there are 1000 gray-scale facial images used for experiment in which 300 images are used as training data and the remaining are used as test images. Each image size is normalized to 64×64 dimensions. The proposed Algorithm have shown good robustness and

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

reasonable accuracy for the test set with low complexity and is suitable for real time facial animation, image surveillance, mood analysis applications.

The success rate for quantification of happy expressions are 94.00%, 95.00% and 96.00% for no happy, bit happy and loud laugh respectively. Therefore, the overall success rate for test images is 95.00% as shown in the below graph. The average recognition time of each test image is 0.30 seconds on a Pentium Quad Core processor with 2 GB RAM.

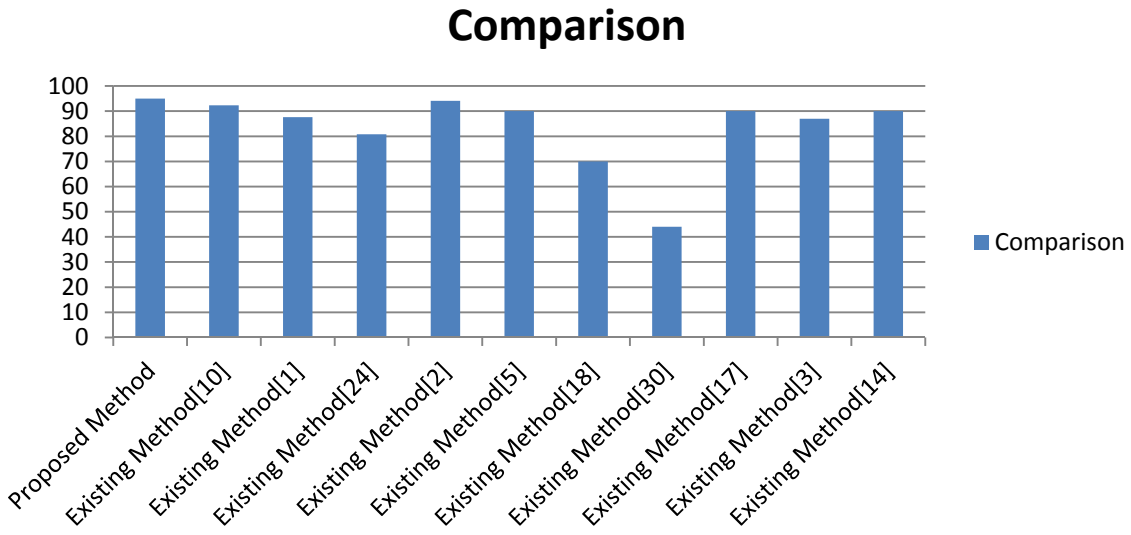


Figure-7 : Comparison of proposed method with other existing methods

As shown in the above graph, the proposed method has 95% of success rate when compared to all others existing methods. The below table shows the details of comparison of the proposed method and the other state of the art techniques and their respective success rates.

Sl.No	Authors	Method used	Success Rate(%)
1	Maedeh Rasoulzadeh [10]	FIS	**92.3**
2	Akanksha Chaturvedi and AlpikaTripathi [1]	Fuzzy Rule-based System	**87.6**
3	The Duy Bui et al [24]	Fuzzy Rule Based System	**80.8**
4	Aleix Martinez and Shichuan Du [2]	Emotion Model	**94.1**
5	Ashutosh Saxena, et al [5]	Geometric model	**90**
6	Jyoti Mahajan and Rohini Mahajan [18]	ANN	**70**
7	Prasad M and Ajit Danti [30]	SUSAN Edge Operator	**44**
8	Jiequan Liet.al [17]	Emotion recognition system	**90**
9	Anissa Bouzalmatet.al [3]	Neural Network and Fourier Gabor Filters	**87**
10	Hiroshi Kobayashi et.al [14]	Neural Network	**90**

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

11	**Proposed Method** **Dileep M R and Ajit Danti**	**FIS**	**95**

Table-1 : Comparison with state of the art methods

However, proposed method fails to detect the side-view faces, occluded faces and partial face images as shown in figure-8.This is due to missing facial features in the process of recognition of facial expressions.

5. CONCLUSIONS AND DISCUSSIONS

In this paper, a fast and efficient quantification of expression system is proposed to classify a facial image into different modes of happy expression groups using Mamdani-type fuzzy system. Mamdani-type fuzzy rule based system recognizes three levels of same happy expression namely No Happy, Bit Happy and Loud Laugh. The proposed method is better in terms of speed and accuracy with success rate of 95% and performance is comparable to other state of the art methods.

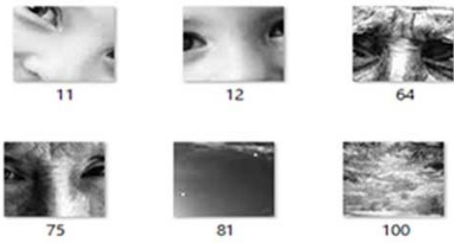


Figure-8 :sample results for mis-detection

In future studies, misclassifications are reduced by further improvement in the proposed system so that it becomes more pertinent to the design of a real-time video surveillance system.

ACKNOWLEDGEMENT

I would like to thank my guide Dr Ajit Danti, Director, Dept of Computer Applications, J N N College of Engg, for helping me to carry out this analysis. I appreciate the helpful comments and suggestions of Dr T Devi, Dept of Computer Applications, Bharathiar University and all the anonymous reviewers.

REFERENCES

[1] Akanksha Chaturvedi, Alpika Tripathi, Emotion Recognition using Fuzzy Rule-based System,International Journal of Computer Applications, Volume 93 – No.11, May 2014, ISSN 0975 – 8887

[2] Aleix Martinez, Shichuan Du, A Model of the Perception of Facial Expressions of Emotion by Humans: Research Overview and Perspectives, Journal of Machine Learning Research 13 (2012) 1589-1608

[3] AnissaBouzalmat, Naouar Beghini, Arsalane Zarghili, Jamal Kharroubi," Face detection and Recognition using base propagation Neural Network and Fourier Gabor Filters" SIPIJ Vol 2, No.3 Sep 2011.

[4] Aruna Chakraborty, Amit Konar, Uday Kumar Chakraborty, and Amita Chatterjee, Emotion Recognition From Facial Expressions and Its Control Using Fuzzy Logic, IEEE Transactions on Systems, Man, and Cybernetics—Part a: Systems and Humans, vol. 39, no. 4, July 2009.

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

[5] Ashutosh Saxena, Ankit Anand, Prof. Amitabha Mukerjee, Robust facial expression recognition using spatially localized geometric model, International Conference on Systemics, Cybernetics and Informatics, February 12–15, 2004
[6] B. K. Bose, "Expert systems, fuzzy logic, and neural network application in power electronics and motion control", Proceeding of the IEEE, vol.82,Aug. 1994.
[7] Dae-Jin Kim and Zeungnam Bien, Fuzzy Neural Networks (FNN) - based approach for Personalized Facial Expression Recognition with Novel Feature Selection Method, the IEEE Conference on Fuzzy Systems,2003.
[8] Dennis Gillette, Ping Zhang, Human-Computer Interaction And Management Information Systems: Applications, Advances in Management Information Systems Series Editor.
[9] S. Dongcheng, J. Jieqing, "The method of facial expression recognition based on DWT-PCA/LDA, " International congress on Image and Signal Processing (CISP), Volume: 4, pp. 1970 – 1974, 2010.
[10] The Duy Bui, Dirk Heylen, Mannes Poel, and Anton Nijholt, Generation of Facial Expressions from Emotion Using a Fuzzy Rule Based System, Springer-Verlag Berlin Heidelberg , M. Brooks, D. Corbett, and M. Stumptner (Eds.): AI 2001, LNAI 2256, pp. 83–94, 2001.
[11]Esau N., Wetzel, E., Kleinjohann, L. & Kleinjohann, B. (2007). Real-Time Facial Expression Recognition Using a Fuzzy Emotion Model. IEEE International Fuzzy Systems Conference, London, England, 1–6.
[12] Francisco Herrera, Luis Magdalena, Genetic Fuzzy Systems: A Tutorial. Tatra Mt. Math.Publ, (Slovakia),(1997).
[13] P. S. Hiremath and Ajit Danti, A fuzzy-rule based method for human face detection, Proceedings of NVGIP-05, 2nd -3rd March 2005, Dept. of CS&E, JNNCE, Shimoga
[14]Hiroshi Kobayashi and Fuimio Haro "Analysis of Neural Network Recognition characteristics at Basic Facial Expression" IEEE International Workshop on Robot and Human Communication 0- 7803-2002-6/94, 1994 IEEE.
[15]A. Jamshidnezhad, "A Learning Fuzzy Model for Emotion Recognition, "European Journal of Scientific Research ISSN 1450-216X Vol.57 No.2, pp.206-211, 2011.
[16] B. Jaychandra, simulation studies on "Speed Sensorless Operation of Vector Controlled Induction Motor Drives Using Neural Networks", Ph.D. Thesis, IIT, Madras, Chennai.
[17] Jiequan Li, Oussalah M, "Automatic Face emotion recognition system" Cybernet Intelligent Systems (CIS) 2010 IEEE 9th International Conference Vol 1,Pg 1-6.
[18] Jyoti Mahajan and Rohini Mahajan, FCA: A Proposed Method for an Automatic Facial Expression Recognition System using ANN, International Journal of Computer Applications (0975 – 8887) Volume 84 – No 4, December 2013.
[19]Khanum, A., Mufti, M. & Javed, M.Y. (2009).Fuzzy case-based reasoning for facial expression recognition. Journal of Fuzzy Sets and Systems, 160(2), 231–250.
[20] G. Klir and B. Yuan, Fuzzy sets and Fuzzy Logic – Theory and Applications, Prentice-Hall, 2010.
[21] G. J. Klir, and B. Yuan, "Fuzzy sets and fuzzy logic," Prentice hall of India, Pvt, Ltd, New Delhi, 2000.
[22] Kyoung- Man Lim, Young- ChulSim and Kyoung – Whan Oh, "A Face Recognition System Using Fuzzy Logic and Artificial neural network", Artificial Intelligence Research Lab, Dept. Of Computer Science, SoGang University, Korea.
[23] S. Y. Lee, Y. K. Ham and R. H. Park, "Recognition of human front faces using knowledge based feature extraction and neuro-fuzzy algorithm," Pattern Recognition, vol. 29(11), pp. 1863-1876, 1996.
[24] Maedeh Rasoulzadeh, Facial Expression recognition using Fuzzy Inference System, International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Issue 4, April 2012, ISSN: 2277-3754
[25] Math Works, Fuzzy Logic Toolbox User's Guide, Jan., 1998.
[26] I. Milki, N. Nagai, S. Nishigama, and T. Yamada, "Vector control of induction motor with fuzzy P-I controller", IEEE IAS Annu. Meet. Conf. Rec., pp. 342-346, 1991.
[27] Muid Mufti, Assia Khanum, Fuzzy Rule-Based Facial Expression Recognition, CIMCA-2006, Sydney, Australia.
[28] Mufti M., & Khanam, A. (2006). Fuzzy Rule Based Facial Expression Recognition, International conference on Computational Intelligence for Modeling, Control and Automation, Sydney Australia, 57.
[29] M. Nasir Uddin, Tawfik S. Radwan and M. Azizur Rahman, "Performances of Fuzzy-Logic-Based Indirect Vector Control for Induction Motor Drive", IEEE TRANSACTIONS ON INDUSTRY APPLICATIONS, VOL. 38, NO.5, SEPTEMBER/OCTOBER 2002, P1219.
[30] Prasad M and Ajit Danti, Classification of Human Facial Expression based on Mouth Feature using SUSAN Edge Operator, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 12, December 2014, ISSN: 2277 128X

[31] Ralescu, A., and Hartani, R., Some Issues in Fuzzy and Linguistic Modeling, IEEE Proc. of International Conference on Fuzzy Systems, 1995.
[32] D. H. Rao, S. S. Saraf, "Study of Defuzzification Methods of Fuzzy Logic Controller for Speed Control of a DC Motor", IEEE Transactions,2007, pp. 782-787.
[33] M.S.Ratliff, E. Patterson, "Emoticon Recognition Using Facial Expressions with Active Appearance Model, "HCI '08 Proceedings of the 3rd IASTED International Conference on Human Computer Interaction, pp.138-143, 2008.
[34] T. A. Runkler, Extended Defuzzification Methods and Their Properties, IEEE Transactions, 1996, pp. 694-700.
[35] M. Schmidt, M. Schels, and F. Schwenker, "A hidden markov model based approach for facial expression recognition in image sequences," ANNPR'10 Proceedings of the 4th IAPR TC3 conference on Artificial Neural Networks in Pattern Recognition, ISBN:3-642-12158-6 978-3-642-12158-6, 2010.
[36] N.Sebe, I. Cohen ; T.S. Huang ; and T. Gevers, ,"Human-computer interaction: a Bayesian network approach, "International Symposium on Signals, Circuits and Systems, ISSCS, 2005.
[37]Starostenko O., Contreras, R. & Alarcon-Aquino, V. (2010). Facial Feature Model for Emotion Recognition Using Fuzzy Reasoning. Advances in pattern Recognition. Lecture Notes in Computer Science, 6256, 11–21.
[38] TakKuen John Koo, Construction of Fuzzy Linguistic Model, Proceedings of the 35th Conference on Decision and Control, Kobe, Japan, 1996, pp.98-103.
[39] T. Takagi and M. Sugeno, "Fuzzy identification of a system and its application to modeling and control", IEEE Trans. Syst. Man and Cybern., vol.15, pp.116-132, Jan./Feb. 1985.
[40] L. H. Tsoukalas and R. E. Uhrig, "Fuzzy and Neural Approches in Engineering", John Wiley, NY, 1997.
[41] Ushida, H., Takagi, T., and Yamaguchi, T., Recognition of Facial Expressions Using Conceptual Fuzzy Sets, Proc. of the 2nd IEEE International Conference on Fuzzy Systems, pp. 594-599, 1993.
[42] M. Usman Akram, Irfan Zafar, Wasim Siddique Khan and Zohaib Mushtaq "Facial Expression Recognition Based On Fuzzy Logic" International Conference on Computer Vision Theory and Applications, P.383-388, 2008.
[43] V. P. Vishwakarma, S. Pandey, and M. N. Gupta "Fuzzy based Pixel wise Information Extraction for Face Recognition," IACSIT International Journal of Engineering and Technology Vol. 2, No.1, ISSN: 1793-8236, February, 2010.
[44] T. Xiang, M.K.H. Leung, and S.Y. Cho, "Expression recognition using fuzzy patio-temporal modeling, "Pattern Recognition, vol. 41, pp. 204-216, 2008.
[45] L.A. Zadeh, "Fuzzy sets", Information and Control, Vol. 8, pp. 338- 353, 1965.

AUTHOR'S PROFILE

Mr. Dileep M R is currently working as Lecturer in the Dept. of Computer Science, Alva's College, a unit of Alva's Education Foundation, Moodbidri, Karnataka, India. He has 4 years of experience in various capacities such as Teaching, Administration and Research. Research interest includes Image Processing, Neural Networks, Fuzzy Inference Systems, Database Applications, Software Engineering, Data Mining and so on. He has presented number of research papers in National and International Conferences and Published number of research papers in the reputed International Journals including SCI, SCOPUS indexed journals and IEEE Xplore digital library which are freely available online. He has Completed Master of Computer Applications (MCA) from Visvesvaraya Technological University, Belgaum, Karnataka, in the year 2013.

Dr. Ajit Danti is currently working as Director and Professor in the Dept. of Computer Applications, Jawaharlal Nehru National College of Engineering, Shimoga, Karnataka, India. He has 26 years of experience in various capacities such as Teaching, Administration and Research. Research interests include Image Processing, Pattern Recognition and Computer Vision. He has published more than 75 research papers in the International Journals and Conferences. He has authored 3 books published by Advance Robotics International, Austria(AU) and Lambert Academic Publishing, German which are freely available online. He has more than 250 citation index in the google scholar and several papers are indexed in DBLP, SCI, Scopus, IEEE Explore etc. He has Completed Ph.D degree from Gulbarga University in the year 2006. He has Completed Masters Degree in Computer Management from Shivaji University, Maharashtra in the year 1991 and M.Tech from KSOU, Mysore in the year 2011 and Bachelor of Engineering from Bangalore University in the year 1988.

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

Privacy Preserving Distributed Association Rule Mining Algorithm for Vertically Partitioned Data

Vadlana Baby
Associate Professor
Department of Computer Science
and Engg., VNR Vignana Jyothi Institute
of Engg. and Technology, Hyderabad, India
Email: baby_v@vnrvjiet.in

Dr. N. Subhash Chandra
Principal and Professor
Department of Computer Science
and Engg. Holy Mary Institute of Technology,
Hyderabad, India
Email: subhashchandra_n@yahoo.co.in

***Abstract*—Data mining over diverse data sources is useful means for discovering valuable patterns, associations, trends, and dependencies in data. Many variants of this problem are existing, depending on how the data is distributed, what type of data mining we wish to do, how to achieve privacy of data and what restrictions are placed on sharing of information. A transactional database owner, lacking in the expertise or computational sources can outsource its mining tasks to a third party service provider or server. However, both the itemsets along with the association rules of the outsourced database are considered private property of the database owner.**

In this paper, we consider a scenario where multiple data sources are willing to share their data with trusted third party called combiner who runs data mining algorithms over the union of their data as long as each data source is guaranteed that its information that does not pertain to another data source will not be revealed. The proposed algorithm is characterized with (1) secret sharing based secure key transfer for distributed transactional databases with its lightweight encryption is used for preserving the privacy. (2) and rough set based mechanism for association rules extraction for an efficient and mining task. Performance analysis and experimental results are provided for demonstrating the effectiveness of the proposed algorithm.

keywords: Security, Association rule mining, Rough sets, Secret key sharing, Distributed computation, Transactional-itemsets, Machine learning.

I. Introduction

Data mining offers the service to extract the meaningful, relevant and useful information in form of knowledge. Today data, which is becoming more and more sensitive and private, contains the sensitive information, to process such data securely become a challenging task. Other most studied problem is to discover frequent itemsets in various transactional databases and mining of association rules. Association rules are utilized in diverse areas e.g. market analysis, weather conditions prediction, risk management, communication networks etc. [5,6,7,8,9].

In distributed transactional database scenario, the databases are present in multiple locations and to mine the results globally, privacy preserving data mining is performed. In this process, the sensitive and private data at each site is preserved through different methods. For the purpose of association rule mining, these parties present at distributed locations need to merge their corresponding databases.

Some significant approaches like secure multiparty computation (SMC) along with other cryptographic techniques [1,2,3] can be employed to acheive privacy preserving database mining. Some threshold cryptography techniques e.g. Shamir's secret sharing [4] are also utilized for this purpose.

The general system model of outsourced data mining on joint database is represented as figure below:-

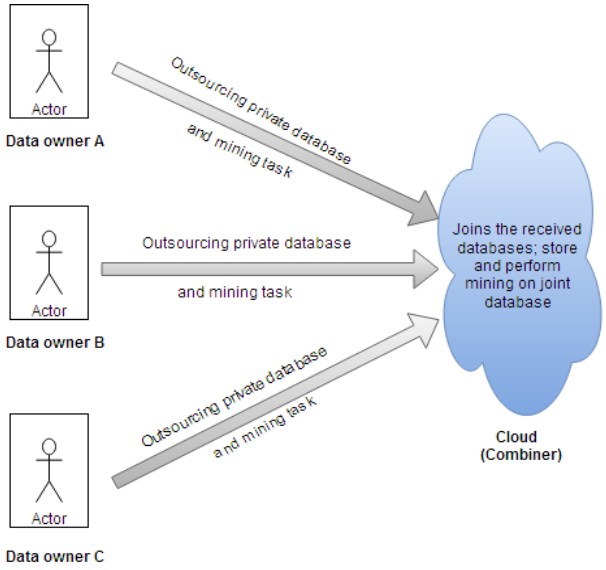


Fig.1

A. Related Work

The domain of privacy preserving data mining has attracted lot of people from various research communities. In this model, private and sensitive data which is stored in distributed sites is collected at a common location for the purpose of association rules mining as well as frequent itemsets mining. In past years, good amount of research has been done to accept this challenge. This section summarizes some of the significant research work which has been proposed in past several years.

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

Fosca Giannotti et. al. [10] has presented the privacy-preserving mining of association rules from the outsourced transaction database. If a customer has a lack of computational resources, then the customer can outsource his mining needs to the third party (server). In this paper, they explained about outsourcing problem of association rule mining task and have proposed the attack model and created the privacy-preserving scheme for outsourcing which is sufficiently effective in nature. Mistery et. al. [11] has employed the heuristic approach in the data distortion technique. We know that association rule mining is one of the important and powerful models among the data mining. The main problem occurs in this, is the security issue while outsourcing the database to the servers. The MDSRRC algorithmic procedure which they have used in this paper is hiding the sensitive association rules with some database modifications.

Li Lichun et. al. [12] mainly focussed on the privacy-preserving mining on vertically partitioned databases. To confirm the data privacy of customer's sensitive data, they have designed two schemes - (i) homomorphic encryption scheme (ii) secure comparison scheme. Later they proposed the cloud aided frequent itemset mining solution, which is used for building association rule mining solution. Compared to all existing solutions, the present solution will leak less information. Yaoan Jin et. al. [13] improved the performance of association rule mining on the distributed databases. They have combined the Apriori algorithm and FP-tree algorithm in present situation. For the purpose of implementation, they used merging FP-tree method. They also utilized homomorphic encryption scheme for comparatively better security purpose.

Kantarcioglu et. al. [14] proposed secure mining of association rules on horizontally partitioned data. They have shown that distributed association rule mining can be performed more efficiently under the preferred security assumptions as shown in secure association rule mining. Vaidya et. al. [15] described privacy preserving association rule mining in vertically partitioned data. With minimum support levels, they have presented the two-party algorithm for discovering frequent itemset. For preserving privacy of individual values, they have given an efficient protocol for computing the scalar product.

Fukasawa et. al. [16] has given a distributed privacy-preserving data mining algorithm. With the proposed algorithm data mining can be performed by a group of sites with the privacy of every site remains protected. In this algorithm, it is characterized by its capability to preserve the privacy without any coordinator site, and specifically its ability to resist the collusion. Performance analysis along with experimental results are given for demonstrating the effectiveness of the proposed algorithm. Cheung et. al. [17] proposed distributed association rule mining algorithm, FDM (fast distributed mining of association rules), which produces a tiny number of candidate sets and extensively reduces the number of messages. It shows that FDM has a superior performance over the direct application of a typical sequential algorithm.

Xinjing et. al. [18] examines the issue of privacy preserving distributed association rule mining in vertically partitioned data among multiple parties. Here, they presented a collusion-resistant algorithm of distributed association rule mining based on the Shamirs secret sharing technique, it prevents effectively collusive behaviors and without compromising their data privacy, it handles the computations across the parties. Moreover, the paper analysed about the security, efficiency, and correctness of the proposed algorithm. Tassa Tamir et. al. [19] have proposed a protocol "secure mining of association rules in horizontally distributed databases" that improve significantly upon the current leading protocol regarding privacy and efficiency. It is based on the Fast Distributed Mining algorithm. The main components of this protocol are two novel secure multiparty algorithms those are, one that estimates the union of private subsets that each of the interacting players holds, and other is testing the inclusion of an element held by one player in a subset handled by another.

B. Motivation and Contribution

With the advent of several security threats over internet and other communication scenarios, customer's data is more sensitive and confidential and to process this data requires the preservation of data privacy. Most of the times, due to lack in the expertise or computational sources from the owner side, it need to outsource its mining tasks to a third party service provider or server, then in this case privacy preserving plays a crucial role.

Main contributions of this paper includes -

- The existing state-of-the-art methodologies and various cryptographic techniques available for secure mining task, which are explored in the literature, are presented here.
- We have proposed our solution, which concerns the association rules mining in a secure manner for distributed transactional databases, located at different sites. Our method assimilates secret sharing based secure outsourcing of transactional databases, which consists of different itemsets and rough set based mechanism for association rules extraction for an efficient and secure mining task.

C. Organization of the paper

Structure of remaining of the paper is as - In section 2, some required preliminaries are discussed. The available privacy preserving data mining techniques are sub-categorized in section 3. In section 4, we have presented our proposed algorithm. Section 5 analyses the security and efficiency analysis of the algorithm. Experimental results are presented in section 6. Finally conclusions along with future research plan are given in section 7.

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

II. Preliminaries

Some of the significant terms and preliminaries required are summarized as below sub-sections:-

A. Data Mining Techniques

Data mining is the mechanism to fetch useful, relevant information by processing of raw data available which may be huge in size, may contains unstructured and irrelevant features set. With the evolution of big data domain, this area in more in demand. The process of data collection, it's analysis and further mine the significant information are the key process in all data mining techniques. Some of the core techniques for data mining are summarized as below-

1) Association-: In this technique, based on patterns and frequent itemsets discovery, the correlation among two or more databases is established. For an example - suppose a person is buying itemsets x and y along with is most of the time he is also buying some third item z, then the association rule set will be:-

$$x \bigcap y \rightarrow z$$

2) Classification-: Classification is the most commonly used technique in data mining. For example one can classify different types of fruits into various catagories such that each decision catagory consists of fruits having same attributes. It involves training phase and later based on knowledge gained at training phase, testing phase is performed.

3) Clustering-: Catagorize the different available objects or itemsets and form clusters based on some particular attribute, is considered as the clustering process. No training phase is involved in the process of clustering.

4) Prediction-: This technique is used to predict the future itemsets association, frequency of various itemsets, confidence and support factors. These predictions are performed utilizing some pre-information available in the form of stochastic, probabilistic and statistical knowledge. This involves association and patterns mining.

5) Decision trees-: As the part of selection criteria or selecting some part in the overall structure, decision trees may be employed. [21] Mostly while predictive systems designing and modeling, decision trees are used. These predictions may be based on past experiences and they help to define the structure of the decision tree.

B. Data Privacy

In today's scenario, when the customer's data is more sensitive and confidential, to process this data requires the preservation of data privacy. Suppose transactional database distributed in different locations are $DB_1, DB_2, DB_3, ..., DB_n$ and task is to mine these databases for rules extraction purpose, analyzation of frequency of itemsets etc. So, these distributed sites need to transfer their private sensitive data in a secure manner through communication channel to any combiner or trusted third party. Then combiner will merge these distributed databases, $DB = DB_1 \cup DB_2 \cup DB_3 \cup .... \cup DB_n$. So, maintaining the data privacy of the distributed transactional databases is considered as a challenging task.

C. Association Rule Mining

Association rule minning is a significant task which comes under the domain of data mining and data discovery. The data mining field and fundamental concepts of association rule mining was first initiated by R. Agrawal. Association rules are nothing but the IF-THEN relational concept among different itemsets. Suppose a person is buying itemsets X, where X is 'laptop' and along with is most of the time he is also buying Y, where Y is an 'antivirus', then the association rule set will be: $X \rightarrow Y$ Association rules have it's applications in diverse fields e.g. image processing, prediction modeling, market business analysis, banking etc.

D. Cryptographic Techniques for Secure Computation

Some significant cryptographic techniques, utilized today for the purpose of association rule mining are as follows:-

1) Secure Multi-Party Computation-: Good amount of work has been done on secure multi-party computation to employ it as an efficient and secure cryptographic technique in various applications [23].

- Consider distributed parties consisting of their transactional databases, who cann't believe each other, nor the channels through which they wish to communicate. Still, suppose the sites need to correctly compute some common function of their local transactional databases inputs, while keeping their sensitive data as private as possible. That, exactly, is the problem of SMPC.
- For example- Consider a case, where multiple users jointly work on a project and utilize the services of the same cloud provider. All of them want to simultaneously evaluate their equations without revealing individual equations to one another. In such scenarios the SMPC solutions can be highly useful to provide privacy.

2) Homomorphic Encryption-: Homomorphic Encryption is the most widely used cryptographic technique today.

- Homomorphic Encryption is an special form of encryption that allows computations to be carried out on ciphertext, thus generating an encrypted result which, when decrypted, matches the result of operations performed on the plaintext.
- Homomorphic Encryption is of two types - Partially homomorphic encryption and Fully homomorphic encryption. Fully Homomorphic encryption gains the advantage, as both the Additive and Multiplicative homomorphic properties are satisfies here. First FHE cryptosystem [22] proposed and developed by Craig Gentry in 2010 was a great breakthrough. So this concept is today utilized in data mining and association rule mining for privacy preserving purpose.

3) Secret Sharing-: The idea of secret sharing was first given by Adi Shamir [4]. In this method, a secret S is divided into n shares and distributed to the n participants, each participant is having one share. Later at the time of secret recovery, at least t shares are required to reconstruct (here t is considered as threshold).

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

So, the secret sharing phenomenon is applied in data discovery and mining to prevent collusive behaviors and tries to maintain the privacy of the user's sensitive data, which may have been distributed in various sites in form of transactional databases.

E. Rough Set Theory

In 1982, Rough Set Theory(RST) [20] was initially proposed by Zdzislaw Pawlak. RST acts such a mathematical model to describe the knowledge and to handle uncertain behaviour. An essential concept in RST is reduct selection. A reduct is the smallest possible set of attributes(features) which can describe an object with same effectiveness, accuracy and precision, as it was represented by the initial set of features. Exclusion of redundant attributes may support to identify the essential, effective and non-redundant classification perspectives.

A decision table (also called Data Table) is a system of the form $S = (U, C \cup \{d\})$ where, C: denotes set of conditional attributes; d: specifies decision attribute. In rough set theory(RST), two terms- Lower Approximation and the Upper Approximation, come under the Approximation terminology of sets.
Let $S = (U, R)$ is an approximation space, X be a concept in that space, then, the lower approximation is

$$R_{lower}X = \{x \in U | [x] \subseteq X\}$$

upper approximation is

$$R_{upper}X = \{x \in U | [x] \cap X \neq \emptyset\}$$

where, $[x]$ is called equivalence class which contains an element e.
Let, $S = (U, C \cup \{d\})$ considered as decision system, consisting the universe of objects, then a subset R of conditional attributes(C) is a reduct if, $POS_R(d) = POS_C(d)$. Computing the reduct is known to be an NP-hard problem, further to process the reduct for huge databases needs extraordinary computational processing environment. We will get Core, by performing intersection of resulting reducts.

$$CORE(C) = \cap RED(C)$$

The reduct is generated from the discernibility matrix. Objects are distinguishable if for some attributes, they are having different attribute values, then they are called discernible. And this property is known as discernibility. Fuzzy Sets involve the membership between set elements of similar class, while RS bothers the association among set of several elements in distinct classes.

III. Privacy Preserving Data Mining Techniques

Due to recent advancements in cloud computing, there evolves a lot of interest which takes place throughout the reserch community for outsourced service based data mining. A database owner or a company, due to limitations like - expertise and computational resources, may outsource their data mining task to any third party or trusted cloud. In below sub-sections we have summarized the horizontal and vertical partitioning techniques for privacy preserving data mining.

A. Horizontal Techniques

Horizontal partitioning of data is one technique for privacy preserving data mining. Yaoan Jin et. al. [13] given this type of technique which has improvement in the performance of association rule mining on the distributed databases. They have combined the Apriori algorithm and FP-tree algorithm in present situation. For the purpose of implementation, they used merging FP-tree method. They also utilized homomorphic encryption scheme for comparatively better security purpose. Kantarcioglu et. al. [14] proposed secure mining of association rules on horizontally partitioned data. They have shown that distributed association rule mining can be performed more efficiently under the preferred security assumptions as shown in secure association rule mining.

B. Vertical Techniques

Vertical partitioning of data is another technique for privacy preserving data mining. Li Lichun et. al. [12] mainly focussed on the privacy-preserving mining on vertically partitioned databases. To confirm the data privacy of customer's sensitive data, they have designed two schemes - (i) homomorphic encryption scheme (ii) secure comparison scheme. Later they proposed the cloud aided frequent itemset mining solution, which is used for building association rule mining solution. Compared to all existing solutions, the present solution will leak less information.

Vaidya et. al. [15] described privacy preserving association rule mining in vertically partitioned data. With minimum support levels, they have presented the two-party algorithm for discovering frequent itemset. For preserving privacy of individual values, they have given an efficient protocol for computing the scalar product.

IV. Proposed Privacy-Preserving distributed Association Rule Mining algorithm in vertically partitioned data

Central thought of our algorithm: In this section, we have presented our proposed algorithm for Privacy Preserving Distributed Association Rule Mining based on Rough Sets. Our model for entire mechanism consists of some main algorithmic phases or sub-modules, which are as below:-

- *Phase 1:* Secure key transfer for distributed transactional databases consisting of itemsets
- *Phase 2:* Privacy preservation of distributed databases
- *Phase 3:* Association rules mining using Rough Sets

The overall functionality of our proposed mechanism is shown as below process flow diagram:-

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

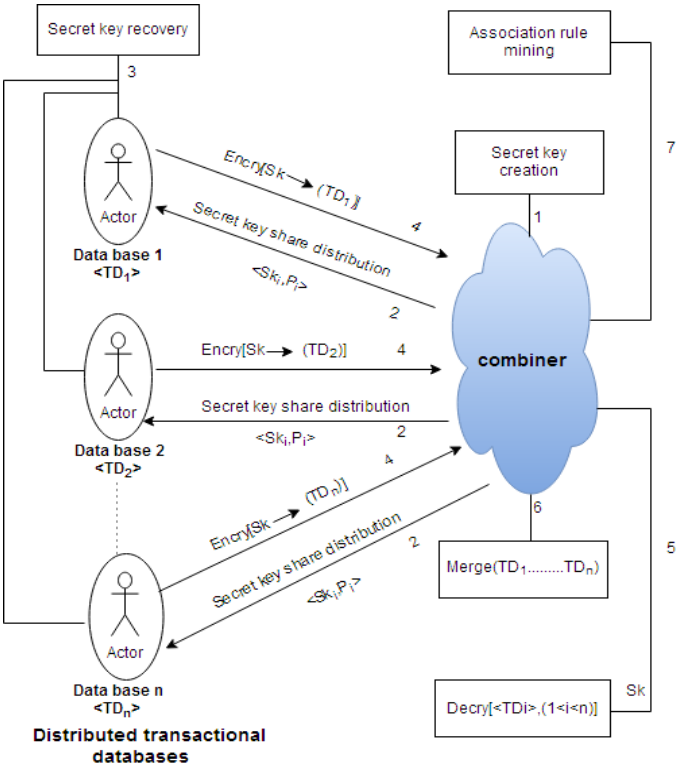


Fig.2

A. Algorithm

Our proposed algorithm for privacy preserving distributed Association Rule Mining consists of mainly three phases. The complete description of these phases is presented as below:-

Phase 1: Secure key transfer for distributed transactional databases consisting of itemsets-

Like distributed computing scenario, various transactional databases or itemsets collection may exist in different locations, but association rules or frequent itemsets need to be mined from these databases. The individual database owners should not reveal their private information to each other due to maintenance of privacy. In this case, some mechanism is needed to securely outsource different transactional databases into combiner.

Let us assume that the transactional database TD is vertically partitioned among N distributed locations.

$$TD = TD_1 \cup TD_2 \cup TD_3 \cup .... \cup TD_N;$$

where, $TD_i \cap TD_j = \emptyset$, $\forall i, j \in N$

The steps involved are presented below:-

Phase 2: Privacy preservation of distributed databases-

In this phase, the distributed data locations use encryption with any symmetric key encryption algorithm *(eg. Block cipher)* and utilizing reconstructed key from above phase 1. Further, the combiner performs decryption using the same symmetric key. Steps involved in this process are presented below:-

Phase 3: Association rules mining using Rough Sets-

After the execution of phase 1 and 2, combiner will perform

1: *Input:* Knowledge representation of distributed itemsets involving transactions along with trusted combiner.

2: *Output:* Securely outsourced different transactional databases into combiner.

3: Combiner, which is assumed as trusted third party in our protocol, generates the sequence of positive integers, $P_1 < P_2 < P_3 < .... < P_N < P_{N+1} < .... < P_{N+t-1}$, which must be coprime with each other; Along with the condition,

$$P_{N-t+2} \times .... \times P_N < P_1 \times P_2 \times .... \times P_t$$

where,

N: total number of distributed locations consists of transactional databases and itemsets

t: represents threshold, such that the secret key can be recovered by utilizing at least t key-shares

P_i: public information

4: Combiner sends the public information P_i to each authorized distributed transactional database location $< TD_i >$.

5: For specified sequence, combiner chooses the secret key S_k in the set $Z_{P_{N-t+2} \times ... \times P_N}, P_1 \times P_2 \times P_3 \times ... \times P_t}$.

6: Combiner generates key shares S_{k_i} for each TD_i in following manner:-

$$S_{k_i} = S_k \bmod P_i$$

where, i = 1, 2, 3,...,N.

7: Each S_{k_i} is sent to each authorized $< TD_i >$ through a secure channel.

8: Each TD_i will prove it's authentication with combiner and requests for $(t-1)$ distinct key shares.

9: After authentication, combiner sends another $(t-1)$ distinct key shares to each TD_i publicly.

10: Now with given t distinct key shares $(S_{k_{i_1}}, S_{k_{i_2}}, ..., S_{k_{i_t}})$, each TD_i can reconstruct the secret key S_k using standard CRT by solving below system of equations-

$C = S_{k_{i_1}} \bmod P_{i_1}$; $C = S_{k_{i_2}} \bmod P_{i_2}$; $C = S_{k_{i_3}} \bmod P_{i_3}$; .... ; $C = S_{k_{i_t}} \bmod P_{i_t}$

Unique solution C is given as -

$$\sum_{r=1}^{t} \frac{n}{P_{i_r}} \times y_{i_r} \times S_{i_r} mod n$$

where,

$n = P_{i_1} \times P_{i_2} \times ... \times P_{i_t}$ and $\frac{n}{P_{i_r}} \times y_{i_r} mod P_{i_r} = 1$.

---

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

1: Consider, M: message space (data); C: ciphertext space (encrypted data); K: keyspace.
2: **Enc (M, k):** It is a probabilistic algorithm that takes input as message and key then outputs ciphertext C.
3: **Dec (C, k):** It is a deterministic algorithm that takes input as ciphertext and key then outputs plaintext or message M.
4: Preferably, we can use any block cipher e.g. TEA(Tiny Encryption Algorithm) or FEAL(Fast data Encryption Algorithm) [24,25], which are designed for fast and efficient encryption and decryption with modest security at software level. Also these posses low memory requirement while processing.
5: Encrypted transactional databases $< TD_i >$ are send to combiner.
6: Finally, decrypted transactional database records are merged into a common knowledge representation system.

the association rules mining algorithm on merged common database. The steps involved in this process are presented below:-

V. Security and Efficiency analysis of the algorithm

Here, we have analyzed the security and complexity involved in our proposed protocol.

Proposition 1: Algorithm 1 is semantically secure against adversaries and communication channel attackers.

Proof: In our proposed algorithm 1 of Secure key transfer for distributed transactional databases, since one secret key share is presented with the individual distributed database owner, so with the publicly available $(t-1)$ key shares, an adversary can not recover the original secret key.

Since the numbers incorporated in the t-threshold range $Z_{P_{N-t+2}\times\ldots\times P_N}, P_1\times P_2\times P_3\times\ldots\times P_t}$ are having upper bound as $P_1 \times P_2 \times P_3 \times ... \times P_t$, which is the smallest product of any t moduli, and lower bounded by $P_{N-t+2} \times ... \times P_N$, and this is the largest product of any $(t-1)$ moduli. The secret key S_k, choosen in this range will always ensure that- (i) the secret can be recovered with any t or more than t key - shares. (ii) the secret cannot be obtained with fewer than t shares.

Proposition 2: Modern cryptology functions on the Kerckhoff's principle - attacker is aware of everything about algorithm, only the keys are assumed to be secret..

Justification: One must make sure that the exhaustive key-search should not be succeed. Two desired things for any symmetric key cipher are as -

- For given few plaintext/ciphertext pairs, search must be performed through all possible keys until the correct secret key found.
- Key must be large enough.

1: *Input:* Merged common knowledge representation system for transactions databases i.e. $TD = TD_1 \cup TD_2 \cup TD_3 \cup ....\cup TD_N$. This knowledge information can be considered in the form of information table (IT).
Information System - An information system in RST is represented as pair (U, A), where -
U: denotes non-empty finite set of database transactions or scenarios.
A: denotes non-empty finite set of attributes.
2: *Output:* Extracted association rules.
3: $begin\{procedure\}$
4: *i/p:* the set X of all attributes, partition $\{d\}^\star$ on U;
5: Compute reduct set partition $X^\star$;
6: P := X;
7: S := Null;
8: if $X^\star \leq \{d\}^\star$
9: then
10: begin
11: for each attr. x in X do
12: begin
13: $Q := P - \{x\}$
14: compute partition $Q^\star$;
15: if $Q^\star \leq \{d\}^\star$ then P := Q
16: end for loop
17: S := P
18: $end\{then\}$
19: $end\{procedure\}$
20: extracted Association Rules are then send back to the individual database owners or transactional database locations $< TD_i's >$.

Block size

In generic way, block size (S) must be reasonably large, $S > 64$ bits, for avoiding-

1) *Text dictionary attacks:* pairs of plaintext-ciphertext for fixed key.
2) *Matching ciphertext attacks:* uncover the patterns in plaintext.

Proposition 3: Privacy for distributed data-owner's private data is securely preserved during association rule mining in Algorithm 3.

Analytical justification: Each distributed transactional database $< TD_i >$ is sending the corresponding encrypted database vectors consisting itemsets to the combiner. Since the further mining of association rules is taking place in combiner, so individual $< TD_i >$ will be unaware of any others' private itemset vector's information.

After merging the itemset vectors, which were partitioned vertically, suppose the dimensions of knowledge information table is $(m \times n)$, where -

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

m: represents total different transactions and n: is the total number of itemset attribute vectors,
then the computational complexity involved in association rule mining algorithm is $O(m^2 \times n)$.

