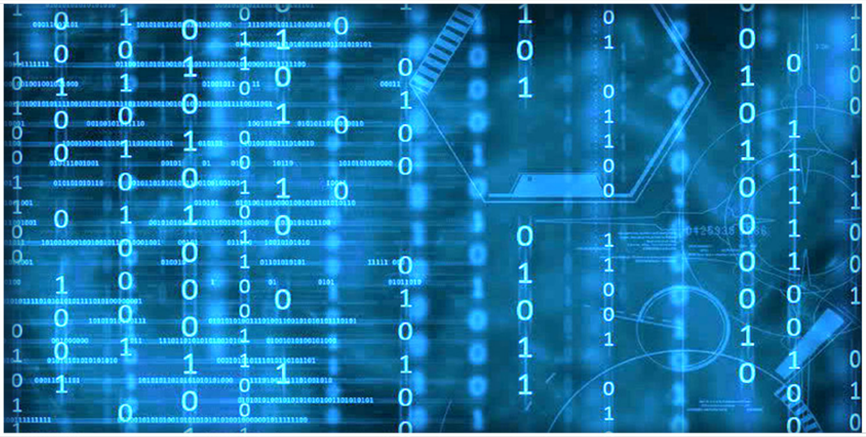
IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 9 Issue 7

July 2018



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

SAI

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 9 Issue 7 July 2018
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

Hesham G. Ibrahim

Hikmat Ullah Khan

Himanshu Aggarwal

Hongda Mao

Hossam Faris

Huda K. Kadhim AL-Jobori

Hui Li

Hüseyin Oktay ERKOL

Ibrahim Adepoju Adeyanju

Ibrahim Missaoui

Ikvinderpal Singh

Ilayaraja Muthalagu

Imad Zeroual

Imed JABRI

Imran Ali Chaudhry

Imran Memon

IRFAN AHMED

ISMAIL YUSUF

iss EL OUADGHIRI

Iwan Setyawan

Jabar H Yousif

Jacek M. Czerniak

Jafar Ahmad Alzubi

Jai Singh W

JAMAIAH HAJI YAHAYA

James Patrick Henry Coleman

Jamil Abdulhamid Mohammed Saif

Jatinderkumar Ramdass Saini

Javed Anjum Sheikh

Jayapandian N

Jayaram M A

Jerwinprabu A

Ji Zhu

Jia Uddin Jia

Jim Jing-Yan Wang

John P Sahlin

JOHN S MANOHAR

JOSE LUIS PASTRANA

José Santos Reyes

Jui-Pin Yang

Jungu J Choi

Jyoti Chaudhary

Jyoti Gautam

K V.L.N.Acharyulu

Ka-Chun Wong

Kamatchi R

Kamran Kowsari

KANNADHASAN SURIIYAN

KARTHIK MURUGESAN

KASHIF MUNIR

Kashif Nisar

Kato Mivule

Kayhan Zrar Ghafoor

Kennedy Chinedu Okafor

KHAIRULLAH KHAN KHAN

Khaled Loukhaoukha

Khalid Mahmood

Khalid Nazim Sattar Abdul

Khin Wee Lai

Khurram Khurshid

KIRAN SREE POKKULURI

KITIMAPORN CHOOCHOTE

Kohei Arai

Kottakkaran Sooppy Nisar

kouki Mohamed

Krasimir Yankov Yordzhev

Krassen Stefanov Stefanov

Krishna Kishore K V

Krishna Prasad Miyapuram

Labib Francis Gergis

Lalit Garg

LATHA RAJAGOPAL

Lazar Vojislav Stošic

Le Li

Leanos A Maglaras

Leon Andretti Abdillah

Lijian Sun

Liming Luke Chen

Ljubica B. Kazi

Ljubomir Jerinic

Lokesh Kumar Sharma

Long Chen

M A Rabbani

M. Reza Mashinchi

M. Tariq Banday

Madihah Mohd Saudi

madjid khalilian

Mahdi H. Miraz

Mahmoud M Abd Ellatif

Mahtab Jahanbani Fard

Majharoddin Kazi Kazi

majzoob kamal aldein omer

Malack Omae Oteri

Malik Muhammad Saad Missen

Mallikarjuna Reddy Doodipala

Man Fung LO

Manas deep

Manisha Gupta

Manju Kaushik

Manmeet Mahinderjit Singh

Manoharan P.S.

Manoj Manoj Wadhwa

Manpreet Singh Manna

Manuj Darbari

Marcellin Julius Antonio Nkenlifack

Marek Reformat

Maria-Angeles Grado-Caffaro

CONTENTS

Wavelet/PSO-Based Segmentation and Marker-Less Tracking of the Gallbladder in Monocular Calibration-free Laparoscopic Cholecystectomy

Haroun Djaghloul
University of Ferhat Abbes - Setif 1, Algeria

Jean-Pierre Jessel
IRIT, Institut de Recherche en Informatique de Toulouse, France

Mohamed Batouche
University of Constantine 2, Algeria

Abdelhamid Benhocine
University of Ferhat Abbes - Setif 1, Algeria

Abstract—**This paper presents an automatic segmentation and monocular marker-less tracking method of the gallbladder in minimally invasive laparoscopic cholecystectomy intervention that can be used for the construction of an adaptive calibration-free medical augmented reality system. In particular, the proposed method consists of three steps, namely, a segmentation of 2D laparoscopic images using a combination of photometric population-based statistical approach and edge detection techniques, a PSO-based detection of the targeted anatomical structure (the gallbladder) and, finally, the 3D model wavelet-based multi-resolution analysis and adaptive 2D/3D registration. The proposed population-based statistical segmentation approach of 2D laparoscopic images differs from classical approaches (histogram thresholding), in that we consider anatomical structures and surgical instruments in terms of distributions of RGB color triples. This allows an efficient handling, superior robustness and to readily integrate current intervention information. The result of this step consists in a set of point clouds with a loosely gradient information that can cover various anatomical structures. In order to enhance both sensitivity and specificity, the detection of the targeted structure (the gallbladder) is based on a modified PSO (particles swarm optimization) scheme which maximizes both internal features density and the divergence with neighboring structures such as, the liver. Finally, a multi-particles based representation of the targeted structure is constructed, thanks to a proposed wavelet-based multi-resolution analysis of the 3D model of the targeted structure which is registered adaptively with the 2D particles generated during the previous step. Results are shown on both synthetic and real data.**

Keywords*—*Medical image segmentation; monocular laparoscopic cholecystectomy; deformable structures tracking; gallbladder segmentation and tracking; markerless augmented reality; wavelets; particles swarm optimisation; minimally invasive surgery (MIS); computer aided surgery (CAS)

I. Introduction

Medical augmented reality consists in a set of techniques that allow the visualization in transparency of anatomical and pathological structures reconstructed pre-operatively using medical images (IRM, CT-Scan) in the surgeon's field of view. Augmented reality provides contextual information in an intuitive and easily implemented display [1], [2]. However, one of the major challenges remains in the limits augmented reality use in clinical laparoscopic abdominal surgery is the difficulty of marker-less immediate and precise registration of preoperative deformable 3D models of digestive organs reconstructed using medical images such as MRI or CT-Scan on the intra-operative laparoscopic view.

Augmented reality allows the enhancement of perceptual capabilities of surgeons during the intervention directly on their filed of view of the intervention or projected on the patient. Because of the rigidity of manipulated anatomical structures, many augmented reality systems have been integrated in the operative rooms especially for orthopedic and neurosurgery. However, in the case of highly deformable anatomical structures as in the case of abdominal surgery, many challenges have been encountered due to the difficulty to precisely track and register the targeted anatomical structures. On the other hand, the massive adoption of minimally invasive surgery techniques even with their advantages have introduced other drawbacks such as the lack of tactile sensation, the limitation of the intervention field of view and the inversion of surgical instruments gesture orientations. These problems can be overcome, thanks to the use of augmented reality.

A. Motivation and Proposition

In this study, we focus on the segmentation and marker-less tacking of the gallblader during monocular minimally invasive laparoscopic cholecystectomy intervention with no camera calibration parameters available. Cholecystectomy is the standard procedure for surgical treatment of gallbladder diseases mainly for symptomatic cholelithiasis (gallstones). It consists in the ablation of the gallbladder and its extraction from the abdomen of the patient. In the case of minimally invasive surgery or laparoscopic surgery, a set of special surgical instruments are inserted into the abdominal cavity of the patient through a small incisions.

Cholecystectomy is the first surgical intervention in the United States with more than a half million operations done each year. Since the first cholecystectomy of Langenbuch [3], [4]. Indeed, cholelithiasis is an extremely common gallbladder condition, generally reaching the quart of the population beyond 50 years with one of three women and one of five men that have or will have it. Cholecystectomy consists in the complete removal of the gallbladder with different techniques such as open, laparoscopic [3], SILS or NOTES [5], [6] procedures. The video-assisted laparoscopic cholecystectomy is currently the gold standard technique with more than 98% of performed interventions [7]–[10].

Digestive organs are highly deformable leading to certain displacement of any physical radio opaque markers between preoperative and intra-operative acquisitions. Moreover, inter and intra-patient geometric and anatomical variability and the great complexity of intra-abdominal surgical environment caused by dissection and bleeding. The preoperative patient specific 3D model can only be used as initial solution for anatomical and pathological structures detection and tracking. Therefore, the surgical intervention cannot be safe and precise in this such complex context of monocular laparoscopic cholecystectomy without handling and tracking the deformation of the primary manipulated anatomical structure which is the gallbladder.

Here, we propose a method for tracking digestive organs on the view of previous medical augmented reality systems in laparoscopic cholecystectomy. Thus, the proposed method can be used as a priori step to markers-based registration systems to align coarsely the patient CT/MRI model reconstructed before the surgical intervention. In other side, the method can be used intra operatively to track targeted organs or surgical instruments with partially occluded or totally invisible markers. Therefore, we propose a new nearly-automatic statistical color model construction method and its application to the pixel-wise anatomical structures detection and tracking in the context of the laparoscopic cholecystectomy.

B. Paper Organisation

The rest of the paper is organized as follows. In Section 2, we outline the fundamentals of a standard laparoscopic cholecystectomy procedure and its operative workflow. Then, we review state of the art of the methods used to handle deformable objects detection and tracking mainly used in laparoscopic medical augmented reality systems and discussing their capabilities and limitations. In Section 4, we describe materials and provide the mathematical formula and the necessary background and implementation of the proposed method. In Section 5, we present experimental results and the overall characteristics of the method. Section 6 presents our conclusions, potential applications of the proposed methods and perspectives.

II. Medical Background

According to the common and standard ports installation and intervention workflow providing optimal results [3], [11], [12], the basic laparoscopic cholecystectomy intervention according to the European operating technique begins with achieving a perfect exposure of the right sub-hepatic region. Then, the surgeon uses the inserted laparoscopic camera to detect and identify all anatomical structures in the abdominal cavity. The next major steps are mainly the dissection of Callot's triangle, dissection/clipping and division of the cystic artery and duct. Finally, a complete removal of the gallbladder is achieved by dissecting the gallbladder bed with the liver. In most cases, computer-assisted surgery workflow models are created manually, which is a time consuming process that might suffer from a personal bias. In their work [13], Blum et al. presented a graphical user interface based on an approach for automatic workflow mining using ten process logs, each describing a single instance of a laparoscopic cholecystectomy, to build a Hidden Markov Model (HMM) with embody statistical information concerning aspects like duration of actions or tool usage during the surgery. In Fig. 1, we outline our proposition for standard laparoscopic cholecystectomy workflow model based on six coarse main steps.

From a topographical anatomy point of view, the principal anatomical structures of the right upper quadrant that have to be explored, during the intra-operative detection and identification step, in the operating field of view are the liver, the gallbladder, the round ligament, the stomach, the duodenum, the transverse colon, the lesser omentum, the hepatic flexure and the greater omentum. In our context, the most important anatomical structure is the gallbladder and its vascular supply. The major gallbladder anatomical structures are the fundus, the body, the infundibulum, the cystic duct, the common hepatic duct and the common bile duct. The vascular supply elements are mainly the cystic artery, the Mascagni lymph node, the proper hepatic artery, the abdominal aorta, the portal vein and gastro-duodenal artery. More details about sub-hepatic anatomical structures can be found in medical literature. In this paper we base our work on the Foundational Model of Anatomy (FMA) ontology to guide the anatomical structures modeling and recognition.