VI. Experimental Results

This section presents our experimental analysis which has been performed on real time databases. We have mainly used JAVA environment for our implementation purpose. Other software and hardware specifications are given below:-

A. System Specifications

Our system specifications are as below:-

- *Software Specifications* -
  OS - Ubuntu 16.04 LTS, 64 bit
  Java version - '1.8.0_111'
  NetBeans product version - NetBeans IDE 8.2

- *Hardware Specifications* -
  RAM size - 4 GB
  Processor - Intel core i3 4030U CPU @$1.90GHz \times 4$

B. Input and Setup

In our experiment, we have performed simulations with various test cases for different key sizes for encryption and decryption process and numbers of distributed transactional database locations. We have considered that the transactional databases are vertically partitioned. We have mainly used JAVA environment for our experiments.

Datasets description -
In our experiment, we have employed Reuters-21578 standard dataset as an input for association rule mining purpose. The documents in the Reuters-21578 collection [26] appeared on the Reuters newswire in 1987. The documents were assembled and indexed with categories by personnel from Reuters Ltd. (Sam Dobbins, Mike Topliss, Steve Weinstein) and Carnegie Group, Inc. (Peggy Andersen, Monica Cellio, Phil Hayes, Laura Knecht, Irene Nirenburg) in 1987. In 1990, the documents were made available by Reuters and CGI for research purposes.

C. Our Results

Our experiment results are summarized as below in which we have taken different parameters e.g. key sizes, no. of different distributed transactional database locations etc.

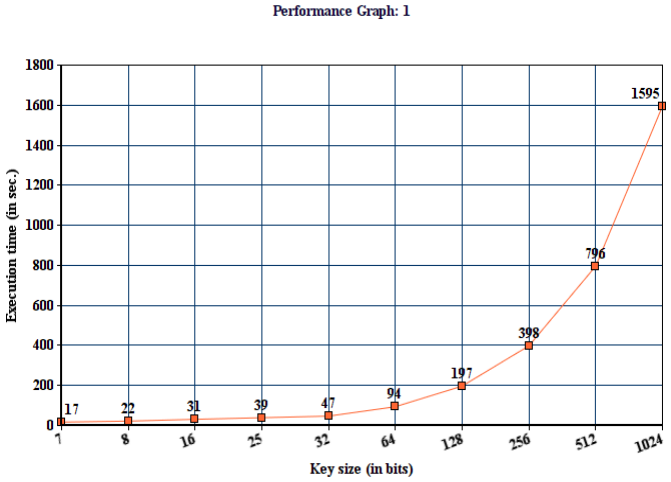
- **Experimental results by utilizing different key sizes -** The end results of execution performance for varying key sizes and key reconstruction is presented as table-1 below:-

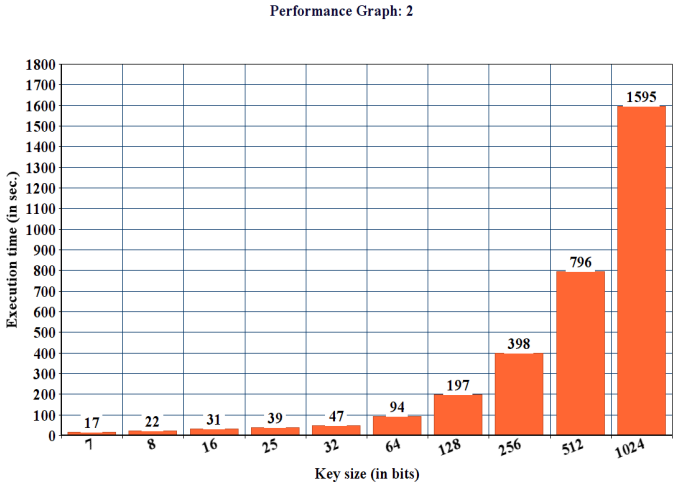
max width= 14 cm

Table for execution time utilizing varying key size	
Key size(bits)	Execution time(in sec.)
7	17
8	22
16	31
25	39
32	47
64	94
128	197
256	398
512	796
1024	1595

Table-1

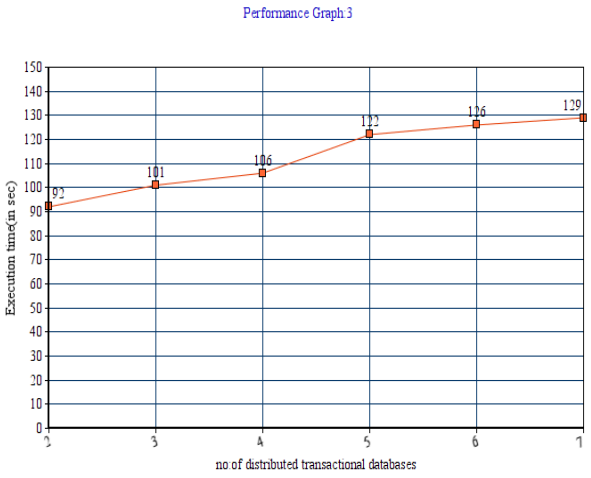
The performance graph with various key sizes taken in our experiments and their corresponding key reconstruction time is presented as graphs below:-

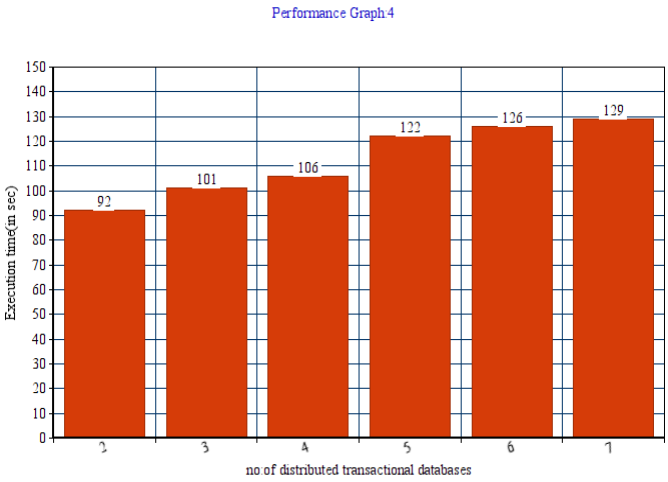




https://sites.google.com/site/ijcsis/
ISSN 1947-5500

- **Experimental results with various test cases by considering different parameters(no. of different distributed transactional database locations) -**
  We have performed experiments by taking varying number of distributed transactional database locations. Each transactional database consists of various itemsets. Reuters-21578 standard dataset is taken as input here. In each case, the corresponding association rules are mined and optimization performance is analyzed in below tables. The end results are shown as tables below:-
  In table-2 below, we have considered the no. of different distributed transactional database locations $n[TD] = 2$. Further the association rules are mined and execution time is given.
  We have first analyzed the execution time taken in the process of secret key reconstruction at client locations. We have performed experiments by taking varying key sizes (in bits). Simulation results are presented as table-1 above.
  Later, we have performed experiments for further association rules mining. For this, the standard Reuters-21578 dataset [26] is utilized. The association rules are mined at combiner side. We have performed experiments by taking varying number of distributed transactional database locations. Further the combiner will send the mined association rules to distributed clients locations.
- **The execution performance analysis is presented as graphs below-** The execution performance analysis graphs for varying number of distributed transactional databases locations are given as below:-





Above graphs shows the execution time for association rules mining process for varying cases of number of distributed transactional database locations.

VII. Conclusion

This section summarizes the conclusion in two aspects. First, we have presented our contribution in this paper and secondly we have discussed the limitations, which exist in current scenario and our future efforts to work towards this research direction.

A. Contribution

Most of the times, due to lack in the expertise or computational sources from the owner side, it need to outsource its mining tasks to a third party service provider or server, then in this case privacy preserving plays a crucial role. In this paper, we have proposed our solution, which concerns the association rules mining in a secure manner for distributed transactional databases, located at different sites. We have also presented the critical security and efficiency analysis for our proposed algorithmic mechanism along with the experimental results.

B. Limitation and future work

Our algorithm should be utilized with some care in practical scenarios where some assumptions taken in our model are, that all the distributed database owners should perform the computation honestly. In some rare practical scenarios, where prover need to prove the verifier, therefore, zero-knowledge protocol can be utilized to prevent the hostile behaviour. So, our future research focus is to discover some more efficient mechanism designed for adversarial computational model in this privacy preserving distributed association rule mining scenario.

References

1) A. C. Yao. ”Protocols for secure computations,” Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science,IEEE Press, New York, 1982.

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

2) P. Paillier. "Public-key cryptosystems based on composite degree residuosity classes," Advances in Cryptography-EUROCRYPT'99, pp.223-238, Prague, Czech Republic, 1999.
3) F. Wu, J. Q. Liu and Sh. Zhong. "An efficient protocol for private and accurate mining of support counts," Pattern Recognition Letters, vol.30(1), pp.80-86, 2009.
4) A. Shamir. "How to share a secret," Communications of the ACM, vol.22(11), pp.612-613, 1979.
5) T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets. "Using association rules for product assortment decisions: A case study," in Proc. SIGKDD, 1999, pp. 254-260.
6) S. E. Brossette, A. P. Sprague, J. M. Hardin, K. B. Waites, W. T. Jones, and S. A. Moser. "Association rules and data mining in hospital infection control and public health surveillance," J. Amer. Med. Inform. Assoc., vol. 5, no. 4, pp. 373-381, 1998.
7) B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. "Effective personalization based on association rule discovery from Web usage data," in Proc. WIDM, 2001, pp. 9-15.
8) C. Creighton and S. Hanash. "Mining gene expression databases for association rules," Bioinformatics, vol. 19, no. 1, pp. 79-86, 2003.
9) X. Yin and J. Han. "CPAR: Classification based on predictive association rules," in Proc. SIAM SDM, 2003, pp. 1-5.
10) Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang. "Privacy-preserving mining of association rules from outsourced transaction databases." IEEE Systems Journal 7.3 (2013): 385-395.
11) Mistry, Bhoomika R., and Amish Desai. "Privacy preserving heuristic approach for association rule mining in distributed database." International Conference on Innovations in Information Embedded and Communication Systems-(ICIIECS-2015) IEEE.
12) Li Lichun, Rongxing Lu. "Privacy-preserving-outsourced association rule mining on vertically partitioned databases." IEEE Transactions on Information Forensics and Security 11.8 (2016), 1847-1861.
13) Jin Yaoan, Chunhua Su, Na Ruan, and Weijia Jia. "Privacy-Preserving Mining of Association Rules for Horizontally Distributed Databases Based on FP-Tree." International Conference on Information Security Practice and Experience. Springer International Publishing, pp. 300-314, 2016.
14) Kantarcioglu, Murat, and Chris Clifton. "Privacy-preserving distributed mining of association rules on horizontally partitioned data." IEEE transactions on knowledge and data engineering 16.9 (2004): 1026-1037.
15) Vaidya, Jaideep, and Chris Clifton. "Privacy preserving association rule mining in vertically partitioned data." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.
16) Fukasawa Takuya, Jiahong Wang, Toyoo Takata, and Masatoshi Miyazaki. "An effective distributed privacy-preserving data mining algorithm." International Conference on Intelligent Data Engineering and Automated Learning, Springer, pp. 320-325, 2004.
17) David W. Cheung, Jiawei Han, Vincent T Ng, Ada W. Fu, Yongjian Fu. "A fast distributed algorithm for mining association rules.", Fourth International Conference on Parallel and Distributed Information Systems, IEEE, (1996).
18) Xinjing Ge, Li Yan, Jianming Zhu, Wenjie Shi. "Privacy-preserving distributed association rule mining based on the secret sharing technique." 2nd International Conference on Software Engineering and Data Mining (SEDM), (2010), IEEE.
19) Tassa Tamir. "Secure mining of association rules in horizontally distributed databases." IEEE Transactions on Knowledge and Data Engineering 26.4 (2014): 970-983.
20) Pawlak, Z. (1982). Rough sets. International Journal of Computer and Information Sciences, 11(5), 341-356.
21) S.R. Safavian, D. Landgrebe. "A survey of decision tree classifier methodology", IEEE Transactions on Systems, Man, and Cybernetics, Vol: 21, Issue: 3, p.660-674, 06 August 2002.
22) Craig G. "Fully homomorphic encryption using ideal lattices", STOC. Vol. 9. 2009.
23) O. Goldreich, "Secure multi-party computation", Sept. 1998, [Online]. Available: http://www.wisdom.weizmann.ac.il/ - oded/pp.html
24) D. Wheeler and R. Needham "TEA, a Tiny Encryption Algorithm" (1994), Cambridge University Press, $http : //www.ftp.cl.cam.ac.uk/ftp/papers/djw - rmn/djw - rmn - tea.html$.
25) Akihiro Shimizu, Shoji Miyaguchi. "FEAL - Fast Data Encipherment Algorithm", Systems and Computers in Japan, Wiley Periodicals, Volume 19, Issue 7, (1988), Pages 20-34.
26) http://archive.ics.uci.edu/ml/databases/reuters21578/

Using BIG DATA implementations onto Software Defined Networking

Mohammed Najm
DEPT. of comput.engineering
Technology University
Baghdad-Iraq
mustafamna@Yahoo

abeer Tariq
DEPT.of comp.science
Technology University
Baghdad-Iraq
abeer28003@Yahoo

Mohamed Qasim
DEPT.of comp.science Technology
Technology University
Baghdad-Iraq
mohammad.qassim2002@gmail

Abstract

An elastic , effective, activety or intelligent ,graceful networking architecture layout be desired to make processing massive data. next to that ,existent network architectures be considerably incapable for cleatting the huge data. massive data thrusts network exchequers into border it consequence with in network overcrowding ,needy achievement, then permicious employer exprtises. this offered the current state-of-the-art research affronts ,potential solutions into huge data networking notion. More specifically, present the state of networking problems into massive data connected intrequirements,capacity,running , data manipulating also will introduce the architectures of MapReduce , Hadoop paradigm within research requirements, fabric networks and software defined networks which utilized into making today's idly growing digital world and compare and contrast into identify relevant drawbacks and solutions.

Keywords:Big Data; MapReduce; Hadoop; SDN;

1. INTRODUCTION

at present times, implementations be fermantly growing a rate such data be mountng within the world.Legal Studies be displayed for 2020 the world be growing 50 times a sizes of data had into 2011, whom currently 1.8 zetta bytes or 1.8 trillion gigabytes of data [18]. A natural cause into a acute mounting into data making a storage along years merely decline into fees from a store.

IT industry making a fees from store become low then implementations be able from provision data exponential rates. That will get an affront from making present network buildings in what way running then execute these massive data to be used to utilizing information [1, 5 - 6, 16]. numerous huge data implementations action or requirements work in a good-time. implementations required into make, storage then operates a big quantities from information whom decrease big accord from quantity ,need for a network. When looking at data from a networking perspective, multi-level regions be required to be reconnoiter which contain network formatting , parallel structures then huge data progressing algorithms, the data recovery then confidentially problemss [13]. The topic massive data as yet afresh rousing parts from research among the Information Technology staff , being demand on extremely regard into the years onto derive applying onto network theory,huge data being described as any aggregate or data-in-motion and many of its applications have real-time requires into making effective especially in an education/research environment [8]. the industry, common implementations utilize to massive data contain for analyzing the data to come up with efficient conclusions in implementation domains like astrophysics, particle physics (e.g., CERN research), biological science (e.g., healthcare), geological science, social networking (e.g., Facebook), trend prediction (e.g., google flu), optimizing route delivery(e.g., UPS package delivery, fuel tracking) [13].

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

A typical organization has a limited network infrastructure and resources which able be catching that sizes into traffic flows whom cause standard services (e.g., Email, Web browsing, video streaming) being strained. these making come down network execution affecting bandwidth and exposing hardware requirements from tools like a firewall progressing be making over whelmed [13]. This paper presents a comprehensive survey on the network theory of big data. This work startsby introducing the concept of big data and networking theory by giving some background information on the state of networking issues related to capacity, management and data processing in Section II. In Section III briefly presents the architecture of Big Data which includes the MapReduce paradigm and Hadoop distributed architecture, fabric network infrastructure and software defined networks (SDN). We present architectures, designs techniques using theorized solutions to handle today's idly growing digital world and compareand contrast them to identify relevant problems and solutions. Finally, we discuss, evaluate, and state conclusions on the analysis of the defined Big Data networking concepts in Section IV.

2. BIG DATA NEEDS

this introduces making network, a different data types, data flow needs and implementations from massive Data.

2.1. meet the requirements for the massive Data Network

Huge data requires a good execution request from a networks buildings implementations that network needs being to extra efficient, flexible then have some form into implementations knowledgs. Huge data network architecture making be deals with a distributed architecture so as to planning , making accessibility from distributed resources when together making in parallel on a single job [7]. Massive data implementations needs starts a progressing big sizes from the information makes a bigger as data propagated towards network. Table 1 shows the comparison of traditional and massive data kinds 11].

Table 1: Traditional vs. Big data types

Feature	Traditional Data	Big Data
Data type	Structured	Partial structure or unstructured
Data Volume	Giga/Terabytes	Peta/Exabytes
Data Relationship	Simple and Known	Complex and Unknown
Data Model	Used fixed schema	No schema
Network Model	Centralized/Distributed	Only Distributed

2.2.Needs from different data kind

Features into massie data kinds being so diverse of the classic data patterns. Huge data implementations which needs a good -time realization needs the big data sets being defective of a little growing volums onto progressing . The system can not able to executed onto perfect online transaction processing , while utilizing the common SQL analysis models [11]. Massive data requires a further agile flat horizontally scaled data base environment being be able to catching differents unstructured datakinds within complicated , unknown data relationships amongst each other in distributed network architecture. a little older data bases modules making efforted to making a mutate onto building a NewSQL, unless generality have abandoned SQL for new models such as NoSQL [8]. These new databases making difficult or

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

hard to use huge data's analytical devices ,responsibilities. Figure 1 shows the throughput comparison of RDMS and No SQL data base over the volume of data.

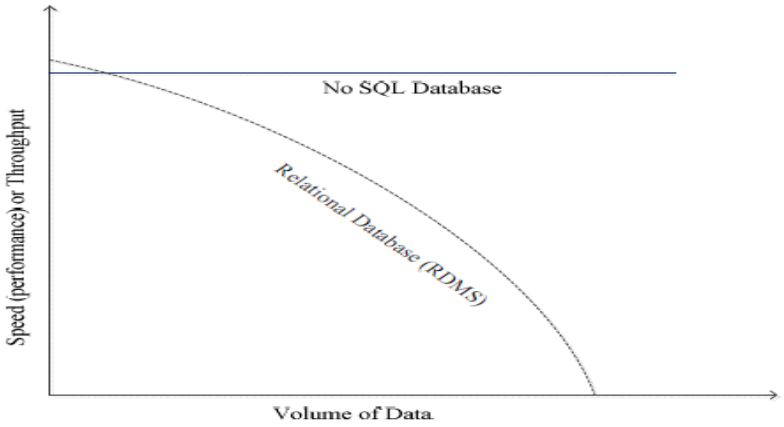


Figure 1: Comparison of scalability of RDMS and NoSQL

2.3. Network Flow needs

network infrastructure a long the years has been adapted in a 3-tier architecture contains computing, storage, networking. the trends of network traffic is used to as "north-south" which means it usually starts at the end user then goes down the integrated stack to the server than to the database. With the needed of massive data in networks, the traffic pattern requires to change. Massive data structure depending on a horizontal scale ,utilizes a distributed approach among the nodes, the traffic demands between the server and storage nodes are much higher than the typical data flow between servers and end users. Data is no longer just being created by external sources, generated from a number of devices and applications. This type of traffic flow is referred to as machine-to-machine/virtual machine network traffic or "east-west" and it needs to be supported when creating a high-execution scalable huge data network.

3. BIG DATA ARCHITECTURE

we present architecture and research challenges of MapReduce and Hadoop [2,3,4]that are used to huge data manipulating. introducint a fabric network technology [19] utilized into massive data infrastructure over within characteristics ,requirements. briefly introduced software defined networks (SDN), that more popular onto huge data execution.

3.1. Hadoop and MapReduce

MapReduce a core component of the Hadoop Apache software framework , a kind from a programming module whicht used to implemented into different onto languages (e.g., Java, C++) which utilizing to progressing massive data [17]. these kinde from the software device be applications into smallerfragments or blocks which are then sent out to nodes in a cluster or map. It uses a map function that will able to filter, sort, and distribute jobs to various nodes and also uses a reduce function to collect the results from those jobs so they can resolved into a single value to be used for efficientanalysis (Figure 2) The MapReduce consists of a job tracker, task trackers, and sometimes a job history server [23].The job

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

tracker is used as the master node that is in charge of managing resources and jobs. The task tracker is used to be deployed to each node in order to run the map and help with some of thecluster task load. The job history server is used to track finished jobs and can be deployed as an independent function. MapReduce operates in parallel across vast cluster sizes, while jobs can bedivided across many different servers [17]. MapReduce has fault-tolerance where each nodesends status updates to the master node, who can re-assign jobs to functioning nodes in cases of node failure.

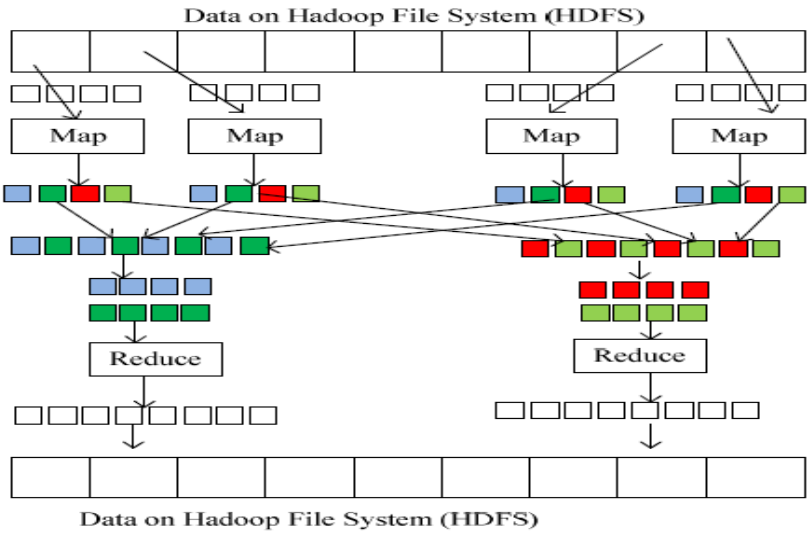


Figure 3: Hadoop that reduces data and processes in parallel

On the other hand, Hadoop is a management software framework that plays an important role inbig data analytics. Hadoop is capable of cataloging, managing, distributing, and queryingunstructured large data sets rapidly across many nodes within a distributed network environment.It uses a Hadoop distributed file storage system (HDFS) for storage that divides data into blockswhich is distributed and stored on multiple nodes. In order to process the data, Hadoop usesMapReduce to break down data for the nodes to process and sort in parallel (Figure 3). The mapprocedure is responsible for filtering and sorting and the reduce procedure focuses on summarytasks. It supports high speed transfer rates and is capable of resilient uninterrupted operation insituations when there is node failure [20]. The infrastructure divides up the nodes into groups orracks. Hadoop is an excellent framework for applications using large search engines such asGoogle

Together Hadoop and MapReduce provide the current popular choice for implementation of bigdata infrastructure [10]. A typical Hadoop network structure consists of a slave (data node, tasktracker), master (name node, job tracker) and a client [9]. The client is basically the user interfaceor query engine. The data nodes are used as storage for the data that contain smaller databasessystems and are horizontally distributed across the network. The task tracker is used to process the broken down fragments of the task that has been distributed to a node. The name node maintains a location index of all the other nodes found in the network, so it knows where the specific data is located in which data node. The job tracker is the software job tracking mechanism that is used to transfer and aggregate request search queries (tasks) through to the tasktracker nodes so the end user can perform information analysis on the result.

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

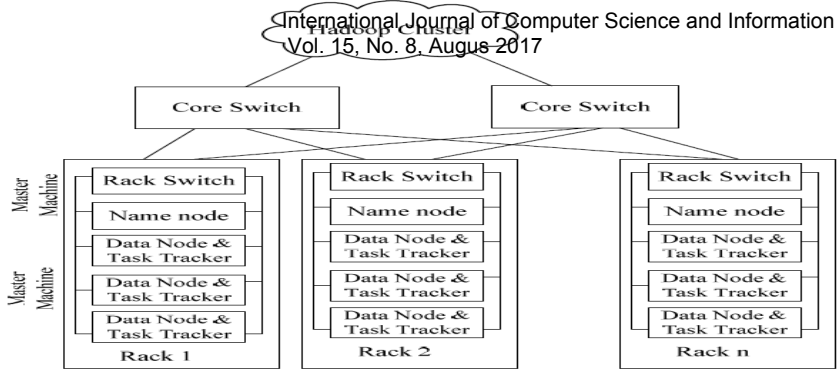


Figure 4: Hadoop cluster using rack switches

3.2. Making a Network substructue onto massive Data

Atopology depending on a concatenation of switches Known as a leafs which take the access layer and these switches be moer meshed onto spine switches directly building a “fabric” network (Figures7 - 8). Every access-layer switch be one hop away onto each other, basically decreasng a likelihood of bottlenecks then reducing the time latency [22].

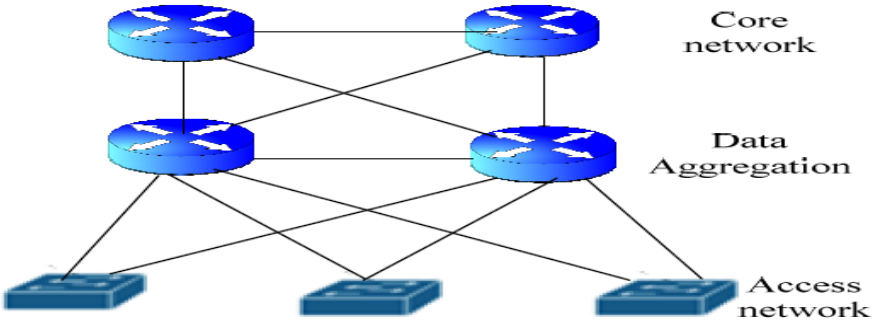


Figure 5: Basic Hierarchical network architecture

A leaf-spine topology can function in layer 2 and 3 that leads switching, routing between the leaf and spine layer. Fabric is said into grown connectivity ,flexibility in networks which executet dynamic virtual environments , prepared for a converged scalability ofdevices. The fabric network topology is defined as “when network nodes are connected into each other using one or many network switches” [19] .

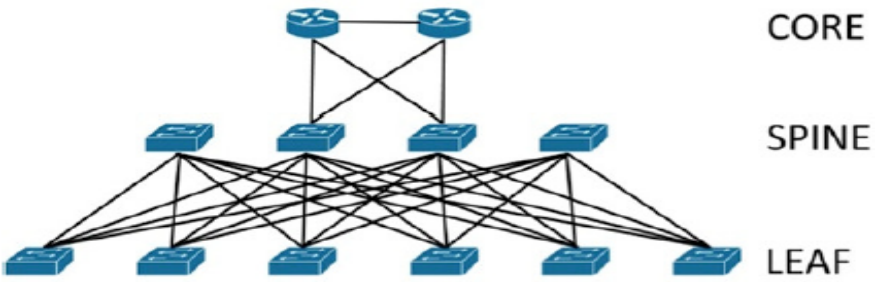


Figure 7: Basic Fabric network architecture

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

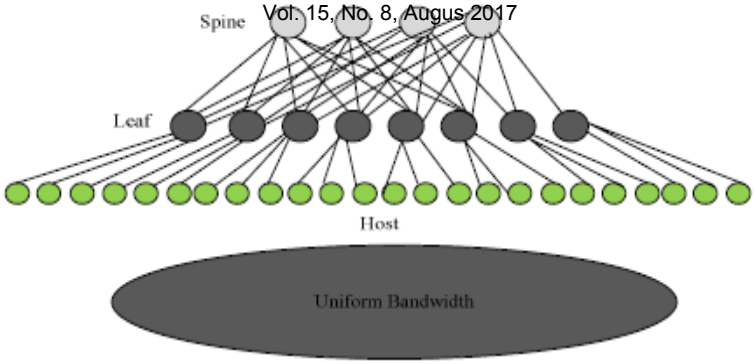


Figure 8: Big Data Fabric topology capable of uniformed bandwidth

Fabric networks are based on spreading the flow of network traffic across many physical links,which outputs an increase in higher throughput, bandwidth and performance over typicalbroadcast networks [15]. A flatter converged network is simpler and cheaper to install, while growing traffic efficiency ,decreasing congestion. Fabric networks have the benefit upon a tree topology because they use layer 2 network paths which help with load balancing and reducingredundancy [21]. A fabrics convergence feature allows multiprotocol communication andtransparency extending to all locations in the network under a single environment. This allows theimplementation of consistent services and a greater network intelligence management policy thatcan provide efficient and secure communication [21].

different characteristces into fabric above a structural tree based topology. first into a fabric topology depending on staffing a point-to-point conexion between nodes utiliziing a single hop distance in relation to the switching process, that greatly decreases latency in the network. Second, may be tolerable to perform a switching fabric utilizing virtualization that permit to diverse networking combinations into perform as one then support to enhance switching execution. likely into make a pool of virtual switches which decrease manual work from the executivor.Third, via having a flat architecture of a single extended environment, it is easier supplier extenination into multi uses areas of the network, when makinging a data center fabric topology[15]. Expansion and enhancement done easily by creating virtual domains or by making point-to-point cconexions.

3.3. Software Defined Networking substructure

SDN [22, 26] were progressing into a network arrangement to make network administrators runs manipulating planes. The massive Data connections uses to run one to many connections of client to databases to supply services and manage a high end implementation s so as to load massive amounts of dynamic unstructured data that traditional networks hard to cure and treat. meaning time a data-planes tools link and react on a standard patterns. detaching the control plane and data plane (packet forwarding) make changes who time locked system to an open , action it imaginable to disposition network tools working. detachs to 3-layers: application, control and infrastructure/data (Figure 9).

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

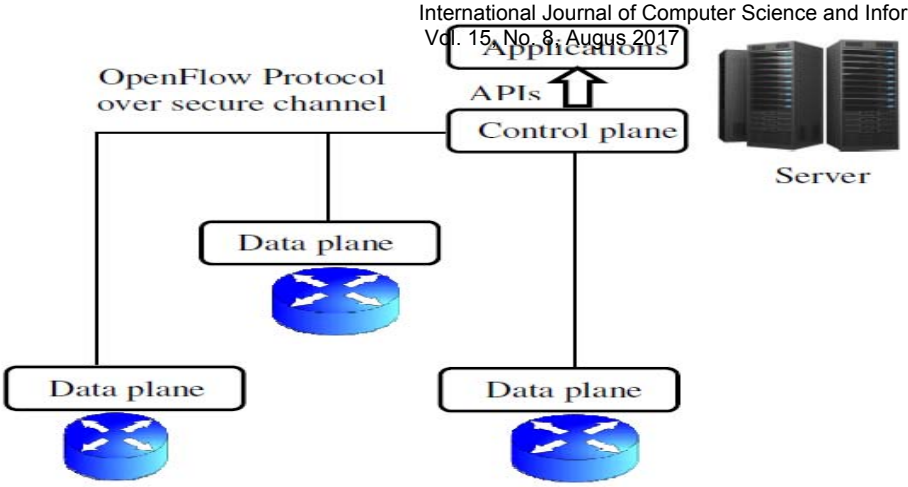


Figure 9: SDN Architecture layer breakdown

With a view to performance employ SDN architecture at supplying programmable network communicate one be fit run these practicability ,employment about a network by software programs. right away running planes grants an originative trend to leverage adjust the control plane to tool up a elastic and adaptable topology which be used to work with multi uses networked applications and services. Open Networking Foundation supplies the mostly gets the SDN:

“Software-Defined Networking becomming the network building wherever the control plane separated of forwarding plane , then fair to programmable” [22]. Figure 10 explain a structure about massive data which programming operations , huge data assure, programming , scheduling and triggering.

3.3.1. parcticabilities about SDN

goal from Software Defined Network making the SW which has been used to manage the network [27], to layer of abstraction the goal to use an elastic network running system to mode huge Data demands performance and flexible. The Switches at most forwarding tools s controlling routing decisions which gain processed through central controller.

A concept from the programmable networks , decoupled data and control plan exposes a rule to whom a SDN controller performance treats. A control plane connects the channel utilized to replacing indexes (messages) amongst logical network tools for forwarding and running . make the most of the available products of SDN set it up with a North Bound plane who employers utilized a programmability by an (API)onto re-attampt the real time statics , service utilizing . data plane also known as the forward plane which all the information set out to switched in a network which traffic evaluted by then measured to service utilizing. Utilizing these network in an application layer for whom custom-made applications utilized.

The Virtualized networks [27] utilized for getting an elastic and resillient . then clients beginning thrust accordingly you utilized a service to holder virtual network iunderlying that utilizing the Virtual Private Server (VMs) to build the customized topologies simplifying the dynamic network re-arrangment utilizing the centralized control plane executed within a software.

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

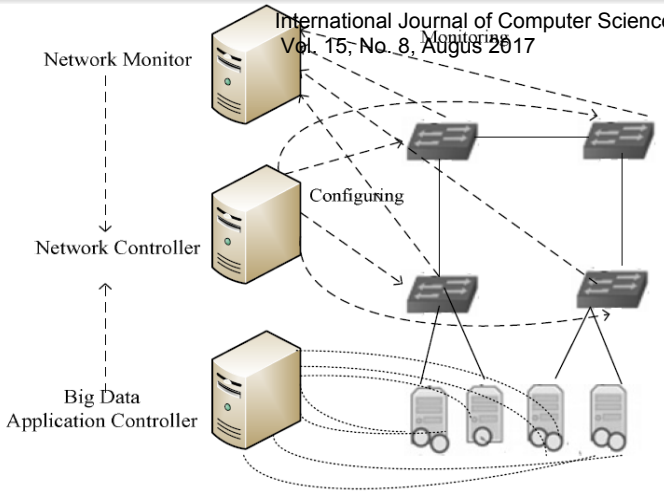


Figure 10: Hybrid awareness operations monitoring, big data controlling, reporting, scheduling, and triggering

3.3.2. Traffic-Aware Architecture

A Northbound API connecting SDN controller utilizes the additional implementation to relocating. Open Flow be used into conform the callbration from a network tool programming , arrangement in the midst controllers , switches. operation Open Flow contains of 2 - rules that identify the fields, running the packet headers then the instructions make to competitions. Through crossing controller forwarding a switch rule then path from a packet. Effectiveness the controller supples established in advance identify the principles which be executed over the static or dynamic SDN control software. the administrator will decide to use a flow to traffic utilizing models, applications , cloud re-resources. During the Open Flow considered a strongly techniques which existence utilized the operations SDN control plane data which has deiciency while taking into consideration a real-world hardware requirements. to be further essential flow a confiqurations which it has suggested by Curtis[28] the authors of Networking to a huge Data moduler Curtis scanning the delineation Devo Flow to “devolve control by cloning rules whenever flow built utilizing the wild cards.” work done in [13] scaling the overhead by Open Flow so that possess further enhancement to flow running to massive Data scanning applications networking utilizies local jobs to switches owns local re-locate decisions does not need to contact a controller. Figure 11 shows SDN controller and flow table connections to Open Flow.

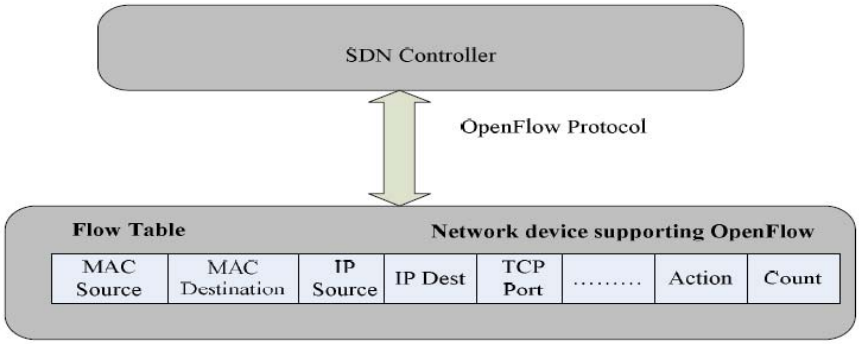


Figure 11: SDN Controller and Flow Table communications with Open Flow.

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

3.3.3. Application-Aware Architecture

Aim of massive data networks being permanently enhancing effectiveness onto a networks during network arrangements , programmability. doing this impotance from network infrastructure be aware of implementation deficiencies capable of service its requires. A technigues onto service implementation requires to utilize switches pre-rendered flow information into control plane so as to locate , arrange scheduled flows. the flows implement-level information and disorder be contain at the transport layer from huge Data implementations. that participate multiple data transfers called as shuffles from mappers and flow tables connecting into specific sorting of flows. enhancement the flows both hardware and application requirements deficiency to orchestrate a correlation for seamless reconfiguration of resources and upper-layer needs. Due to weak implementations by Hadoop in heterogeneous clusters optimizers have been proposed. OF Scheduler runs MapReduce jobs, asses network traffic, and load-balances traffic on the links decreasing the time , takes jobs to finish based on the demand of MapReduce tasks to enhance the execution [29]. Heavy loaded links are off loaded by preference of load balancing flows especially duplicated jobs, larger flows can affect global cost so offloading happens to large flows to maximize rendering showing and prevent the re-scheduling [13].

REFERENCES

[1] Han, Jing, et al. “Survey on NoSQL database”. Pervasive computing and applications (ICPCA), 20116th international conference on. IEEE, 2011..

[2] Lam, Chuck. Hadoop in action. Manning Publications Co., 2010.

[3] M. R. Jam, L. M. Khanli, M. S. Javan and M. K. Akbari, “A survey on security of Hadoop”, 2014 4thInternational Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, 2014, pp.716-721.

[4 Borthakur, Dhruba. “The hadoop distributed file system: Architecture and design”, Hadoop ProjectWebsite 11, No. 21, 2007.

[5] X. W. Chen and X. Lin, “Big Data Deep Learning: Challenges and Perspectives,” in IEEE Access,vol. 2, pp. 514-525, 2014.

[6] Sagiroglu, Seref, and DuyguSinanc. “Big data: A review.” Collaboration Technologies and Systems(CTS), 2013 International Conference on. IEEE, 2013.

[7] Bushong, Michael. “Six considerations for big data networks”, searchsdn.techtarget.com, Accessed Web on March 2016

[8] Kvernvik Tor and Matti Mona. “Applying big-data technologies to network architecture”, EricssonReview, 2012

[9] Guohui Wang, T.S. Eugene Ng, and AneesShaikh, “Programming your network at run-time for bigdata applications”, In Proceedings of the first workshop on Hot topics in software defined networks (HotSDN '12). ACM, New York, USA, pp. 103-108, 2012

[10] Idiro Analytics, “Hadoop”, http://idiro.com/about-idiro/our-technology/ Accessed Web in May 2016

[11] Introduction to big data: infrastructure and networking considerations, Juniper Networks White Paper2012.

[12] Jain, Raj. “Networking issues for big data”, Class Lecture, Washington University in St. Louis 2013

[13] Lin Xiaodong, MisicJelena, ShenXuemin, and Yu Shui, “Networking for big data”, Boca Raton:CRC Press, 2015 Print.

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

[14] Bertolucci, Jeff. “Big data development challenges: talent, cost, time”. Informationweek.com. Web Aug 18, 2012.

[15] Borovick, Lucinda and Villars and L. Richard, “The critical role of the network in big dataapplications”. Cisco White Paper, April, 2012.

[16] Chen, Min, Shiwen Mao, and Yunhao Liu. "Big data: a survey." Mobile Networks and Applications19.2 (2014): 171-209.

[17] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters."Communications of the ACM 51.1 (2008): 107-113.

[18] Mearin, Lucas. “World’s data will grow by 50x in next decade, IDS study predicts”.

Computerworld.com, Web Accessed on May 2016.

[19] Eneboe, Michael K., and Andrew D. Hospodor. “Isochronous switched fabric network”, U.S. Patent

[20] Jeffrey Dean and Sanjay Ghemawat, “MapReduce: simplified data processing on large clusters”, TheACM Communication Magazine, Vol. 51, Issue. 1, pp. 107-113, January 2008

[21] Borovick Lucinda and Villars L. Richard, “The critical role of the network in big data applications”.Cisco White Paper, April 2012.

[22] Banks, Ethan. “Data center network design moves from tree to leaf”.

Search data center.tech target.com, Packet Pushers Interactive, Accessed on June 2016

[23] Jeffrey Dean and Sanjay Ghemawat, “MapReduce: simplified data processing on large clusters”. TheACM Communication Magazine, Volume 51, Issue 1, pp. 107-113, January 2008.

[24] IBM, “Bringing big data to th enterprise”,https://www01.ibm.com/software/in/data/bigdata/ Accessed Web on June 2016.

[25] McDougall, Richard. “Is your cloud ready for big data?” Strata Conference – Co-presented by O’Really Cloudera, New York, Oct 28- 30, 2013

[26] Open Network Foundation, “Software-defined networking: The new norm for networks,” White Paper, April 2012

[27] Open Networking Foundation, “SDN Architecture Overview”, White Paper V1.0, December 2013.

[28] Curtis, A. R., Mogul, J. C., Tourrilhes, J., Yalagandula, P., Sharma, P., and Banerjee, S., “Devoflow:Scaling flow management for high-performance networks”, ACM SIGCOMM Computer Communication Review, 41(4): 254-265, July 2011

[29] Zhao Li, Yao Shen, Bin Yao, MinyiGuo, “OF Scheduler: A Dynamic network Optimizer for MapReduce Heterogenous Cluster”, International Journal of Parallel Programming, Volume43, Issue 3, pp. 472-488, June 2015.