III. Related Works

Tracking real objects is an important topic in computer vision. Many methods for tracking real objects have been proposed in the literature. In this paper, we are mainly interested to the visual tracking of non-rigid objects. Several methods have been proposed with applications to different domains [14]. However, few clinical results exist for deformable abdominal organs tracking in laparoscopic cholecystectomy. Medical tracking systems for digestive surgery can be classified into two categories, namely, optical and hybrid systems. In their work [15], Nicolau et al. proposed a low cost and accurate guiding system for laparoscopic surgery with validation on abdominal phantom. The system allows real time tracking of surgical tools and registration at 10 Hz of the preoperative patient CT/MRI reconstructed model with accuracy tracking of instrument tip close to 1.5 mm and endoscopic overlay error under 1.0 mm. The system is totally based on the AR-Toolkit [16], [17] markers and patterns use for both patient model registration and surgery instruments tracking. To register abdominal markers the method minimize the Extended Projective Points Criterion (EPPC) instead of the Standard Projective Points Criterion (SPPC) because of its error support either of 2D image or 3D CT-Scan data [18], [19]. Validation of the criterion has mainly been made for radio-frequency surgery without abdominal gas insufflation. Accurate tracking and registration of such markers in real intra-abdominal laparoscopic surgery is very difficult because of the digestive organs deformation and the pneumoperitoneum establishment [3], [20]. Feuerstin et al. [21] use multiple optical and electromagnetic tracking systems to determine the position and orientation of intra-operative imaging devices (mobile C-arm, laparoscopic camera and flexible ultrasound) allowing direct superimposition of acquired patient data in minimally invasive liver resection. To our knowledge, there is

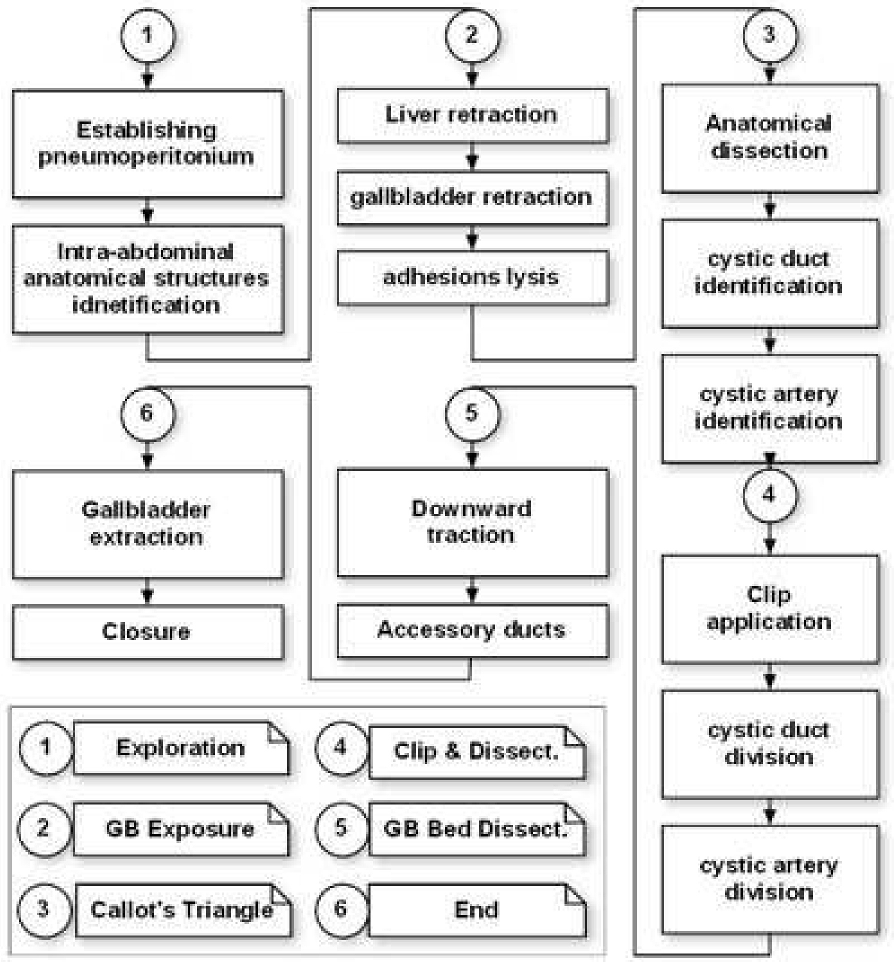


Fig. 1. Standard laparoscopic cholecystectomy procedure.

no clinically approved automatic marker-less tracking systems of the gallbladder during laparoscopic cholecystectomy.

IV. Proposed Method

In this section, we describe the proposed method for the segmentation and tracking of the gallbladder using photometric features. Therefore, we need to build a color model of gallbladder as well as the other neighboring anatomical structures. To achieve this goal with minimal user interaction, we propose the following straightforward method that can take into account the patient anatomy variability and the standard intervention workflow according to the European operative room set-up and common standard installation of the patient in the operative room during the intervention.

In Fig. 2, we show the global architecture of the proposed system for the application of the pre-operative statistical color model in anatomical and pathological structures detection and tracking tasks. Indeed, the system allows on-line enhancement of the initial pre-operative color model using the intra-operative laparoscopic intervention video.

In the following sub-sections we describe the off-line statistical anatomical color model construction. We first build a general histogram density using a set of high quality captured videos and photos of standard laparoscopic cholecystectomy interventions available at the World Wide Web. The images have 240 x 320 RGB coded pixels with 256 bins per channel (24 bits per pixel). Each video sequence is acquired at a frame rate of 30 Hz. We remove manually from the video training dataset all frames that are not relevant, such as tutorial text and operative room presentation, focusing only on inner abdominal laparoscopic camera photos. We have a final set of 16735 colored laparoscopic images, resulting in a training dataset containing more than one billion pixels.

A. Pre-operative Anatomical Color Model

This first step allows to automatically extract both important anatomical structures and surgical instruments blobs that are directly visible using the laparoscope camera. The first laparoscopic image of each training intervention video is captured and segmented manually into four main regions (liver, gallbladder, surgical instruments and other). Other region contains the pixels of remaining anatomical or pathological structures that are not segmented manually and with no impact on the following 3D model registration step. According to the cholecystectomy intervention workflow step (t), we construct for each anatomical region (i) a statistical color model using a histogram with 256 bins per channel in the RGB color space.