Proposal of a Transparent Relay System with vNIC for Encrypted Overlay Networks

Satoshi Kodama
Tokyo University of Science
Department of Information Science
2641 Yamazaki, Noda-shi, Chiba-prefecture, JAPAN
kodama@is.noda.tus.ac.jp

Rei Nakagawa, Toshimitsu Tanouchi
Tokyo University of Science
Department of Information Science
2641 Yamazaki, Noda-shi, Chiba-prefecture, JAPAN
j6316627@gmail.com, j6316625@ed.tus.ac.jp

***Abstract*— New generations of applications call for new demands that are totally different from previous uses of the Internet (e.g., cross-layer and network function virtualization), and the existing networks are not optimized for these new demands due to being overwhelmed by enormous numbers of external network protocols. Overlay network technologies aim to respond to such future network demands. Systems on overlay networks mitigate this protocol overload by exploiting the unlimited programmability of the overlay nodes comprising the system. This paper proposes an overlay node that works as a transparent proxy server and router for encrypted communication over overlay networks. This overlay node acts as a virtual switch over multiple layers of the OSI reference model (the datalink, network, transport, and session layers) using general-purpose components (a personal computer, physical network interface card, and virtual network interface card, developed using the C language). The ideas behind this proposal derive from the effectiveness of software-defined networks and network function virtualization. Finally, we examine the performance of the overlay node experimentally and suggest possible designs for future overlay networks.**

Keywords Cross-Layer; Network Design; Software-Defined Management; Transport Layer Security; Transparent Proxy Server; Overlay Networks;

I. INTRODUCTION

The remarkable success of TCP/IP networks over the past five decades highlights the importance of the traditional address resolution system. However, the new generations of applications make new demands, totally different from the current uses of the Internet (e.g., cross-layer and network function virtualization (NFV)), and the existing network has been unable to meet these new demands due to the requirements of enormous numbers of external network protocols. Therefore, deployment of customized network protocols to satisfy such new demands has recently become an essential issue.

To solve this problem, overlay network technologies constitute a new research area with the potential to provide a flexible foundation for the demands of new applications [1–3]. Our understanding of these studies is that overlay network technology is a powerful framework that breaks the current end-to-end principle by placing resources and intelligence for customized policies in the middle of the network. For theoretical implications, the core features of deploying overlay networks include scalability and cost-effectiveness. Scalability provides extensions to the network infrastructure that are specialized for innovative ideas, such as defining new network applications in terms of existing network applications and defining effective support protocols for existing network protocols, all while ensuring backward compatibility with external network protocols. This is cost-effective because the number of overlay network nodes required is minimal. Overlay nodes must be virtualized for backward compatibility, and it then becomes possible to receive the various benefits of network virtualization. In particular, the cost-effectiveness derives from the benefits which can be provided with a minimal number of overlay nodes placed discreetly over the Internet.

In engineering terms, the core feature of an overlay network is the software-defined system, i.e., the daemon, which enables cooperation among the overlay nodes. The daemon software installed on each of the overlay nodes works as both a service provider for the overlay network and a virtualized switch for multiple layers in the network protocol stack. The service providers coordinate the global state of the cloud solution to provide new applications or services that can be accessed by external clients. The virtualized switches provide packet processors that refer to the header information from each layer of the network protocol stack and rewrite this information as necessary, working as routers between overlay nodes. This is the architecture of software-defined networks (SDNs) on the overlay network, which have made a remarkable contribution to network virtualization research. Based on these considerations, overlay network technologies will be able to respond to the demands of future networks.

This paper proposes a software-defined architecture for a transparent proxy server (an overlay network infrastructure technology) in an overlay node. This server can access application data encrypted in the session layer of the OSI reference model. Overlay nodes with this new functionality will lead to overlay network systems based on a highly secure content cache mechanism (discussed in Section 6). The architecture of the transparent proxy server is composed of physical network interface cards (pNICs) and virtual network interface cards (vNICs) for the packet processor corresponding

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

to the datalink, network, transport, and application layers of the OSI reference model (discussed in Section 3).

First, however, in the following section, we discuss some of the related research that inspired the proposed system.

II. Related Work

While designing the architecture of the proposed system, we were inspired by some related works. We used an NFV/SDN architecture [4–11]. Using an SDN mitigates network complexities caused by having too many protocols installed on the network's infrastructure as it separates the data and control planes. The data plane refers to the data processing systems (e.g., IP, TCP, and Ethernet), whereas the control plane refers to the systems that determine how and where packets are forwarded (e.g., routers, traffic management, and firewalls). Separating the data and control planes is an effective method for making network control flexible and scalable not only within a network device but also over the entire network. NFV provides network services, such as firewalls, traffic caching, intrusion detection system, network address translation, multicast routing, and redundancy, in a virtualized infrastructure. The principle of NFV is to deploy the new network service architecture on the virtualized infrastructure provided by the SDN.

We aim to take advantage of the effectiveness of the NFV/SDN by incorporating a packet processor into the data plane and the new virtualized network services on the SDN (including a packet processor) into our software-defined overlay node architecture. In terms of a detailed blueprint for the software-defined overlay node architecture, there have been several helpful discussions related to the engineering of network virtualization architectures [6, 10, 11]. In particular, Jain and Paul have suggested that open application delivery networks (openADNs) [6] will represent the future of cloud computing architectures, the idea being that most applications can easily obtain computing and storage facilities from cloud services by multiple providers distributed across the Internet. They discussed an architecture for virtualized NICs, assuming the existence of virtual service providers (vSPs) in the cloud and highlighting the lack of scalability affecting pNICs. In other words, they insisted that each vSP needs its own NIC and proposed three virtualization designs based around NICs. We believe that these network virtualization architectures are closely related to the software-defined architecture of overlay nodes acting as service providers. Our architecture aims to provide vNICs as fundamental software components via the supervisor node, which has been proposed by virtual machine software vendors and uses a virtual Ethernet bridge (VEB).

Looking at overlay networks from a conceptual standpoint, there have been several theoretical discussions about the applicability of overlay networks [2–4]. In particular, Chowdhury and Boutaba discussed network virtualization environments (NVEs) in detail [4]. In their discussion of underlying overlay network concepts, they investigated the applicability of overlay networks, which provide infrastructures for innovative technologies such as cloud services. NVEs are composed of several client–server systems, which is virtualized over a physical network infrastructure that provides the network resources needed to offer end-to-end services to clients, wherein each client–server system connects to clients through a virtual network (VN). Our overlay network system shares the following design goals of NVE architectures: flexibility, manageability, scalability, isolation, and programmability.

Flexibility indicates that there is freedom in every aspect of networking in the NVE: network virtualization must offer services without having to coordinate with any other parties (e.g., consensuses, pledges with other protocols, or interfaces). Manageability indicates that there is complete end-to-end control in the NVE: network virtualization must modularize network management tasks and introduce accountability at every networking layer. Scalability indicates that the NVE must scale to support an increasing number of coexisting VNs without affecting their performance. Isolation indicates that network virtualization must ensure isolation between coexisting multiple networks to improve fault tolerance, security, and privacy. Programmability indicates that customized protocols and diverse services can be deployed on multiple networks through network virtualization.

Up to this point, we have presented the essential references that shaped our study of an overlay network system and its components. The following sections describe the details of our overlay network and its nodes. An overview of the overlay network system and schematics of its architecture are presented in Section 3.1. Then, a detailed explanation of the software engineering aspects of the overlay node proposed in this paper is given in Section 3.2, together with illustrations of the workflow and architecture.

III. System Architecture

A. Overlay Network System

Before we present the architecture of the overlay node proposed in this paper, it will be useful to discuss the basis of the overlay node and the underlying overlay network system used to deploy the customized policy. The overlay network system uses a cloud foundation to offer the services requested by clients and is composed of overlay nodes that together comprise the global state. Fig. 1 shows a schematic of the overlay network system. In Fig. 1, there are three components that are split between the supervisor and service provider sides. On the service provider side, the overlay network is based on the Internet and provides a service network using overlay nodes to offer a new service. The resources distributed across the overlay nodes can be used to provide this new service.

The service network is composed of overlay nodes, managed by the overlay control node using the underlying overlay service control protocol (OSCP). The overlay control node is the first node that clients communicate with to make service requests, and its IP address represents the service network to the client. The OSCP separates the overlay communication flow of the specified service from the other communication flows and routes it between the overlay nodes in the service network using sequences of queries to maintain the most appropriate state for the service. For example, the OSCP manages the routing table based on amount of delay to offer a real-time service.

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

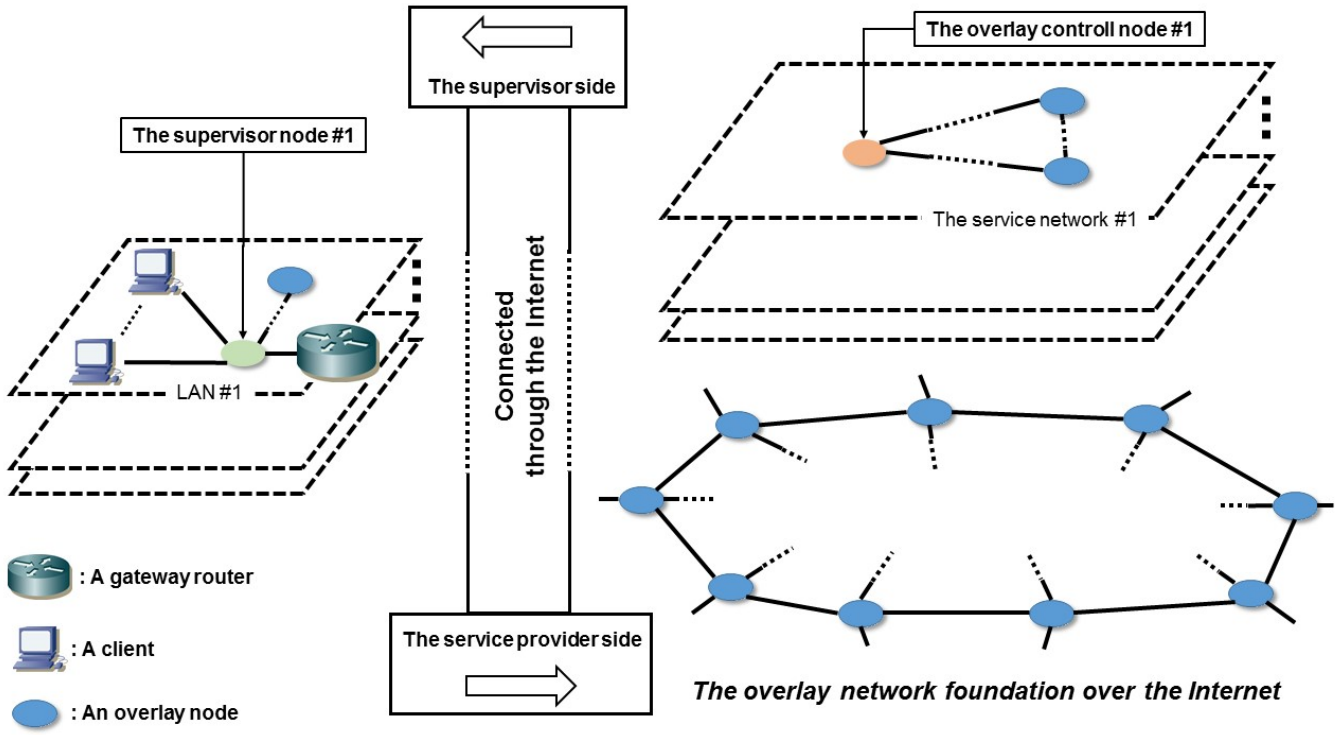


Figure 1. Overview of the overlay network system

On the supervisor side, the supervisor node manages all communication from clients and the overlay nodes in the LAN using the underlying overlay link control protocol (OLCP). The supervisor node is the overlay node that is closest to the LAN's gateway router. The OLCP transparently separates overlay communications from the overall communication flow in the LAN and ensures the quality of the overlay communications between the supervisor and service provider sides (e.g., manages the order of sequential data for the new service and encrypted communications over transport layer security (TLS)/IPSEC). Moreover, if a service network needs to extend its scope to the LAN, the OLCP acts as a relay server, handling address resolution between the internal and external overlay nodes in the service network.

When clients receive services provided by the overlay network system, they benefit from the work described so far. The client can just see a simple client¬–server system involving the overlay control node and cannot see the work done by the supervisor node and the overlay nodes in the service network. This overlay network system architecture achieves the design goals previously described in Section 1.2, namely, flexibility, manageability, scalability, isolation, and programmability. The question asked in this paper is how to ensure manageability and programmability over encrypted communication links. The following subsection describes in detail the use of an overlay node as a transparent proxy server wherein communication quality is guaranteed by TLS.

B. Transparent Proxy System over TLS Communication

Proxy servers are well-known intermediaries between endpoint devices such as personal computers and servers from which the clients request services. The advantage of a proxy server is that it can store frequently requested results in a cache, from which they can be served to clients directly. For example, when a proxy server receives a request for an Internet resource (such as an HTTP request) from a client, it first looks in its local cache of previously served pages. If it finds the requested content, the proxy server can return it to the client without needing to forward the request to the main server. If the page is not in the cache, the proxy server, acting on behalf of the client, uses one of its IP addresses to request the page from the main server. When the page is returned by the server, the proxy relates it to the original request and forwards it to the client.

A transparent proxy server is a particular type of proxy server wherein clients are not aware of its presence [12, 13]. The proposed overlay node architecture is a type of transparent proxy server over TLS, proposed as a function of the supervisor node, thus providing an approach to engineering an overlay node that forms part of the overlay network system described previously in Section 3.1. Fig. 2 shows the architecture of the proposed transparent proxy system.

In Fig. 2, the supervisor node lies immediately between the LAN's gateway router and its cluster of network devices. In other words, the supervisor node handles all communication from the network device cluster. The network device cluster is composed of the clients and overlay nodes accessed using TLS. Note that the supervisor node acts as a transparent proxy server

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

over TLS. The purpose of this node is to transparently monitor the communications of the given TLS service and manage its content while understanding it semantically.

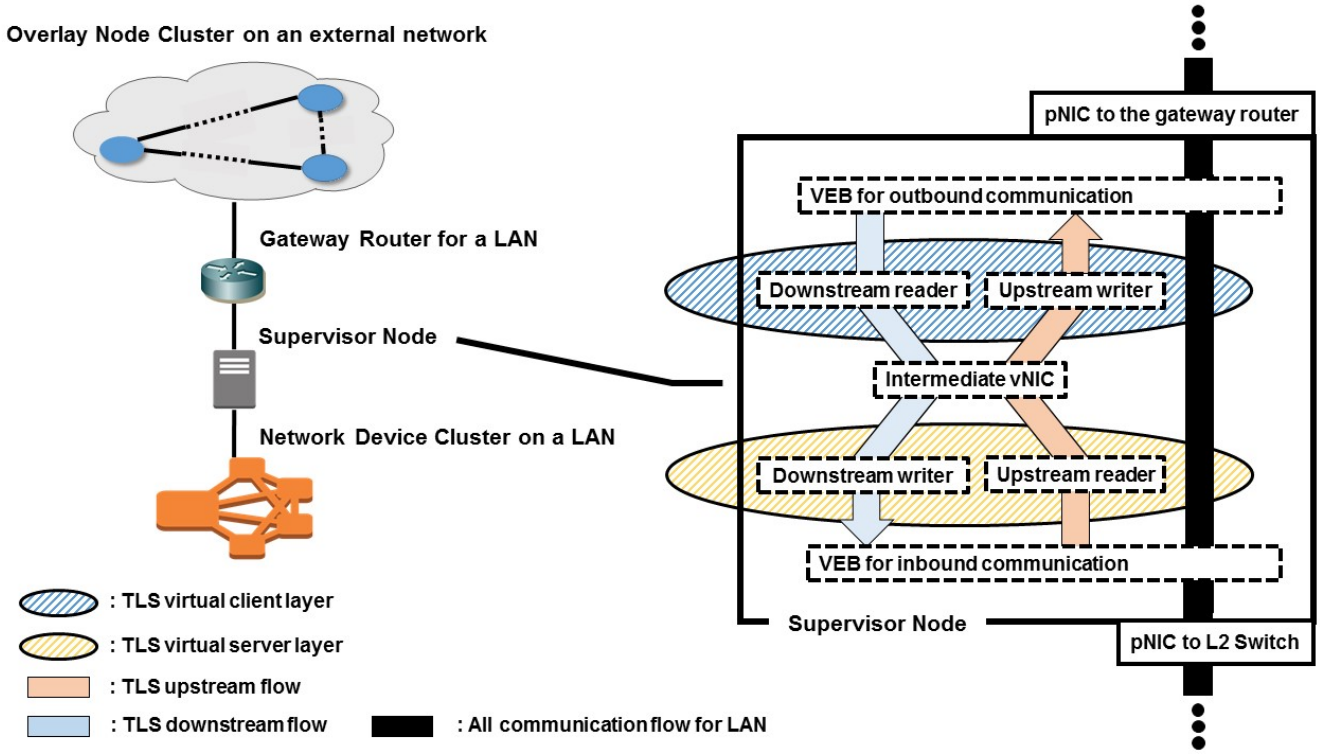


Figure 2. *Architecture of the proposed transparent proxy system*

The design comprises physical NICs (pNICs), VEBs, virtual NIC (vNIC), and virtual client and server layers for the TLS service. The pNIC to datalink layer (L2) switch accepts communication requests for the L2 switch managing the network device cluster. The pNIC to gateway router link accepts communication requests for overlay node clusters on external networks. The VEB for inbound communication handles the TLS flow for inbound communication to the LAN and connects it to the intermediate vNIC. Similarly, the VEB for outbound communication handles the TLS flow for outbound communication from the LAN and connects it to the intermediate vNIC.

The intermediate vNIC's IP address is that of the LAN and represents the endpoint for the TLS service, providing the virtual TLS server layer that actually accepts and handles TLS communications from clients and the virtual TLS client layer that connects to the overlay control node for the server (the server node) in the TLS service network and requests content from TLS service resources. The virtual TLS server layer provides the upstream reader that receives the sequence of TLS communications from the client and transmits the content requested in the communication, together with its semantic interpretation, to the intermediate vNIC. In the virtual TLS server layer, the upstream reader receives the sequence of TLS communication requests from the client and transmits the requests to the upstream writer in the virtual TLS server layer. Likewise, the downstream writer receives the content originally requested by the client from the TLS service resources from the downstream reader. In the virtual TLS client layer, the upstream writer receives the client request from the upstream reader and transmits the request to the server node, whereas the downstream reader receives the requested content from the server node and transmits it to the downstream writer in the virtual server layer. Fig. 3 shows the detailed processing flow for each of the components around the intermediate vNIC.

Fig. 3 shows the processes involved in one session for the virtual TLS server and client. First, the client initiates the TLS session with the TLS server node, and the virtual TLS server handles it instead. If this TLS session initialization is successful, the virtual TLS client initiates a TLS session with the TLS server node. Second, after TLS session initialization, the client sends the appropriate content request to the TLS server node. The virtual TLS server receives the request packet instead, extracts and decrypts it using the upper application header, and then transfers it to the transmission buffer of the upstream writer in the virtual TLS client. After that, the virtual TLS client re-encapsulates the content request into a packet for the TLS session between the virtual TLS client and the TLS server and transmits the content request packet to the TLS server. Third, after the content request has been transmitted, the virtual TLS client receives the requested content packet from the server, extracts, and decrypts the packet and then transfers it to the transmission buffer of the downstream writer in the virtual TLS client. Finally, the virtual TLS server transmits the requested content to the client.

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

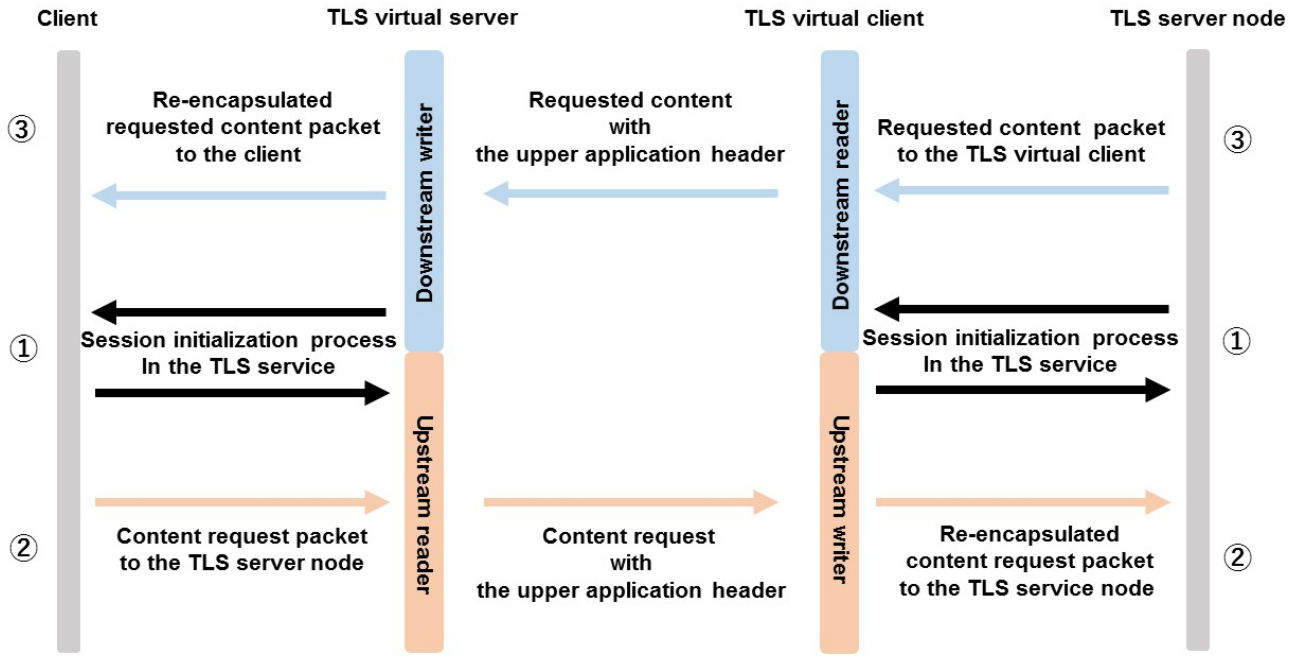


Figure 3. *Architecture of the transparent proxy system around the intermediate vNIC*

What needs to be emphasized at this juncture is how the virtual TLS server handles the process in place of the main server and how we manage both the sessions of the virtual TLS client and server. The virtual TLS server needs to rewrite the destination IP address of the packet using the IP address of the intermediate vNIC, which is handled by the VEB for inbound communication. The supervisor node therefore needs to obtain the IP address of the TLS server in advance. Care must be taken while programming the management of the TLS session between the virtual TLS server and client because subtle dependency problems can creep in to both of them, causing session errors. The essential management policy is to have one virtual TLS client session for each virtual TLS server session. The process sequence of first receiving a content request from the client and transmitting it to the TLS server node and then receiving the requested content from the TLS server and transmitting it to the client must be used for both the synchronized virtual TLS server and client sessions. Therefore, the start and termination conditions of each session must be maintained by a highly robust sequential process, which cannot be handled using a concurrency control scheme such as multithreaded programming or multi-processing.

The following section describes the experiment that we conducted over an HTTPS service to evaluate our engineering approach.

IV. EVALUATION

Our engineering approach, though simple in terms of architecture and functionality, must prove that the complete software model can function as a transparent proxy over TLS. We therefore conducted an experiment wherein we used our engineering approach for a client–server system over HTTPS using general-purpose equipment. Table 1 and Fig. 4 show the experimental setup of the supervisor node.

TABLE I. EXPERIMENT SETUP

	Supervisor node PC	**TLS server PC**
CPU	Intel Core i7-4790 CPU at 3.60GHz	Intel Core i5-6500 CPU at 3.20GHz
Memory	8GB DDR3 SDRAM	
OS	CentOS Linux release 7.3.1611 (core)	
Networking equipment	LUA3-U2-ATX (Buffalo) and RTL 8111/8168/8411 (Realtek)	RTL 8111/8168/8411 (Realtek)

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

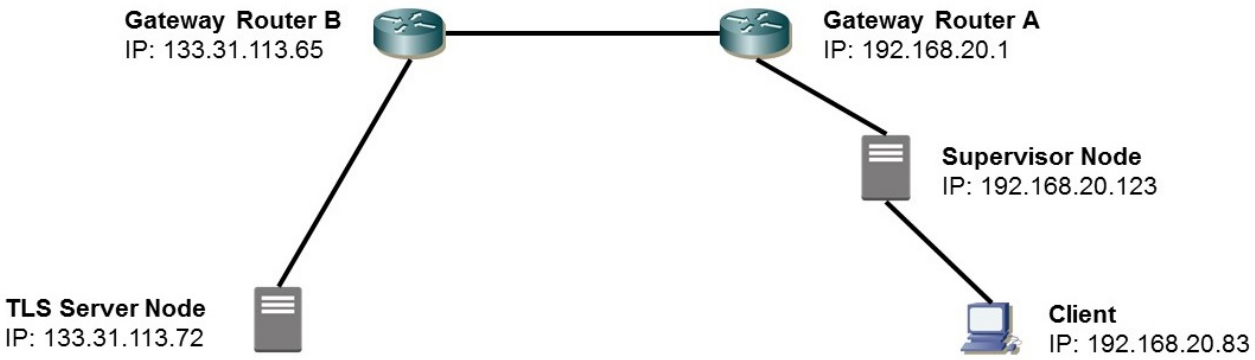


Figure 4. *Experimental topology*

The client was a normal PC, which did not include any additional components of our system and could use browsers such as Firefox and Google Chrome. The TLS server was a PC running CentOS with an HTTPS server implemented using Apache and OpenSSL. The supervisor node included an intermediate vNIC with local IP address 192.168.20.123 for gateway router A, each VEB was implemented using the specified vNIC with no IP address, and the virtual TLS client and server in the supervisor node were connected to the intermediate vNIC and developed using OpenSSL in the C language. In the experiment, the supervisor node analyzed two streams provided by the TLS server. Fig. 5 shows the HTTP GET requests representing these streams, as decrypted by the transparent proxy system.

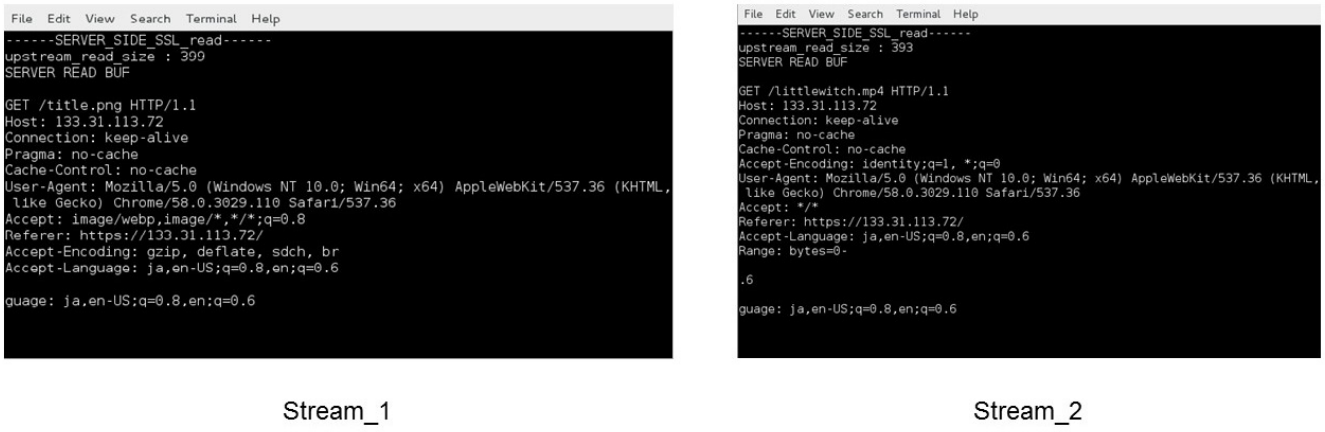


Figure 5. *The two streams used for the experiment, as decrypted by the transparent proxy*

In Fig. 5, stream_1 requests “title.png,” which was 2218 kB in size, and stream_2 requests “littlewitch.mp4,” which was 33 MB in size. The following subsection shows the results of this experiment, wherein the performance of the supervisor node was evaluated in the following terms: its usability as a software model on general-purpose equipment; Round Trip Time (RTT), representing the Quality of Experience (QoE) of the client; and CPU utilization, to see how easily it could coexist with other processes. The client measured the RTT as an endpoint of the HTTPS service, whereas the supervisor node measured the CPU utilization, accounting for the software model of our transparent proxy system.

A. Results

Table 2 shows the results of the experiments described in the previous section, which were conducted five times for each measurement. We used the TCPDUMP CentOS command to measure the RTT on the client, and the TOP CentOS command to measure CPU utilization on the supervisor node. For comparison, Table 3 shows the RTTs measured without the

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

supervisor node. Fig. 6 compares the RTT results with and without the supervisor node.

TABLE II. RESULTS FOR THE HTTPS SERVICE WITH THE SUPERVISOR NODE

Number of trials	Stream type	RTT(s)				CPU utilization (%/s)
		Median	Average	Max	Min	
1	Stream_1	0.000120	0.000161	0.001393	0.000044	0.90
	Stream_2	0.000096	0.000606	0.051331	0.000014	
2	Stream_1	0.000081	0.000940	0.000446	0.000033	0.81
	Stream_2	0.000098	0.000670	0.051136	0.000016	
3	Stream_1	0.000080	0.000096	0.00038	0.000027	0.84
	Stream_2	0.000096	0.000712	0.051237	0.000015	
4	Stream_1	0.000091	0.000115	0.000790	0.000032	0.87
	Stream_2	0.000096	0.000651	0.051663	0.000017	
5	Stream_1	0.000116	0.000130	0.000514	0.000056	0.83
	Stream_2	0.000098	0.000667	0.050910	0.000016	

TABLE III. RESULTS FOR THE NORMAL HTTPS SERVICE WITHOUT THE SUPERVISOR NODE

Number of trials	Stream type	RTT(s)			
		Median	Average	Max	Min
1	Stream_1	0.000129	0.000166	0.001291	0.000041
	Stream_2	0.000102	0.000643	0.051330	0.000016
2	Stream_1	0.000132	0.000169	0.001341	0.000042
	Stream_2	0.000106	0.000542	0.051868	0.000016
3	Stream_1	0.000121	0.000168	0.004285	0.000034
	Stream_2	0.000103	0.000493	0.050956	0.000018
4	Stream_1	0.000123	0.000168	0.006150	0.000042
	Stream_2	0.000110	0.000647	0.051421	0.000018
5	Stream_1	0.000127	0.000170	0.004688	0.000035
	Stream_2	0.000113	0.000565	0.051489	0.000019

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

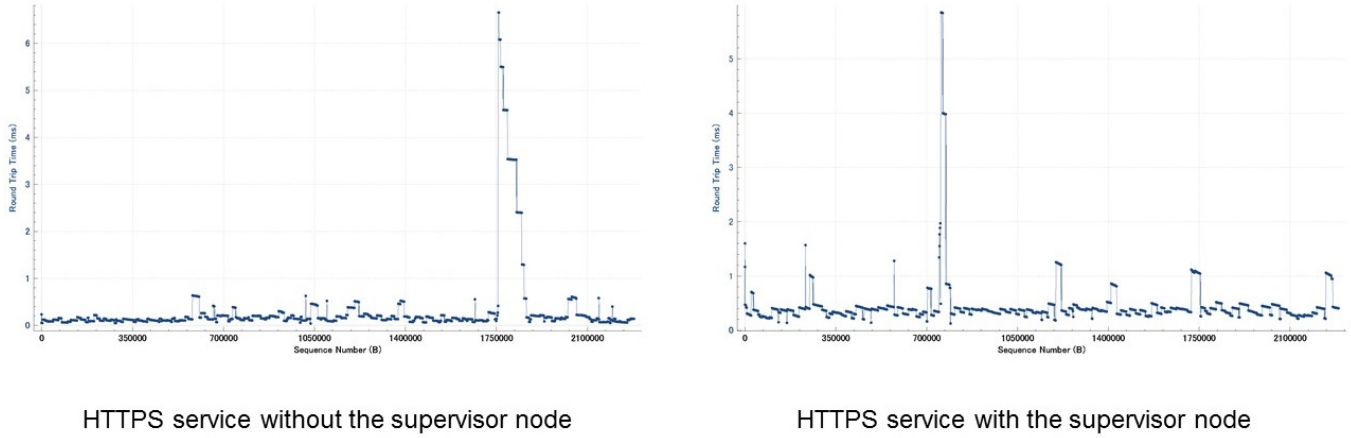


Figure 6. *Comparison showing the extent of the RTT change between the HTTPS flows with and without the supervisor node*

In Tables 2 and 3, stream_2 always shows a very high maximum RTT compared with stream_1 because its larger size caused network bandwidth congestion. An interesting result is that the RTT for stream_1 with the supervisor node is smaller than that without the supervisor node. We believe that this is because the small delay, which can be seen on the graph of the HTTPS service with the supervisor node in Fig. 6, prevented any significant congestion control from happening, which was not the case without the supervisor node. While the change in RTT for stream_1 with the supervisor node was very variable, the RTT for stream_2 with the supervisor node was uniformly higher by around 0.1 ms. These results indicate that there is no difference from a QoE standpoint for long-term communication. In addition, the CPU utilization results for the transparent proxy show that it has very good coexistence characteristics.

V. CONCLUSION

We have proposed an approach to engineering a transparent proxy over TLS. The purpose of this node is to transparently monitor the communications of the TLS service and manage and semantically interpret the content served by it. The architecture of the transparent proxy benefits from the effectiveness of the NFV/SDN architecture. The proposed transparent proxy is a function of the supervisor node in the overlay network, which is designed for manageability, flexibility, scalability, isolation, and programmability. The transparent proxy was evaluated in an experiment to measure its performance as a software model on general-purpose equipment, showing that the transparent proxy is preferable over TLS. The experimental results show good performance and ability to coexist with other processes.

VI. DISCUSSIN AND FUTURE WORK

From the standpoint of the transparent proxy server's caching and forwarding functionality, it is clear that it is helpful for strategic management based on the content of the service. It has various applications, such as securing the confidentiality of the content, monitoring users' traffic usage, and adapting the routing strategy based on the content. In particular, a content-adaptive routing strategy will be an essential application for future content-based networking, such as information-centric networking, content-centric networking, and named data networking. There have been several interesting discussions about content-based networking [14–16], with the aim of implementing a robust content discovery mechanism using a content cache system with the OpenFlow architecture over SDN. We believe that our overlay network system is a possible engineering solution for the content cache system and will be helpful for deploying transparent content cache systems using TLS-encrypted communication.

ACKNOWLEDGMENT

The authors would like to express their heathy thanks to the referee who pointed our several typological errors and made a very suggestive proposal to revise the manuscript of this paper.

REFERENCES

[1] Amy Babay, Claudiu Danilov, John Lane, Michal Miskin-Amir, Daniel Obenshain, John Schultz, Jonathan Stanton, Thomas Tantillo, and Yair Amir, "Structured Overlay Networks for a New Generation of Internet Services," International Conference on Distributed Computing Systems, 2017.

[2] Paolo Medagliani, Stefano Paris, Jérémie Leguay, Lorenzo Maggi, Xue Chuangsong, and Haojun Zhou, "Overlay Routing for Fast Video Transfers in CDN," IFIP/IEEE International Symposium on Integrated Network Management, 2017.

[3] Joe Touch, "Dynamic Internet Overlay Deployment and Management Using the X-Bone," International Conference on Network Protocols, 2000.

[4] N.M. Mosharaf Kabir Chowdhury and Raouf Boutaba, "Network Virtualization: State of The Art and Research Challenges," IEEE Communication Magazines, Vol. 47, No. 7, 2009, pp.20–26.

[5] Bo Han, Vijay Gopalakrishnan, Lusheng Ji, and Seungjoon Lee, "Network Function Virtualization: Challenges and Opportunities for

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

Innovations," IEEE Communications Magazine, Vol. 53, No. 11, 2015, pp.90–97.

[6] Raj Jain and Subharthi Paul, "Network Virtualization and Software Defined Networking for Cloud Computing: A Survey," IEEE Communications Magazine, Vol. 51, No. 11, 2013, pp.24–31.

[7] Stuart Clayman, Lefteris Mamatas, and Alex Galis, "Efficient Management Solutions for Software Defined Infrastructures," Network Operations and Management Symposium (NOMS) IEEE/IFIP, 2016.

[8] Albert Greenberg, Gisil Hjalmtysson, David A. Maltz, Andy Myers, Jennifer Rexford, Geoffrey Xie, Hong Yan, and Jibin Zhang, "A Clean State 4D Approach to Network Control and Management," ACM SIGCOMM Computer Communication Review, Vol. 35, No. 5, 2005, pp.41–54.

[9] Hyojoon Kim and Nick Feamster, "Improving Network Management with Software Defined Network," IEEE Communication Magazine, Vol. 51, No. 2, 2013, pp.114–119.

[10] Jon Matias, Jokin Garay, Nerea Toledo, Juanjo Unzilla, and Eduardo Jacob, "Toward an SDN-Enabled NFV Architecture," IEEE Communication Magazine, Vol. 53, No. 4, 2015, pp.187–193.

[11] Myung-Ki Shin, Ki-Hyuk Nam, and Hyoung-Jun Kim, "Software-Defined Networking (SDN): A Reference Architecture and Open APIs," International Conference on ICT Convergence (ICTC), 2012.

[12] Mark O'Neill, Scott Ruoti, Kent Seamons, and Daniel Zappala, "TLS Proxies: Friend or Foe?," Proceedings of the 2016 Internet Measurement Conference, 2016.

[13] Jianxin Wang, Anupama Sundaresan, Vijaya Bharathi Kaza, and Dario Calia, "Transparent Proxy of Encrypted Sessions," US20080126794 A1, 2008.

[14] Abhishek Chanda and Cedric Westphal, "A Content Management Layer for Software-Defined Information Centric Networks," Proceedings of the 3rd ACM SIGCOMM workshop on Information-centric networking, 2013.

[15] Alex F. R. Trajano and Marcial P. Fernandez, "ontentSDN: A Content-Based Transparent Proxy Architecture in Software-Defined Networking," Advanced Information Networking and Applications (AINA), 2016.

[16] Panagiotis Georgopoulos, Matthew Broadbent, Bernhard Plattner, and Nicholas Race, "Cache as a Service: Leveraging SDN to Efficiently and Transparently Support Video-on-Demand on the Last Mile," Computer Communication and Networks (ICCCN), 2014.

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

Conceptual Modeling for Control of a Physical Engineering Plant: A Case Study

Sabah Al-Fedaghi
Computer Engineering Department
Kuwait University
Kuwait
sabah.alfedaghi@ku.edu.kw

Abdulaziz AlQallaf
Instrument Maintenance Department
Ministry of Electricity and Water
Kuwait
Alqallaf.AQ@gmail.com

***Abstract*—We examine the problem of weaknesses in frameworks of conceptual modeling for handling certain aspects of the system being modeled. We propose the use of a flow-based modeling methodology at the conceptual level. Specifically, and without loss of generality, we develop a conceptual description that can be used for controlling the maintenance of a physical system, and demonstrate it by applying it to an existing electrical power plant system. Recent studies reveal difficulties in finding comprehensive answers for monitoring operations and identifying risks as well as the fact that incomplete information can easily lead to incorrect maintenance. A unified framework for integrated conceptualization is therefore needed. The conceptual modeling approach integrates maintenance operations into a total system comprising humans, physical objects, and information. The proposed model is constructed of (abstract) machines of "things" connected by flows, forming an integrated whole. It represents a man-made, intentionally constructed system and includes technical and human "things" observable in the real world, exemplified by the study case described in this paper. A specification is constructed from a maximum of five basic operations: creation, processing, releasing, transferring, and receiving.**

Keywords-conceptual model; engineering system; diagrammatic representation; physical plant

I. INTRODUCTION

The use of models is an important aspect of engineering disciplines because of their essential role in understanding and engaging with the world [1]. Models can take many forms:

> We can use words, drawings or sketches, physical models, computer programs, or mathematical formulas. In other words, the modeling activity can be done in several languages, often simultaneously. [2]

Accordingly, models can be classified as different types: conceptual, physical, or mathematical [3]. In this paper, we focus on conceptual models used to capture "conceptual structures of a domain" [4].

"A model is an abstract view of portion of reality that assists developers to concentrate on relevant aspects of the system and discount needless complications" [5]. ISO/IEC/IEEE 42010 (2011) [6] defines a model as follows: "M is a model of S if M can be used to answer questions about S." In principle, a model is anything that can describe a system, and in this sense, all kinds of typical engineering work products that are created to specify or describe a system are models [7]. The major advantage of modeling is that models are expressed in terms of concepts bound much less to the underlying implementation technology and more closely to the problem domain [1].

ISO/IEC/IEEE 15288 (2015) [8] defines a *system* as a combination of interacting system elements organized to achieve one or more stated purposes. In this paper, we view a system as an (abstract) machine of "things" (to be defined later) connected by flows to form an integrated whole. The machine represents a man-made, intentionally constructed system (hence, it has a purpose) and includes technical and human "things" observable in the real world, as we exemplify in a case study in this paper. "Things" can be pipes, valves, structures, events and happenings, procedures, or materials, e.g., water, chlorine, and heat. A machine is constructed from at most *five* basic operations: creation, processing, releasing, transferring, and receiving. In this paper, we focus on the control and tracking of flows of "things" through machines for maintenance, operations, and management.

Conceptual modeling is a phase of system development that usually occurs after requirements analysis and precedes the design phase in the life cycle of "things". The conceptual model is constituted of a structure that reflects the composition of the physical elements of the system, and behavior that specifies the operational scenarios and functions of the system [7]. It facilitates understanding and communication among stakeholders and serves as a base for consequent phases. Valued features in conceptual models include completeness, faithfulness to realization of the system, understandability, and susceptibility analysis.

Most current conceptual modeling techniques use object-oriented methodology (e.g., UML, SysML), because their main foundation requires breaking system behavior into several pieces and then further decomposing those into other diagrams. Many claims have been made regarding the benefits of an object-oriented model, such as simulating the modeler's way of thinking [9] and contributing to "reducing complexity in the representation of technical systems and design processes" [10]. Researchers have examined and proposed extending the use of object-oriented languages such

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

as UML, but Evermann [11] notes that "UML is suitable for conceptual modeling but the modeler must take special care not to confuse software aspects with aspects of the real world being modeled." The problem with extending UML is that "[UML] possesses no real-world business or organizational meaning; i.e., it is unclear what the constructs of such languages mean in terms of the business" [11]. The object-oriented modeling domain deals with objects and attributes, whereas the real-world domain deals with things and properties. According to Mordecai [12], there is a "significant inability of common conceptual modeling frameworks to appeal to practicing designers and analysts." These frameworks have an "inherent limitation and even fixation to handling the nominal view of the system being modeled. ... A unified framework for integrated, multipurpose, robust, and disruption-accommodating modeling and management is therefore urgently needed" [12].