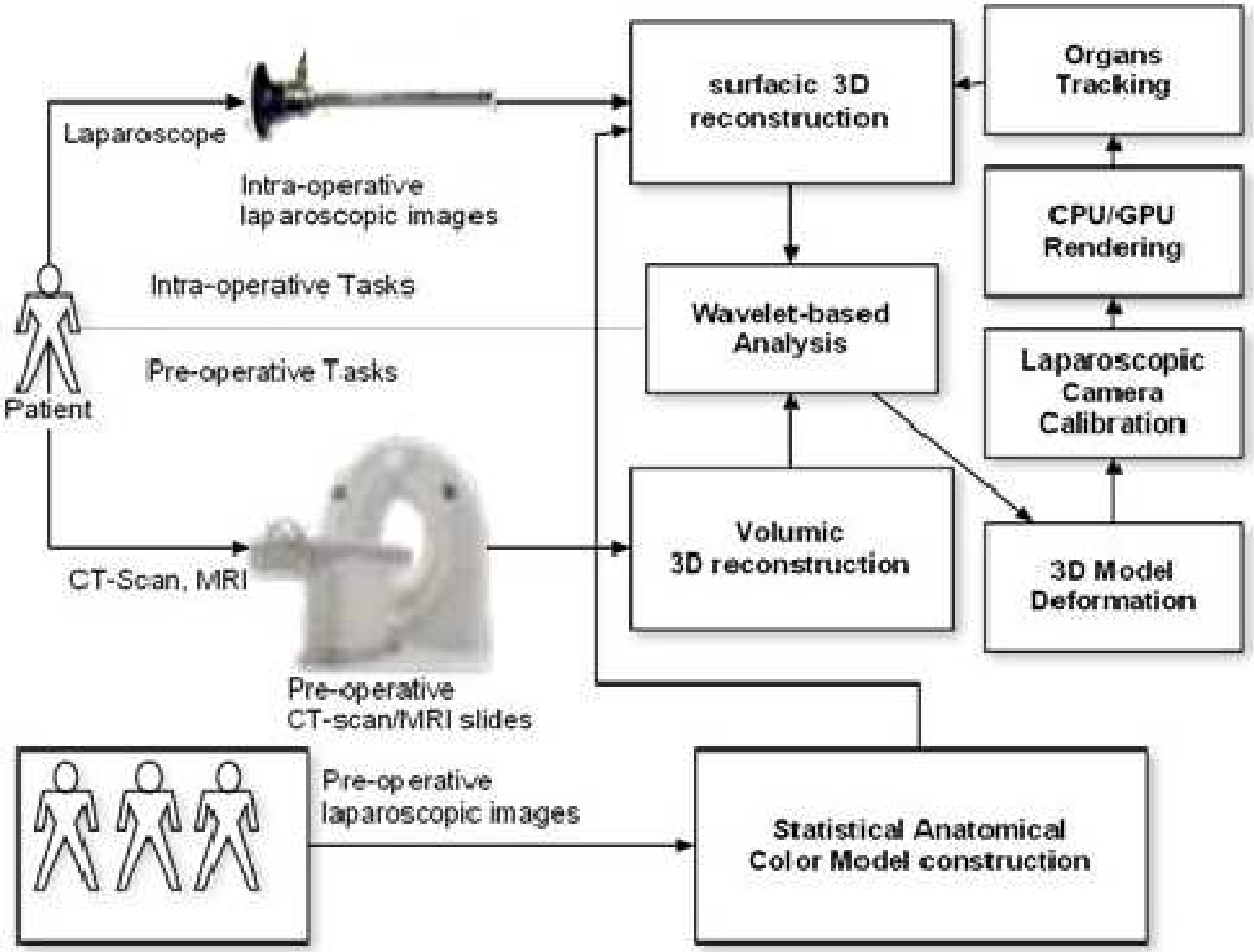


Fig. 2. A schematic illustration of the proposed method.

Each color vector (x) is converted into a discrete probability distribution in the manner:

$$P_{i,t}(x) = \frac{c_{i,t}(x)}{\sum_{j=1}^{N_{i,t}} c_{i,t}(x_j)}, t = t_1 \dots t_6, \, i = 0 \dots S_t \,. \quad (1)$$

where $c_{i,t}(x)$ gives the count in the histogram bin representing the *rgb* color triplet (x) and $N_{i,t}$ is the total count of the *rgb* histogram entries returned by the histogram bins number of the structure region (i) during the intervention step (t). the number of structures (anatomical, pathological or surgical instruments) S_t varies according to the procedure step and the priori knowledge-based patient-specific data. To ensure a generic color model construction, it is important to fix the steps and structures number for the whole training data set. In this study and according to the European standard and common laparoscopic cholecystectomy installation and intervention workflow described in section 2, the number of structures classes is limited to four ($S_t = 4$) and surgical steps to six ($t_1 = $ '$exploration...$'). The class structure $i = 0$ contains the histogram bins with corresponding *rgb* triples which are not included in construction of the previous color model. In practice, the step (t) denotes a time interval represented by the sequence of laparoscopic images of the same intervention $t = \left[I_{v,1}^t \dots I_{v,n}^t\right]$ in the different videos (v) that compose the training data set.

Several mathematical morphological operators are used to eliminate even noise small regions. The determination of major blobs can be performed thanks to the application of connected component labeling. For each anatomical blob we compute statistics such as centroids coordinates, blob area and probabilities of each RGB triple associated to each anatomical structure.

Once the laparoscopic images of the intervention were prepared, different statistical features and properties are extracted. First, we scan each image in the training dataset for all color model features. In Table I, we give some statistical properties of one the laparoscopic cholecystectomy intervention videos used in our study.

Fig. 3 shows the evolution of the RGB histogram bins count over the 16735 frames of the video training dataset.

TABLE I. GLOBAL TRAINING DATASET HISTOGRAM BINS STATISTICS

Feature	*RGBTriplet*	*Red*	*Green*	*Blue*
Mean	1997	31	63	31
Median	2024	32	64	32
Min	61	21	41	20
Max	3245	32	64	32
Standard deviation	499	0	1	0
Total count	10017	32	64	32

We can observe that each laparoscopic image can contains at most 10017 RGB color triplets over all the video sequence. therefore, the RGB histogram is mostly empty with 99,94% of the RGB triples that are not used.

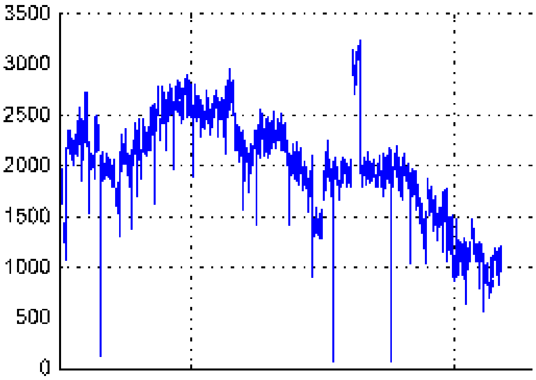


Fig. 3. Evolution of RGB bins count in video training dataset.

In Fig. 4, we can see the spatial distribution of a given color model feature within the laparoscopic image of the gallbladder. One can observe that the same is located over different regions that can belong to different anatomical structures or surgical instruments.

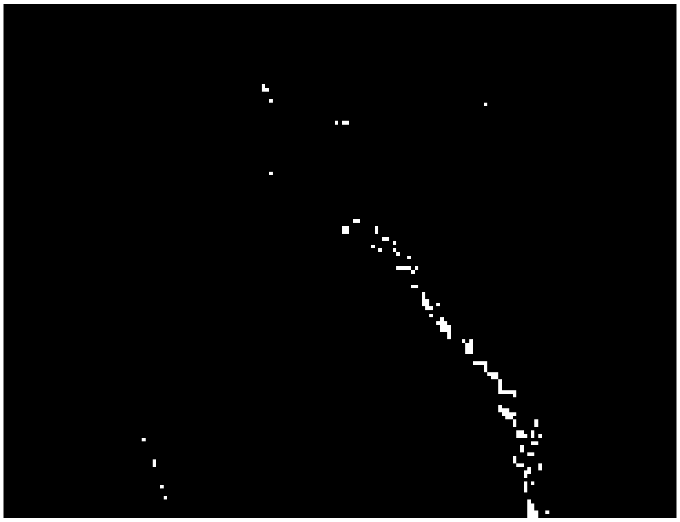


Fig. 4. Spatial distribution of an RGB color triplet in a single laparoscopic image.

The only application of the anatomical color and spatial model using the criteria given above can lead to a course segmentation of the laparoscopic image with a considerable number of artifacts as shown in Fig. 5.

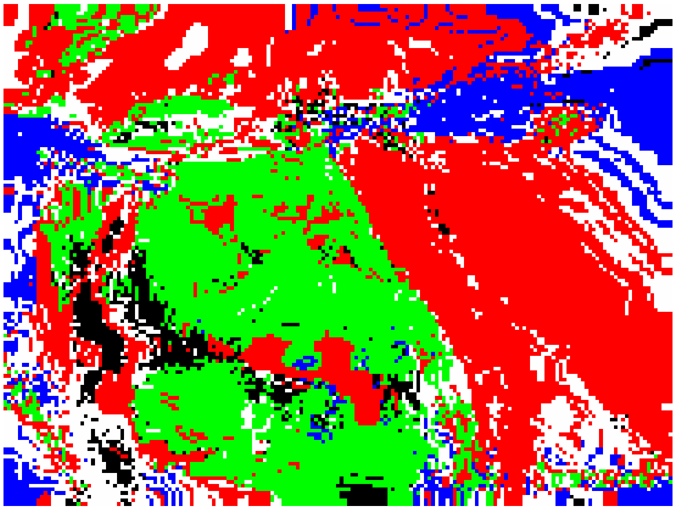


Fig. 5. Detection of anatomical structures using the proposed color model.

The result shown in Fig. 5 confirms the need of additional steps to enhance the segmentation and thus detection result of the gallbladder and surrounding anatomical structures such as the liver. These necessary steps are described in the following subsections given below.

B. Proposed Wavelet for Multi-resolution Tree of Spheres Modeling of Anatomical Structures

In this section, we propose a new multi-resolution analysis of 3D objects modeled as a set of elementary non intersected particles defined by their centers and rays. The virtual model of the anatomical structure is subdivided into a set of spheres. We call this representation the tree of spheres (TOS) model. Here, the closest greatest sphere to the preoperative 3D model gravity center represents the TOS model root or simply the TOS root. The TOS root is used during the first step of the 2D/3D registration between the preoperative reconstructed TOS model and the PSO-based gallbladder detection particles (PSO-DP). The establishment of a correspondence between the TOS root and the PSO-DPs allows to maintain a certain level of stability during deformable registration along the whole laparoscopic intervention video. The TOS root represents the coarsest resolution level of the virtual model of the tracked anatomical structure (the gallbladder).

We suppose that S^j is the TOS at the resolution level (*j*). We have:

$$S^j = \left[S_{j,1} S_{j,2} \dots S_{j,i} \dots S_{j,n_j}\right]^t \qquad (2)$$

where $S_{j,i}$ is the i^{th} sphere of the virtual model at the resolution (*j*) and n_j is the length of the spheres chain at the resolution level (*j*) denoting its number of spheres. The initial resolution level is S^0 and the coarsest one is S^r corresponding to the sphere chain root.

The relation between two successive resolution levels is given by:

$$\begin{aligned} S^{j+1} &= A^{j+1} S^j \\ D^{j+1} &= B^{j+1} S^j \end{aligned} \qquad (3)$$

with D^j represents the wavelet detail coefficients of the resolution level (*j*):

$$D^j = \left[D_{j,1} D_{j,2} \ldots D_{j,i} \ldots D_{j,n_j}\right]^t \tag{4}$$

The A^j and B^j matrices are called the analysis filters of the resolution level *j*.

To reconstruct the superior resolution level we use two matrices P and Q called synthesis filters. The initial resolution level is given by:

$$S^j = P^{j+1} S^{j+1} + Q^{j+1} D^{j+1} \tag{5}$$

The relation between the analysis and synthesis filters is formulated by:

$$[A|B]^t = [P|Q]^{-1} \, then \, [A|B]^t \, [P|Q] = I \tag{6}$$

In order to make a multi-resolution analysis of the spheres based subdivision of the anatomical structure virtual model, we have to compute the filters A^j, B^j, P^j and Q^j for each resolution level (*j*). In the simplest case, the transformation to an inferior resolution level (*j*) of the 3D decomposition of the volumetric model consists in replacing two spheres of the resolution level (*j-1*) by a representative one containing both of them.