In contrast to the object-oriented paradigm, according to Dori [13], models of complex systems should conveniently combine *structure* and *behavior* in a single model. Object-Process Methodology (OPM) [13] was developed for multidisciplinary, complex, and dynamic systems and processes [12]. OPM is chartered as ISO/PAS 19450 for system and process modeling [14]. It is considered "a state-of-the-art methodology and paradigm" in both the conceptual modeling domain [15] and the model-based systems engineering domain [16]. OPM [13] is a holistic approach to modeling, studying, and developing engineering systems.

> The OPM paradigm integrates the object-oriented, process-oriented, and state transition approaches into a single frame of reference. *Structure and behavior coexist in the same OPM model without highlighting one at the expense of suppressing the other* to enhance the comprehension of the system as a whole. [17] (Italics added)

In this paper, we introduce an alternative to object-oriented and object-process methodologies, a conceptual modeling methodology based on *flows*, and also present a different conceptualization of such notions as processes, things (objects), and events. To show the viability of the proposed methodology, and without loss of generality, we develop a conceptual description that can be used for control of maintenance and operations of a physical system; as an example, we use the flow of operations within an existing electrical power station. *Maintenance* here refers to "actions taken to prevent a system structure or component from failing or to repair normal equipment degradation experienced with the operation of the device to keep it in proper working order" [18]. *Operations* ensure [19] the implementation and control of activities and safe and reliable processes, as well as recognition of the status of all equipment and operators' knowledge and performance; this aspect supports safe and reliable plant operation.

Recent studies reveal difficulties in finding comprehensive answers to problems inherent in monitoring of operations in physical systems, such as identifying risks and difficulties related to incomplete data, which can lead to incorrect maintenance and operations [20-21]. According to Vieira and Marques [22], "The definition of policies and strategies and the understanding of the efficiency and effectiveness of the maintenance department continue to present opportunities for improvement" [22].

A unified conceptual framework (a single diagram) seems to provide many benefits, e.g., completeness, understandability, and simplified analysis, and is a potential solution to the problems mentioned in the previous paragraph. The conceptual modeling approach integrates maintenance and operations into a total system that comprises humans, physical objects, and information. In our study case, as a result of the complexity of maintenance management of the electrical power plant, operations such as maintenance and technical management are no longer considered mere technical matters. Hence, operations must be integrated into the total management of the system, and a system for future possible online control must be developed. To this end, a conceptual description of the site is needed to provide a holistic overview of the various processes in the system.

II. FLOWTHING MACHINE

For the sake of a self-contained paper, in this section, in subsection A, we briefly review our proposed methodology, which forms the foundation of the theoretical development in this paper called the Flowthing Machine (FM). It involves a diagrammatic language that has been adopted in several applications [23-31]. In subsection B, we provide a new example to explain the approach more completely.

A. Basic Model

The FM modeling language is a uniform method for representing "things" that flow, called "flow things". Flow in the FM refers to the exclusive (i.e., being in one and only one) transformation among five states (also called stages): *transfer*, *process*, *create*, *release*, and *receive*. A flow thing (hereafter a *thing*) cannot be in two stages simultaneously. A thing is defined as what is created, released, transferred, received, and processed. Things in stages are analogous to molecules of water being in one of three states while in Earth's atmosphere: solid, liquid, or gas.

Each *stage* can be expressed by many words:

- *Create:* generate, appear (in the system), produce, make . . .
- *Transfer:* transport, communicate, send, transmit ...
- *Process:* millions of English verbs that change the form of a thing without creating a new one, e.g., paint, package, categorize . . .

Notions in FM can be described as follows.

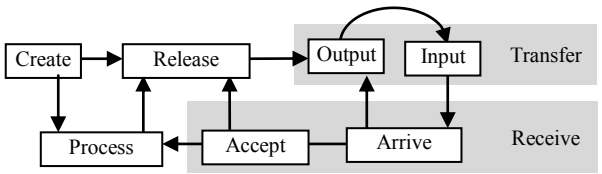


Fig. 1. Flow machine.

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

A flow machine (hereafter *machine*) is depicted in Fig. 1, which shows the internal flow of a system along with the five stages and transactions among them. The machine displayed in Fig. 1 is a generalization of the typical input-process-output model used in many scientific fields.

- *Spheres and subspheres:* These are the network environments and relationships of machines and submachines. The FM model represents a web of interrelated *flows* that cross the boundaries of intersecting and nested *spheres*. A particular static model is the space context for *happenings*, as will be explained later.
- *Triggering:* Triggering is a transformation (denoted by a dashed arrow) from one flow to another; e.g., a flow of electricity triggers a flow of air.

B. Example

Rahim et al. [32] proposed a transformation to derive a modular Petri net from SysML activities to formalize and verify SysML requirements. They present a case study of the operation of a ticket vending machine (TVM):

> The behaviour of the machine is triggered by passengers who need to buy a ticket. When a passenger starts a session, the TVM will request trip information from commuter. Passengers use the front panel to specify their boarding and destination place, details of passengers (number of adults and children) and date of travel. Based on the provided trip info, the TVM will calculate payment due and display the fare for the requested ticket. Then, it requests payment options. Those options include payment by cash, or by credit or debit card. After that, the passenger chooses a payment option and processes to payment. After a successful payment, the TVM prints and provides a ticket to the passenger.

Of interest in the present paper is the type of diagram used by Rahim et al. [32]. The activity diagram utilizes a composite activity concept that incorporates other activities, as shown partially in Fig. 2.

The purpose of scrutinizing this figure is not to present a fair description of Rahim et al.'s [32] study; rather, the aim is to visually contrast their activity diagrams with FM diagrams without a detailed comparison to demonstrate that the latter can be appreciated for their simple visual appearance and understandability.

C. Static Description

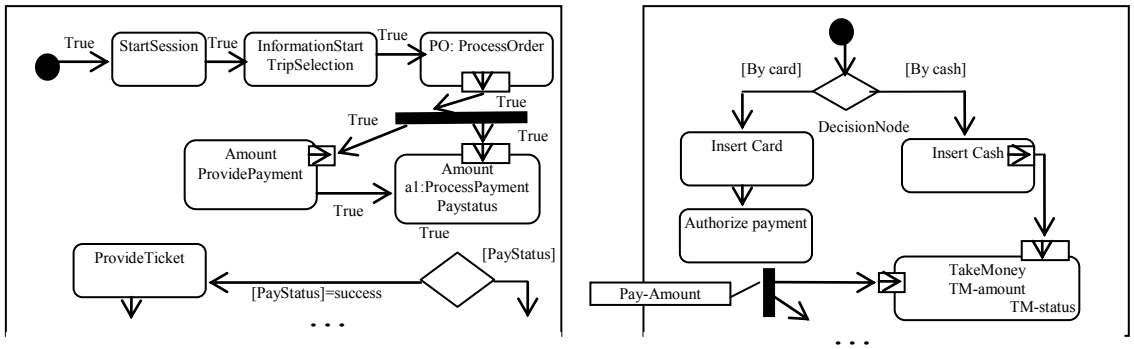
Fig. 3 shows the FM representation of TVM activities. It comprises two main spheres: the passenger (number 1 in the figure) and the TVM (2). The passenger creates a request to start (3) that flows to the machine (4), where it is processed (5) to trigger (6) the generation of a message to input information (7). The message flows to the customer (8) to be processed to create the requested information (9-10), which then flows to the TVM (11). There, the information is processed (12) to trigger a payment transaction that includes

- creating the amount of payment (13) and
- creating the selection of payment options (14).

The payment amount and options flow to the traveler (15) to be processed (16) and to trigger selection of a payment method (17).

The selection flows to the TVM (18), where it is processed (19); depending on the type of payment,

- if the selection is for a cash payment, this triggers (20) the creation of a message to insert cash that flows to the passenger (21) to trigger the passenger to "create" (22; produce) cash that flows to the TVM (23). The cash is processed (24) as follows:
  (a) If the cash is not sufficient, then the TVM creates a message (25) to complete the amount and sends it to the passenger (26).
  (b) If the cash is correct, then the TVM triggers the creation (27) of tickets and sends them to the passenger (28).
  (c) If the passenger decides to cancel the transaction and generates a signal (29) to refund the cash, then the TVM releases (30) the cash back to the passenger.



(a) A main activity diagram for TVM. (b) A subactivity diagram of the payment process.

Fig. 2. Main and subactivity diagrams for TVM (redrawn, partial from [32]).

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

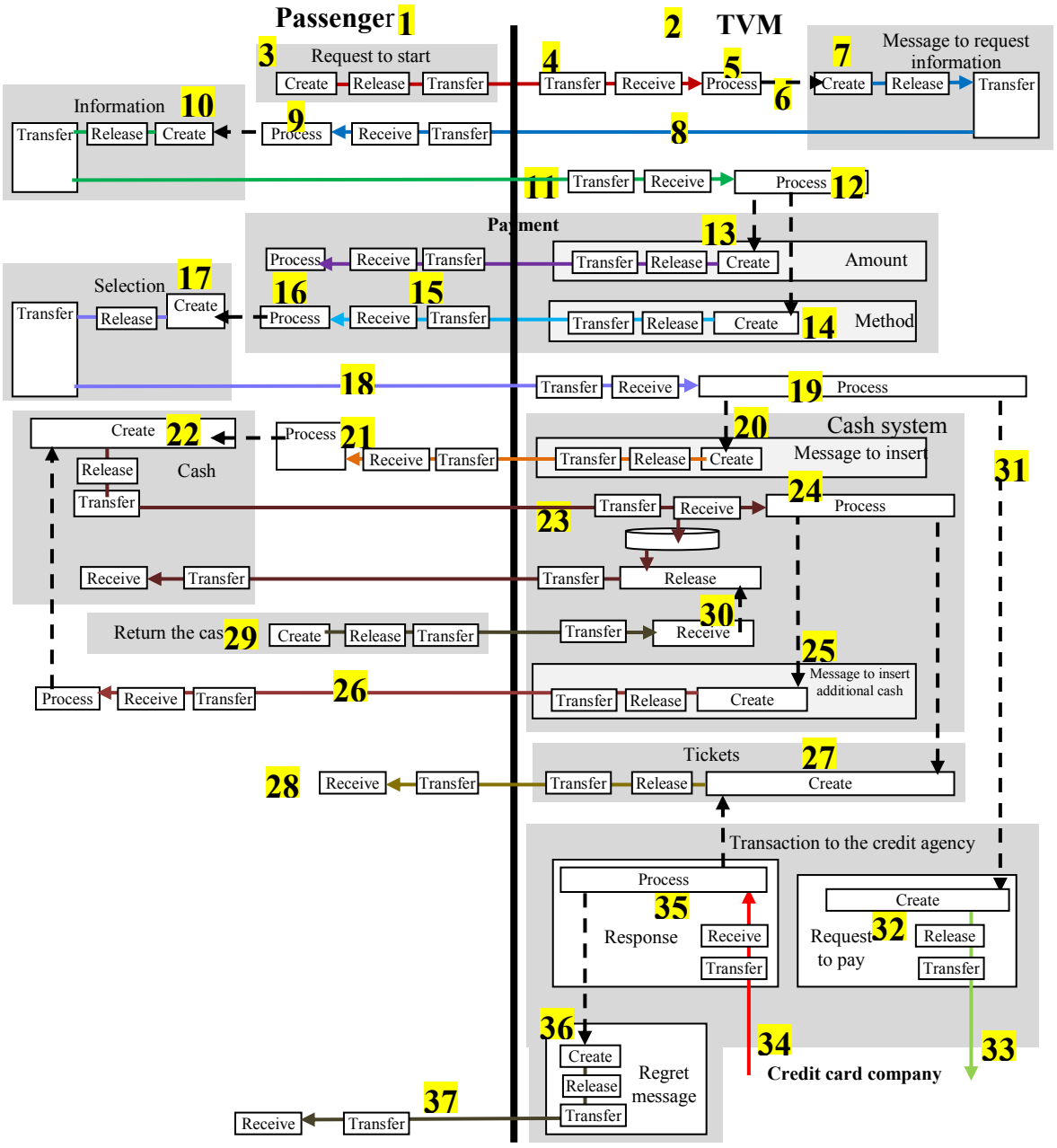


Fig. 3. FM representation of the TVM.

- If the selected payment method is credit card, then this triggers (31) the creation of a request for payment (32) that is sent to the credit card company (33). The TVM waits for a response (34); when it comes, the TVM processes (35) it as follows:
  (a) If the response to the request for payment is positive, then the TVM creates the tickets (27) and sends them to the passenger (28).
  (b) If the response to the request for payment is negative, then the TVM creates (36) a payment declined message and sends it to the passenger (37).

D. *Behavior Description*

Note that Fig. 3 shows a *static schema* that does not embed dynamic behavior. It is a frame that constitutes the region in which events occur, "a possibility of fact—it is not the fact itself" [33]—in which a certain event is mapped to a subdiagram of the network of machines.

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

Behavior description is defined as the entire set of events that a system can perform and the order in which such events can be executed [34]. In system modeling, with FM methodology, behavior is modeled in a phase that occurs after the structural description is complete (e.g., Fig. 3) and involves modeling the "events space." Here, *behavior* involves the behavior of things during *events* when the system framework shown in Fig. 3 is acted upon. The chronology of events can be identified by orchestrating the sequence of events in their interacting processes.

In FM, an *event* is a *thing* that can be created, processed, released, transferred, and received. A *thing* becomes active in events. An event is specified by (1) its spatial area or subgraph, (2) its time, (3) the event's own stages, and (4) other possible qualities, e.g., intensity. For example, Fig. 4 shows the event of a passenger starting a transaction. Note that the region of the event is a subdiagram of Fig. 3. Note also that this event is not itself an elementary event because it is constituted of elementary events such as Create and Release.

E. Control

Accordingly, the entire static representation of Fig. 3 is "event-ized", and the resulting events are utilized to control and manage the system.

For example, to save space, only *selection to pay in cash* of Fig. 3 is event-ized in Fig. 5, with the following events included:

- Event 1 (V_1): The TVM displays the instruction to insert cash.
- Event 2 (V_2): The passenger inserts cash that is received by the TVM.
- Event 3 (V_3): The TVM processes the cash.
- Event 4 (V_4): The TVM displays the instruction to insert more cash.
- Event 5 (V_5): The TVM creates tickets and sends them to the passenger.
- Event 6 (V_6): The passenger sends a request to cancel, hence to withdraw the cash.
- Event 7 (V_7): The TVM returns the cash to the passenger.

Accordingly, control of the chronology of the seven events can be developed as shown in Fig. 6. Going from left to right according to the flow of time,

- V_1, V_2, and V_3 (circles 1, 2, and 3) occur in sequence.
- This sequence is followed by *either* (circle 4) V_4 *or* V_5 (circles 5 and 6).
  - If V_5, then this is the end of the transaction.
  - If V_4, then it triggers (7) the creation of a repetition event (8), i.e., repeating V_2 and V_3. Note that this event has the attribute of possibility (9), that is, it may never occur.

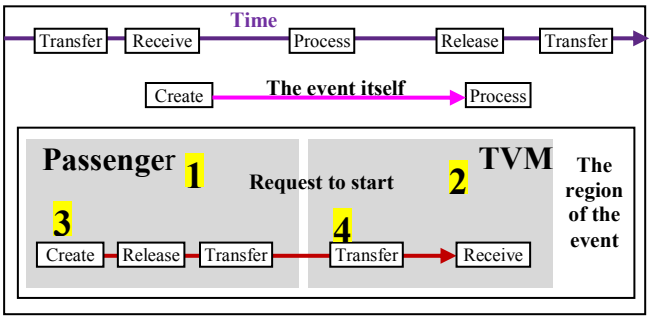


Fig. 4. Event of the passenger starting a transaction.

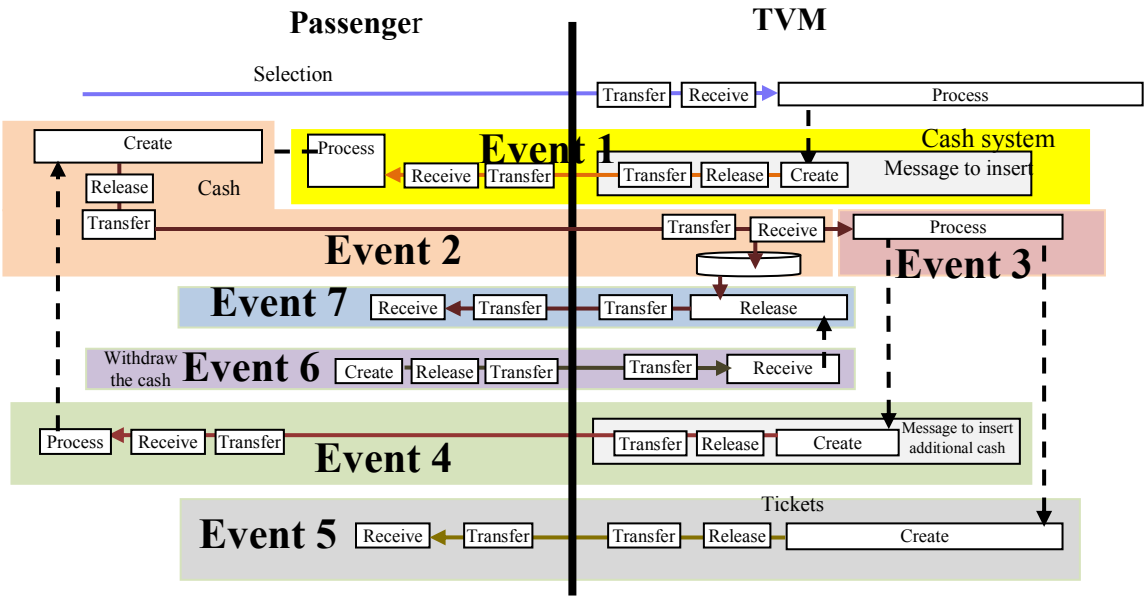
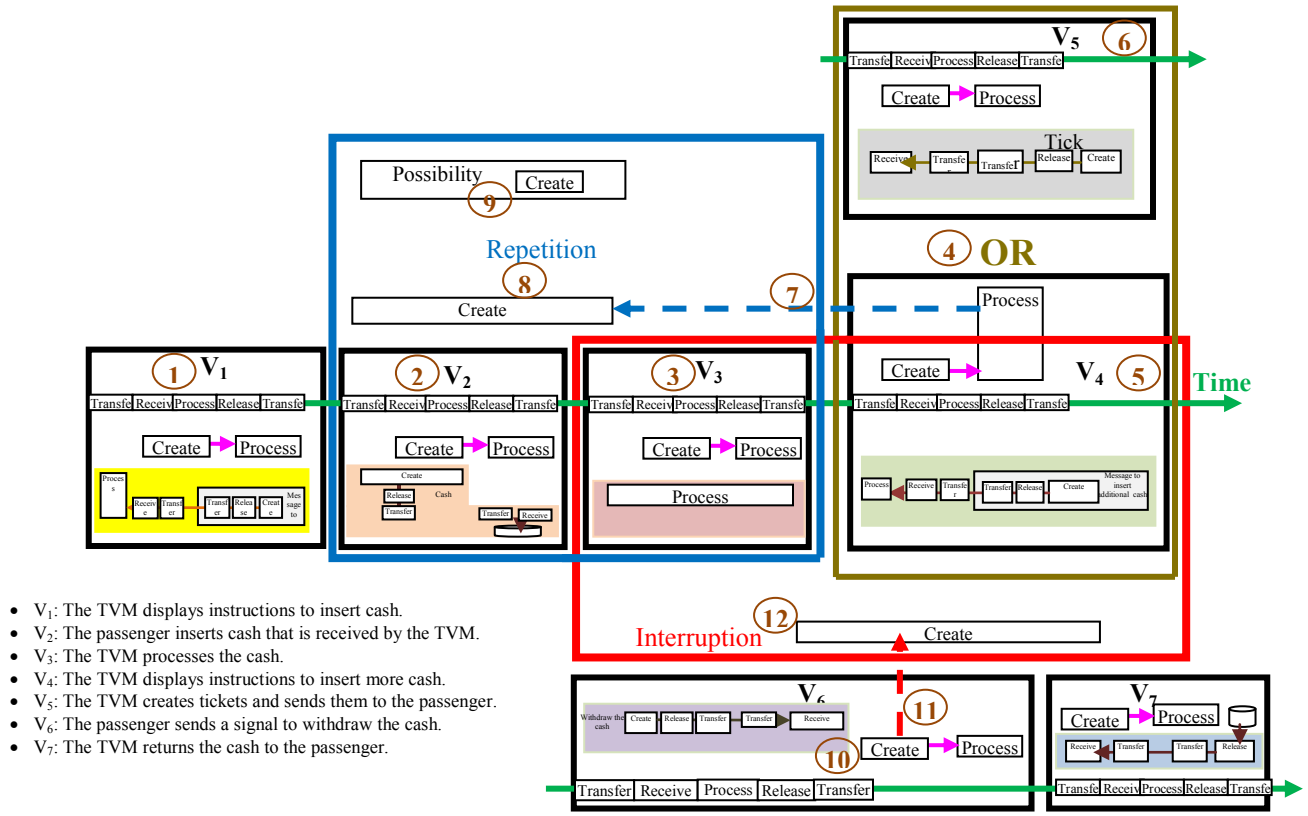


Fig. 5. Some events in the FM representation of the TVM.

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



- V_1: The TVM displays instructions to insert cash.
- V_2: The passenger inserts cash that is received by the TVM.
- V_3: The TVM processes the cash.
- V_4: The TVM displays instructions to insert more cash.
- V_5: The TVM creates tickets and sends them to the passenger.
- V_6: The passenger sends a signal to withdraw the cash.
- V_7: The TVM returns the cash to the passenger.

Fig. 6. FM representation of the control of the chronology of events.

- When the cash is received, V_6 (in parallel with V_3 AND (V_4 OR V5)) is activated (10); thus, when the passenger signals to withdraw the cash, V_6 triggers (11) the interruption (12) of whatever flow machine (V_3, V_4, or V_5) is being executed at that moment, followed by V_7.

Fig. 7 simplifies this control specification by using an extension (e.g., adding a triggering interruption) of the classical specification of a chronology of events. Fig. 8 shows an FM simplification of Fig. 6.

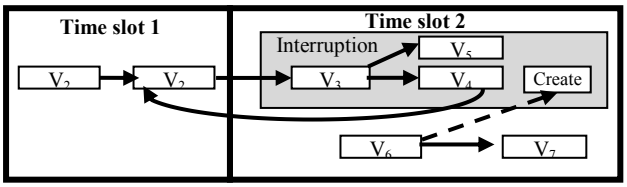


Fig. 7. Classical methods of representing the chronology of events.

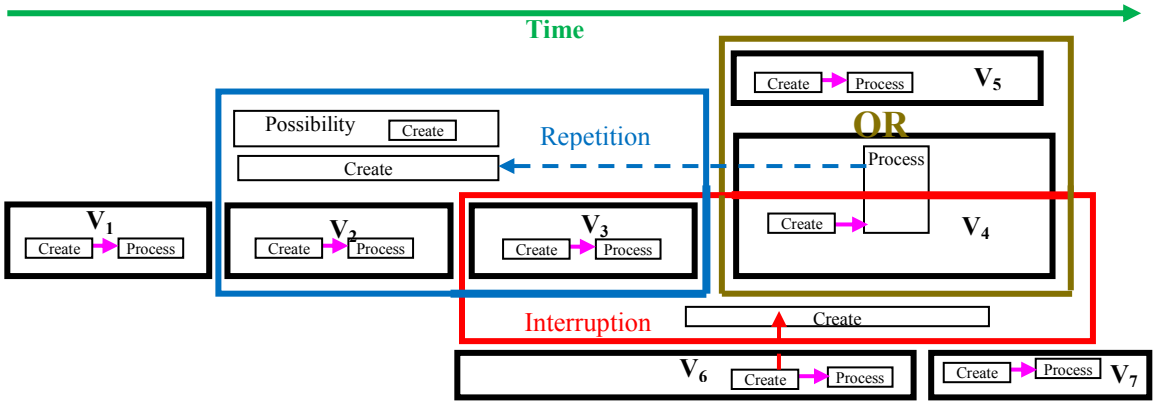


Fig. 8. FM simplification of the representation of control of a chronology of events.

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

F. General Comments

Note that FM modeling encompasses cross-world spheres of physical, digital, and social domains (cf. [4]). It can be used in the so-called cyber-physical system to orchestrate computers and physical systems to control physical processes that include feedback in both directions. It can provide a common model and methods for mechanical, environmental, civil, electrical, chemical, and industrial engineering.

Note that, in general, the control module of a physical system is formed from physical things; e.g., wires carry basic measurement signals, and network components transfer messages between controllers, ports or terminals, and sensors. This also includes computer systems, message broadcasting, and services with request and reply messages.

III. Case Study: Electrical Power Plant

A model can be developed to serve many purposes, e.g., prediction and design. In our case, since the system to be modeled already exists, our purpose is to produce a conceptual representation of the system and its macroscopic behavior that can be used for many purposes, such as (physical and informational) control, maintenance, monitoring, management, and communication. The conceptual model can also help process engineering teams investigating a plant operational crisis, e.g., mysterious pipe vibration issues or problematic pieces of equipment or sections, by using it to simulate operational scenarios and for decision-making.

The system is an electricity-generating plant called Shuaiba South Power and Water Production Station (SSPWPS). The total compound electrical power generated by the plant can reach 804 MW (more details in [35]). An engineering schema of the modeled portion is shown in Fig. 9.

A. First-Level Model

The process of electrical power generation is modeled in Fig. 10, which shows fresh water flowing (circle 1) from the distillation station (not shown in this diagram) to two destination water tanks (2A and 2B).

- The water flows from the two-tank system (2A) to a *pipes/valves assembly* (3), then to a *common header valve system* (4). The *pipes/valves assembly* is a complex of pipes and valves used to control the rate of flow through several pipes. The *common header valve system* is used to unite flows from different sources. Thus, if there is only a single inflow, then other inflows in the figure would not be shown. Hereafter, to simplify the diagram, the interior structure of the *pipes/valves assemblies* and *common header valve system* will not be shown.
- The water also flows to the 2B water tank (6) through the pipes/valves assembly (7), then to the common header valve system (4).

Accordingly, the water in the common header valve system flows (5) to pumps (7) to increase flow pressure to reach another pipes/valves assembly (8), then another common header valve system (9). Simultaneously, the water in water tank 2B (6) flows through pumps (6A) then to a pipes/valves assembly (6B) to join other water in the common header valve system (9).

The mixed water in the common header valve system flows to the following:

(i) The demineralization plant (10)

(ii) The intake expansion tank (11), used to cool down the turbine (36)

(iii) The station water tank (12), where it is stored for firefighting purposes.

The water reaches another pipes/valves assembly (13) inside the demineralization (DM) plant, where it branches, as follows:

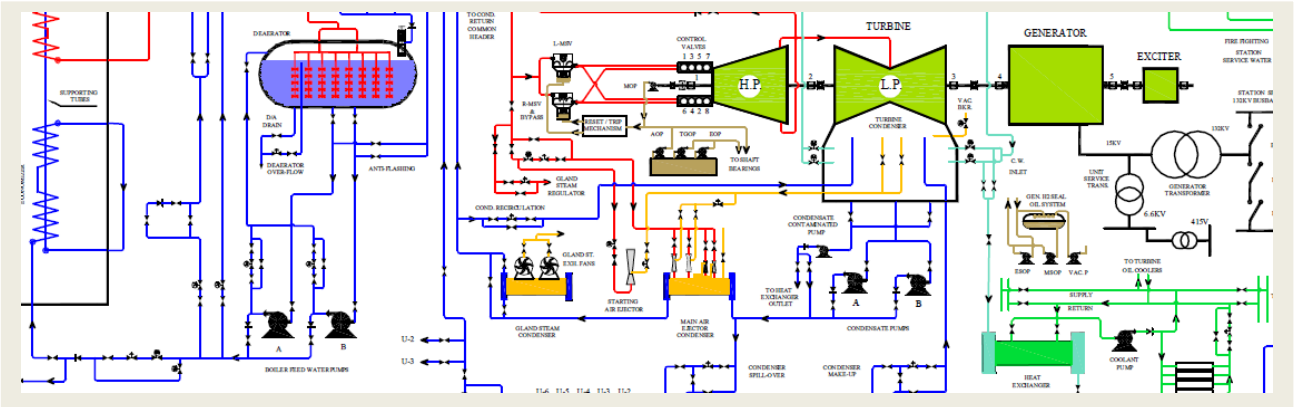


Fig. 9. Partial engineering schemata of the system to be modeled with FM.

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

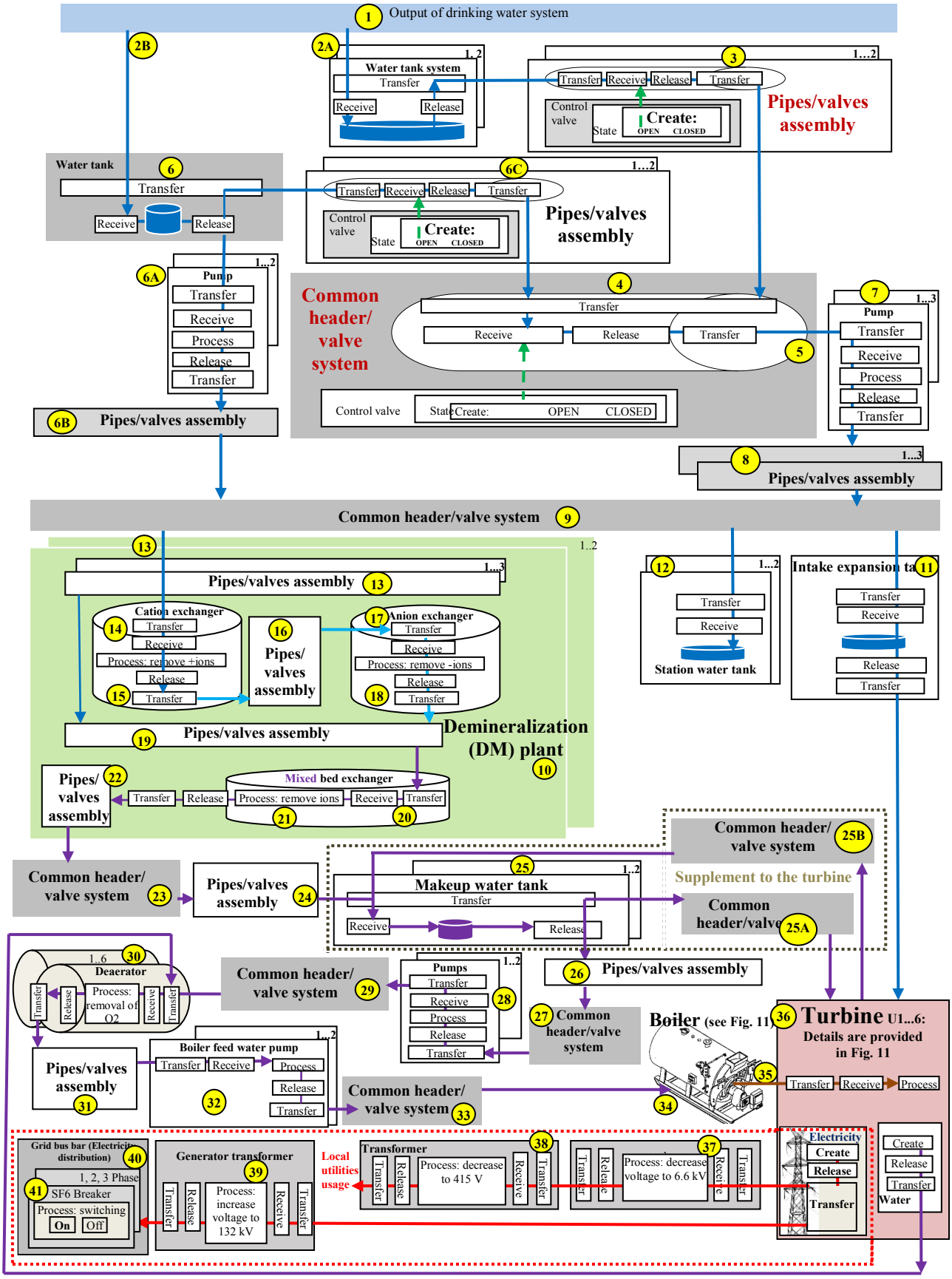


Fig. 10. FM representation of the electrical power plant system.

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

- The water flows to the cation exchanger (14), where the positive ions (cations) are removed (15), after which the water flows to a pipes/valves assembly (16) to reach the anion exchanger (17), where negative ions (anions) are removed (18).
- The resulting deionized water flows through another pipes/valves assembly inside the DM plant (19) before reaching the mixed bed exchanger (20), where *both* cations and anions are removed (21). The deionized water then reaches the pipes/valves assembly (22).

1) Flow of demineralized water: From (22), the water flows through a common header valve system (23), then through another pipes/valves assembly (24) to reach the two makeup water tanks (25). These are used to supplement water needed for the turbine and boiler. Accordingly, the water flows in two directions:

- Supplement component to the turbine
- Supplement component to the boiler

2) Supplement to the turbine: The water flows from the makeup water tank (25) through a common header valve system (25A), then to the turbine (36). The turbine diverts excess water to a common header valve system (25B) that then returns it to the makeup water tank (25).

3) Supplement to the boiler: The water flows from the makeup water tank (25) through a pipes/valves assembly (26) and then to a common header valve system (27) connected to two water pumps (28). From the pumps, the water flows through a common header valve system (29) to the deaerators (30), which remove excess oxygen molecules from the water, removing the bubbles.

Water from the deaerators flows through a pipes/valves assembly (31) to the two boiler feed water pumps (32). The water then flows through a common header valve system (33), arriving at the boiler (34; a detailed FM subdiagram of the boiler will be shown), which produces high-pressure steam (35) that flows to the turbine (36) to lose its energy in running the turbine and converts to water that flows back to the deaerator (30).

Additionally, the high-pressure steam used to generate electricity in the turbine is then processed through a unit stepdown transformer (37) to reduce the voltage from 15 kV to 6.6 kV and send the electricity to another stepdown transformer (38) to further reduce the voltage to 415 V for local utilities usage.

Finally, the electricity generated by the turbine flows to another unit step-up transformer (39) to increase the voltage from 15 kV to 132 kV to be transported to the grid bus bar (40), passing by three circuit breakers (41).

To show that this modeling process can be applied to any level of description using the same technique, the turbine (34-35) will now be described in its own diagram.

B. The Turbine

Fig. 11 shows an FM representation of the turbine. It can be explained as follows:

1) Heating the water: The water from the common header (1; 33 in Fig. 10) reaches the first part of the turbine, the economizer (2). The economizer's function is to reduce the amount of energy needed to convert the water to steam, as follows:

- The water is heated (3) using the surrounding heat generated from the operation of the furnace (4) in order to convert it completely to steam with less fuel than would be necessary with low-temperature water.
- The furnace (4) is connected to 9 burners (5) that are fueled with gas (6). The ignition gun (7) is used to create a spark (8) to ignite the burners (5). Note that the heat generated by the furnace flows to the economizer (3).

In addition, the furnace receives atmospheric air (9). The air (top left of diagram) is sucked from the atmosphere by a forced draft fan (10) in the Boiler Air fuel gas system (11) to flow to the Air Preheater (12). The heated air flows to the damper (13) then (14) to the furnace, keeping the flame burning in the furnace (5). As a result, the burner produces exhaust gases (15) that flow to the damper (16), then to the air preheater (17), which heats the inlet air. The exhaust gases then flow to the chimney (18) to be released to the atmosphere (19).

Creating steam:

The high-temperature water in the economizer (2) flows to a boiler drum (20), where it is converted by heat (21) to wet steam (22) that flows to the primary super heater (23). This steam (24) flows though the attemperator (25) to reach the secondary super heater (26). The attemperator controls the temperature of the steam with water received from the attemperator spray water valve (27) originating from the common header (1).

3.3 Further consideration

Diagrams such as Figs. 10 and 11 can be applied in many areas, with the simplest being documentation, where "Documents are a means to present information instead of being containers of information" [7]. However, here we suggest that the FM diagram is an important tool for maintenance/operations and management.

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

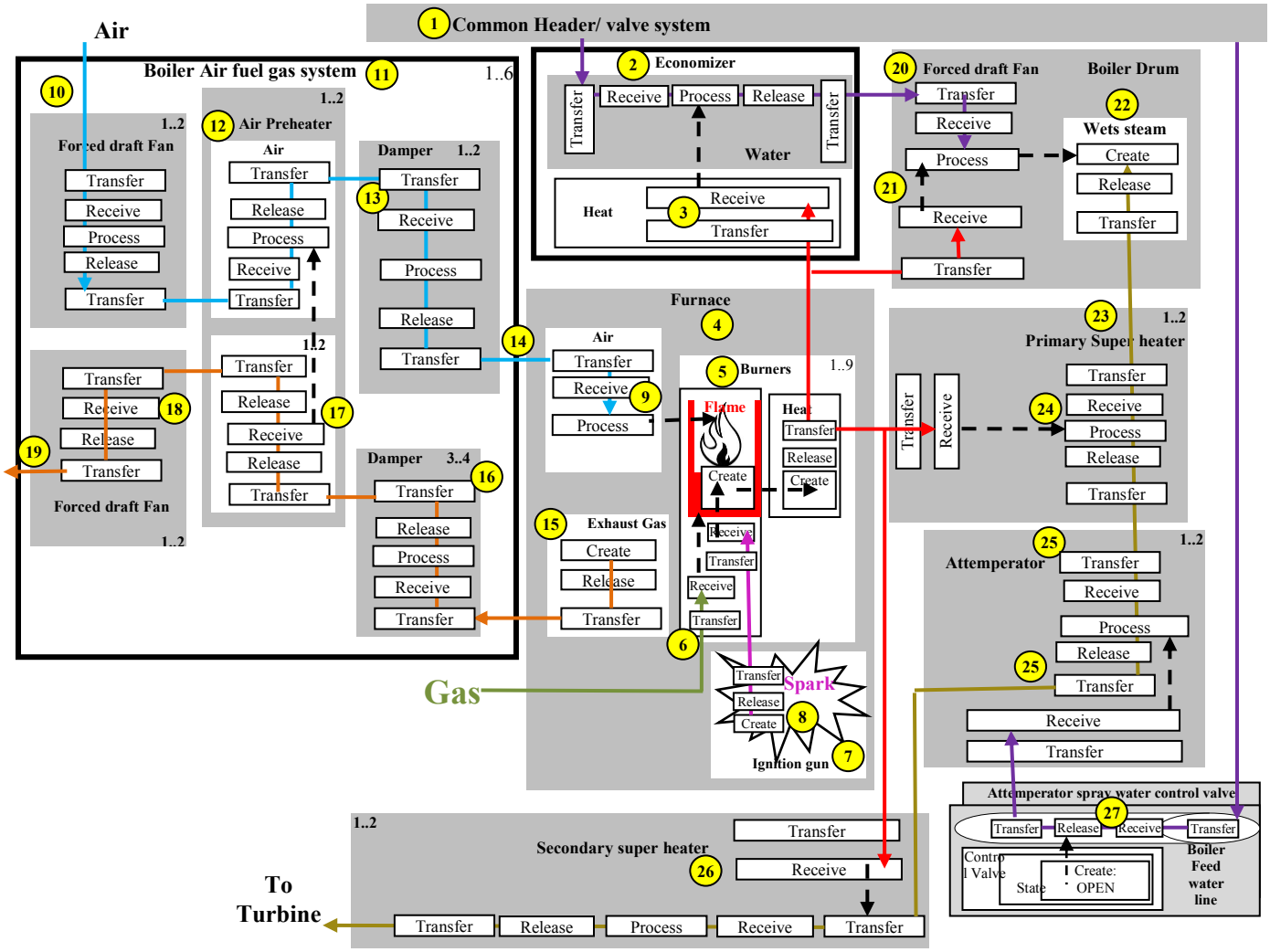


Fig. 11. FM representation of the turbine.

The flows shown in Figs. 10 and 11 can be event-ized as discussed in Section II, according to meaningful events, in order to impose control over different components of the system. Because of space limitation, we focus here on a sample case: the situation of keeping track of parts replacement over time. Equipment needs to be regularly maintained or replaced, and equipment history is a major issue for situations such as scheduled maintenance. In addition, from a conceptual point of view, difficulties arise "When equipment is scheduled for maintenance, it is looked at on an individual basis without evaluating its impact on a system" [36].

There is a need for holistic views and systems thinking in the planning of service and maintenance activities… more efforts are desired to support the development in this direction and to quantify the benefits of being more holistic and flow-oriented [in] the planning of service and maintenance activities. [37].

Clearly, the FM approach with its holistic representation of systems can help with this type of problem. In this section, we briefly demonstrate how FM can be used to conceptualize the situation of "changing parts" over time. According to Tommila and Alanen [7],

Elements of a system may be changed over time without the system losing its identity. Therefore, the elements of a system can be understood as place holders for actual component individuals that, in many cases, are instantiations of a commercial product or device type and have a manufacturer's serial number. For example, [Fig. 12] shows a functional pump object P101.

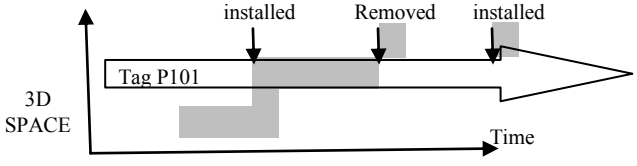


Fig. 12. Replacing a pump (redrawn, partial from [6]).

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

It is distinct from the individual "pump 1" that was first installed as P101 and later replaced by a spare part item "pump 2". Including the time dimension seems useful not only for design and system modeling but also for configuration management and traceability during system operation.

Let us assume that Tommila and Alanen's [7] pump object P101 is one of the two pumps shown at (28) in Fig. 10, shown again in Fig. 13. Fig. 14 shows the FM representation of the history of replacing pump object P101 over time. Note that the same FM notations are used to represent this history.

In the figure, the sphere of pump P101 (circle 1) includes the pump machine itself (2) and the water machine (3) as part of the description of the total system. Event 1 (3) is a "happening" that occurs to that pump during a certain period of time beginning at (5) and ending at (6). The event involves receiving the pump (7) and installing it (8).