The A^j filter is used to select elements of the next inferior resolution level and B^j to extract wavelet coefficients of each level. Hence, the analysis process is formulated by:

$$\begin{aligned} S^r &= A^r S^{r-1} = A^r A^{r-1} \ldots A^2 A^1 S^0 \\ D^r &= B^r S^{r-1} = B^r B^{r-1} \ldots B^2 B^1 S^0 \end{aligned} \tag{7}$$

Assuming that the initial spheres based decomposition is composed of (2^r) spheres. We have, $S^{(j=0)} = \left[S_{0,2^0} S_{0,2^1} \ldots S_{0,i} \ldots S_{0,2^r}\right]^t$ with $n_{(j=0)} = 2^r$ and *r* gives the number of levels to reach the coarsest representation corresponding to the spherlet root.

In the case where $r = 3$ (2^3 spheres), we have the TOS approximation and detail vectors given in Table II.

TABLE II. Analysis Process of a $8 = 2^3$ Spheres Chain

$S^0, D^0 = \emptyset$	S^1, D^1	S^2, D^2	$S^3 = root, D^3$
ϖ_1	ϖ_1	ϖ_1	ϖ_1
ϖ_2	ϖ_2	ϖ_2	$d_2 = \varpi_2 - \varpi_1$
ϖ_3	ϖ_3	$d_3 = \varpi_3 - \varpi_1$	$d_3 = \varpi_3 - \varpi_1$
ϖ_4	ϖ_4	$d_4 = \varpi_4 - \varpi_2$	$d_4 = \varpi_4 - \varpi_2$
ϖ_5	$d_5 = \varpi_5 - \varpi_1$	$d_5 = \varpi_5 - \varpi_1$	$d_5 = \varpi_5 - \varpi_1$
ϖ_6	$d_6 = \varpi_6 - \varpi_2$	$d_6 = \varpi_6 - \varpi_2$	$d_6 = \varpi_6 - \varpi_2$
ϖ_7	$d_7 = \varpi_7 - \varpi_3$	$d_7 = \varpi_7 - \varpi_3$	$d_7 = \varpi_7 - \varpi_3$
ϖ_8	$d_8 = \varpi_8 - \varpi_4$	$d_8 = \varpi_8 - \varpi_4$	$d_8 = \varpi_8 - \varpi_4$

C. Deformable Structures Detection

As in our previous work [22], the first step of the method consists in the precise detection of the gallbladder in the laparoscopic view. First, a deformable particles based model is constructed for each anatomical structure. This is performed by using pre-operative surfacic model of the anatomical structures concerned with the surgical intervention of the gallbladder ablation that have been generated from pre-operative medical image of the gallbladder and surrounding structures. Then, a registration scheme begins with a coarse 2D TOS root detection in the laparoscopic image using the color (C) and spatial (S) models that we have described above. The C-model is used to segment the laparoscopic cholecystectomy images and build the points cloud associated to the anatomical structure (i) which is visible in the step (t) according the surgical workflow of a standard laparoscopic intervention procedure as described in Section 2. The result is the construction of world frame using the relationship between the greatest and most stable parts of anatomical structures such as the liver and the gallbladder. "(1)" gives a pixel wise segmentation of the different structures visible according to the intervention step:

$$P_{C(i,t)}(x_{rgb}) \geq \theta_{rgb} \tag{8}$$

Therefore, we have an initial segmentation of initial of surgical instruments and anatomical structures. However, we observe generally a high correlation between organs *RGB* colors in the case of abdominal laparoscopic surgery. As already described above, an *RGB* triplet can be found in different regions with low spatial concentration and connectivity. Thus, we propose to use particles swarming to detect and track anatomical structures using the color as well as the spatial distribution.

Particles Swarm Optimization (PSO) is a global search strategy for optimization problems. The first version has been proposed by Kennedy and Eberhart [23] in 1995 and it is based on the social evolution simulation of an arbitrary swarm of particles based on the rules of Newtonian physic. Assuming that we have an N-dimensional problem, the basic PSO algorithm is formulated by position $x_m(t)$ and velocity $v_m(t)$ vectors representing the time evolution of *M* particles with random affected initial positions. Hence, we have:

$$x_m(t) = [x_1(t) \, x_2(t) \ldots x_N(t)]^T \tag{9}$$

$$v_m(t) = [v_1(t) \, v_2(t) \ldots v_N(t)]^T \tag{10}$$

The evolution of the swarm particles in the classical algorithm is done by the following equations:

$$\begin{aligned} v_m(t+1) &= f_{m_i} \, v_m(t) + f_{m_c} \, [D_c]_N \, (x_m(t_c) - v_m(t)) \\ &\quad + f_{m_s} \, [D_s]_N \, (x_{opt}(t_s) - v_m(t)) \end{aligned} \tag{11}$$

Thus, the new position of the particle *m* is given by:

$$x_m(t+1) = x_m(t) + v_m(t+1) \tag{12}$$

Where $v_m(t)$ and $v_m(t+1)$ are, respectively, the past and the new velocity vectors of the particle m. f_{m_i} is the inertia factor of the particle m, f_{m_c} is its the cognitive factor and f_{m_s} is the social factor. $[D_c]_N$ and $[D_s]_N$ are the N-dimensional diagonal matrices composed of statistically independent normalized random variables uniformly distributed between 0 and 1. t_c is the iteration where the particle m has reached its best position given by x_m. t_s is the iteration where the population has found its best global value given by the coordinates of the particle x_{opt}. It is obvious that particles reach their best local values before that one of them becomes the global best.

The particles swarm optimization method is a meta-heuristic used in combinatorial optimization problems. Its independence from the continuity and gradient information allows it superior behavior especially in cases where it is impossible to rely on the gradient descent because of discontinuity or hard gradient changes. For this reason, we base our method on the PSO method for the segmentation and tracking of the point clouds generated by the first step of the segmentation process based primarily on knowledge based pixel wise anatomical geometric and color model. As the result of the first step presented previously is a set of disconnected point clouds, it is very difficult if not impossible to apply traditional method such as histogram thresholding, edge detection and even deformable models based segmentation and tracking. In particular, the proposed method is used for the detection of the gallbladder in the video-based laparoscopic cholecystectomy intervention. In the laparoscopic cholecystectomy intervention, the endoscope is focused on the gallbladder so that it is always at the center of the laparoscopic endoscopic image.