This event is followed at a later period of time by event 2, which comprises removal of the pump (9). Note that the occurrence of this sequence of events is represented perpendicularly over the static description of Fig. 10, as reflected by the downward right-angled arrows connecting the flow of time. Similarly, during a later period of time, events occur until event *n*.

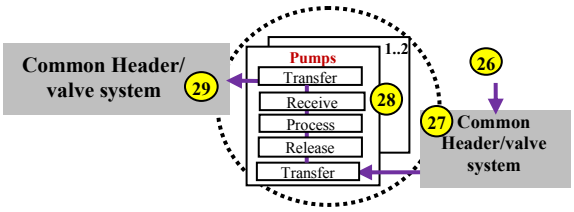


Fig. 13. Portion of Fig. 10 that includes the replaced pump.

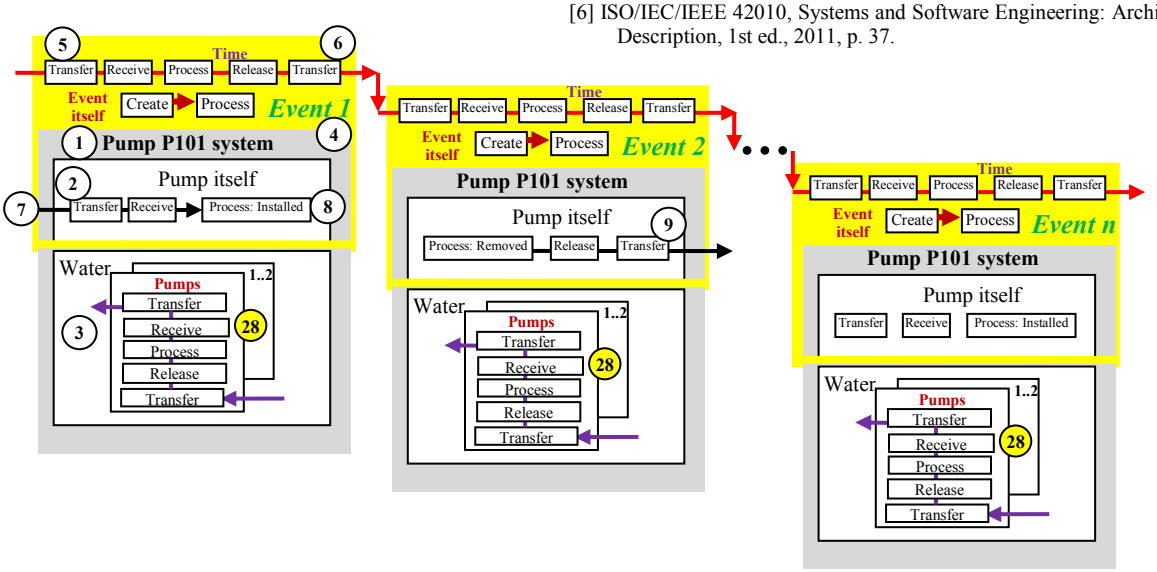


Fig. 14. FM representation of the events of replacing the pump 1 P101.

Each event can include additional information such as who performs the work and name of the maintenance contractor. Thus, the FM diagram "grows" vertically to represent time and to register changes of different parts in the system description. The result is a clear conceptual and orderly foundation of the operations of the system and its changes over time. Of course this foundation would have to be translated into a practical informational and control scheme.

IV. Conclusion

The FM model can be utilized uniformly to describe physical engineering systems and their behavior for purposes of integrating maintenance and operations into a total system that comprises humans, physical objects, and information. Conceptual complexity is resolved through simple, uniform notations applied across macro- and micro-levels of detail. FM diagrams become more complex as specifications become more complete. It is possible to utilize granularity levels, refinement, and zooming to reduce the appearance of complexity.

Still, a great deal of work is needed to apply the FM approach in practical situations. Nevertheless, the FM model seems promising and merits further development in diverse engineering applications.

References

[1] B. Selic, "The pragmatics of model-driven development," IEEE Softw., vol. 20, no. 5, pp. 19–25, 2003.

[2] C. L. Dym, Principles of Mathematical Modeling, 2nd ed. Elsevier Academic Press, 2004.

[3] B. Acock and M. Acock, "Potential for using long-term field research data to develop and validate crop simulators," Agron. J., vol. 83, pp. 56–61, 1991.

[4] R. Wieringa, Conceptual Modeling in Social and Physical Contexts, Centre for Telematics and Information Technology, University of Twente Report, 2008.

[5] J. Mukerji and J. Miller, MDA Guide, version 1.0.1. OMG, 2003. http://www.omg.org/docs/omg/03-06-01.pdf.

[6] ISO/IEC/IEEE 42010, Systems and Software Engineering: Architecture Description, 1st ed., 2011, p. 37.

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

[7] T. Tommila and J. Alanen, Conceptual Model for Safety Requirements Specification and Management in Nuclear Power Plants, vol. 238, VTT, 2015.

[8] ISO/IEC/IEEE 15288, Systems and Software Engineering: System Life Cycle Processes, 1st ed. Geneva: International Organization for Standardization, 2015, p. 108.

[9] A. Kusiak, E. Szczerbicki, and R. Vujosevic, "Intelligent design synthesis: An object-oriented approach," Int. J. Prod. Res., vol. 29, pp. 1291–1308, 1991.

[10] Y. P. Khanal, "Object-oriented design methods for human centered engineering," Ph.D. thesis, Ontario, Canada: University of Western Ontario, 2010.

[11] J. Evermann, "Thinking ontologically: Conceptual versus design models in UML," in Ontologies and Business Analysis, M. Rosemann and P. Green, Eds. Idea Group Publishing, 2005.

[12] Y. Mordecai, "Cyber-physical disruption modeling, analysis, and management: An evolutionary object-process model-based robust systems engineering approach," Ph.D. thesis, Israel Institute of Technology, February 2016.

[13] D. Dori, Object-Process Methodology: A Holistic Systems Paradigm. Berlin: Springer, 2002.

[14] International Organization for Standardization, ISO/PAS 19450:2015, Automation Systems and Integration: Object-Process Methodology. https://www.iso.org/standard/62274.html

[15] D. Embley and B. Thalheim, Eds., Handbook of Conceptual Modeling: Theory, Practice, and Research Challenges. Springer, 2011. doi: 10.1007/978-3-642-15865-0

[16] A. L. Ramos, J. V. Ferreira, and J. Barceló, "Model-based systems engineering: An emerging approach for modern systems," IEEE Transl. Syst. Man Cybern. Part C Appl. Rev., vol. 42, pp. 101–111, 2012. doi: 10.1109/TSMCC.2011.2106495

[17] D. Dori, "Modeling knowledge with object-process methodology," http://esml.iem.technion.ac.il/wp-content/uploads/2011/08/Object-Process-Methodology.pdf

[18] Federal Energy Management Program, "Chapter 5: Types of maintenance programs," in Operations and Maintenance (O&M) Best Practices Guide: Release 3.0 (no date). https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwju4OT_3bHVAhWCzhoKHeTkC38QFggnMAA&url=https%3A%2F%2Fwww1.eere.energy.gov%2Ffemp%2Fpdfs%2FOM_5.pdf&usg=AFQjCNFWokFZLFCHaJIiJ2gz9WxjI6UpgQ

[19] Federal Energy Management Program, "Chapter 3: O&M Management," in Operations and Maintenance (O&M) Best Practices Guide: Release 3.0., (no date). https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=5&ved=0ahUKEwitx9DrwLLVAhVH7xQKHZDoD7cQFgg5MAQ&url=https%3A%2F%2Fenergy.gov%2Fsites%2Fprod%2Ffiles%2F2013%2F10%2Ff3%2FOM_3.pdf&usg=AFQjCNFE0-X4uGh28n7q_8p7SWmAjG-lpQ

[20] B. Al-Najjar, A. Ingwald, and M. Kans, "Maintenance in real estate and manufacturing industries: Differences, problems, needs and potentials: Four case studies," in Proc. 10th World Congress on Engineering Asset Management, K. T. Koskinen et al., Eds., 2015, lecture notes. doi: 10.1007/978-3-319-27064-7_2

[21] S. K. Bhavnani, "The retrieval of highly scattered facts and architectural images: Strategies for search and design," in Automation Construction, vol. 14, 2005, pp. 724–735.

[22] A. C. V. Vieira and A. J. Marques, "Maintenance conceptual models and their relevance in the development of maintenance auditing tools for school buildings' assets: An overview," lecture notes in Proceedings of Maintenance Performance Measurement and Management Conference, 2014.. doi: 10.14195/978-972-8954-42-0_1

[23] S. Al-Fedaghi, "How to create things: Conceptual modeling and philosophy," Int. J. Comput. Sci. Inform. Sec., vol. 15, no. 4, April 2017.

[24] S. Al-Fedaghi, "Context-aware software systems: Toward a diagrammatic modeling foundation," J. Theor. Appl. Inform. Technol., vol. 95, no. 4, 2017.

[25] S. Al-Fedaghi, "Flow-based provenance," Informing Sci., vol. 20, 2017.

[26] S. Al-Fedaghi, "Securing the security system," Int. J. Sec. Appl., vol. 11, no. 3, pp. 95–108, 2017.

[27] S. Al-Fedaghi, "Business process modeling: Blueprinting," Int. J. Comput. Sci. Inform. Sec., vol. 15, no. 3, pp. 286–291, 2017.

[28] S. Al-Fedaghi, "Toward a philosophy of data for database systems design," Int. J. Database Theory Appl., vol. 9, no. 10, 2016.

[29] S. Al-Fedaghi, "Function-behavior-structure model of design: An alternative approach," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 7, 2016.

[30] S. Al-Fedaghi, "Heraclitean ontology for specifying systems," Int. Rev. Comput. Softw. (IRECOS), vol. 10, no. 6, 2015.

[31] S. Al-Fedaghi, "Toward flow-based semantics of activities," Int. J. Softw. Eng. Appl., vol. 7, no. 2, pp. 171–182, 2013.

[32] M. Rahim, M. Boukala-Ioualalen, and A. Hammad, "Petri nets based approach for modular verification of SysML requirements on activity diagrams," International Workshop on Petri Nets and Software Engineering, Tunis, Tunisia, June 23–24, 2014, a satellite event of Petri Nets [The 35th International Conference on Application and Theory of Petri Nets and Concurrency].

[33] G. Deleuze, Cinema I: The Movement-Image, Transl. H. Tomlinson and B. Habberjam. Minneapolis: University of Minnesota Press, 1996.

[34] J. C. M. Baeten, "A brief history of process algebra," Theor. Comput. Sci., vol. 335, no. 2–3, pp. 131 & 146, 2005.

[35] MEW (Ministry of Electricity and Water), accessed June, 15. http://www.mew.gov.kw/en/?com=content&id=73&act=viewhttp://www.mew.gov.kw/en/?com=content&id=73&act=view

[36] L. Obispo and M. W. Gage, "Equipment maintenance and replacement decision making processes," project report, Industrial and Manufacturing Engineering, San Luis Obispo: California Polytechnic State University, 2013.

[37] M. Gopalakrishnan, A. Skoogh, and C. Laroque, "Simulation-based planning of maintenance activities in the automotive industry," [Proceedings of the 2013 Winter Simulation Conference].

AUTHORS PROFILE

Sabah Al-Fedaghi holds an MS and a PhD in computer science from the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, Illinois, and a BS in Engineering Science from Arizona State University, Tempe. He has published two books and more than 270 papers in journals and conferences on software engineering, database systems, information systems, computer/ information privacy, security and assurance, information warfare, and conceptual modeling. He is an associate professor in the Computer Engineering Department, Kuwait University. He previously worked as a programmer at the Kuwait Oil Company and headed the Electrical and Computer Engineering Department (1991–1994) and the Computer Engineering Department (2000–2007).

Abdulaziz Alqallaf holds Bachelor's and Master's in computer engineering from the Department of Computer Engineering, Kuwait University. He has been working since 2015 as a computer engineer in the Instrument Maintenance Department, Ministry of Electricity and Water, Kuwait. His interests include computer networks, security and software engineering.

Design and Simulation of a Bio-inspired Hyper-heuristic Generic Model for Solving Combinatorial Optimization Problems

Sangeetha Muthuraman
Research Scholar,
Department of Computer Science and Engineering,
Manonmaniam Sundaranar University
14, V.O.C. Street,
Kamaraj Nagar, Puducherry - 605011
00918903864543
sangsasi@yahoo.com

V. Prasanna Venkatesan
Professor and Head,
Department of Banking Technology,
School of Management,
Pondicherry University, Kalapet,
Puducherry - 605014
00919486199939
prasanna_v@yahoo.com

***Abstract*—The approaches used in literature for solving combinatorial optimization problems have applied specific methodology or a specific combination of methodologies to solve it. However, less importance is attached to modeling the solution for the given problem systematically. Modeling helps in analyzing the various parts of the solution clearly, thereby identifying which part of the methodology or combination of methodologies applied is efficient or inefficient. In order to find how efficient the different parts of the applied methodology is or methodologies are, it may be better to solve the given problem using the notion of hyper-heuristics. This can be done by solving the different parts of the given problem with many different methodologies realized, implemented and benchmarked, enabling to choose the best hybrid methodology. A theoretical model or representation of the problem's solution may facilitate clear proposal and realization of the different methodologies for the various parts of the solution. The literature reveals that there is a need for a generic model which could be used to represent the solution for combinatorial optimization problems. Therefore, inspired by the basic problem solving behavior exhibited by animals in their day to day life, a new bio-inspired hyper-heuristic generic model for solving combinatorial optimization problems has been proposed. To demonstrate the application of this generic model proposed a problem specific model is derived that solves the web services selection/composition problem. This specialized model has been realized with a trip planning case study and the results are discussed.**

Keywords-hyper-heuristics; model for solving combinatorial optimisation problems; generic model; bio-inspired model; hyper-heuristic model

I. INTRODUCTION

Numerous approaches are available to solve combinatorial optimization problems. This includes applying exact, heuristic, meta-heuristic or hybrid meta-heuristic [7], [11]-[18], algorithms or techniques. All these approaches use some specific algorithm or combination of algorithms to solve the given problem. That is, the solution to the given problem is frozen. However, it may be better to solve the given problem by appropriately modeling its solution and implementing the modeled solution using the notion of hyper-heuristics. Using a hyper-heuristic notion enables the various parts of the model to be implemented using different algorithms, each of which can then be benchmarked. Benchmarking helps in comparing the performances of the algorithms implemented for every part of the model and identifying which algorithm is most efficient for this part of the model. This may enable in choosing an optimal combination which solves the problem best for almost all its instances, from the different combinations of efficient algorithms. To achieve the above said goal systematically it is essential to represent the solution to a problem in the form of a model, which may enable the understanding of the solution as well as the various parts that form this solution better. Going one step further and making everything general, it may be better to have a generic model, which could be used to represent the solution for combinatorial optimization problems, in an abstract manner from which problem specific models could be derived.

The literature reveals that there is a need for such a generic model, which can clearly represent the problem's solution and increase the understanding of the strategies used to solve the problem in these hybrid arrangements. Therefore, inspired by the problem solving behavior exhibited by animals in their day to day life, by selecting appropriate actions which result in appropriate behavior during appropriate situations, a new bio-inspired hyper-heuristic generic model for solving combinatorial optimization problems has been proposed. As this proposed model is a generic model it only gives a general representation of the concept; Different problem specific models could be derived by using the concept of this generic model as a base. In this work, a specific model to solve the web services selection/composition problem has been derived from the generic model that is proposed and its correctness is demonstrated using a case study.

The rest of this paper is organized as follows. Section 2 presents the background for the proposed work and Section 3 presents the motivation behind this proposed work. Section 4 discusses the newly proposed bio-inspired hyper-heuristic generic model for solving combinatorial optimization problems. Section 5 describes the application of the proposed generic model to get a specific model for the web services selection/composition problem. Section 6 presents the implementation details of the specific model by considering a trip planning web services selection/composition scenario as case study. Section 7 discusses the results of the case study. Section 8 concludes the paper and suggests some future enhancements.

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

II. Background

The literature reveals that there are a variety of techniques used by researchers to solve combinatorial optimization problems [7], [11]-[18]. Since these problems are NP-Hard, recently researchers have adopted hybrid meta-heuristic approaches to solve such problems. Therefore, an extensive study has been made on using hybrid meta-heuristic approaches for solving combinatorial optimization problems [7] by collecting research resources between the years1990 - 2015. These research resources using hybrid meta-heuristic approaches for solving combinatorial optimization problems seen in literature could be categorized into three different categories [7] based on the work presented in the research resources studied.

The first category is the proposal of a hybrid meta-heuristic technique or methodology that can be used to solve combinatorial optimization problems. Only a few papers proposed belong to this category and in most of these papers the authors have proposed cooperative schemes between meta-heuristics and exact, heuristic or meta-heuristic methods [32], and then some authors have analyzed, every scheme proposed with one or more examples [30]. The idea behind is to solve the given problem in the best possible manner, which has led to such cross-fertilization produced by combining different optimization techniques. Some authors have categorized these hybridizations [21].

The advantage of such hybrid meta-heuristic technique or methodology proposal is that it demonstrates how the various techniques can be combined, serving as an initial point for researchers planning to develop hybrid meta-heuristics. It also illustrates such co-operations in a detailed manner which may help in further discussion and implementation in future papers.

The limitation of this category is that it is only technique or methodology proposal which was not implemented. Therefore, there seems to be only a theoretical approach which constitutes mixing of different techniques or methodologies, whereas practical implementation and the results are hypothetical. Also, no attempt is made to identify the contribution of different techniques to the overall strategies performance.

The second category of resources had both a proposal and implementation of a hybrid meta-heuristic algorithm. Most of the research resources studied in the survey belongs to this category. A few papers that belong to the second category include the planning of medium-voltage power distribution systems [29], Hybrid meta-heuristics algorithms for task assignment in heterogeneous computing systems [24], GELS-GA: Hybrid Metaheuristic Algorithm for Solving Multiple Travelling Salesman Problem [16], A two-phase hybrid meta-heuristic for the vehicle routing problem with time windows [23]. To solve such problems efficiently, the authors have proposed hybrid meta-heuristic algorithms which combines guided local search and large neighborhood search [29] as in case of medium-voltage power distribution systems, a Hopfield neural network (HNN) is combined with Genetic algorithm (GA) and simulated annealing (SA) as in case of the task assignment in heterogeneous computing systems [24], a hybrid algorithm called GELS-GA is used in which of the global search capabilities of GA with that of the local search of GELS algorithm has achieved optimality as in Multiple Travelling Salesman Problem [16]. In addition, some problems had more than one objective functions as in case of the vehicle routing problem with time windows which has two objective functions [23]. The first one is the minimization of the number of vehicles (primary criterion) is achieved by means of an evolution strategy in the first phase and the second one is the total travel distance (secondary criterion) which is minimized using a tabu search algorithm in the second phase.

The advantage of this algorithm based approach is that since combinatorial problems were solved by implementing a specific hybrid meta-heuristic algorithm, its performance could be analyzed when compared to using any exact, heuristic or meta-heuristic approach which is used to solve the same problem. Most of the time significant improvements in performance were noted [16], [23], [24], [29]. Efficient behavior and higher flexibility were significant benefits obtained by cleverly combining the different algorithms [31] from both the meta-heuristic as well as outside the meta-heuristic field.

The limitations of this category include that there seems to be no modeling of the solution seen. Some of the authors recommend that in future the researchers should make an attempt to have a sound scientific experimental methodology consisting of theoretical models for describing the hybrid meta-heuristic algorithm used, as done in natural sciences [21]. Such models would bring more clarity in terms of solution representation especially when different combinations of hybrid meta-heuristics were used to solve a given problem.

The third category presented an implementation of a framework for solving the problem. Only a few papers proposed belong to this category. Examples of papers that belong to the third category include Hybrid Metaheuristics for Generalized Network Design Problems [25], Hybrid Meta-heuristic Frameworks: A Functional Approach [15], which is a project in computational biology which makes use of meta-heuristic methods to identify homology between proteins etc. The Generalized Network Design Problems solved in this work includes the Generalized Minimum Spanning Tree Problem (GMSTP), the Generalized Traveling Salesman Problem (GTSP), the Generalized Minimum Edge Biconnected Network Problem (GMEBCNP), and the Railway Traveling Salesman Problem (RTSP). To solve the above said problems the authors have combined linear and integer programming techniques like branch and bound, branch and cut, etc. with that of meta-heuristics which compute approximate

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

solutions but usually require significantly less runtime. In order to identify the homology between proteins [15] five major types of meta-heuristic algorithms have been selected for hybridization. A framework and library of combinators to allow the implementation and hybridization of these algorithms has been created.

The advantage of this approach is that it helps in achieving the benefit of hybridization to some extent. That is since a number of libraries of algorithms or toolkits were created in the framework, it fastens the process of identifying or testing the different combinations of hybrid meta-heuristics for the given problem with different combinations of parameter settings which is otherwise a time consuming task.

The limitation of such framework based approaches include specificity in terms of the problem or problems that this framework could be used to solve, limitations of the specific exact, heuristic and meta-heuristic algorithms implemented in the framework, the language chosen for implementation etc. In addition, in framework based approaches there were no concrete modeling of the solution for the given problem seen.

III. Motivation

Although a significant growing interest has been seen in the field of hybrid meta-heuristics it is still in its basic stages and has sufficient space for new developments and research [32]. In general, the literature comprised of either a theoretical methodology proposal or some form of implementation (algorithm or framework). It is necessary to do a detailed examination of the problem to solve, before the selection of an appropriate solution technique that could be used to solve it [22]. One clear and concise way of examining the given problem and understanding its various parts is by depicting the problem in the form of a model. Therefore, there is a need for having an experimental methodology consisting of theoretical models [21] of the solution which is then implemented. This modeling gives better clarity of the proposed solution and may help in enhancing the understanding of the different parts which constitute the proposed solution. The different parts of this solution could then be implemented using different exact, heuristic, meta-heuristic or hybrid meta-heuristic algorithms, each of which could be benchmarked, thus enabling the identification of an optimal algorithm for each part. Therefore, this enables in choosing the optimal hybrid combination that solves the problem best for almost all its instances [7].

Thus, by using such a hyper-heuristic notion of solving each and every part of the problem with different implementations and by combining these implementations to form several hybrid solutions, the full benefit of hybridization could be captured. The biggest advantage of using a model is that it would correctly portray the different hybrid methodologies used to solve the same problem from an architectural perspective. Therefore, inspired by the basic problem solving behavior used by animals in their day to day problem solving, by selecting appropriate actions which result in appropriate behavior during appropriate situations, a new bio-inspired hyper-heuristic generic model for solving combinatorial optimization problems has been proposed. This is described in the next section.

IV. Bio-inspired Hyper-heuristic Generic Model for solving Combinatorial Optimisation Problems

The core concept based on which this newly proposed bio-inspired hyper-heuristic generic model functions is animal behavior. Behavioral ecology is the scientific study of animal behavior on the ecological as well as evolutionary basis particularly in natural environments [4]. Any behavior is said to be a combination of one or more actions [10]. When an animal decides to exhibit a particular type of behavior in a particular situation, then it has to decide the actions that it needs to exhibit in order to achieve the behavior. So, behavior can be defined as a set of one or more actions made by individuals, organisms, systems or artificial entities in conjunction with the other systems or organisms around as well as the physical environment [10]. The most important decision that an animal has to make in a day to day basis is to choose the apt behavior for the apt situation. This in turn involves choosing the appropriate actions that compose the behavior as well. Therefore, choosing an apt behavior and an apt set of actions that composes it is basic and important for the survival of any animal.

The new bio-inspired hyper-heuristic generic model proposed uses the methodology followed by animals in enabling an accurate decision making of choosing the apt behavior and an apt set of actions for the current situation. Accordingly, the model is divided into three major components. They are the nervous system, motor system and the memory or information/knowledge store. The nervous system is once again divided into the nature-nurture function and, allothetic and idiothetic controls [2].

- Idiothetic control: If the behavior or selection of actions of an animal for a particular situation depends purely on the internally registered or stored information then this behavior is driven by idiothetic control.
- Allothetic control: If the behavior or selection of actions of an animal for a particular situation depends purely on the external information then this behavior is driven by allothetic control.

Therefore, the behavior can either be driven by pure idiothetic control or allothetic control or a combination of both, depending on the situation at hand and the problem to be solved. Both the allothetic and the idiothetic controls receive information from the external environment. The allothetic control passes the external information as it is to the actions selector

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

and the behavior selector. The idiothetic control uses a nature–nurture function to select the appropriate internal information from the brain or memory of the animal with respect to the current situation. The nature function is in charge of providing information with regard to the genetic contribution in molding the behavior. The nurture function provides information with regard to the learning and experience that an animal has gained from the environment [8], in modeling the behavior.

There are a lot of behavioral differences, seen from one animal to another, for the same situation, even if they belong to the same species. This results due to the difference in experience. So, learning from experience is a major element that brings behavioral flexibility, which is an important aspect which needs to be taken into account. Therefore, learning can be classified into three types. They are environment based learning, learning of innate behavior and knowledge or experience based learning. The animal acquires environment based learning from the external information that it gets from its environment. Social learning [1], interactive teaching learning [3] and spatial learning and cognition [4], [5], could be categorized as part of environment based learning. Innate behaviour is usually fixed and has strong genetic influence. Imprinting [4] could be categorized as part of innate behavior learning.

When learning is based on the knowledge or experience that an animal has [6] and the understanding of what would happen by selecting a particular behavior and set of actions, then this is termed as knowledge or experience based learning. Associative learning [4] could be categorized as part of knowledge or experience based learning. Most animals learn all their ways of living, communicating, searching, and foraging etc. as part of, environment based learning, learning of innate behavior and knowledge or experience based learning. All these types of learning are stored in the animal's memory or information/knowledge store which in turn guides the nature-nurture function of the nervous system in choosing the appropriate information required or relevant to the current scenario or situation. Thus, the external information from the allothetic control and the internal information from the idiothetic control are passed to both the actions and behavior selector of the motor system. The actions and behavior selector does the necessary processing and applies a suitable ranking algorithm on the internal or external information or a combination of both to arrive at a set of actions and a corresponding behavior. Thus, the proposed model depicts how an animal makes fast and accurate decisions with regard to appropriate actions and behavior selection. This logic followed by animals in apt behavior selection could be used to solve several combinatorial optimization problems. The Fig. 1 below depicts a bio-inspired hyper-heuristic generic model for solving combinatorial optimization problems.

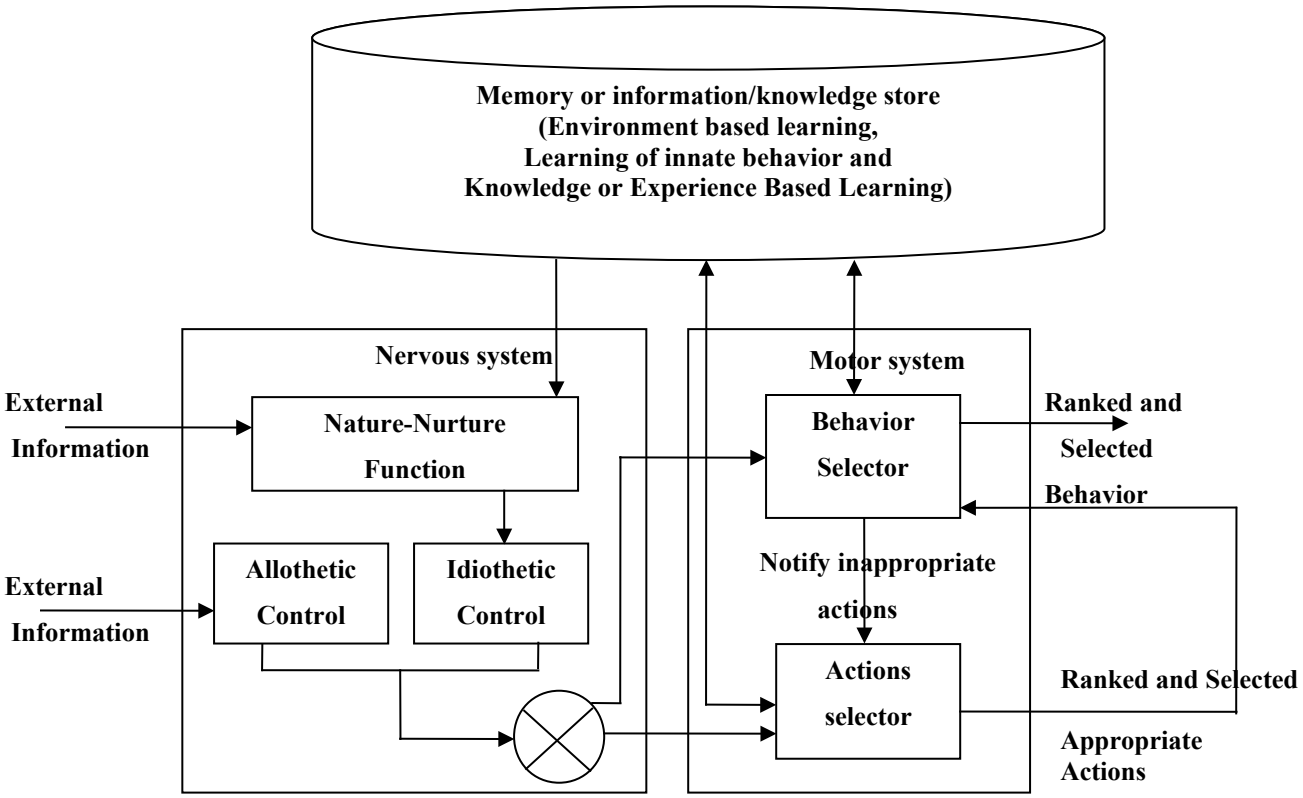


Fig. 1. Bio-inspired Hyper-heuristic Generic Model for solving Combinatorial Optimization Problems

This is a generic model which has been adopted by animals. Depending on how complex the behavior of an animal is the complexity of the algorithms used by the allothetic control, idiothetic control, nature-nurture function and combined actions/behavior selector will also increase or decrease. Ranking algorithms are used in appropriate actions and behavior

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

selection, which ranks the various actions and behaviors and selects the best actions and behavior for the current situation. Thus, the more accurate the ranking algorithms are, the more perfect will be the actions/behavior selection.

The next section details on the application of the proposed generic model to web services selection/composition problem.

V. Web Services Selection/Composition Model

An architecture that uses various mechanisms to solve the problem at hand with the help of a collection of services is basically called as service-oriented architecture (SOA). The fundamental elements for developing applications in SOA are services. These services converse with each other to solve the given problem. This conversation can entail either simple data passing between the services or it could entail some kind of co-ordination of two or more services which thereby harmonize some activity. Therefore the services are connected to one another by some means or the other. This architecture is actually based on two primary actors. They are the service provider and the service requester or customer. The service provider, based on his assignment and available resources, puts together pieces of executable software applications that provide some functionalities and non-functionalities; the service requestor based on his service description, calls and executes such pieces of software that were put together as a whole in order to solve the problem in hand. In real life, there are a lot of services available. Examples include train ticket booking, flight reservation, banking, shopping, paying of household bills etc. It is very important for a service requestor to find and choose the most apt and absolutely necessary services according to their requirements in case of an SOA based environment.

Today, there are a number of services available with the same functionality and this number is going to increase further in the years to come. With this increase in number, the services with the same functionality are selected based on their non-functional properties that suit the customer's request [20]. Most of the time any task requested by a service customer is achieved by using a combination of two or more services. For example train ticket booking may require the processing of ticket enquiry, ticket payment and ticket reservation or booking services to co-ordinate with one another in order to achieve the task. Hence, the process of web service composition is done by selecting the required services based on their functional and non-functional properties and compositing them. Here one apt service is chosen from a set of available services for each activity and these selected services are composed in such a way that the composition meets the consumer's request. Therefore, one of the most important problems that need to be dealt with is the web service selection/composition problem in case of service oriented architecture. To solve this problem of web service selection/composition a specific web service selection/composition model [19] is derived from the proposed bio-inspired hyper-heuristic generic model. The Fig. 2 below gives the web services selection/composition model derived from the proposed generic model.

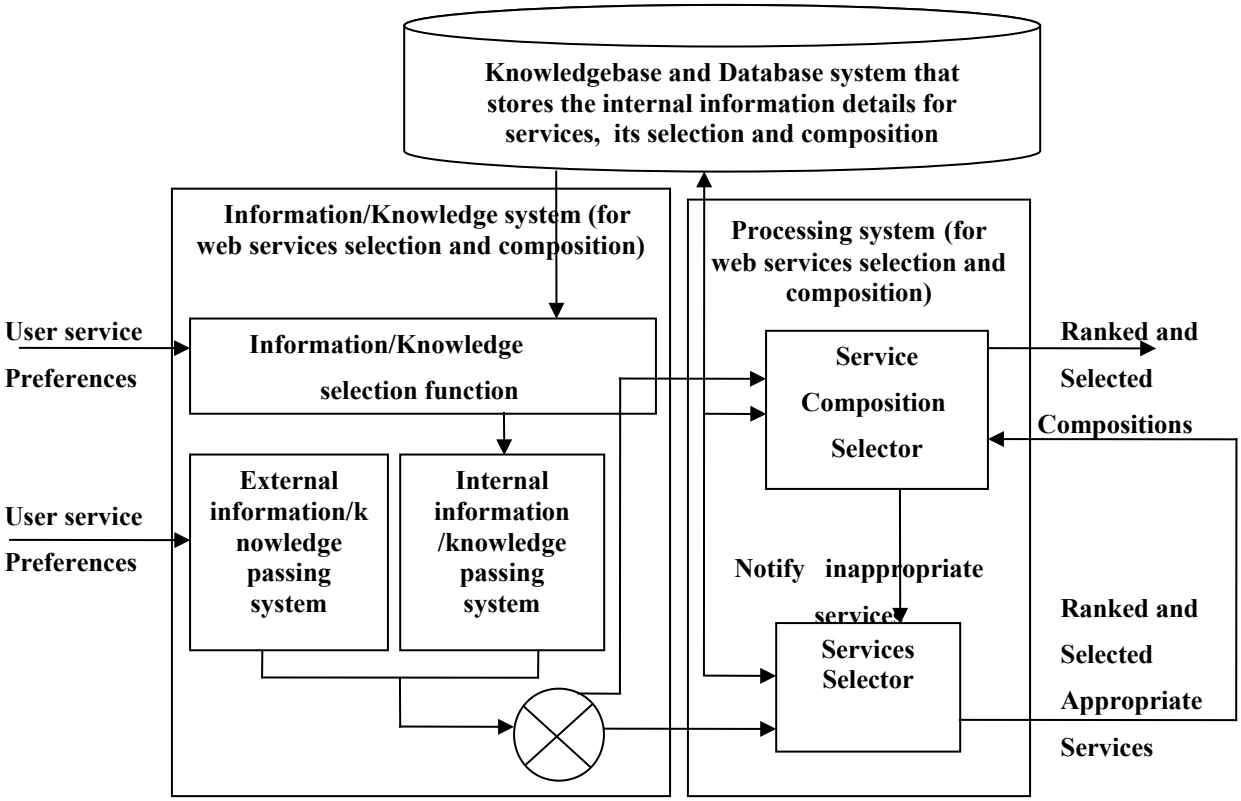


Fig. 2. Web Services Selection/Composition Model derived from the Proposed Generic Model

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

The architecture of this web services selection/composition model comprises of information/knowledge system, knowledgebase and database system and the processing system. The information/knowledge system in turn consists of external information/knowledge passing system, the internal information/knowledge passing system and information/knowledge selection function. The customer's preferences regarding the needed service are accepted from the customer by means of both external information/knowledge passing system and the internal information/knowledge passing system.

1. Internal information/knowledge passing system: This system receives the customer or service user's request. In order to make appropriate decisions for web services selection and composition according to the current request this system uses the Information/Knowledge selection function.

    - Information/Knowledge selection function - The appropriate information/knowledge with regard to the current user's request is retrieved from the database and the knowledge base respectively by this Information/Knowledge selection function.

2. External information/knowledge passing system: This system receives the customer or service user's request. This input is nothing but the customer or user preferences about the various functional and non functional properties of the services requested, that is required for selection and composition processing. This system passes the current state of art information and knowledge as it is to the services selector and composer.

Knowledgebase and Database system: This is used for storing web services related information, knowledge related to previous web services selection and composition that could be reused etc.. The web services related internal information includes the various functional and non functional properties of the services provided by the various service providers which are stored in the service database or registry. The knowledge related to previous web services selection and composition is obtained from the processing of previous similar or dissimilar customer requests. This previous processing results are thus stored in a knowledgebase for reuse whenever necessary. This knowledge is used by the services selector and composer. Additionally, customer feedback and rating for the services provided by different service customers' is also valuable internal information that could be stored and processed during service selection and composition.

Thus, the external information/knowledge passing system passes the request as it is and the internal information/knowledge passing system helps in providing the various services related QOS information, foraged information or information based on experience about the services and compositions stored in the database and knowledgebase. The appropriate information with regard to the current request is obtained from the knowledgebase and database by the information/knowledge selection function. Therefore the information/knowledge system helps in passing appropriate external and internal information required for further processing to the processing system. The processing system consists of services selector and composition selector. These services selector and composition selector selects the appropriate set of services and their corresponding composition by employing some exact, heuristic, meta-heuristic or hybrid meta-heuristic algorithms. Any selection or composition decision made by the selector or the composer depends on the information and knowledge passed by both the external as well as the internal information/knowledge passing system. Under special circumstances either the internal information/knowledge passing system or the external information/knowledge passing system may dominate, but in normal circumstances any decision made by the processing system is a reflection of a combination of both.

VI. Case study and Implementation

The specific web services selection/composition model described above is implemented by considering a trip planning scenario as a case study for service composition [19]. This scenario consists of five different services with functionalities such as hotel booking, airline reservation, train reservation, bus reservation and car rental services. The customer's request any combination of these services for planning a trip, by specifying the functionality and non-functional attributes as their user preferences. The non-functional or QOS (Quality of service) attributes considered in this implementation includes price, execution time, availability, reliability and the reputation of services. These functional and non-functional qualities are received as input from the customers by both the internal and external information/knowledge passing system. The services selection and composition is done by considering these functional and non-functional qualities.

The external information/knowledge passing system passes these inputs as it is to the services selector and composer whereas the internal information/knowledge passing system uses the information/knowledge selection function to select the relevant information/knowledge corresponding to the customer's request from the information/knowledge store and passes this to the processing system. In our implementation, the information/knowledge selection function uses simple database queries to select the appropriate information/knowledge from the information/knowledge store. To do the service and composition selection according to the customer's request, the services selector has been implemented by using a reference score and trust based services selection algorithm and the composition selector is implemented using a strategic tree based composition algorithm [19].

These algorithms are implemented on the Windows XP platform using Microsoft Visual Studio .NET development environment and Microsoft Visual C# as the programming language. The service and QOS registries are implemented as a web service which interacts with the Microsoft SQL Server 2005 database, which is used to store both the functional and non-functional values, service customer requests and provider details. However, the limitations of this implementation include using hard customer constraints and objective QOS values only. When using hard customer constraints, the services do not get selected even if there is a minor deviation in the constraints requested by the customer. Therefore, it may be better to implement the same algorithms by applying soft customer constraints. For simplicity, we have used only objective QOS values. It may be better to use subjective QOS along with this objective QOS as this enables receiving soft customer's constraints in a detailed manner, helping customers to specify the level of requirement for all these QOS attributes and the level of relaxation for every QOS attribute when the level of requirement cannot be achieved.

VII. Results and Discussion

Although importance is attached to testing based web services ranking and reputation based web services ranking [28], so far there seems to be no paper found in literature which has compared the selection algorithm's performance based on the service provider's trust and customer satisfaction as an important factor. With the enormous increase in the number of web services it is necessary to analyze the genuineness of the performance of services (in turn the genuineness of the service providers) which may increase the satisfaction of customers. Hence, the parameters that were considered to evaluate the selection algorithm were the number of successful service selection with and without considering trust, and customer satisfaction on the selected services when selecting it with and without considering trust. Therefore, the selection algorithm was executed by both considering and without considering the service providers trust factor [19]. That is the service providers were ranked based on the genuineness of the QOS values provided by them for their services. The QOS values provided by the service providers were compared with the monitored QOS values. The providers were ranked based on the amount of deviation. Lesser deviation indicates more genuineness. The Fig. 3 below depicts the number of successful service selections with and without considering providers trust for 30, 50 and 100 service requests.

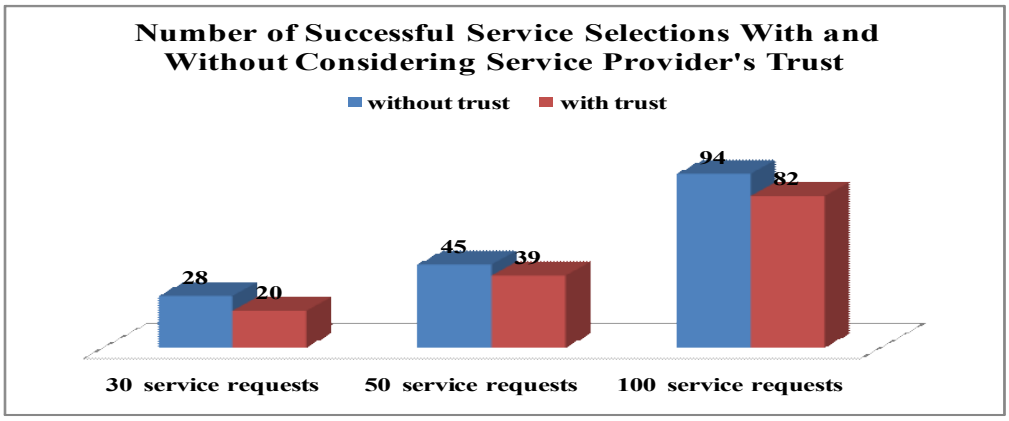


Fig. 3. Successful selection of services with and without considering service provider's trust

For 30, 50 and 100 service requests 28, 45 and 94 services are selected without considering trust and 20, 39 and 82 services are selected with considering trust. Therefore, from the above figure it is clear that number of services selected is more when not considering service provider's trust as an important factor in selecting services and that number of services selected is less when considering service provider's trust as an important factor, in selecting services. But, the number of selected services that customers were satisfied are 18 out of 28, 27 out of 45 and 56 out of 94 when not considering trust and 17 out of 20, 35 out of 39 and 74 out of 82 when considering trust. Therefore, although the number of services selected are said to be more the customer satisfaction was as less as 59% in considering selection without trust. But, the customer satisfaction was 85% or more when trust was considered, as almost the quality values provided by the service providers have matched the monitored quality values. Thus, the Fig. 4 below depicts the percentage of customer satisfaction with and without considering providers trust for 30, 50 and 100 service requests.

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

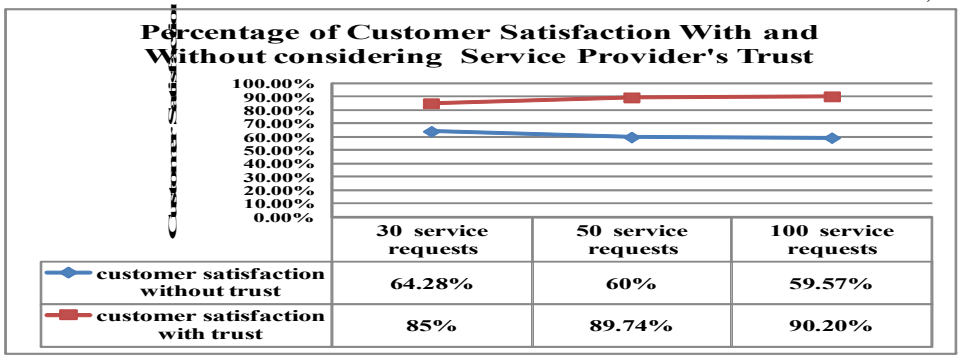


Fig. 4. Percentage of Customer satisfaction with and without considering service provider's trust

The composition algorithm was evaluated in terms of the number of enumerative compositions generated for a single request while selecting the best composition. In general, the composition algorithm proposed performed better when compared to other algorithms seen in literature [26, 27]. When the proposed composition algorithm was compared with a traditional algorithm [26], in terms of the number of enumerative compositions generated for a single composition request of five abstract services as considered in the case study. The proposed algorithm performed well as the number of enumerations it generated was only a maximum of 243 compositions in the worst case [19], whereas the other algorithm generated a minimum of 3125 compositions or even more in the worst case. In the best case only one single composition was generated and on an average analysis the number of compositions generated was less than 100 most of the time. The Fig. 5 below depicts the comparison of the composition algorithm based on number of enumerations for a single composition request in the worst case.

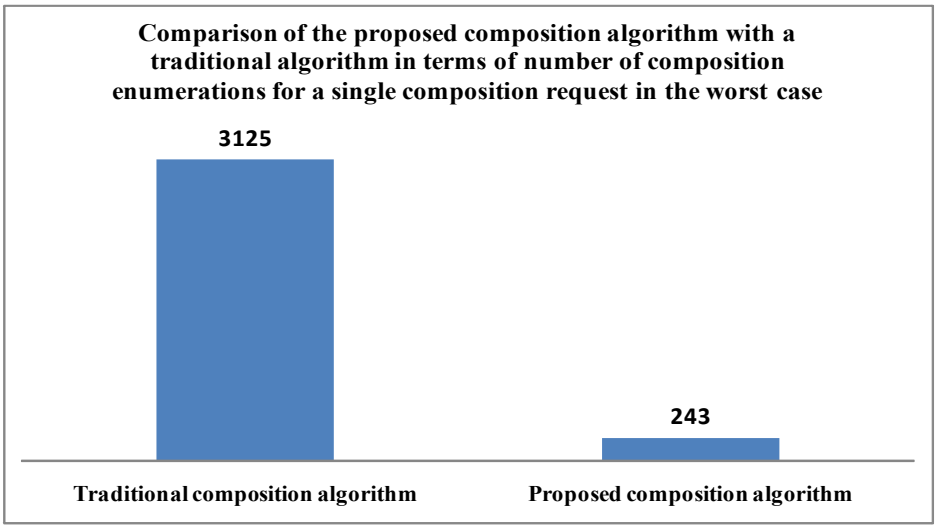


Fig. 5. Comparison of number of enumerations made for a single composition request by the proposed composition algorithm with other traditional algorithms

VIII. Conclusion and Future enhancements

Any classic approach taken in modeling begins with a proposal of a theoretical model, with many new predictions based on observation [9]. Therefore, in this paper a new bio-inspired hyper-heuristic generic model for solving combinatorial optimization problems has been proposed based on the observed problem solving behavior of animals. As this model is a generic model it could be realized into a specific model, according to the problem that needs to be solved. In this work the generic model has been applied to obtain a specific model for the web services selection/composition problem. This specific model has been implemented for a trip planning web services selection/composition case study and the results have been analyzed. Further, in future different algorithms could be applied for the various elements or parts of the web services selection/composition model, thus identifying the most optimal methodology for every part of the model, from which the most optimal hybrid methodology which solves the problem efficiently for almost all its instances could be identified. In addition, the proposed bio-inspired hyper-heuristic generic model could be applied in future to solve various other complex combinatorial optimization problems as well. The biggest advantage of using a model is that it would correctly portray the different hybrid methodologies used to solve the same problem by using the notion of hyper-heuristics from an architectural perspective, when the same problem is solved using different hybrid methodologies.

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

REFERENCES

[1] Lawrence Williams, Martha Glasgow, "Squirrel Monkey Behavior in Research", J. ILAR, vol.41, no.1, pp. 26-36, 2000. http://ilarjournal.oxfordjournals.org/content/41/1/26.full

[2] J.H.Visser, "Host –Plant Finding by Insects: Orientation, Sensory input and Search Patterns", J. Insect Physiol, vol.34, no.3, pp. 259-268, 1988.

[3] Wikipedia, the free encyclopaedia, "Ant - Behaviour and Ecology", http://en.wikipedia.org/wiki/Ant#Behaviour_and_ecology, (accessed 1-9-2015).

[4] Erin Barley, Fathleen FitzPatrick, "Animal Behavior", www.dublin.k12.ca.us/cms/lib5/.../192/Chapter%2051%20**Animal**%20**Behavior**.**ppt** (accessed 1-9-2015).

[5] Susanne Åkesson, Rüdiger Wehner, "Visual navigation in desert ants Cataglyphis fortis: are snapshots coupled to a celestial system of reference?" The Journal of Experimental Biology, vol. 205, pp. 1971–1978, 2002.

[6] eHow Contributor, "Knowledge Based Learning", http://www.ehow.com/about_5403738_knowledge-based-learning.html, (accessed 1-9-2015).

[7] Sangeetha Muthuraman, Prasanna Venkatesan, "A Comprehensive Study on Hybrid Meta-Heuristic Approaches Used for Solving Combinatorial Optimization Problems", in Proceedings of the IEEE Xplore 2016 World Congress on Computing and Communication Technologies (WCCCT 2016), 7th to 9th December 2016, Tamil Nadu, India. pp. 185-200, DOI 10.1109/WCCCT.2016.53.

[8] Andre Ferran, Anthony F.G. Dixon, "Foraging behaviour of ladybird larvae (Coleoptera: Coccinellidae)", European Journal of Entomol., vol. 90, pp. 383-402, 1993.

[9] Christian Blum, Andrea Roli, "Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison", ACM Computing Surveys (CSUR), vol.35, no.3, pp. 268-308, 2003. http://kursinfo.himolde.no/forskningsgrupper/optimering/phdkurs/Metaheuristics%20in%20Combinatorial%20Optimization.pdf

[10] G¨unther R. Raidl, Jakob Puchinger2, "Combining (Integer) Linear Programming Techniques and Metaheuristics for Combinatorial Optimization", Springer Chapter : Hybrid Metaheuristics Volume 114 of the series Studies in Computational Intelligence 2008, pp. 31-62.

[11] Jorg Homberger, Hermann Gehring, "A two-phase hybrid metaheuristic for the vehicle routing problem with time windows", European Journal of Operational Research, Elsevier, vol. 162, pp. 220–238, 2005.

[12] Sancho Salcedo-Sanz,Yong Xub, XinYao, "Hybrid meta-heuristics algorithms for task assignment in heterogeneous computing systems", Journal of Computers & Operations Research, Elsevier, vol. 33, pp. 820–835, 2006.

[13] Bin Hu, "Hybrid Metaheuristics for Generalized Network Design Problems", Ph.d thesis, Vienna University of technology, 2008, https://www.ac.tu**wien**.ac.at/files/pub/**hu**_08.pdf (accessed 1-9-2015).

[14] Pieter Vansteenwegen, Wouter Souffriau, Kenneth S¨orensen, "Solving the mobile mapping van problem: A hybrid metaheuristic for capacitated arcrouting with soft time windows", Journal of Computers & Operations Research, Elsevier, vol.37, no.11, pp.1870-1876, 2010.

[15] Richard James Senington, "Hybrid Meta-heuristic Frameworks : A Functional Approach", Ph.d thesis, The University of Leeds School of Computing, February 2013, e**theses**.whiterose.ac.uk/4847/1/Richard%20Senington's%20**thesis**.pdf

[16] Ali A. R. Hosseinabadi, Maryam Kardgar, Mohammad Shojafar, Shahaboddin Shamshirband, Ajith Abraham, "GELS-GA: Hybrid Metaheuristic Algorithm for Solving Multiple Travelling Salesman Problem", in Proceedings of the 14th Int. Conf. Intelligent Systems Design and Applications (ISDA), pp. 76-81, 2014.

[17] Thibaut Vidal, Maria Battarra, Anand Subramanian, Gunes Erdogan, "Hybrid Metaheuristics for the Clustered Vehicle Routing Problem", Journal of Computers and Operations Research, vol.58 no.C, pp. 87-99, 2015. http://arxiv.org/pdf/1404.6696.pdf

[18] Sherif A Masoud, Scott J Mason, "A bi-criteria hybrid metaheuristic for analysing an integrated automotive supply chain", Journal of the Operational Research Society, advance online publication, 30 September 2015, vol.67, no.3, pp. 516-526, 2016.

[19] Sangeetha Muthuraman, Prasanna Venkatesan, "Design of QOS based Web Service Selection/Composition Hyper-Heuristic Model", in Proceedings of the International Conference on Informatics and analytics (ICIA'16), 25th – 26th August 2016, Pondicherry Engineering College, Pondicherry, India, 2016, ACM, ISBN: 978-1-4503-4756-3.

[20] Sangeetha Muthuraman, Prasanna Venkatesan, "Qualitative and Quantitative Review of QOS based Web Services Selection and Composition Techniques", in Proceedings of the International Conference on Informatics and analytics (ICIA'16), 25th – 26th August 2016, Pondicherry Engineering College, Pondicherry, India, (2016), ACM, ISBN: 978-1-4503-4756-3.

[21] Christian Blum, Jakob Puchinger, Gunther R. Raidl and Andrea Roli, "Hybrid metaheuristics in combinatorial optimization: A survey", Journal of Applied Soft Computing, vol. 11, pp. 4135–4151, 2011.

[22] Van Valdhuizen D.A., Lamont G.B., "Multiobjective Evolutionary Algorithms: Analyzing the State-of-the-Art", Evolutionary Computation, vol. 8, no. 2, pp. 125-147, 2000.

[23] Jorg Homberger, Hermann Gehring, "A two-phase hybrid metaheuristic for the vehicle routing problem with time windows", Elsevier, European Journal of Operational Research, vol.162, pp. 220–238, 2005.

[24] Sancho Salcedo-Sanz,Yong Xub, XinYao, "Hybrid meta-heuristics algorithms for task assignment in heterogeneous computing systems, Elsevier, Journal of Computers & Operations Research, vol. 33, pp. 820–835, 2006.

[25] Bin Hu, "Hybrid Metaheuristics for Generalized Network Design Problems", Ph.d thesis, Vienna University of technology, 2008, https://www.ac.tu**wien**.ac.at/files/pub/**hu**_08.pdf

[26] Yuan-sheng Luo, Yong Qi, Di Hou, Lin-feng Shen, Ying Chen and Xiao Zhong, "A novel heuristic algorithm for QOS-aware end-to-end service composition", Journal of Computer Communications, vol.34, no.9, pp. 1137-1144, June 2011.

[27] Ping Wang, Kuo-Ming Chao and Chi-Chun Lo, "On optimal decision for QoS-aware composite service selection" Journal of Expert Systems with Applications, vol. 37, no.1, pp.440–449, Jan 2010. DOI= https://ir.nctu.edu.tw/bitstream/11536/6100/1/000271571000050.pdf

[28] Vuong Xuan Tran, Hidekazu Tsuji and Ryosuke Masuda, "A new QOS ontology and its QOS-based ranking algorithm for Web services" Journal of Simulation Modelling Practice and Theory, vol. 17, no. 8, pp. 1378-1398, September 2009.

[29] Xiaohu Tao, Hans-Jürgen Haubrich, "A HYBRID METAHEURISTIC METHOD FOR THE PLANNING OF MEDIUM-VOLTAGE POWER DISTRIBUTION SYSTEMS", 15th PSCC, Liege, pp. 22-26, August 2005.

[30] L. Jourdan *, M. Basseur, E.-G. Talbi, "Hybridizing exact methods and metaheuristics: A taxonomy", European Journal of Operational Research, vol. 199, pp. 620–629, 2009.

[31] Christian Blum, Andrea Roli and Michael Sampels, "Hybrid Metaheuristics: An Emerging Approach to Optimization" Studies in Computational Intelligence, Springer 2008 edition, ISBN-10: 354078294X, ISBN-13: 978-3540782940.

[32] Christian Blum, "Hybrid Metaheuristics in Combinatorial Optimization: A Tutorial", A.-H. Dediu, C. Mart´ın-Vide, and B. Truthe (Eds.): TPNC 2012, LNCS 7505, pp. 1–10, Springer-Verlag Berlin Heidelberg 2012.

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

Literature Review: Cloud Computing Security Issues and Techniques

Pawan Kumar
Research Scholar, Department of Computer Science and Engineering
Institute of Technology
Gopeshwar, Uttarakhand, India
Pawankmrarya.2010@gmail.com

Dr Ashutosh Bhatt
Assistant Professor, Department of Computer Science and Engineering
Birla Institute of Applied Science
Bhimtal, Uttarakhand, India
Ashutoshbhatt123@gmail.com

Abstract— **Cloud computing environment is a new way in which web base enable applications provide as a services for the users with low computational cost through internet. As we store data and it also provide services in distributed environment. Cloud ease its users by providing virtualization technology of resources through internet. Cloud computing is the emerging field, due to this reason the various new techniques are still developing. At current scenario new security challenges were increases for cloud professionals. Due to lack of security in cloud computing environment user of cloud lost it trust in cloud. Multi-tenancy, elasticity, Security Performance and Optimization, etc are various security issues in cloud computing. In this paper we will discuss some of the issue in cloud. This paper also discuss some of the existing security technique for securing a cloud and help researchers and professionals to know about various security threats.**

Keywords- Cloud Computing, Security Issues, Security Techniques.

I. INTRODUCTION

The computing undergoes many changes through grid computing to cloud computing. A new computing model proposed by the researchers in computer industry is known as "cloud computing" [1], which commercialize its previous models [2]. Cloud computing environment, is the major achievement of computing, which can bring reform in IT industry. This make the IT industry more attractive and useful to the users and creating the way to designed and purchase in the IT industry [3]. It would also changing the people livelihood and work style. One of the definition of Cloud computing is "a mix approach of grid and utility computing which together form a collection of dynamically interconnected computers. They presented as more unified computing resources. Which is built on service-level agreements (SLA).

As cloud computing is still a new and evolving field it provide new technology for industries. PAAS (platform as a service) and IAAS (infrastructure as a service) types of application are defined in cloud computing. Platform as a service it's provide servers configuration and reconfiguration. Physical/virtual machine is use as a server. On the other side, cloud computing describes application is accessible via internet and for this reason very big data centers and powerful servers are required. Major difference between Cloud computing from tradition computing as it is elasticity, scalability and where the resources are easily provisioned by its users for scaling. It's also provides various level of services to its users.

The paper concentrates on study of cloud computing with several security risk, and its counter measure.

The rest of the paper is organized as follows: Section II Cloud service Model. Section III Cloud deployment model. Section IV Cloud security issues. Section V Technique to secure data in cloud computing. Section VI. Risks and security consideration. Finally, the paper was concluded in section VII

II. CLOUD SERVICE MODELS

A. Software as a Service

Software as a Service sometime referred as "on-demand", is software delivered model in which user can individually provision its resources as requirement without any interaction with cloud service provider. SaaS is typically accessed by customer using a web browser. Saas application are often updated more frequently as compare with traditional software. SaaS has become delivery model for various business applications, likes Payroll Processing, CRM (Customer Relationship management), MIS (Management information System), ERP (Enterprise resource planning) and HRM (Human Resource management and Service).

B. Platform as a Service

Its provide a computing platform and a solution stack as a service. In this service model, the costumer creates the software using tools and libraries from the provider. The service delivery model also provides virtualized servers and associated services for running existing application. The provider provides the server, hardware, storage and networking. The main advantage of PaaS that it allows higher level programming and multiple developers are work simultaneously on a single project.

C. Infrastructure as a Service

Its provides virtualized computing resources over internet and also provide capability to the consumer by which, it can provision processing, storage, hardware, servers and network and other fundamental computing resources where the

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

consumers can deploy and run the software(i.e. operating systems, applications)

III. CLOUD DEPLOYMENT MODEL

A. Public Cloud

It is the computing model based on the standard computing model in which utility computing is available to the general public over internet in payment bases. The main benefits are scalability, resources are properly utilize and inexpensive.

B. Private Cloud

This type of the cloud is dedicated to a single organization. It also provide scalability and self-service.

C. Community Cloud

Community cloud is a multi-tenant infrastructure. In which, the infrastructure of the cloud is shared among several organizations and supports a specific community with common computing concerns.

D. Community Cloud

The cloud infrastructure that is a composition of at least one public and one private cloud.

IV. CLOUD SECURITY ISSUES

Above models and services has various cloud security issue. In most applications, confidential data is stored at servers. Securing data is always vital importance. So many challenges regarding security. Leakage of confidential data fatal many computing systems today. For example, last year marks a peak in data breaches about 740 million records were exposed, the largest number till now.

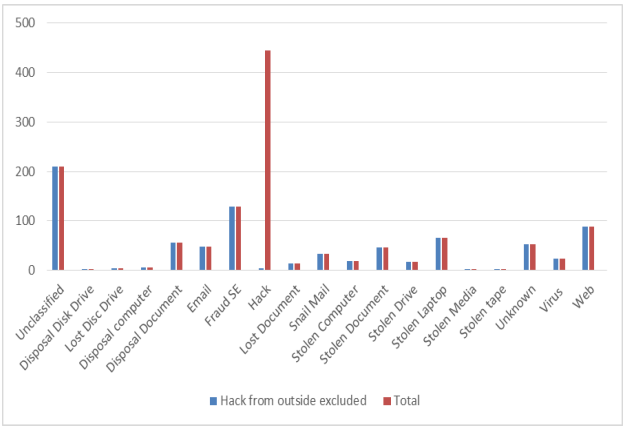


Fig 1. Distribution of data breaches types reported in 2014

A. Multi Tenancy

Multi tenancy is built for reasons like allocation of resources, sharing of memory, storage and distributed computing. It's provide effective utilization [9] of hardware components, and maintain cost is very low. It gives distribution of resources, services and application with other components residing on same physical/logical platform at service providers. Thus it breaches the confidentiality of data and leakage of information and this causes the possibility of attacks.

B. Insider Attacks

Cloud computing is a multitenant based model that is provided by the service provider. So the threat of leakage of information arises within the organization. There is no rules for hiring cloud employees. So an organization can easily hack by the third party vendor, due to this the data of one organization cannot be safe. It's leads loss of information of user, confidentiality, integrity and security. This attack is difficult to defend and the solution of this attack is no found yet [17].

C. Outsider Attacks

This is also one of the major issue in an organization. Data are resided in server and this confidential data of an organization in open to other. In Clouds there many interfaces, so cloud is differ from a private network. One of the disadvantage is that hackers and attackers to exploiting the API, weakness and this result breaking in connection.

D. Elasticity

When a system is adaptable to changing environment. In this resources are provisioned by the user as there requirement. In this synchronization of available resources and current demand occurs. It implies scalability, and users are able to scale up and down as requirement. Due this scaling tenants use a reusable resource.

E. Security Performance and Optimization

The system adopt Security Measures which may affect the performance of underlying services badly. So while applying this security measures we should have check the system performance parameter also. So we should try to make a proper balance between both.

F. Information Integrity and Privacy

In a cloud environment, various organizations put their data on server but some flaws in the security of cloud infrastructure occurs. There is breaches of information privacy, integrity and authentication issues come up.

G. Network level attacks

During resource pooling process all data or services flow over the network needs to be secured from attacker to prevent the breaching of sensitive information or other susceptibilities [10].

a) Man in the Middle attack: It is also a category of eavesdropping. The attacker set up the connection between both victims and makes conversation. Attacker making believe that they talk directly but infect the conversation between them is controlled by attack.

b) Brute force attack: In this attack when attacker want to find the password it will try all possible combination of password until correct password not found.

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

c) Reply attack: In this attack valid data transmission is repeated or delayed due to malicious or fraudulent activity.

d) Distributed denial of service attack: In this attack, servers is down due to huge amount of network traffic. This attack is classified into two broad categories based on protocol level which they targeted one is Network level attack and another is application level attack.

e) Byzantine failure: It is a malicious activity which done at a server or a set of server to degrade the performance of cloud.

f) Network probe: It is used to find out the possible topology of the network which contain IPs and server. Its used to attack for a sub group in the network.

H. Hardware Based Attack

It is one of the most frequently discovered vulnerabilities in cloud which direct result of language and programmes that are as follows.

a) Trojan horses/Malware: They are the unauthorized program that are contained or injected by malicious user within valid program to perform unknown and unwanted function. Unlike viruses it does not replicate themselves.

b) XML Signature wrapping Attack: Protocol like SOAP that use XML format to transfer the request for services are attack by this types of attacks. In this, attack moves the original body of SOAP message to newly inserted wrapping element writing within SOAP header attack perform in new body.

V. TECHNIQUE TO SECURE DATA IN CLOUD COMPUTING

A. Encryption Algorithm

We that cloud service provider encrypt user's data using a strong encryption technique [11] but in some circumstances encryption accidents can make data completely useless and on the other side encryption it also complicated. As this task is challenging cloud provider must provide proof that encryption technique were design and properly tested by knowledgeable and experience authority.

B. Authentication and Identity

The most common method of authentication of users is cryptography. Through cryptography, authentication is provide between communicating systems [13]. Passwords is one of most common form of authentication of users individually. Other form authentication is security token, or in the form a biometric like fingerprint etc. This traditional identity approaches is not sufficient respect to cloud environment. When the enterprise uses multiple cloud service providers (CSPs). In this synchronizing of identity information not scalable. Infrastructure is also one of major concern when we shifting toward traditional approach to cloud-based.

C. Scrutinize Support

Checking of illegitimate activities is a difficult task. When users store their data in the provided cloud they store data in server and they don't have the information where the data is stored. Therefore cloud service provider must provide inspection tools to the users to scrutinize and control various policy implementation.

VI. RISKS AND SECURITY COSIDERATION

As the IT industry more attractive and useful to the users, if implementation of a cloud computing is not managed properly, can present a number of risks to the enterprise. Many of these risks can have a direct impact on business operations, so it is important to take appropriate mitigating in this process. Figure 1 provides a list of the operational risks related to the implementation of Cloud computing.

. Table 1. A comprehensive study on cloud threats and its solutions

Threats	Effects	Affected Cloud Services	Mitigation Strategy
Insecure API and interfaces	Improper authentication and authorization, wrong transmission of content.	SaaS, PaaS and Iaas	Data transmission is in encrypted form, Strong access control and authentication mechanism.
Insider Intruder	Penetrate organizations resources, damage assets, loss of productivity, affect an operation.	SaaS, PaaS and Iaas	Use agreement reporting and breaching notification, security and management process transparency.
Data loss and leakage	Personal sensitive data can be deleted, destructed and corrupted.	SaaS, PaaS and Iaas	Provide data storage and backup mechanism.
Identity theft	Intruder get identity of valid user to access the resources and other benefits of user	SaaS, PaaS and Iaas	Use strong multi-tier passwords and authentication mechanisms
Risk profiling	Internal security operations, security policies, configuration breach, patching, auditing and logging	SaaS, PaaS and Iaas	Acknowledge partial logs, data and infrastructure aspect, to secure data use monitoring and altering system
Shared technology issues	Interfere one user services to other user services by compromising hypervisor	Iaas	Audit configuration and vulnerability, for administrative task use strong authentication and access control mechanisms
Abusive use of cloud computing	Loss of validation, service fraud, stronger attack due to unidentified sign-up	PaaS and IaaS	Observe the network status, provide robust registration and authentication technique

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

VII. CONCLUSION

Cloud computing is the effective technology which depend on cost, time and performance. It gives benefit to the users of cloud and of course the practice of cloud computing will surely will increase more in next few years. In this paper we have discussed and examine the basic of cloud computing and issues regarding securities in the cloud computing. Some security issues are the very crucial in the cloud computing. Privacy and integrity of data are the especially key concern security issues. In the cloud as data is stored in server and we don't know the exact location of the data resided, due to this data stored in the cloud has a threat of being accessed or theft by unauthorized person during transmission.

REFERENCES

[1] I. Foster, Y Zhao, I. Raicu, and S. Lu, "Cloud Computing and Grid Computing 360-degreecompared[C]", in *Grid Computing Environments Workshop*, 2008, pp. 1-10.

[2] Rich Wolski, Daniel Nurmi, Chris Grzegorczyk, Graziano Obertelli, Sunil Soman,Lamia Youseff, Dmitrii Zagorodnov, "The Eucalyptus Open-source Cloudcomputing System ", *2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, CCGRID 2009, pp: 124-131.

[3] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy H. Katz, Andrew Konwinski, Gunho Lee, David A. Patterson, Ariel Rabkin, Ion Stoica, Matei Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing", *Technical Report No. UCB/EECS-2009-28*, 2009.

[4] "NIST Cloud Computing Definition", NIST SP 800- 145.

[5] Enrique Jimenez Domingo and Minguel Lagares Lemos, CLOUDIO: A Cloud Computing-oriented Multi-Tenant Architecture for Business Information Systems In Proc. of the 23rdInternational Conference on Cloud Computing pages 532-533.IEEE, 2010.

[6] D.G. Cameron, R. Carvajal-Schiaffino, A.P. Millar, C. Nicholson, K.Stockinger, F. Zini, Evaluating scheduling and replica optimisation strategies in OptorSim, in:*Proceedings of the Fourth International Workshop on Grid Computing (Grid2003)*, IEEE CS Press, Los Alamitos,CA, USA, Phoenix, AZ, USA, 2003.

[7] Rajkumar Buyya, Rajiv Ranjan, Rodrigo N. Calheiros, "Modeling and Simulation of Scalable Cloud Computing Environments and the CloudSim Toolkit: Challenges and Opportunities", in *The 2009 International Conference on High Performance Computing and Simulation, HPCS 2009*, pp:1-11.

[8] Juefu Liu, Peng Liu, "Status and Key Techniques in Cloud Computing", in *Proceedings of 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)* , pp: V4-285–V4-288.

[9] Chang Jie Guo, Wei Sun, Ying Huang, Zhi Hu Wang, Bo Gao , "A Framework for Native Multi-Tenancy Application Development and Management''2007 9th IEEE International Conference on Ecommerce Technology and The 4th IEEE International Conference on Enterprise Computing, E-Commerce and E-Services..

[10] C. Hong, M. Zhang, and D. Feng, AB-ACCS: A cryptographic access control scheme for cloud storage, (in Chinese), Journal of Computer Research and Development, vol. 47, no. 1, pp. 259–265, 2010.

[11] William Stallings, Cryptography and Network Security Principles and Practice, fifth Edition, Pearson Publication

[12] Enrique Jimenez Domingo and Minguel Lagares Lemos, CLOUDIO: A Cloud Computing-oriented Multi-Tenant Architecture for Business Information Systems In Proc. of the 23rdInternational Conference on Cloud Computing pages 532-533.IEEE, 2010.

[13] D. Feng, Y. Qin, D.Wang, and X. Chu, Research on trusted computing technology, (in Chinese), Journal of Computer Research and Development, vol. 48, no. 8, pp. 1332–1349, 2011.

[14] H. Zhang, L. Chen, and L. Zhang, Research on trusted network connection, (in Chinese), Chinese Journal of Computers, vol. 33, no. 4, pp. 706–717, 2010.

[15] G. Wang, F. Yue, and Q. Liu, A secure self-destructing scheme for electronic data, Journal of Computer and System Sciences, vol. 79, no. 2, pp. 279–290, 2013.

[16] S. Qamar, N. Lal and M. Singh. Deelman, G Singh (2010). Internet Ware Cloud Computing: Challenges. (IJCSIS) International Journal of Computer Science and security, Vol. 7, No. 3, March 2010.

[17] Naresh vurukonda and B.Thirumala Rao, in 2nd International Conference on Intelligent Computing, Communication & Convergence, ICCC 2016,

AUTHORS PROFILE

Pawan Kumar is an assistant professor in deptt. of Computer Science Engg. Institute of Technology at Gopeshwar (India). He received his B. Tech. and M. Tech degrees from Kumaon Engg. College Dwarahat (India) 2010 and 2012, respectively and pursuing Phd from UTU. He once worked as teaching person in Department of IT in Pantnagar University. His research interests include network architecture, computer security, and data analysis.

Dr Ashutosh Bhatt is an assistant professor in deptt. of Computer Science Engg. Birla institute of applied science (India). He completed his Phd in 2009. His work area reasearch was artificial neural network. He has more fourteen year teaching and research experience in various organisations of repute for PG and UG courses of Computer Science and IT. He also associated with many renowed national/international publication as reviewer/editorial board member. More than 22 research publication credited to him.

Intellectual Person Identification Using 3DMM, GPSO and Genetic Algorithm

A.Vijaya Kumar [*1], Dr. R. Ponnusamy [#2]

[*] Research Scholar, Bharathiar University, Coimbatore, India.
E-mail: a.vijayakumar@hotmail.com

[#] Professor, Department of CSE, Sri Lakshmi Ammaal Engineering College, Chennai, India.
E-mail: r_ponnusamy@hotmail.com

***Abstract*—Three-dimensional multimodal models of objective classes are a great tool in modeling and recognition. The multimodal involuntary emotion recognition during a mentally challenged-based communication is presented. We have easily found the mentally disorder people without a doctor. The features are built upon the emotion, motion and frequency to identifying the percentage of mentally disorder peoples. Using Different categories of an image, video, audio and emotions can be discriminated. An image using an algorithms for classification is 3DMM (Three-dimensional morph able models) used to fit the model to images, and a framework for face emotion recognition. GPSO (Guided Particle Swarm Optimization) the emotion finding problem is basically an exploration problem, where at every point; we are pointed to recognize which of the thinkable emotions ensures the current facial expression denotes and GA (Genetic Algorithm) has the virtues of overflowing coding, and decoding, assigning complex information flexibly. GA is calculating the percentage of mental disorder. We proposed using different algorithm to identify the mentally challenged persons.**

***Index Terms*— Genetic Algorithm, Guided Particle Swarm Optimization, Image registration, Multimodal emotion recognition, Three-dimensional morph able models.**

I. INTRODUCTION

An Image processing process is a learning of any algorithm that proceeds an image as input and yields an image as output. Image processing includes as an image displaying, editing the image and enhancing the image, detecting the Feature and image compression. Applications of processing the Image are Medicine, Astronomy, Biology, and Satellite Imagery. Image processing examples are removal of noise, adjusting the contrast value, edge detection, detecting the region, and image compression. A mental retarded is also named a mental illness. The meaning and ordering of mental retarded is key problems for investigators, service workers and the person who may be identified. The clinical terms mentally "retarded", "disorder", “illness”, “challenged are common. A mentally retarded is a disease that reasons, slight to severe disorders in thought and/or activities, causing in an incapability to handle with life's normal stresses and habits. There are other than 200 categorized as the mentally retarded persons. In addition the other common illnesses are sadder, dementia, bipolar sickness, and schizophrenia and nervousness disorders. Indications may include variations in mood, character, individual habits and/or community withdrawal. Psychological health problems may be connected with extreme stress due with an exact situation or sequence of events. As with cancer, heart syndrome and diabetes, mental disorders are often fleshly, sensitivity and mental. Psychological sicknesses may be affected by a rejoinder to ecological pressures, hereditary problems, biochemical variances, or a mixture of these. With proper conservation and handling many folks learn to manage or improve from a mental sickness or sensitive disorder. An image processing is used to find easily the mental disorder, using some features to identify the Mentally Retarded Persons. The features are built upon the emotion means we sense many different feelings every time, like fear, enjoyment and unhappiness. Motion is the process of moving something or varying places, or even just varying position and Frequency defines the quantity of waves that pass a stable place in a given volume of time. Feature extraction is a one kind of dimensionality decline that efficiently signifies interesting fragments of the image as a compressed feature direction. This technique is very much useful when image dimensions are enormous and a reduced feature demonstration is required to fast complete jobs such as image corresponding and retrieval. Detection of feature, extracting, and matching the features are commonly collective techniques that are used to give solution for common laptop vision difficulties namely as detection and recognition of objects, content-based retrieval of image, detection and recognition of faces, and classification of texture. Most Common method of extracting the features that includes Histogram of Oriented Gradients (HOG), Haar wavelets, Speeded-Up Robust Features (SURF), Local Binary Patterns (LBP), and color histograms. For details, see Image Processing Toolbox and Laptop System Toolbox. These toolboxes are used in MATLAB. This programmer can identify emotions of the human from an image. It takings an image, then by skin color separation, it identifies human skin color, then it identifies human face. Then it parts of the eyes and lip of the face. Then it attractions Bezier arc for eyes and lips. Then it relates the Bezier arc of eyes and lips to the Bezier arcs of eyes and lips that are reserved in the data base. Then it treasures the adjacent Bezier arc from the record & gives that record kept Bezier arc emotion as this image sentiment. It based upon extracting the features and emotion detection which is used to identify the percentage of mental disorder and normal person. An image processing using different algorithm to recognize the mentally challenged persons. The various types of algorithms are 3DMM, GPSO and GA. 3DMM (Three-dimensional morph able models). These can be useful in their own exact as an origin for 3D face

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

recognition and investigation involving 3D face records. However, their dominant use terminated the last period has been useful as a tool in 2D face recognition to standardize pose, brilliance and look of 2D face images. A 3DMM has the prerogative capacity that increases the training and trial databases for numerous 2D face processing connected tasks. The method can be beneficial to extend the gallery usual for pose-invariant face matching. For any particular 2D facial image it can deliver corresponding information, in relation to its 3D face outline and feel. It can also support several frames merging by providing the means of action a set of 2D images. A main supporting knowledge for this flexibility is 3D face model into fitting of 2D face image. An improved Particle Swarm Optimization algorithm is also termed as called Guided Particle Swarm Optimization (GPSO) which can performs recognition of facial emotion. A directed Particle Swarm Optimization Algorithm which is another PSO algorithm in improved version is planned in this paper. The method of the algorithm includes tracking the arrangements of 10 Action Units (AUs) which are to be placed at suitable points on the face of an individual. All time two elements are taken in an active unit and by computing the current location and velocity. Genetic Algorithms (GAs) are adaptive empirical search algorithm founded on the evolutionary concepts of natural collection and genetics. As such, they characterize an intellectual manipulation of a random search used to describe optimization problems. Although randomized, gas is by no revenues random, instead they exploit ancient material to direct the exploration into the area of well act within the search space. The Frame of the paper is separated into the following units. Unit II contains the Related Work illustrating the previous methods executed for high resolution of face shape and pattern recognition. In Unit III we have defined our Earlier Work and our Improved Work has been described and analysis of our detection algorithm. Unit IV covers the implementation results. And finally, Section V discuss the Conclusion.

II. RELATED WORK

Ankur Patel and William A. P. Smith [1] in this paper we reconsider the process of constructing a high resolution 3D morph able model of face shape difference. In this Paper how the statistical tools of thin-plate Projections and Produces analysis can be used for hypothesis a morph able model that is both more effective and simplifies to novel face surfaces more accurately than earlier models. We also reformulate the probabilistic preceding that the model delivers on the dissemination of parameter vector lengths. This sharing is determined exclusively by the various model dimensions and can be utilized as a regularization Limitation in suitable the model to data without the need to empirically choose a restriction controlling the trade-off between believability and feature of fit.

Patrik Huber, Zhen-Hua Feng, William Christmas, Josef Kittler [12] A novel fitting method that practices local image structures to fix a 3D Morph able Model to 2D images. To come out of the difficulty of enhancing a charge function that covers a non-differentiable feature extraction operative, we use a learning-based spills regression method that absorbs the incline direction from records. The technique allows to instantaneously solve for figure and position parameters. Our technique is methodically estimated on Morph able Model produced records and first outcomes on real records are presented. Related to old fitting methods, which use humble raw features like pixel color or control maps, local procedures have been exposed to be much extra robust against differences in imaging circumstances. Our method is exclusive in that we are the major to use limited features to appropriate a Morph able Model.

Yas Abbas Alsultanny∗, Musbah M. Aqel [19] the genetic algorithm executed by neural network to decide routinely the appropriate network architecture and the set of limitations from a constrained section of space. The neural-genetic algorithm of multilayer was functional in image processing for pattern recognition, and to decide the object orientation. The collection to cover the opinions of the object was constructed from actual images of (10×10) pixels. Which is the minutest image size can be used in this algorithm to distinguish the kind of aircraft with its way. The multilayer perceptron neural network joined with the genetic algorithm, the outcome exposed good optimization, by tumbling the various secreted nodes essential to train the neural network.

Mr. R. Balakrishnan, Mr. U. Karthick Kumar [10] various heuristic algorithms have been implemented for the grouping problem, which is known as NP Hard. In this research paper explains a Genetic Algorithm for grouping of images. Genetic algorithms have been proposed in an extensive variety of fields to achieve grouping, however, the method normally has an extended consecutively time in terms of contribution size. A heuristic method based on Genetic Algorithms (GA) is adopted to mechanically determine the amount of group centroids during unverified classification. Effective time methods are used as an act measure for grouping of image records.

Rabab M. Ramadan and Rehab F. Abdel – Kader [13] Feature selection (FS) is a global optimization trick in machine learning, which decreases the volume of features, removes unrelated, noisy and completed data, and results in appropriate credit accuracy. It is the most important step that moves the act of a pattern recognition system. In this research paper provides a novel feature selection algorithm made on particle swarm optimization (PSO). PSO is a computational form based on the indication of shared behavior encouraged by the social behavior of fish schooling or bird gathering. This algorithm performs as removal of coefficients by two feature extraction methods are the discrete cosine transforms (DCT) and the discrete wavelet transform (DWT).

Yang Chen*, Yuri Owechko, and Swarup Medasani [18] the image registration is very hard process as a search for optimum process restoration parameters, and demonstrates how to practice Particle Swarm Optimization (PSO) to improve the objective function in a multi-dimensional limitation space professionally. Associated with old-style image registration methods, which are “open-loop” algorithms, the planned method uses registration excellence feedback twist to drive the limitation search while avoiding possible problems shared to open-loop algorithms, specific feature detection, and equivalent and change parameter estimation.

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

III PROPOSED ARCHITECTURE

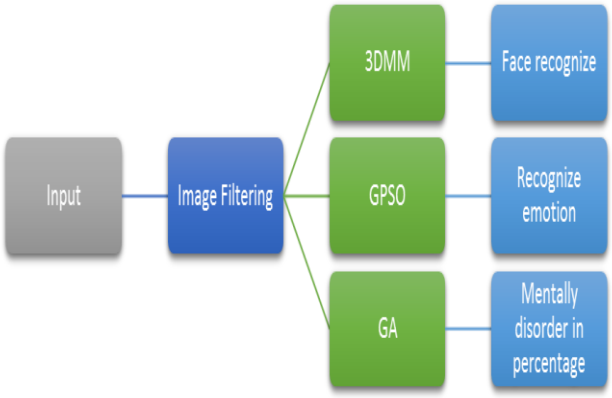


Fig. 1. Overview Architecture

The proposed system the input has been given as a raw image. In image pre-processing we are going to remove a background from a raw image using a slice timer, realignment, normalization, background removal, smoothing. Using a pre-processed image we are going to filter face based images. From the filtered image using a 3DMM we are going to recognize featured elements of the faces like eye, eyebrow, nose, lips, teeth, cheeks, chin, ear, and forehead. Using a position of a feature and the point of the changes is detected using GPSO. Using the MMI Facial Expression dataset we will compare the positions and attitude of the normal and abnormal person images using GA. While we compare we will get the mentally retarded and normal percentages.

Fig. 2. Extensive Architecture

A. *IMAGE FILTERING*

An image is frequently corrupted by chance differences in passion, illumination, or have unfortunate contrast and can't be used perfectly. Filtering is a convert pixel passion, values to reveal certain image features. Enhancement is progressing contrast, Smoothing eliminates noise, Template matching is identifies recognized patterns. The several algorithms for filtering the image and methods used for image filtering and image smoothing. Image smoothing is one of the best significant and broadly used task in image processing. We have described many algorithms and methods for filter the images and which algorithm is the be the greatest for leveling and filtering the images, particularly as we have generally focused on non-linear filtering algorithms i.e. intermediate filtering is very significant in edge preserving. The image may be ruined by chance differences in intensity, differences in illumination or unfortunate contrast that may be distributed within the initial stages of vision processing.

B. *3DMM (Three-dimensional morph able models)*

The propagative model in this learning effort on the 3D Morph able Model (3DMM), and the corresponding implication procedure is termed as the fitting algorithm. 3DMMs have been generally used for face evaluation because the fundamental things of 3D faces offer an ideal illustration which is protected to differences in the face look introduced by the imaging method such as perspective, illumination and occlusion. The 3DMM consists of a distinct face outline and surface models educated from a group of 3D standard faces. The shape and surface models convert interpersonal variant. In accumulation, 3DMM can representation intra-personal variations such as position, clarification, and expression. Given an only facial involvement image, a 3DMM can improve both interpersonal (3D shape and surface) and intra-personal belongings (position, clarification) via a fitting algorithm. It is very inspiring to accomplish an effective and perfect fitting. The experiments are twofold. First, the 3DMM efforts to improve the 3D face outline that has been vanished through the prediction from 3D into 2D. Second, it is problematic to isolate the contributions of face surface and clarification from a single effort image. 3DMMs can improve the invariant facial things, the 3DMM is useful in a variety of presentations in computer graphics and visualization. The presentations of 3DMMs contain face recognition, 3D face modernization, face tracking, facial quality edition. The 3DMM is functional to face recognition and modernization. We take efforts on 3DMM supported face recognition. Face recognition has established major attention over the last insufficient decades. At least two motives account for this movement. First, face recognition has a varied range of presentations such as law implementation, surveillance, smart cards, and access mechanism. Second, some offered face recognition Methods are possible for the applications under precise environments. 3DMM are recognized features of the faces like eye, eyebrow, nose, lips, teeth, cheeks, chin, ear, and forehead.