The proposed scheme is inspired from the behavior of a swarm of predating eagles. In nature, social eagles swarm construct a circle in the sky around their prey. This is due to the anatomy of their eyes which are symmetric to the axis of the head allowing simultaneous visualization of the prey and the environment around it. In our PSO scheme, each particle in fact represents the left and the right eye of an eagle. For segmentation purposes, we can describe them as the in-eye and out-eye. the In-eye maximizes a set of features concerning the segmented and tracked organ and the Out eye represents the outer organs such as the liver in the case of the laparoscopic cholecystectomy. The In-Eye maximizes the density of ACM color bins defined in the previous of the targeted organ which is in our case the gallbladder. On the other hand, the Out-Eye maximizes the density of the bounding organ which is in this case the liver. Each of the In-Eye and Out-Eye particle are represented by a circle or a square delimiting coarsely the pixels of the respective organs in the endoscopic image. Thus, the two particles are defined using the upper left and the down right pixels in the images. The fitness function has the role of maximizing the density of the targeted organ (gallbladder), minimizing the density of the bounding organ (liver) and reducing the distance between the two particles (In-eye, Out-Eye) all without allowing any collision between them. Thus, this fitness function for the gallbladder as it is the targeted organ becomes

$$F_{gallbladder} = \frac{H}{D} \tag{13}$$

with

$$H = \frac{In - Eye_{density}}{Out - Eye_{density}}, \tag{14}$$

where the density of each particle is given by the ratio between the number of ACM rgb bins and the surface of the particle for each organ (gallbladder, liver). and

$$D = \frac{\|InEye_{center} - OutEye_{center}\|}{InEye_{radius} + OutEye_{radius}}, \tag{15}$$

Here, ($\|.\|$) denotes the Euclidean distance between the centers of the In-Eye and the Out-Eye particles. This allows to maximize the minimal variance between the targeted structure (gallbladder) and the surrounding structures such as the liver and the covering surgical instruments parts.

The determination of the the points number of the PSO tracking particle of the gallbladder in the laparoscopic view (without calibration) is given with the same manner given in our previous work [22]. Here, we present it for more clarity. the difference is in that, here, we are using a couple of bi-particles one for the inside and other for the outside in extension to our previous work which lies on only one internal particle for hte tracked structure. The number of pixels in each particle is given from the preoperative 3D model based on our previous method [22]. Here, we use the surfacic 3d model of the gallbladder reconstructed using pre-operative images such as ct scan or mri. Assuming the distance between the laparoscope tip (L_{tip}) and the pic of the gallbladder surface (G_{pic}) and assuming that the silouhette of the gallbladder is completely visible in the laparoscopic image, we propose to approximate the gallbladder point cloud by considering the ratio between the half of the surfaci corporal model area (Ω_{gal}) and that of an elementary surface projected into a camera CCD pixel sensor (ω). This ratio is given thus by,

$$\alpha = \frac{\Omega_{gal}}{2 * \omega}, \tag{16}$$

with,

$$\omega = \left| L_{tip}\vec{G}_{pic} \right| \frac{\Omega_{pixel}}{f}, \tag{17}$$

where (Ω_{pixel}) is the area of the pixel in the camera ccd matrix and (*f*) is the focal length which is the distance between the central point and the image place.

Now, if we consider that the gallbladder is modeled by a set of polygons P_i, we get

$$\Omega_{gal} = \sum_i \Omega_{P_i}, \tag{18}$$

By combining "(17)" and "(18)" in "(16)", α is given so that:

$$\alpha = \frac{f * \sum_i \Omega_{P_i}}{2 * \left| L_{tip}\vec{G}_{pic} \right| * \Omega_{pixel}}, \tag{19}$$

By taking P_i as small as a millimetric surfacic unit, we obtain:

$$\alpha = \frac{f}{2 * \rho^2 * \left| L_{tip}\vec{G}_{pic} \right|} * \nu, \tag{20}$$

where ν is the number of elementary surfaces (surfels) and ρ^2 is the metric area of ccd pixel. Here, we consider it constant during the intervention.

From "(20)", the only measure that is varying during the intervention is $\left|\overrightarrow{L_{tip}G_{pic}}\right|$. This is due to the cardiac and the respiratory activities. However, is is generally maintained by the surgical staff as constant and invariant as possible during the whole intervention.

By considering the preoperative medical images (MRI or CT-Scan) exist with millimetric precision, the parameter ν is computed as the length of the segmented gallbladder contour in each tomographic medical image. Assuming that for each preoperative image (i), the gallbladder contour length is given by (Γ_i). Then, "(20)" becomes:

$$\alpha = \frac{f}{2 * \rho^2 * \left|\overrightarrow{L_{tip}G_{pic}}\right|} * \Gamma, \tag{21}$$

with

$$\Gamma = \sum_i \Gamma_i, \tag{22}$$

Here, The principal wavelet sphere S_i^0 is projected on the gallbladder area in the 2D laparoscopic image which has the same perimeter as that of the contour of the gallbladder (Γ_i) in the slide (i). Then:

$$\Gamma_i = 2 * \pi * r_i, \tag{23}$$

Given (n) medical imaging slides that cover the target organ and by replacing Γ_i in "(22)" from "(23)", we have:

$$\Gamma = 2 * n * \pi * \sum_{i=1}^{n} r_i, \tag{24}$$

By replacing Γ from "(24)" in "(21)", we get:

$$\alpha = \frac{n * \pi * f}{\rho^2 * \left|\overrightarrow{L_{tip}G_{pic}}\right|} * \sum_{i=1}^{n} r_i, \tag{25}$$

By putting

$$\kappa = \frac{n * \pi * f}{\rho^2 * \left|\overrightarrow{L_{tip}G_{pic}}\right|}, \tag{26}$$

and

$$P = \sum_{i=1}^{n} r_i, \tag{27}$$

Then, α is given by:

$$\alpha = \kappa * P \tag{28}$$

The internal parameters of the laparoscopic camera are assumed to be invariant during the intervention. the distance between the laparoscope tip and the gallbladder pic can be determined using distance estimation techniques and devices.

V. Experimental Results

To assess the performance of the proposed PSO-based gallbladder detection method in laparoscopic images, we first conduct an experiment on the synthetic image using the method defined in the previous section. Our first synthetic image (Fig. 6) consists of a set of point clouds differing in volume and density. These point clouds represent the possible result of the photometric and textural segmentation step of the gallbladder during the laparoscopic cholecystectomy intervention. Here, the true positive points are those belonging to the greatest cloud positioned in the middle of the image. The points belonging to smaller point clouds or simply black areas around the primary targeted greatest point cloud are either false positive points or true negatives which represent in fact anatomical structures other than the targeted deformable structure which is the gallbladder in our case.