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

Fig. 3. Feature Extraction

C. 3DMM Feature Based Fitting Algorithms

This explanation adds the regularization continuous to the transverse of the particular values of the SVD factorization of p. It answers the difficulty of valuing shape factors from a sparse usual of 2D structures while also declaring a plausible result. However, the explanation is dependent on the collection of the regularization term η and an exact estimation of the camera matrix, M. As rises the 3D shape approximation inclines to the mean 3DMM shape, or slightly, the norm of the valued constants approach zeroes. Where η is improved to determine its normalizing effect. However, presenting a regularization familiarizes a trade-off between a perceptually even 3D shape approximations and a slight re-projection error between the 3DMM and the mined 2D image features. Where U to reduce with the mapping matrix, α are shape coefficients. This shows an explanation of shape constant valuation from limited information as a pseudo inverse and it is exposed to be plausible even when face simulations were significantly blocked. However, an imperative omission of the technique is that the shape factors are not enforced to be effective shape model parameters. To realize this, the parameters must be limited in some manner, which a pseudo-inversion does not realize. Now SVD (singular value decomposition) can be valuable to calculate the inverse, absolutely in the case PTP, where P =Z • WUT and PT=U.SZT. y = r − LUs. The cost function form can be explained in a straight forward fashion

$$\alpha = \underline{U}\ \mathrm{diag}\left(\frac{w_i}{w_i^2 + \eta}\right) Z^T \mathbf{y} \qquad (1)$$

D. Fitting Morph able Models to n-Views

An important problem with the method offered in is that it is individual resulting for a single set of 2D structures and accordingly can only be useful to a solo image. Opportunely, it is direct to extend the method to approximation shape constants given many images which allows 3DMM fitting to cover to various views of the matching face over time and across pose variations. In the first case the difficulty of estimating 3DMM shape factors using n views is shown. This can be possible to answer for the coefficients α in a frank manner. Unfortunately, the elegance of the SVD (singular value decomposition) characteristics cannot be exploited, because of various instances of the P matrix. Instead a result is derived:

$$\propto = (P_0^T P_0 + P_1^T P_1 + 2I_\tau)\ (P_0^T y_0 + P_0^T y_1) \qquad (2)$$

E. GPSO (Guided Particle Swarm Optimization)

The detection of emotion is essentially a search problem, where at all points; we are mentioned to find which of the possible emotions does the present facial expression denotes. Thus, visibly emotion detection offers itself as an imaginable candidate for PSO application, since PSO is mostly a search algorithm. However, in order to relate PSO to explain the emotion detection problem, we want to first describe the various limitations of the algorithm in relative to the problem. In particular, we want to describe the following:

(1). what is the exploration space and what is its measurement?

(2). How do we characterize an element in the emotion-detection setting?

(3). How do we characterize the location and velocity of an element?

(4). what is the detached function being reduced by the PSO?

A memory that in unit , we have definite our method to the emotion detection problem, which is fundamentally to display the changes in the locations of the action units, located on the face of a question over a period of interval, from which we can then decide the emotion communicated at each point in stage. With this in mind, we define the limitations of the PSO as tracks:

Definition 1: Search space and its measurement

Let the exploit Units (AUs), be denoted by, $r_1, r_2 \ldots r_n$. Let $S_1, S_2 \ldots S_n$ signify the domains of the AUs, $r_1, r_2 \ldots r_n$ respectively. That is S_j represents a 2-dimensional rectangular skylight consisting of the likely points that q_j can be assigned to. And then the search room is a n-tuple,

R_n, given by: $R_n = (S_1, S_2 \ldots S_n)$ (3)

The search space measurement is n, where n is the figure of action units being experiential.

Definition 2: Particle, Position and Velocity

Constituent part, location and speed A subdivision P is a conceptual thing in the R_n search gap that has a point and a quickness and represents a promising solution. The place, $x_{i(t)}$ of an atom, Pi at time t, is a inclusive project of principles $(wal_1, wal_2, \ldots, wal_n)$, where walj ϵ sj. Thus, $x_{i(t)}$ is a vector, $(wal_1, wal_2, \ldots, wal_n)$.The speed, $e_{i(t)}$ of atom i at time t is an n-tuple $(e_1, e_2, \ldots, e_n)$ where e_j represents the speed of the subdivision in dimension s_j. There are two odd issues which make the feeling detection difficulty a less diverse than regular problems to which PSO is practical.

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

First, in ordinary PSO problems, there is frequently one target that all particles in the cloud are tiresome to reach. In our exacting case, though, there are a figure of likely emotions and any one of them might be encountered at any time. In arranging to explain this multi-target problem, we tender to have many swarms, one for all possible sentiment. Since each horde has a diverse goal to reach, the purpose meaning of each crowd must be defined in a different way. We classify the point job of each cloud as the Euclidean reserve among its recent best top and its target. For example, the following is the definition of the objective function for the swarm that is targeting the happy emotion.
Let T = (t_1, t_2… t_n) represents the happy emotion. (Note: this is derived through the training session). Then the objective function for the swarm, $f_s: R^n \rightarrow R$ is defined as,

$$f_s(Z_i(t)) = |Z_i(t) - n| = \sqrt{(l_1 - n_1)^2 + (l_2 - n_2)^2 + \cdots + (l - n_n)^2} \quad (4)$$

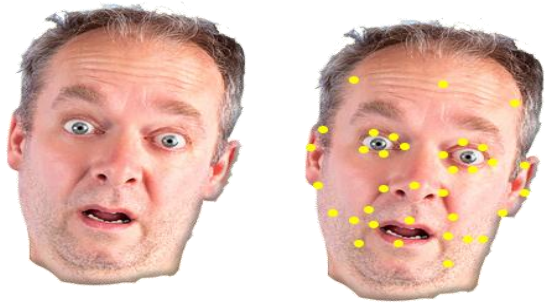


Fig. 4. Emotion detection

F. Genetic Algorithm

To determine the degree of equivalent between two feature sets Q and R, the fitness function is finished by the partial bidirectional Hausdroff distance. The yield of GA is the feature established which has the largest value of the fitness function.

F1. Chromosome Code

The unique population is generated arbitrarily. However, the generating range of chromosomes is not random, but imperfect within the image size. Each chromosome's detailed structure method is certain of the feature collections and coding them into binary codes.
Here binary coding is used because the image resolution is always inadequate. The feature collections can be denoted by integers in the possibility of the image resolution.

F2. Fitness Function

Openly, if the equal degree between Q and R can be measured, it is equal to estimating the fitness of the chromosome. Since the partial bidirectional Hausdorff distance between feature sets Q and R, the reduced the distance is, the equal degree between Q and R is bigger. So, the fitness function can be particular as the converse of the partial bidirectional Hausdorff distance

$$fitness = \frac{1}{a + I_{ML}(Q,R)} \quad (5)$$

Where *a* is a positive, persistent. In order to circumvent the denominator is a zero, the partial bidirectional Hausdorff distance is combined.

F3. Genetic Operators

A set of chromosome is arbitrarily chosen from the population and is used as the parents to repeat the offspring. The selection attitude is the additional chromosome amount of its new offspring to the next group, the larger fitting function value G_i with the greater probability Q_i.

$$Q_S = \frac{G_i}{\sum_{j=1}^{i} G_j} \quad (6)$$

The nature best phenomenon of the biosphere is replicated by Crossing operator is achieved after changing the parent's facts which are developed the selection process. Crossing processing is achieved according to a definite probability, which is called crossing probability Q_c. The crossing consequence is to yield better chromosome after the arrangement of the producing materials of the parents. Here, a particular point crossover operator is accepted. Mutation operates on each binary bit of a chromosome in alternative predefined probability Q_m, the mutation operator is recognized by reversing the significance of the present binary bit, i.e. 1-0, 0-1.

F4. Genetic Algorithm Based Feature Match

For two features sets Q and R, define population size O, crossing probability Qc, mutation probability Qm, fractions gL and gK of the partial bidirectional Hausdorff distance and the determined iterative steps Hmax. Randomly create O feature sets in the check image and then exchange them into chromosomes for first generation.

1. Estimate the qualification of each chromosome in recent population and then construct an original population by repeating next steps until the original population is complete.
2. Use new generated population for a further run of the algorithm.
3. If the end condition is fulfilled, stop, and yield the best solution S best in current 4.
4. Population, where best be the feature sets defined by the best chromosome, and feature sets best (Q) and R according to the unassuming nearest neighbor rule. If the determined iterative step Hmax is not extended, go to Step 2.

IV EXPERIMENTAL RESULT

The set of 80 associated faces was arbitrarily divided into 4 subsets. In harmony with the cross-validation algorithm permutations, these subsets were used to afford implementation and testing data to build 3DMMs for succeeding Shape-Update experiments. Since the Shape-Update attitude to 3DMM suitable from addresses the problem of fitting given a sequence of measurements, the challenging data were used to provide perfect sets of measurement vectors V and camera matrices N. In 3DMM Using Feature Based Face Detection in 3D View with the Help of the 3D Dataset which has been created by us 80 images are divided into 4 subset 50% of the images are

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

Mentally Retarded and 50% of the images are Normal Persons. The test set consisted of 200 pictures of individuals under various illumination and pose conditions. A test set was completed by enchanting images of the six people in the database. The subjects were instructed to rotate their faces in penetration and the lighting conditions were changed by moving a light source around the subject. The component-based face recognition model was compared to a global face recognition model; both the models were instructed and tested on the same images. In testing, we get 90% of the Accuracy of Face Detection using 3DMM.

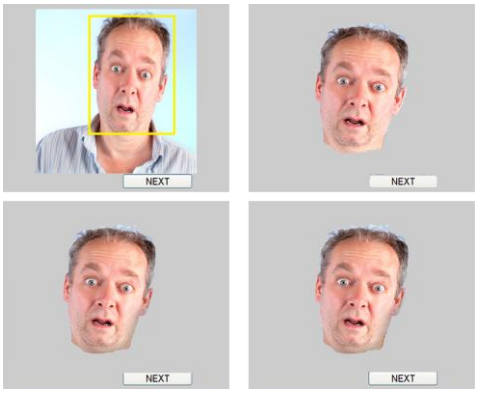


Fig. 5. Output for 3DMM

In GPSO the output of the 3DMM is given as the input into the algorithm. It Detect the Features Position in axis like Right eye (x axis:-20,y axis: 2 and z axis:25) and in the dataset it will have three kinds of data's like Normal Abnormal and Mentally Disorder Position of eyes, lips, nose, forehead, chick etc. GPSO use to detect the emotion of the person in the image. In GA it will compare with the dataset and it will do the crossover, verification and mutation process and it lists all the possible ways of mental disorder and normal person in percentage and give you as output.

.

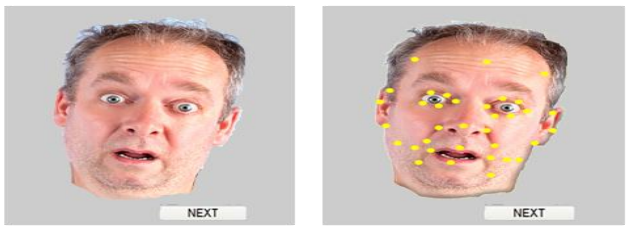


Fig. 6. Output for GPSO

Iterations	True Rate	False Rate
I1	97.0%	3%
I2	97.5%	2.5%

TABLE I : Rate Value of GA Output

V CONCLUSION AND FUTURE WORK

This paper projected a new expansion in component-based facial acknowledgement by the integration of a 3D morph able model into the training method. Medley allowed the training of a facial emotion recognition system which mandatory only three facial images of each person. From these 3 images, 3D face models were figured and then used to extract a numerous amount of artificial images under fluctuating poses and lighting conditions. In GPSO It will recognize the Spot and position of the each feature in the 3 Dimension method. It predicts whether the person is abnormal or normal and detect emotion of the person in the image. Using GA it will chromosome, mutate and find the percentage of the mental disorder and normal. The true rate of the testing Data is 97% true rate and the false rate is 3% in the First Iteration. While in the second Iteration we got 97.5% true rate and false rate is 2.5%. In the Future work we have decided to research this topic with video processing based and we can find the result in high accuracy by motion based detection.

REFERENCES

[1] Ankur Patel and William A. P. Smith, "3D Morph able Face Models Revisited," Department of Computer Science, The University of York, 978-1-4244-3991-1/09/$25.00 ©2009 IEEE.

[2] A. S. Tolba, A.H. El-Baz, and A.A. El-Harby, "Face Recognition: A Literature Review," International Journal of Signal Processing, vol. 2, no. 2, pp. 88-103. 2006.

[3] Bashir Mohammed Ghandi, Ramachandran Nagarajan, Hazry Desa, "Classification of Facial Emotions using Guided Particle Swarm Optimization I," School of Mechatronics Engineering Universiti Malaysia Perlis (UniMAP) 02600 Jejawi, Perlis, Malaysia, Int. J. Computer and Communication Technology , Vol. 1, No. 1, 2009.

[4] Cohen, I., Sebe, N., Garg, A., Chen, L.S. and Huang, T.S (2003), "Facial expression recognition from video sequences - temporal and static modelling," Journal of Computer Vision and Image Understanding, vol. 91, pp. 160-187.

[5] D. Snyers a, y. P6tillot h, "Image processing optimization by genetic algorithm with a new coding scheme," LIASC, Ecole Nationale Sup~rieure des Telecommunications de Bretagne, BP 832, 29285 Brest Cedex, France, b Departement d'Optique, Ecole Nationale Sup~rieure des Tdl~communications de Bretagne. BP 832, 29285 Brest Cedex, France, Received 21 December 1994; revised 3 April 1995.

[6] Jennifer Huang1, Bernd Heisele1, and Volker Blanz, "Component-based Face Recognition with 3DMorphable Models," Center for Biological and Computational Learning, M.I.T., Cambridge, MA, USA,Honda Research Institute US, Boston, MA, USA, Computer Graphics Group, Max-Planck-Institut, SaarbrÄucken, Germany.

[7] Joaquim Jose Furtado1* , Zhihua Cai1 & Liu Xiaobo1, "DIGITAL IMAGE PROCESSING: SUPERVISED CLASSIFICATION USING GENETIC ALGORITHM IN MATLAB TOOLBOX," China University of Geosciences, 388 LuMo road, Wuhan, Hubei, P.R. China. Zip code 430074, Report and Opinion 2010; 2(6).

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

[8] L. Zhang and D. Samaras, "Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics," IEEE Trans. Pattern Anal. Mach. Intell., 28(3):351–363, 2006.

[9] Miss. Komal R. Hole1, Prof. Vijay S. Gulhane, Prof. Nitin D. Shellokar, "Application of Genetic Algorithm for Image Enhancement and Segmentation," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 4, April 2013.

[10] Mr.R.Balakrishnan, Mr.U.Karthick Kumar, "An Application of Genetic Algorithm with Iterative Chromosomes for Image Clustering Problems," Department of MCA & Software Systems, VLB Janaki Ammal College of Arts and Science, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012 ISSN (Online): 1694-0814.

[11] Omran, M.G., Engelbrecht, A.P. and Salman, A. (2005), "Particle Swarm Optimization Method for Image Clustering," International Journal on Pattern Recognition and Artificial Intelligence, vol. 19, no. 3, pp.297-322.

[12] Patrik Huber, Zhen-Hua Feng, William Christmas, Josef Kittler, Matthias, "Fitting 3D Morphable Models using Local Features," Centre for Vision, Speech and Signal Processing University of Surrey, Guildford GU2 7XH, UK, R¨atsch Reutlingen University D-72762 Reutlingen, Germany, 1https://github.com/patrikhuber arXiv:1503.02330v1 [cs.CV] 8 Mar 2015.

[13] Rabab M. Ramadan and Rehab F. Abdel – Kader, "Face Recognition Using Particle Swarm Optimization-Based Selected Features," Electrical Engineering Department, Faculty of Engineering - Port-Said, Suez Canal University, Port Fouad 42523, Port-Said, Egypt, International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 2, No. 2, June 2009.

[14] S. Romdhani and T. Vetter, "Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior," In Proc. CVPR, volume 2, pages 986–993, 2005.

[15] Sousa, T., Silva, A. and Neves, A. (2004), "Particle Swarm Based Data Mining Algorithms for Classification Tasks," Parallel Computing, vol. 30, no. 5-6, pp. 767-783.

[16] van der Merwe, D.W. and Engelbrecht, A.P. (2003), "Data Clustering using Particle Swarm Optimization," Proceedings of IEEE Congress on Evolutionary Computation, vol. 1, pp. 215-220.

[17] Volker Blanz, Thomas Vetter, "A Morphable Model For The Synthesis Of 3D Faces," Max-Planck-Institut f¨ur biologische Kybernetik, T¨ubingen, Germany, MPI f¨ur biol. Kybernetik, Spemannstr. 38, 72076 T¨ubingen, Germany.

[18] Yang Chen*, Yuri Owechko, and Swarup Medasani, "A Multi-Scale Particle Swarm Optimization (PSO) Approach to Image Registration," Information Science and Systems Laboratory HRL Laboratories, LLC, Malibu, California, USA, Author contact information: ychen@hrl.com.

[19] Yas Abbas Alsultanny∗, Musbah M. Aqel, "Pattern recognition using multilayer neural-genetic algorithm," Computer Science Department, College of Computer and Information Technology, Applied Science University, P.O. Box17, Amman 11931, Jordan, Received 2 July 2001; accepted 8 May 2002.

[20] Zhang, Y. and Ji, Q. (2005), "Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 5, pp. 699-714.

BIBLIOGRAPHY OF AUTHORS

A. Vijaya Kumar received the B.Sc. in Computer Science, MCA, and M.Phil. in Computer Science From Bharathidasan University, Tamil Nadu, India in 1998, 2001 and 2004 respectively and M.Tech. in Computer Science & Engineering from SRM University, Tamil Nadu, India in 2015. He is a Ph.D. Research Scholar in the field of Computer Science at Bharathiar University, Tamil Nadu, India. He has also qualified in TNSET 2016.
He has presented many research papers in SCOPUS Indexed Journals, International Journals, International and National Conferences. His research areas are Image, Video, Signal Processing and Artificial Intelligence.
He is having 15 Years of Experience in various Academic institutions. He is a cherished teacher with ability to train and counsel students. Presently he is an Assistant Professor in Department of Computer Science, Hindustan College of Arts & Science College, Chennai, Tamil Nadu, India.

Dr. R. Ponnusamy received the B.Sc. and M.Sc. in Computer Science from Bharathidasan University, Tamil Nadu, India in 1994 and 1996 respectively and M.Tech. in CSE from Pondicherry University, Pondicherry, India in 2000. He also has a Ph.D. in Computer Science & Engineering from College of Engineering, Anna University, Chennai in 2008.
His areas of interests are Distributed Artificial Intelligence, Soft-Computing, E-Governance, Information Retrieval, Human-Computer Interaction and Higher-Education. He has presented/ published 70 papers in various international conference and journals. He is having 20 Years of Experience in various Academic institutions. Presently he is a Professor in the Dept. of Computer Science & Engineering, Sri Lakshmi Ammaal Engineering College, Chennai, Tamil Nadu.
He has organized and edited the proceedings of the IEEE International Conference on Intelligent Agent & Multi-Agent Systems 2009 and is serving as an Editorial Board Member of the Journal of the World University Forum, World Academy of Science, Engineering and Technology and the International Journal of Computer Applications.

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

Scalable Statistical Detection of Tunnelled Applications

Ghulam Mujtaba

Electrical Engineering Department,

Comsats Institute of Information Technology,

Abbottabad, Pakistan.

Corresponding Author: gmujtaba@ciit.net.pk.

Absract- **In protocol tunnelling, one application protocol is encapsulated within another carrier protocol possibly to circumvent firewall policy. Application-layer tunnels are a significant security and resource abuse threat for networks. The existing techniques for identification of applications running across the network, for example packet data analysis techniques are not always successful, especially for applications which use encrypted tunnels. This work describes a statistical approach to detect applications which are running using application layer tunnels. A fast machine learning algorithm, k-Nearest Neighbours is demonstrated to be able to perform the detection of tunnelled applications based on statistical features obtained from the network applications running inside protocol tunnels. The scalability of the mechanism is also investigated.**

Keywords- **Protocol tunnelling, application detection, firewalls, http-tunnels, network security**

INTRODUCTION

Tunnelled Applications are those that are run inside another protocol to circumvent firewall policies. These are quite common to observe in networks where some restrictions are applied for usage of the network. Some applications are forbidden over certain networks depending on the Acceptable Usage Policies of the particular organisation. In some countries, there is restriction and censorship on the internet usage. However some people are inclined to still run the same applications but using a protocol tunnel or a VPN (Virtual Private Network). Those applications which are restricted by firewalls such as high data-rate games, peer-to-peer file sharing, video and audio streaming, and chat are carried through via allowed protocols like HTTP, HTTPS and the firewall security policy is thwarted. Protocols such as HTTP and HTTPS are indispensable today for any network which has to be connected to the Internet; hence these become a high value target for running restricted

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

applications via tunnelling. Since the carrier protocol could be encrypted one, hence it would be much harder to discover the application being run from the packet traces, because that would require the capability to decrypt the data first. The decryption itself without the knowledge of the decryption key is computationally intensive even when it is possible in real time.

In such scenario, this work presents a statistical approach to identify the applications based on the patterns of the packet related statistical features. The applications' packets do observe a certain statistical behaviour as to provide insight into identifying the application. The network traffic is captured in promiscuous mode, and from the captured packets statistical parameters are obtained. Once sufficient amounts of training features are extracted, then it is possible to identify an application even inside encrypted tunnel with reasonable accuracy. Distance based algorithms are relatively faster algorithms than Artificial Neural networks or Bayesian algorithms. The kNN algorithm is a fast and efficient algorithm that can perform application detection with reasonable accuracy even in tunnelled applications.

RELATED WORK

In network monitoring, application detection is performed mostly by "Deep Packet Analysis" or Deep packet Inspection, a form of computer network packet filtering that examines the data part (and possibly also the header) of a packet as it passes an inspection point, searching for protocol non-compliance, viruses, spam, intrusions, or defined criteria to decide whether the packet may pass or if it needs to be routed to a different destination, or, for the purpose of collecting statistical information [13] . The purpose of Deep Packet Inspection is the implementation of network management services, network data mining, monitoring, and internet usage policy compliance management. This method attempts to look at the data held in the packets, and usually it is based on searching certain unique identification patterns or signatures held in the data. Mostly Deep Packet Analysis requires capturing the packets of session establishment and session initiation phase, because these are the packets which are likely to contain the unique signature pertaining to the application protocol. This technique is also used in Intrusion Detection Systems, for example Snort and Bro systems. The captured packets are searched for unique identifiers and keywords which are kept in a fairly large database. The detector in such a case would fail to detect if it is initiated after the session has been established, or the network is heavily loaded with capturing becoming difficult. Deep Packet Analysis or Deep Packet Inspection requires more processing power too as it has to operate with faster networks and the process itself is computationally quite intensive. Then in case of encryption, this technique cannot work, because it would require decryption of the packet data first, which is another challenging task.

There have been several attempts for statistics based classification of network data in general and a few about tunnelling application detection in particular. Network Intrusion detection is performed from statistical data analysis in several papers of the like of [14]. Five learning algorithms are compared based on their performance in the IP traffic flow classification problem in [16]; the authors have given a performance comparison of five machine learning algorithms for IP traffic flow classification. The C4.5 tree algorithm was computationally found to be faster than other four. In a significant piece of work 248 flow features were extracted and used in the

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

supervised Naïve Bayes classification algorithm to differentiate between different types of applications [6]. These included packet size, inter- arrival times, some features derived by transforms, some derived from TCP header. Correlation-based feature analysis was used to find the stronger features which showed that only fewer than 20 features were required for accurate classification. These works however have experimented on plain network traffic and not tunnelled traffic. Another important contribution is the system called "Tunnel Hunter" [4], in which the Naïve Bayes algorithm is used to identify protocol tunnelling in POP3, CHAT, SMTP protocols. The statistical features utilized for classification are packet inter-arrival time and packet size. This system performs the detection of when these protocols are being used inside tunnels, but doesn't tell what application is run from behind the tunnel. In [5], the same algorithm is utilized for encrypted tunnels detection. The work in [15] has attempted to formulate a statistical approach for identification of yet unknown protocols on the network as there is an increase in the protocols and applications on the internet.

In [12], the author presented a different approach to the problem of detecting non-registered UDP based applications, which is based on packet statistics. The author showed empirically that the packet size distribution follows an application specific profile, and this statistical property can be employed as detection metric. Hence the process takes a sample of the data stream generated by an application, and performs detection based on or over that sample, rather than an individual packet. The Packet Size Distribution based detection mechanism can also operate on encrypted applications. The scheme uses significantly less information than the previous Deep packet analysis techniques, reducing the storage and processing requirements. Then, [2] worked on the identification of TCP embedded applications based on similar packet size distributions analysis. TCP covers a majority of internet applications today. In the paper [9], the Packet Size Distribution feature of [12] and [2] , and other statistical features from [16] , [7] and [14] are combined to address the problem of Tunnelled Application detection, and several machine learning techniques were compared. In this work, we have singled out the classifier that was best performing in the work by [9], and for that classifier presented the results of previously unseen instances of the data set. The scalability of the same methodology is explored by extending the number of applications used from 10 to 15and using the fast and simple classifier algorithm, k-Nearest Neighbours. It is argued that empirically the results indicate that the technique is scalable for practical purpose.

THE METHODOLOGY

The network applications which were considered for detection by the system were run inside the packet tunnelling tools. Then using a network sniffer, like Wireshark, these applications' packets were collected in packet trace files. The packet traces were captured for the applications for several minutes of run time per application. Ten popular network applications were selected and their packets were captured.

Depending on the source address, destination address, source port, destination port and protocol, the packets were assigned into connections. A connection is a tuple of these five parameters. From the captured file, any connections with very few packets were filtered out and discarded, because packet size distributions using these connections would not be reliable. The other connections were saved for further processing as they contain the packets belonging to a particular application being run inside a protocol tunnel. The features obtained for the connections of each tunnelled application trace file were Data rate (bytes/sec) for upstream direction, Data rate

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

(bytes/sec) for downstream direction, Data packet ratio: ratio of the data packet number downstream to upstream, ByteRatio: ratio of the total bytes transferred in the downstream (remote to local) to upstream (local to remote) direction, Ratio of large and small packets: The threshold of small packets is arbitrarily set to 300 bytes, time spent idle downstream: the collection of time periods of 2s or greater duration in which there was no packet sent downstream given as a percent of the total time so that the value is normalized for various length packets, time spend idle upstream and Packet Size Distributions with 15 bins since PSD is an effective identifier for tunnelled applications. This has been shown in the work [12] .The Packet Size Distribution bins are derived based on the work done by [10].

THE KNN ALGORITHM

Weka's IBK* is an instance-based classifier which implements the k Nearest Neighbor algorithm. The predicted class of a test instance is based upon the class of those training instances nearest to it, as determined by a distance function. The distance function used here is an entropy-based distance function. This algorithm was chosen because it is fast and accurate compared to other machine learning algorithms for this kind of prediction problems [8].

The K* distance between two data points is obtained from the complexity of transforming one data point into the other. To begin with a finite set of transformations which map data points to data points is defined. "A "program" to transform one instance (*a*) to another (*b*) is a finite sequence of transformations beginning at *a* and ending at *b*. The complexity of a program is the length of the shortest string representing the program, and a Kolmogorov distance between two instances is defined to be the length of the shortest string connecting the two instances. The result is a distance measure from which K* distance is calculated by summing over all possible transformations between two instances." [3] . "WEKA" [6] software has IBK* algorithm which implements the kNN. In the next section, the results of experiments with the kNN i.e. WEKA's IBK* implementation for different sets of data are given.

RESULTS OF THE KNN CLASSIFICATION FOR THE DATA

The data collected and obtained from the network applications in tunnelled mode has 120 instances, 12 instances pertaining to each of the 10 applications. Each instance corresponds to packet trace file taken from the network while the application is live and running. The default implementation of WEKA of the kNN algorithm is used with k = 1.The experiments are performed in training mode, 66 % split mode, 10 fold cross validation mode and separate test set mode. The various modes of the experiments are explained as:

Training Mode: In this mode, the complete data set is used for training the ML algorithm and the same complete dataset is the used to test the algorithm.

66 % split Mode: In this mode, two thirds of the data is used for the training and one third is used for testing the ML algorithms. The split is random, but approximately equal contribution of each class is taken in both test set and training set.

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

10 fold cross validation: According to [17], the standard way of predicting the error rate of a learning technique given a single, fixed sample of data is to use stratified 10-fold cross-validation. In this method, the available data set is divided randomly into 10 sections such that the application classes have similar representation in each section as in full dataset. In first run, one of the 10 parts is used for testing, and the remaining 9 parts are used for training. Then similarly in 10 runs, all 10 parts are respectively used for testing set and other 9 used for training, and its error rate is calculated on the test set. Thus the learning procedure is executed a total of 10 times on different training sets (each of which have a lot in common). Finally, the 10 error estimates are averaged to give an overall error estimate [17].

Separate test set mode: in this mode the test set is completely unseen by the algorithm and is obtained from separate packet files.

Table 1 summarises the results of the three modes for the 22 attributes based classification for each instance, with 15 attributes from the Packet Size Distribution bins and 7 other statistics described in the previous section. For comparison, the results for just the 15 bin PSD based classification is given in **Table 2** also, but it is observed that inclusion of PSD with the other statistical attributes is better than using the PSD alone.

Test Mode	Total test instances	Total training instances	Time taken to build model	Correct Predictions	Incorrect Predictions
Training data	120	120	<0.01 s	120(100 %)	0(0%)
10 fold cross validation	120	120	<0.01 s	118(98.8%)	2(1.2%)
66 % split	41	79	<0.01 s	41(100 %)	0(0%)

Table 1: Summary of classifications with 22 attributes

Test Mode	Total test instances	Total training instances	Time taken to build model	Correct Predictions	Incorrect Predictions
Training data	120	120	<0 .01s	106(88.3%)	14(11.7%)
10 fold cross validation	120	120	<0 .01s	100(83.3%)	20(16.7%)
66 % split	41	79	<0 .01s	35(85.4%)	6(14.6%)

Table 2: Summary of classifications with only 15 PSD attributes

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

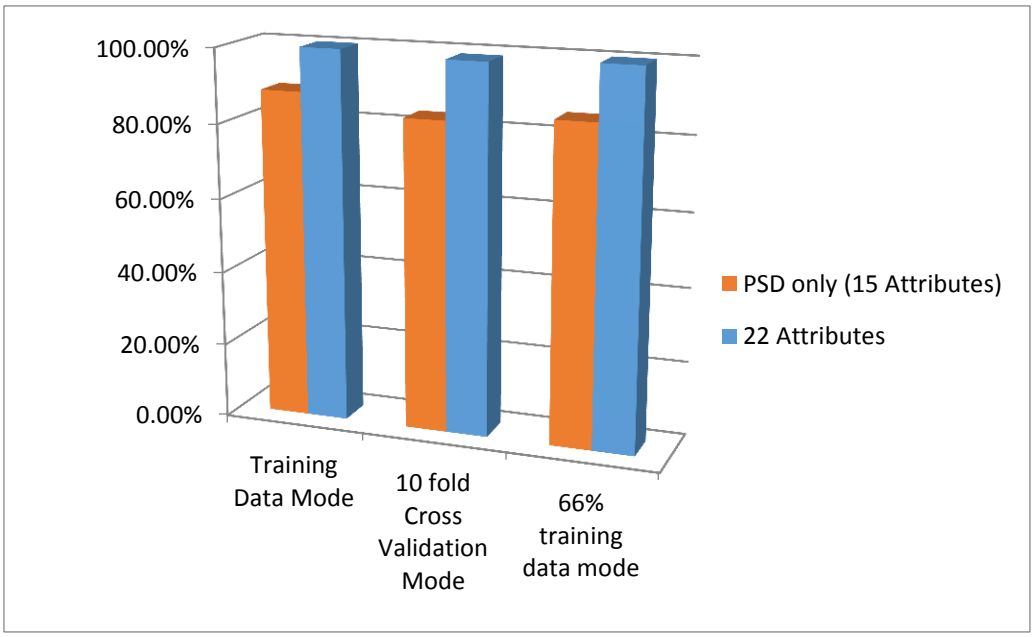


Figure 1: Comparison of classifications using PSD only and PSD with 7 other attributes

Having established the usefulness of the Packet Size Distribution along with a number of other statistics, it was also important to investigate higher resolutions of PSD bins. Hence the experiments were repeated with 37 attributes including 30 bins of PSD with bin size of 50 bytes, and then with 57 attributes including 50 PSD bins with bin size of 30 bytes. These results are given in **Table 3** and **Table 4** respectively.

Test Mode	Total test instances	Total training instances	Time to build model	Correct Predictions	Incorrect Predictions
Training data	120	120	<0.01 s	120(100 %)	0(0%)
10 fold cross validation	120	120	<0.01 s	117(97.5%)	3(2.5%)
66 % split	41	79	<0.01 s	41(100 %)	0(0%)

Table 3: Summary of classifications with 37 attributes including 30 PSD bins

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

Test Mode	Total test instances	Total training instances	Time to build model	Correct Predictions	Incorrect Predictions
Training data	120	120	<0.01 s	120(100 %)	0(0%)
10 fold cross validation	120	120	<0.01 s	117(97.5%)	3(2.5%)
66 % split	41	79	0 .01 s	41(100 %)	0(0%)

Table 4: Summary of classifications with 57 attributes

From these results it is concluded that the 15 bin resolution or 30 bin resolution of Packet size distribution can be appropriate and optimal to be used with the KNN algorithm, because the classification accuracy is not much different both cases. In previous work by [1] and [10] , 30 bins was found a better resulting Packet Size Distribution resolution.

Testing on Unseen Data

In the previous experiments, the testing of machine learning algorithm was performed using the same training data, although with different variants such as splitting 66% into training and 34% into test parts or by 10 fold cross-validation, or even on exactly the same training set. Statistically, it seems that these results would be a good estimate of the actual performance over real test data. Next these training data were used with unseen data for testing, i.e. not from the same dataset. This test data was collected from separate pcap trace files. Hence, it was never used in the training. This would further corroborate the robustness of the methodology.

The dataset prepared with the new tracefiles contains 33 instances to be tested from the same applications used in previous sections. **Table 5** summarises the results of the KNN classifier algorithm tested using the separately applied test data. The new trace files also have 30 bins of packet size distribution in their corresponding attributes files.

Classifier	Test Mode	Total test instances	Total training instances	Time to build model	Correct Predictions	Incorrect Predictions
K Nearest Neighbour	Separate test set	33	120	0.01 s	33(100 %)	0(0%)

Table 5: Summary of classifications of fresh data using 37 attributes

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

The overall results do not show much difference between the fresh data or the 10 fold cross validation case. So the distance based classifiers k-Nearest Neighbours produced accurate predictions not only when tested on the training data in various combinations, but also on fresh test data.

SCALABILITY

The previous results have shown that the kNN algorithm is able to successfully identify the applications in the data set based on the statistical parameters of the application trace file. The data set included 10 applications. The issue of scalability is that would the method perform similarly if the number of applications is increased further. The definite answer is hard to prove, because this work is based on empirical results of the applications' data. However, as an experiment the number of applications was increased from 10 to 15 applications and it was investigated whether the method still is able to differentiate between them. The five new applications used are: Unreal Tournament, Zattoo which is an IP TV application, Real Player, Remote Desktop Application and Guild Wars, a role playing game. The packet size distribution resolution of the trace files is 30 bins. The test mode selected is 10 fold cross-validation, the standard way of predicting the error rate of a learning technique given a single, fixed sample of data. The results of the experiment on using 15 applications are:

Total Number of Instances: 178

Correctly Classified Instances: 171 or 96.0674 %

Incorrectly Classified Instances: 7 or 3.9326 %

These results suggest that addition of new applications would not affect the classification accuracy very much at least, as the 15 applications which have been tested here show an accuracy of 96.1 %. The same case of the kNN algorithm with 10 applications tested with 10 fold cross validation had an accuracy of 97.5% as given in Table 3. Therefore these figures suggest that the method would hold its utility for a considerable number of applications. However it has not been proven that it would work given an extremely large database of applications. Even though the exact number of supportable applications is difficult to predict, it still is a useful approach because it can be utilized to detect the few most malicious applications on the network, albeit they may be finite in number.

CONCLUSIONS

The Tunnelled Network applications can be detected using a simple and fast Machine Learning algorithm kNN. The statistical attributes of each application packet file were a combination of Packet Size Distribution as discrete bins with resolution 15 bins or 30 bins and seven other attributes. The resolution of 15 bins works well with the 10 and 15 applications chosen in this work, however it is predicted that when many more applications are there to be identified, then 30 bins could be a better resolution, because there will be lesser probability of Packet Sizes following same distribution in 50 bytes packet size than in 100 bytes packet size. Even if the technique doesn't work accurately in the number of applications being exceedingly high, it is significant progress to work in limited number of applications. The kNN algorithm being lightweight can be implemented

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

on the network, and a hardware implementation on a network gateway would be able to cope with most networks' traffic. The technique can be fully utilized in those places where the network administrator wants to strictly thwart some selected applications from his network.

REFERENCES

[1] *Bharadia, K. 2001.* Network Application Detection Techniques. PhD Thesis, Loughborough University, UK.

[2] *Bo, L., Parish , D.J., Sandford, J.M. and Sandford, P. 2006.* Using TCP Packet Size Distributions for Application Detection. PGNet 2006 Proceedings, PGNet, Liverpool John Moores University, UK.

[3] *Cleary, John G. and Trigg, Leonard E. 1995*. K*: An Instance-based Learner Using an Entropic Distance Measure. 12th International Conference on Machine Learning.

[4] Dusi, M., M. Crotti, F. Gringoli, L. Salgarelli. 2009. Tunnel Hunter: Detecting application-layer tunnels with statistical fingerprinting. Computer Networks. 53(1).81-97.

[5] *Dusi, M., M. Crotti, F. Gringoli, L. Salgarelli.2008.* Detection of encrypted tunnels across network boundaries. Proceedings of the 43rd IEEE International Conference on Communications (ICC 2008), Beijing, China.

[6] Hall, M., Frank. Eibe, Holmes, Geoffrey, Pfahringer, Bernhard, Reutemann, Peter, Witten, Ian H. 2009. The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1.

[7] *Moore, A. W. and D. Zuev. 2005.* Discriminators for use in flow-based classification. Technical report, Intel Research, Cambridge.

[8] *Mujtaba, G. 2011.* Identification of Networked Tunnelled Applications. PhD Thesis, Loughborough University, UK.

[9] Mujtaba, G. and Parish, D.J. 2015. A statistical Framework for identification of tunneled applications using Machine Learning. IAJIT, vol. 12, p755.

[10] *Mujtaba, G. and D.J. Parish. 2009a.* Detection of Tunnelled Applications Using Packet Size Distributions. Proceedings of the PGNet 2009, Liverpool JMU, Liverpool, UK

[11] *Mujtaba, G. and D.J. Parish. 2009b.* Detection of Applications Within Encrypted Tunnels Using Packet Size Distributions. Proceedings of the International Conference on Internet Technology and Secured Transactions, ICITST 09, London, UK.

[12] *Parish, D.J., Bharadia, K., A. Larkum, A., I. W. Phillips, I. W., and Oliver, M. 2003*. Using Packet Size Distributions to Identify Real-Time Networked Applications. Communications, IEE Proceedings.

[13] *Porter, Thomas. 2005.* The Perils of Deep packet Inspection. *Security Focus.*

[14] Rastogi, Rahul, Khan, Zubair and Khan, M.H. 2012. Network Anomalies Detection Using Statistical Technique : A Chi- Square approach. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 3.

[15] *Wang ,Y., Chen, C. and Xiang ,Y. 2015.* Unknown pattern extraction for statistical network protocol identification. 2015 IEEE 40th Conference on Local Computer Networks (LCN), USA.

[16] Williams , N., Zander, S. and Armitage ,G. 2006. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. Special Interest Group on Data Communication (SIGCOMM) Computer Communication Review, vol. 36, no. 5, pp. 5– 16.

[17] Witten, Ian H., Eibe Frank. 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Series.

Detection of Microaneurysm in Diabetic Retinopathy

Morium Akter
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka, Bangladesh
e-mail: ecs_morium@yahoo.com

Abstract—**Diabetic retinopathy is a severe vision complication of diabetes and it is one of the main causes of blindness all over the world. Early detection of diabetic retinopathy is essential to cope with this adverse effect. Microaneurysm is one of the early signs of diabetic retinopathy. So the presence of microaneurysm detection is a prerequisite for early diagnosis of diabetic retinopathy. In this paper, we have proposed a simple morphological method for the detection of microaneurysm that uses top-hat transform. Our method can detect the faint microaneurysm at low resolution due to contrast enhancement and noise reduction as preprocessing. We also compare the results of different retinopathy detection techniques.**

Keywords-Diabetic retinopathy, microaneurysm, blindness, contrast enhancement, morphological operation.

I. INTRODUCTION

Diabetic retinopathy is the most common cause of blindness. It is one of the consequences of diabetes. Around 7% people who have diabetes for 10 years will have developed diabetic retinopathy. The rate of blindness in global population from diabetic retinopathy will rise to 4.4% at the end of 2030 [1], [2], [3].

Microaneurysm is one of the earliest symptoms of the diabetic retinopathy. If we can detect the microaneurysm at an earlier stage then diagnosis of retinopathy can be done effectively. As a result the treatment of diabetic retinopathy can reduce the threat of blindness by 50% [4]-[8].

Microaneurysm [4], [9]-[10] is a retina lesion which is caused by local swelling of capillary walls and generate small red dots on the surface of the retina shown in Figure 1. It ranges from 25 – 100 microns in size.

The detection of microaneurysm is difficult due inherent low contrast characteristic of eye fundus images. So our aim is to develop a system for the detection of diabetic retinopathy by detecting microaneurysm, helping the patients as well as the doctors for early diagnosis of diabetic retinopathy for reducing blindness of the diabetic patients.

Sopharak et al. [2] and Prakash and K. Sumathi [4] proposed microaneurysm detection methods using mathematical morphology. In this paper, we propose a simple morphological method for the detection of microaneurysm using top-hat transform along with image preprocessing. We compare the results of our method with the results of the above methods.

Besides this, another method was proposed by Pallawala et al. [9] based on eigenvectors of affinity matrix. We also compared the results with our results.

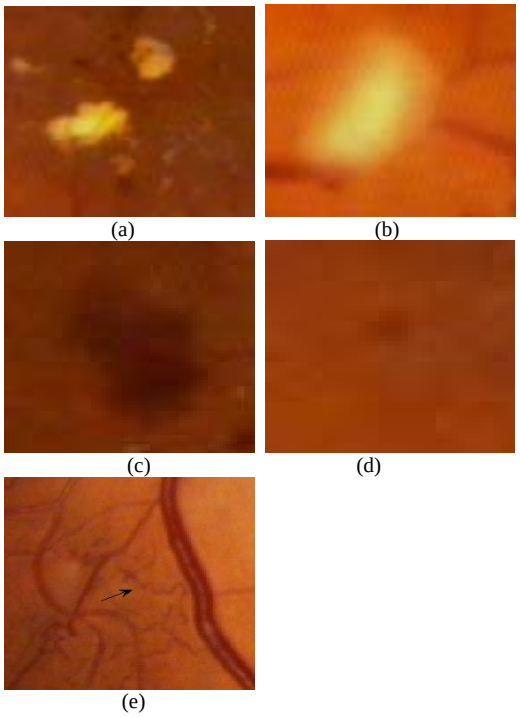


Figure 1. Abnormal findings in the eye fundus images caused by diabetic retinopathy: (a) hard exudates, (b) soft exudates, (c) hemorrhage, (d) microaneurysm and (e) neovascularizations (images taken from references [11]).

Section II describes medical knowledge of diabetic retinopathy, Section III presents our proposed method, Section IV shows the experimental results and discussions and finally, Section V draws the conclusions of our paper.

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

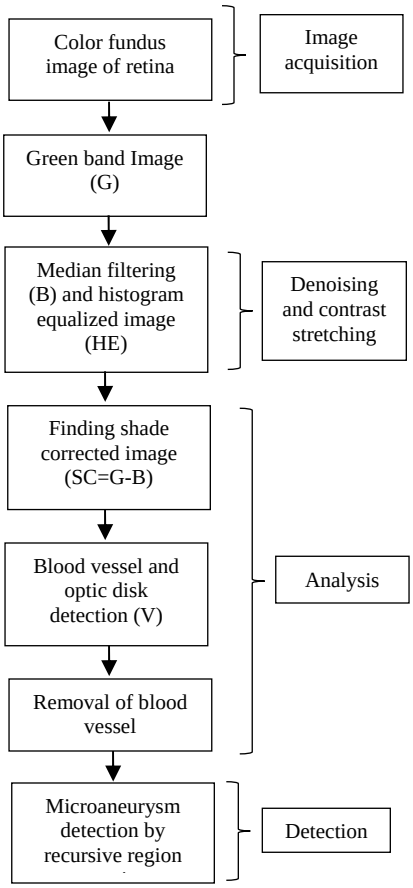


Figure 2. Flow diagram

II. Medical Knowledge of Diabetic Retinopathy

Diabetic retinopathy is a consequence of diabetes. It affects retina's blood vessels and prevents light entering the retina that may create blindness [3], [11] and [12] of diabetic patents. Sometimes these vessels may dilate or close completely. Figure 1 shows the symptoms of diabetic retinopathy. In severe retinopathy, the blood vessels may become diluted which create microinfarcts in the retina. Lack of oxygen in the microinfarcts cause the development of new fragile vessels. This phenomenon is called neovascularization – a serious cause of vision loss.

Microaneurysm is the earliest symptoms of ratinopathy. It looks like as small, dark red spots and causes intra retinal haemorrhages. It's size may vary from 10-100 microns.

III. Proposed Method

The flow diagram of our method is shown in Figure 2. At first we have to take a color fundus image as an input. Then the image is converted into green band image, G. We used median filtering (to get B image) and histogram equalization (HE image) as preprocessing for noise reduction and enhancement purpose. We get shade corrected (SC) image by subtracting B image from the green band image G through equation (1).

$$SC=G-B \quad (1)$$

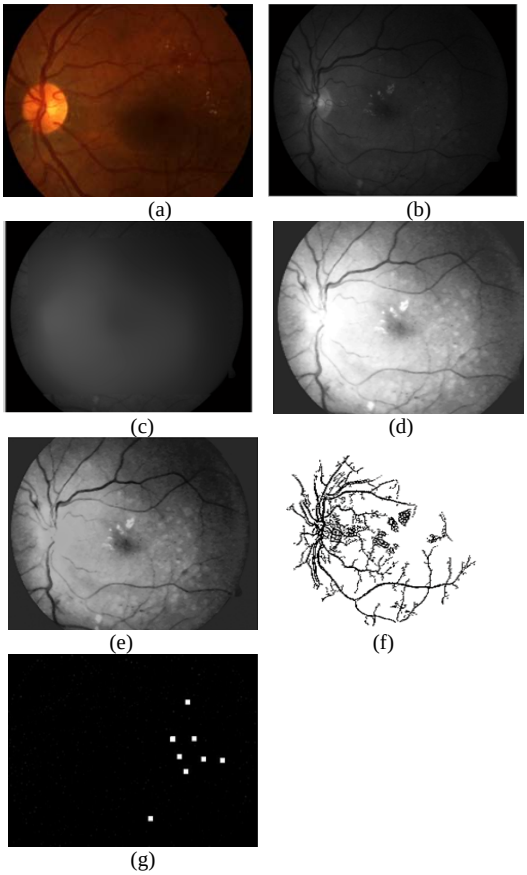


Figure 3. Microaneurysm detection: a) input image (b) green image, G (c) background image after median Filter, B (d) image after histogram equalization, HE (e) shade corrected image, SC= G-B (f) detected blood vessel image, V (g) detected microaneurysm.

The top-hat and binarization operations are applied to the shade corrected (SC) image to get optic disk and blood vessel which is named as V.

Then the difference image (D) is obtained by equation (2)

$$D=G-SC-V \quad (2)$$

Then on D image we apply recursive region growing algorithm to get microaneurysm. The output of the proposed method is shown in Figure 3.

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

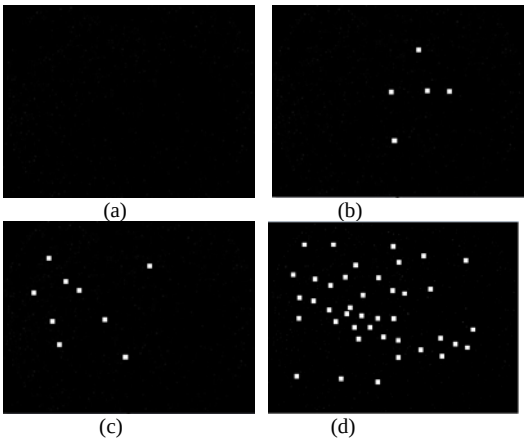


Figure 4. a) Output of a normal image b) output of a mild retinopathy image, c) output of a moderate retinopathy image, and d) output of a severe retinopathy image.

IV. Experimental Result and Discussions

We have tested our system with 100 retinopathy color fundus images from Bangaldesh Eye Hospital, Dhaka. We implemented our algorithm using Matlab 7.1. In the testing phase we calculated the number of clusters of microaneurysm. Then we categorize these images into normal, mild, moderate and severe diabetic retinopathy according to the number of detected clusters. Figure 4 shows the output of our method according to the category of detected microaneurysms. Among 100 images the proposed technique finds 40 images are normal, 38 are mild diabetic retinopathy, 6 images are moderately affected, 12 images are severely affected and 4 image gives wrong result which are similar to opthamologist hand-drawn ground truths. Our method gives 96% correct result which is better compared to the method described in [2], [4] and [9]. The accuracy of the method shown in reference [2] is higher than us as it uses only 15 retinal images. The method shown in reference [4] have used top hat transform but it did not specify the performance of the method. The method shown in reference [9] cannot detect the blurred microaneurysm clusters and hence the performance is somehow lower. Table I shows the comparison of different method's performances.

TABLE I. Comprision

Authors	Accuracy
Akara Sopharak et. al. [2]	99.99%
Prakash and K.Sumathi [4]	Not specified
Pallawala et. al [9]	93%
Our method	96%

V. Conclusions

Microaneurysm is the earlier sign of diabetic retinopathy. In this paper we described a microaneurysm detection method for early diagnosis of diabetic retinopathy through a simple morphological operation along with top hat transform. We have tested our method using practical diabetic retinopathy images. Experimental results confirmed that our method is effective and will be helpful for diagnosis of retinopathy diagnosis.

Acknowledgment

We are grateful to Dr Omor Jafarullah, Medical officer, Bangladesh Eye Hospital, Satmasjid Road, Dhaka, Bangladesh for giving retinopathy color fundus images and helping us in doing the research.

References

[1] Akara Sopharak, Bunyarit Uyyanonvara and Sarah Barman, "Automated microaneurysm detection algorithms applied to diabetic retinopathy retinal images", Maejo International Journal of Science and Technology, 2013, pp. 294-314, ISSN 1905-7873.

[2] Akara Sopharak, Bunyarit Uyyanonvara and Sarah Barman, "Fine Microaneurysm Detection from Non-dilated Diabetic Retinopathy Retinal Images Using a Hybrid Approach", Proceedings of the World Congress on Engineering 2012 Vol II , WCE 2012, July 4 - 6, 2012, London, U.K.

[3] Atsushi Mizutani, Chisako Muramatsu, Yuji Hatanaka, Shinsuke Suemori, Takeshi Hara, and Hiroshi Fujita, "Automated microaneurysm detection method based on double-ring filter in retinal fundus images", Medical Imaging 2009: Computer-Aided Diagnosis, edited by Nico Karssemeijer, Maryellen L. Giger Proc. of SPIE Vol. 7260, 72601N · © 2009 SPIE · CCC code: 1605-7422/09/$18 · doi: 10.1117/12.813468

[4] Prakash and K.Sumathi, "Detection and Classification of Microaneurysms for Diabetic Retinopathy", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 National Conference on Advanced ommunication & Computing Techniques (NCACCT-19 March 2013, pp.no 31-36.

[5] C. Sinthanayothin, J.F. Boyce, T. H. Williamson, A. TT. Elliot, "Automated Detection of Diabetic Retinopathy on Digital Fundus IImage ", International Journal of Diabetic Medicine, vol. 19, pp. 105-112, 2002

[6] The Royal College of Ophthalmologists 2012, "Diabetic Retinopathy Guidelines", December 2012, available online at http://www.rcophth.ac.uk/core/core_picker/download.asp?id=1533 (accessed on 112.5.17).

[7] Iqbal, M.I Aibinu, A.M Gubbal, N.S and Khan, A "Automatic Diagnosis Of Diabetic Retinopathy Using Fundus Images", available online at http://paper.ijcsns.org/07_book/200812/20081230.pdf (accessed on 112.4.16).

[8] S. Kavitha, K. Duraiswamy, " Automatic Detection of Hard and Soft Exudates in Fundus Images Using Color Histogram Thresholding", European Journal of Scientific Research, ISSN 1450-216X Vol.48 No.3(2011), pp.493-504.

[9] P M D S Pallawala, Wynne Hsu, Mong Li Lee and Say Song Goh, "Automated Microaneurysm Segmentation and Detection using Generalized Eigenvectors", available online at http://www.math.nus.edu.sg/~matgohss/microfin.pdf (accessed on 12.5.17).

[10] G. Yangy, L. Gagnonz, S. Wangy and M.-C. Boucher, "Algorithm for detecting micro-aneurysms in low-resolution color retinal images", available online at https://crim.ca/Publications/2001/documents/plein_texte/VIS_YanGLals_VI01.pdf (accessed on 112.5.17).

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

[11] Pavle Prentaˇsi´c, “Detection of Diabetic Retinopathy in Fundus Photographs”, available online at
https://www.fer.unizg.hr/_download/repository/KDI_Prentasic_Pavle.pdf (accessed on 112.5.17).

[12] Seema Garg, and Richard M. Davis, “Diabetic retinopathy screening update, available online at
http://clinical.diabetesjournals.org/content/27/4/140.full.pdf (accessed on June 28, 2013).

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

Study Egyptian Students' Perception of Using Social Media in Learning

Abeer A. Amer

Computer Science & Information System Department

Sadat Academy for management and Sciences, Alexandria - Egypt

abamer_2000@yahoo.com

Abstract

Social networking sites like Facebook, YouTube, Instagram, Google+ Twitter and etc; are becoming an integral part of students' lives in Egypt. This study attempts to investigate the student's perception of social networks as a learning tool. A survey was conducted by 757 questionnaires given to a sample of students of different ages and genders representing various colleges in Egypt during the academic year 2016/2017. SPSS is used to analyze the collected data. The results show that most of students are using social networks in their learning, moreover the result support the advantages of social networks in learning and don't show any apparent disadvantages.

Keywords: Social networks sites, social media tools, student perception, academic performance, Egyptian colleges.

1. INTRODUCTION

Social media has become prevalent, the most popular social media sites are Facebook, Twitter, MySpace, LinkedIn, YouTube, Google + and Skype. They allow all users to communicate and share information with each other and also allow users to build relationships [1,2,3]. Furthermore social media can help in group discussion, resource sharing and entertainment. College students considered as a large proportion of users on social media networks. Many of these students use social media networks to communicate with family, friends, and others. Social media sites have created new ways for students to interact with others and they have taken advantage of this type of technology [4,5].

There are many advantages and disadvantages for universities and colleges that can be gained through connecting students with social media sites, the advantages of using social media in learning are facilitate communication, acquire knowledge, share information, allows students to discuss ideas, and facilitate informal learning for students. Disadvantages include social media distract students from learning, time spent by students on non-academic activities, negative relationship between GPA and time spent on social media, and students delay their meals and sleepless [2,6,7].

Although many studies have investigated the impact of social media on college students' perception as a learning tool, few of them have focused on Egypt. The main contribution of this paper is exploring the Egyptian student's perceptions of social media as a learning tool. The study was applied on undergraduate students in three different colleges in Egypt and also identified the differences and similarities on students in the different colleges [1,7, 8].

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

The rest of the paper is organized as follows: section 2 discuses related works, section 3 identifies the research methodology of the study. Section 4 shows the sampling and analysis, section 5 talks about results and discussion and finally section 6 gives the conclusion, recommendation and future work.

2. RELATED WORKS

A number of studies have focused on advantages and disadvantages of social media and the effect of the social network sites on students. A study by Al-Sharqi,Hashim and Kutbi (2015) was conducted in King Abdelaziz University's. It investigated the use of different categories of social media for academic and non- academic purposes. The study found that a category entertainment was the highest category, then information searching category and the third one was learning. This study concluded that students agree on the advantages of social media but they did not sure of the disadvantages[9]. Mingle and Adams (2015) concluded that although the benefits of social media like sharing information and improving reading skills there is a negative impact on the students' academic performance [1].

Ramprathap (2016) identified that social media have a great positive and negative impacts and it is important to put rules and regulations of using the social media [10]. Another study was conducted by Mehmood and Taswir (2013) to investigate using social networks for academic purposes like search for information and look for career opportunities[7]. Sponcil and Gitimu (2013) revealed that Facebook and twitter are considered as the most popular websites and all people using them at increasing rates and also investigated the effect of social media on college students either they believe that social media tools affect their self-esteem positively or neither positively or negatively [4].

Kamal, Tariq, Ishtiaq, Nawab and Idrees (2015) concluded that social media sites like Facebook have effect on the individual 's life because more time of using social media cause more negative effect on their overall life[11]. A survey conducted by Bagget and Williams (2012) to investigate the use of social media as means to share common interests. According to this survey social media sites are useful tools for communication and education, and also giving an opportunity for networking in any profession. They also assist students to do multitask[12].

In [13] the basic purpose of this study was to see the academic outcome of student who spend most of their time on such interacting sites, the result of the study by Mansoor and Heshmet (2016) found that students spend time of their day activities on social media sites. Shambare, R et al. (2012) Conducted a study on Social networking among students, the study indicated that social Media is the most commonly used by young students and Facebook is most widely using by a large number of communities and effecting the student's life[14]. Chen andBryer (2012) investigated the perceptions of faculty members for using social media in formal and informal learning they revealed that the Facebook is used for nonacademic activities and linkedln for academic activities.

3. RESEARCH METHODOLOGY

To study the student’s perception of social networks as a learning tool a questionnaire was developed. The questionnaire used likert scale which is the most widely used scale in survey, the approach of five level likert items are strongly agree, agree, undecided,

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

disagree and strongly disagree this approach was ranking from strongly agree (5) to strongly disagree (1) and it was used to measure each item of the questionnaire. The study was conducted in three educational institutions Faculty of Engineering, specialized scientific programs (SSP), Sadat Academy for Management and Sciences (SAMS), computer and information systems department (CIS) in a Faculty of Commerce; in Alexandria, Egypt. The questionnaire was distributed to a number of students in the three colleges; the students were from different majors. The sample of undergraduate male and female students (n=757) are from different age groups (younger than 20 and older than 20 years).

The questionnaire was divided into three parts. The first part, was demographic questions about (gender, age, specialization), the second part included specific questions about using social media and internet, the third part focused on the advantages and disadvantages of social media use in learning.

4. SAMPLING AND ANALYSIS

The questionnaires distributed to sample size of 757 students and they were asked to respond and their answers were confidential, the collected data was analyzed using SPSS software. The sample of undergraduate students and the percentages of the three different colleges are shown in table 1. Faculty of Engineering (SSP) performs (36%), faculty of Commerce (CIS) represents (47%); and Sadat Academy for Management and Sciences (SAMS) exemplifies (17%). The sample 757 consisted of 445 male; its percentage is (%58.8) and 312 female and its percentage is (%41.2), most of students respondent 436 (58%) were aged more than 20 years and 321(42%).were aged 20 or under 20 years. These percentages represent the ratios of the three college's population.

TABLE 1: Demographic Characteristics of the Respondents

College	Frequency	Percent	Gender		Age	
			Male	Female	<20	>20
Sadat Academy For Management Sciences	131	17%	72	59	74	57
Faculty of Commerce	357	47%	222	135	152	205
Faculty of Engineering	269	36%	151	118	95	174
Total	757	100%	445	312	321	436

Table 2 also represents the students' time spent on social media every day; the survey finds that 296 of students (%47.5) spent from 3 to 6 hours on social media, 180 respondent (%28.9) spent from one to 3 hours, 111 of the respondent (%17.8) spent more than 6 hours and 36 of respondent (%5.8) spent less than one hour. Also most of students (53%) use both languages English and Arabic, in addition (51%) of students use social media from more than three years.

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

TABLE 2: Student's Background in Internet

Student's Background in Internet			
Questionnaire Questions	options	frequency	percent
Do you use social networks & have account?	Yes	757	100%
	no	0	0
How many hours do you spend in using social networks /week?	none	0	0
	<3	8	1%
	>3-<6	22	3%
	>6-<10	223	29%
	>10	504	67%
How much time do you spend on social media every day?	<hour	47	6%
	>1-<3	216	29%
	>3-<6	333	44%
	>6	161	21%
Which language do you use in social media	English	85	11%
	Arabic	271	36%
	Both	401	53%
Duration for using social media	<year	20	3%
	1-3years	349	46%
	>3years	388	51%

Figure1 shows the relation between the percent of students and the time they spent in using social media. This figure demonstrates that 21% of the students spend more than six hours daily on social media, 44% of the students spend from three to six hours, and 29% spend more than one hour and less than three hours. It indicates that the students spend much time in using social media every day.

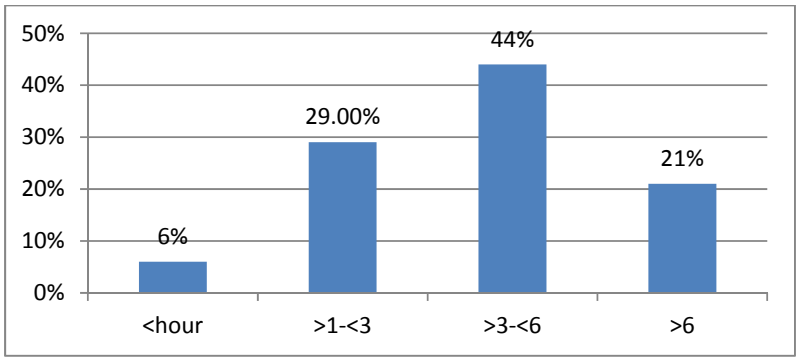


Figure 1: Time Spent on Social Media by Students

In the survey there was a multiple response type about what are the most commonly used social media tools. The top tools are Facebook (90%) as the most popular, followed by WhatsApp(86%), and YouTube(83%), Instagram (56%) as shown in figure 2. These social media tools considered as the top social media tools which are used by colleges' students.

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

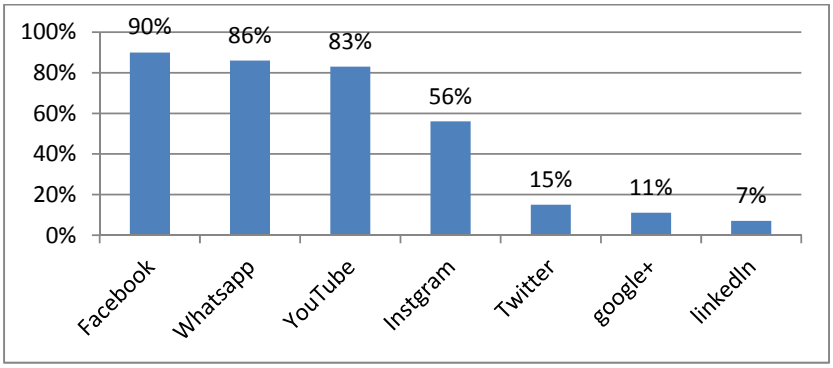


Figure. 2: Top Social Media tools Used by Students

Figure 3 illustrates the percentages of the results of the most common purposes of using social media networks. It shows that the first purpose is "entertainment" which represents (80%) considered as nonacademic purpose, the second is "search for information" (75%), the third is "learning" (62%), and the fourth is "share resources" (53%), These four purposes represents the most common purposes of using social media sites.

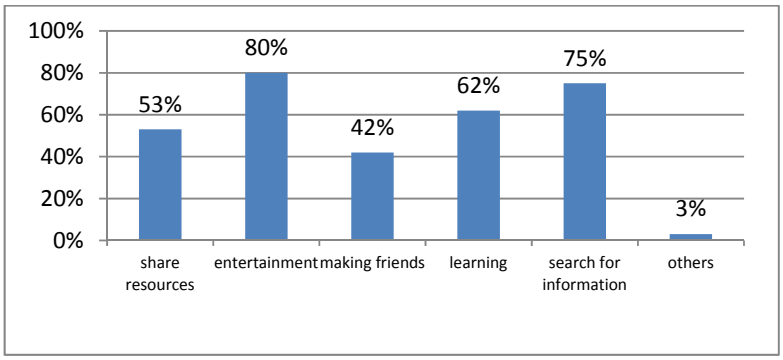


Figure. 3: Purposes of Social Media Used by Students

The data collected for the third part of the questionnaire about the advantages and disadvantages of using social media in learning are represented in tables 3 and table 4. These tables demonstrate the percentages of the descriptive statistics for the advantages and disadvantages using the Likert scale. Table 3 reflects the opinion of students about the advantages of using social media in learning. Most of the responses support the advantages of the social media in learning with high percentages, for example descriptive analysis of the responses found that (37.9%) strongly agree and (39.7%) agree in response to "social media improve my interest in learning" as shown in Figure 4. Another example is "social media help me to communicate with my instructors" (42.8%) strongly agree and (50.6%) agree.

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

TABLE 3: Descriptive statistics of advantages of social media in learning

Item No.	Questionnaire Indicator	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree	Not Responded	Total
	Advantages of using social media in Learning							
1	Social media help me to communicate with my classmates	350 (46.2%)	385 (50.9%)	12 (1.6%)	6 (0.8%)	4 (0.5%)	0 (0%)	757 100%
2	Social media help me to communicate with my instructors	324 (42.8%)	383 (50.6%)	14 (1.8%)	27 (3.6%)	9 (1.2%)	0 (0%)	757 100%
3	Social media give me chance to access new resources	315 (41.6)	342 (45.2)	10 (1.3)	76 (10.1%)	14 (1.8%)	0 (0%)	757 100%
4	Social media assist me to be a good learner	322 (42.5%)	339 (44.8%)	19 (2.5%)	52 (6.9%)	25 (3.3%)	0 (0%)	757 100%
5	Social media improve my leadership skills	268 (33%)	295 (43.6%)	20 (9.2%)	102 (11%)	68 (2.7%)	4 (0.5%)	757 100%
6	Social media makes learning easier between students and teachers	324 (42.8%)	354 (46.8%)	21 (2.8%)	37 (4.9%)	19 (2.5%)	2 (0.2%)	757 100%
7	Social media help me to collaborate with others	333 (44.0%)	367 (48.5%)	12 (1.6%)	34 (4.5%)	11 (1.4%)	0 (0%)	757 100%
8	Social media improve the ability to be creative and innovative	305 (40.3%)	327 (43.2%)	21 (2.8%)	63 (8.3%)	34 (4.5%)	7 (0.9%)	757 (100%)
9	Social media improve my research skills	312 (41.2%)	351 (46.4%)	22 (2.9%)	53 (7.0%)	16 (2.1%)	3 (0.4%)	757 (100%)
10	Social media reduce cost of learning	304 (40.2%)	361 (47.7%)	18 (2.4%)	48 (6.3%)	24 (3.2%)	2 (0.2%)	757 (100%)
11	Social media help me to improve problem solving skills	269 (35.5%)	285 (37.7%)	25 (3.3%)	114 (15.1%)	57 (7.5%)	7 (0.9%)	757 (100%)
12	Social media improve my interest in learning	287 (37.9%)	301 (39.7)	33 (4.4%)	86 (11.4%)	46 (6.1%)	4 (0.5%)	757 (100%)
13	Social media makes learning more interesting	327 (43.2%)	395 (52.2%)	4 (0.5%)	21 (2.8%)	9 (1.2%)	1 (0.1%)	757 (100%)
14	Social media help me to finish my tasks more easier	292 (38.6%)	356 (47.0%)	18 (2.4%)	65 (8.6%)	23 (3.0%)	3 (0.4%)	757 (100%)

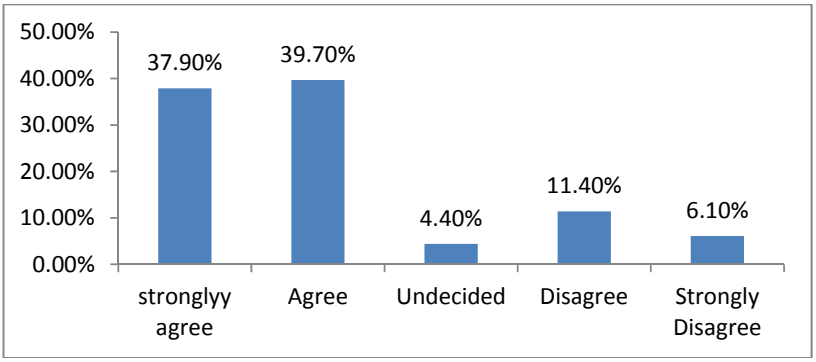


Figure. 4: Social Media Improve My Interest in Learning.

Table 4 shows the questionnaire elements that indicate the student's opinion about the disadvantages of using social media in learning. For example the descriptive analysis of the responses found that 32.4% strongly disagree and 23.4% disagree that Social

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

media has a bad influence on student time spent on non-academic activities. Another disdvantage represents that social media is uneasy to manage learning activities through 42.9% strongly disagree and 30.5% disagree as shown in Figure 5.

TABLE 4: Descriptive statistics of disadvantages of social media in learning

Item No.	Questionnaire Indicator	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree	Not Responded	Total
	Disadvantages of using social media in Learning							
1	Social media can cause misuse	109 (14.4%)	137 (18.1%)	23 (3.0%)	297 (39.2%)	182 (24.1)	9 (1.2%)	757 100%
2	Social media consume more time than the topics are worth	78 (10.3%)	114 (15.1%)	13 (1.7%)	295 (39.0%)	253 (33.4%)	4 (0.5%)	757 100%
3	social media need work and preparation	68 (9.0%)	112 (14.8%)	19 (2.5%)	301 (39.7%)	251 (33.2%)	6 (0.8%)	757 100%
4	Social media distract me from learning	178 (23.5%)	192 (25.4%)	19 (2.5%)	243 (32.1%)	120 (15.8%)	5 (0.7%)	757 100%
5	Social media cause intrusion on my privacy	81 (10.7%)	114 (15.1%)	20 (2.6%)	294 (38.8%)	246 (32.5%)	2 (0.3%)	757 100%
6	Social media has a bad influence on student time spent on non - academic activities	105 (13.9%)	217 (28.7%)	11 (1.5%)	245 (32.4%)	177 (23.4%)	2 (0.3%)	757 100%
7	Social media need training	46 (6.1%)	67 (8.9%)	9 (1.2%)	345 (45.6%)	286 (37.7%)	4 (0.5%)	757 100%
8	Social media, it's uneasy to manage learning activities.	75 (9.9%)	109 (14.4%)	14 (1.9%)	325 (42.9%)	231 (30.5%)	3 (0.4%)	757 100%
9	Social media Raise my financial expenses.	62 (8.2%)	78 (10.3%)	43 (5.7%)	358 (47.3%)	213 (28.1%)	3 (0.4%)	757 100%

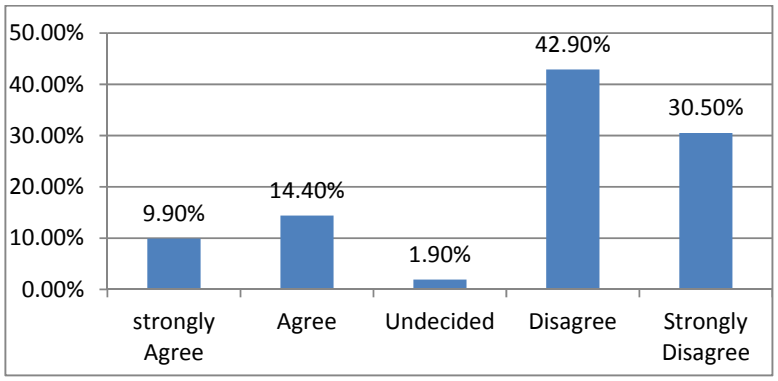


Figure. 5: Social media it's uneasy to manage learning activities.

5. RESULTS AND DISCUSSIONS

Data in Table 5 and 6 illustrate (mean, standard deviation, students' perception) of the advantages and disadvantages of using social media in learning in descending order. Table 5 shows that students are strongly agree the top five of the advantages of using social media as a learning tool and agree the remaining nine of the advantages. The highest mean score (4.41) represents the students responses to the question “social media help me to communicate with my classmates” with 350 strongly agree and 385 agree,

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

whereas the lowest mean score (3.79) of students responses to the question “Social media improve my leadership skills”. However students’ perception of the disadvantages of using social media in learning are (unsure, disagree) about the disadvantages, in table 6 the highest mean score (3.09) for the question “social media distract me from learning”, so the students are neutral to the questionnaire because they are not sure about the influence of the social media about distracting them from learning. While the lowest mean score (1.99) illustrates that students are disagree to the question “social media require training”.

TABLE 5: Mean and Standard Deviation of the Advantages on Using Social Media in Learning.

	Questionnaire indicator	N		Mean	Std. Deviation	Perception
Item No.		Valid	Missing			
1	social media help me to communicate with my classmates	757	0	4.41	0.621	Strongly agree
2	Social media makes learning more interesting	756	1	4.34	0.738	Strongly agree
3	Social media help me to communicate with my instructors	757	0	4.30	0.778	Strongly agree
4	Social media help me to collaborate with others.	757	0	4.29	0.826	Strongly agree
5	Social media makes learning easier	755	2	4.23	0.910	Strongly agree
6	Social media improve my research skills.	754	3	4.18	0.940	Agree
7	Social media assist me to be a good learner	757	0	4.16	0.998	Agree
8	Social media reduce cost of learning	755	3	4.16	0.971	Agree
9	Social media give me chance to access new resources	757	0	4.15	0.989	Agree
10	Social media help me to finish my school tasks more easier	754	3	4.10	1.011	Agree
11	Social media improve the ability to be creative and innovative	750	7	4.07	1.084	Agree
12	Social media improve my interest in learning	753	4	3.93	1.195	Agree
13	Social media help me to improve problem solving skills	750	7	3.79	1.282	Agree
14	Social media improve my leadership skills	753	4	3.79	1.302	Agree

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

TABLE 6: Mean and Standard Deviation of the Disadvantages on Using Social Media in Learning.

Item No.	Questionnaire indicator	N		Mean	Std. Deviation	Perception
		Valid	Missing			
1	Social media distract me from learning	752	5	3.09	1.469	Unsure
2	Social media has a bad influence on student time spent on nonacademic activities	754	3	2.77	1.434	Unsure
3	Social media can cause misuse	748	9	2.59	1.404	Unsure
4	Social media cause intrusion on my privacy	755	2	2.32	1.349	Disagree
5	Social media make me delay my meals, sleepless.	754	3	2.30	1.308	Disagree
6	Social media consume more time	753	4	2.29	1.344	Disagree
7	social media need work and preparation	757	0	2.26	1.306	Disagree
8	Spent more raise my financial expenses	755	3	2.23	1.201	Disagree
9	Social media require training	757	0	1.99	1.140	Disagree

Using Likert-type scales lead derive to calculate Cronbach's alpha which is a statistical analysis to calculate the coefficient for internal consistency reliability for any scales may be using. The generally agreed upon lower limit for Cronbach's alpha is 0.70 [15], which are regarded as acceptable reliability coefficients. The results of the reliability analysis are presented in Table 7. As the table demonstrates that the questionnaire is a reliable measurement instrument.

TABLE 7. Reliability of Measurements

Questionnaire indicators	Number of items	Cronbach's Alpha
Advantages of using social media in learning.	14	0.917
Disadvantages of using social media in learning.	9	0.766

6. CONCLUSION AND FUTURE WORK

This study attempts to obtain students' perceptions if social media considered as a learning tool. In this study a survey was conducted by distributing 757 questionnaires to students from Faculty of Engineering, Faculty of Commerce, and Sadat Academy for management Sciences. Most of Egyptian students are familiar with using social media tools especially Facebook, WhatsApp, and YouTube. They use the social medial tools for different purposes academic like search for information, learning, share resources and non-academic like entertainment, making friends. A high percentage (62%) of students use social media in learning. Majority

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

of students agreed that social media help them to communicate with their classmates, instructors collaborate with others, also make learning more interesting and easier. Moreover they agreed that social media improve their research skills and experiences in learning, also assist them to be good learner, finish their school tasks easier, access new resources and reduce cost of learning. Also most of students are unsure that social media can distract them from learning, has a bad influence on spending time on nonacademic activities, and it also can causes misuse. Furthermore they disagree that social media sites causes intrusion on their privacies, make them delay their meals, and consume more time than the topic is need.

The students should concentrate in using social media for educational purposes; students should use social media to enhance academic activities and accordingly avoid negative impacts. Students also should be observed by their teachers and parents on how they use the sites of social media to create a balance between using social media in entertainment and academic activities they need to control and manage their times to avoid the negative impacts on the students' academic performance. The Egyptian government should adopt the strategies to benefit from social networking sites in education. In future work, replicate the study in different settings in Egypt will allow for comparisons, and study the influence of social media on student's academic performance.

REFRENCES

[1] Mingle Jeffery and Adams Musah. (2015), "Social Media Network Participation and Academic Performance In Senior High Schools in Ghana", Library Philosophy and Practice(e-journal). Paper 1286.

[2] Yang Heng-Li and Tang Jih-Hsin. (2003), " Effectsof Social Network on Students' Performance: A Web- Based Forum Study in Taiwan",JALN, Vol. 7 (3), PP. 93-107.

[3] Al-Tarawneh Heyam A. (2014), " The Influence of Social Networks on Students' Performance", Journal of Emerging Trends in Computing and Information Sciences, Vol. 5 (3), PP. 200-205.

[4] Sponcil M. and Gitimu P.(2013),"Use ofsocial media by college students: Relationship to communication and self-concept", Journal of Technology Research, Retrieved from http://www.aabri.com/manuscripts/121214.pdf

[5] Tayseer M., Zoghieb F., Alcheikh I., and Awadallah Mohammad N.S. (2014), " Social Network: Academic and Social Impact onCollege Students", http://www.asee.org/documents/zones/zone1/2014/Student/PDFs/185.pdf

[6] Velenzuela, S., Park, N., & Kee, K.F. (2008), "Lessons from Facebook: The Effect of Social Network Sites on College Students' Social Capital", Retrieved from https://online.journalism.utexas.edu/2008/papers/Valenzuela.pdf

[7] Mehmood S and Taswir T. (2013)."The Effects of Social Networking Sites on the Academic Performance of Students in Collegeof Applied Sciences", International Journal of Arts and Commerce, Vol. 2 (1).

[8] Al-Rahmi W. and OthmanM. (2013), "The Impact of Social Media use on AcademicPerformance among university students: A PilotStudy", Journal of Information Systems Research and Innovation, http://seminar.utmspace.edu.my/jisri/

[9] Al-Sharqi L., Hashim K. and Kutbi I. (2015), " Perceptions of Social Media Impact on Students' Social Behavior: A Comparison between Arts and Science Students", International Journal of Education and Social Science, Vol. 2(4), PP. 122-131.
Peter O. (2015), "Social Media and Academic Performance of Students In University of Lagos ", https://www.researchgate.net/publication/273765340

[10] Ramprathap K., Rajaram S., Sriram V. and Ahamed S. (2016), "Emerging Trends of ends of Social Networking Sites Towards Professional Students: essional Students: essional Students: Conceptual Review", IJCTA, Vol.9 (21), PP.71-75.

[11] Kamal T., Tariq M., Ishtiaq M., Nawab K., Idrees M. (2015), " An Investigation into the Negative Impacts of Social Media on Academic Performance of Youth", IJSTE, Vol. 34, P-50-56.

[12] Bagget, S.B., & Williams, M. (2012). "Student Behaviors and Opinions Regarding the Use of Social Media, Mobile Technologies, and Library Research", Virginia Libraries, 58(1), 19-22. Retrieved from http://scholar.lib.vt.edu/ejournals/VALib/v58_n1/baggett.html.

[13] Amin Z., Mansoor A., Hussain S. And Hashmat F. (2016), " Impact of Social Media of Student's Academic Performance", International Journal of Business and Management Invention, Vol. 5 (4), P.22-29.

[14] Shambare, R., Rugimbana, R., & Sithole, N. (2012)," Social networking habits among students", African Journal of Business Management, Vol. 6(2), P. 578-786.

[15] Al-Rahimi1 W., Othman M., and Musa M., (2013) "Using TAM Model To Measure The Use Of Social Media For Collaborative Learning", International Journal of Engineering Trends and Technology (IJETT) – Vol 5 No. 2.

[16] Brown J.,(2002) "The Cronbach alpha reliability estimate The Cronbach alpha", Shiken: JALT Testing & Evaluation SIG Newsletter., Vol. 6 No.1, P. 17 – 18.

[17] Ghadi I., Alwi N., Bakar K. and Talib O., "Construct Validity Examination of Critical Thinking Dispositions for Undergraduate Students in University Putra Malaysia Ibrahim", Higher Education Studies, Vol. 2, No.2.
http://www.stata.com/meeting/spain15/abstracts/materials/spain15_alarcon.pdf

[18] Campball D. and Campball S. (2008), "Introduction to regression and data analysis", StatLab Workshop Series.

[19] Rawlings J., Pantula S., and Dickey D. (1998), "Applied Regression Analysis: A Research Tool Methods", Second Edition, Springer-Verlag New York, Inc.

[20] Sharma S. (2015), "Use of Social Networking Sitesby Undergraduates in Relation to Their Academic Achievement", Scholarly Research Journal for Interdisciplinary Studies, Vol.3 (21), PP. 1229-1234.

[21] Roy S. and Chakraborty S. (2015), "Impact of Social Media / Social Networks on Education and life of Undergraduate level students of Karimganj town-A survey" International Research Journal of Interdisciplinary & Multidisciplinary Studies (IRJIMS) , Vol. 1, PP. 141-147

[22] ElGazzar N. ADOLESCENTS' PERCEPTION AND ATTITUDES TOWARDS SOCIAL MEDIA NETOWRKS IN EGYPT – A SURVEY

[23] Zhanga Z. and Xueb Y. (2015), "An Investigation of How Chinese University Students Use Social Software for Learning Purposes", Procedia - Social and Behavioral Sciences, PP. 70-78.

[24] Kamel S. (2015), "The Value of Social Media in Egypt's Uprising and Beyond", The Electronic Journal on Information Systems in Developing Countries (EJISDC), Vol. 60 (5),PP. 1-7.

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

[25] Muhingi W., Mutavi T., Kokonya D., Simiyu V., Musungu B., Obondo A., Kuria M. (2015), "Social Networks and Students' Performance in Secondary Schools: Lessons from an Open Learning Centre, Kenya", Journal of Education and Practice, Vol.6 (21), PP.171-177.

[26] Henseler J, Ringle Ch and Sarstedt M. (2015), "A new criterion for assessing discriminant validity in variance-based structural equation modeling", J. of the Acad. Mark springer, P-115-135

[27] Alshareef, M. (2013). Evaluate Student Satisfaction for Social Learning Network at King Abdulaziz University.Advances in Internet of Things, 3, 41–44.

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

IJCSIS REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia
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 2017-2018
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/