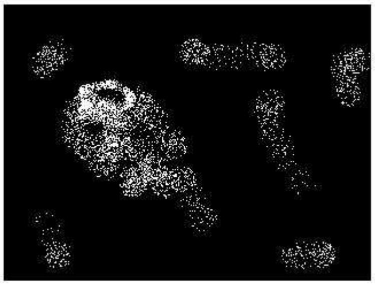


Fig. 6. Synthetic images representing a set of point clouds as the result of photometric and textural segmentation step used to test the PSO-based method.

The PSO-based method is applied to this synthetic image to detect a deformable point cloud representing a gallbladder with no need of explicit initialization of the position of tracking PSO swarm of particles. Initially, the particles are distributed randomly over all the original image. According to the used tracking shapes of the PSO particles (circular or rectangular for instance), each particle is characterized by either the particle's center and radius of the circle or the two points defining the rectangle, namely, the upper left and the lower right corners. In the following experiments, we have used tracking particles with rectangular shapes as this allows to compute efficiently their density by simply comparing the coordinates of the belonging feature points to the rectangle two corners. Next, the intermediate tracking swarm for the synthetic image obtained by applying the PSO-based scheme are shown in Fig. 7 and 8, corresponding to the PSO process application after 10 and 20 iterations, respectively.

The detection result of the major point cloud in the synthetic image using the PSO-based particles is shown in Fig. 9. As expected, we observe that the resulting global particle detects always the major point cloud regardless of the presence of discontinuities due to the large variations of gradient and the existence of neighboring sub-major point clouds of false positives belonging to hypothetically surrounding anatomical structures in the case of a real laparoscopic cholecystectomy image. However, we can observe at this first generation there are some false positives and negatives. this can be heavily

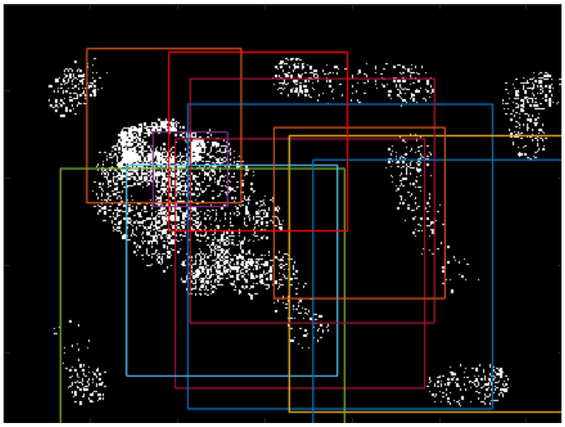


Fig. 7. Intermediate PSO-based tracking swarm of rectangular particles: after 10 iterations.

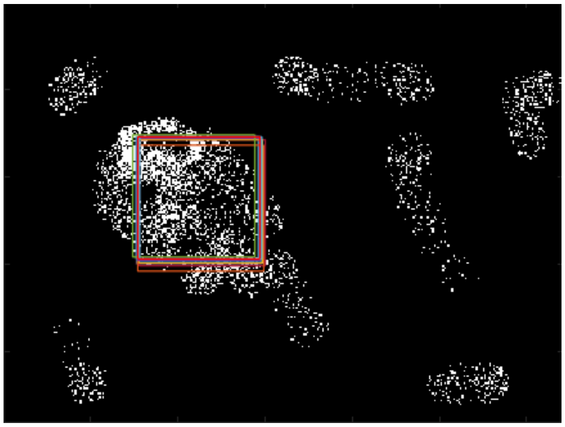


Fig. 8. Intermediate PSO-based tracking swarm of rectangular particles: after 20 iterations.

enhanced by applying a second and third generation detection passes along the boundaries of the first root detection particle.

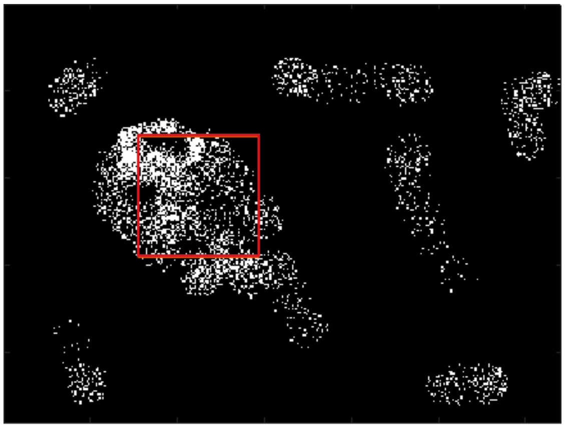


Fig. 9. The detection result of the synthetic image of a deformable structure points' cloud.

The different terms of the PSO-based evolution scheme must be weighted properly to guide the evolving tracking swarm under different image conditions such as those encountered during a laparoscopic cholecystectomy intervention. In the previous experiment on the synthetic image of the points cloud, we have considered the following parameters specific to the PSO-based detection scheme, namely, the size of the tracking swarm expressed by the number of the particles in the population (N=10); The number of the corners of the particle's shape (C=2) as it is a rectangular; the inertial parameters preventing the particles swarm from early collapsing ($W_{min} = 0.3$) and divergence ($W_{max} = 0.9$); the number of iterations of the PSO-based evolution scheme ($It_{max} = 50$); local (c1 = 0.4) and global (c2 = 0.4) PSO parameters which govern the influence of the individual and social terms on the evolution scheme, respectively. In addition, we consider the parameter (α_r) that gives the ratio between the surface of the visible corporal surface of the tacked deformable structure and the size of the image in terms of pixels. Thus, the number of pixels that constraints the size of the particles of the population is given by ($\alpha_p = \alpha_r * \|I\| = 2273$) where $\|I\| = I_L * I_W$ is the resolution of the image in terms of pixels and I_L, I_W are the length and the width of the laparoscopic image, respectively. The graphs represent the position, size and density of the global best particle of the population to segment and detect the deformable structure point cloud along time (PSO evolution iterations, $It_{max} = 50$). As it can be seen (Fig. 10), the tracking swarm stabilizes after only 20 iterations.

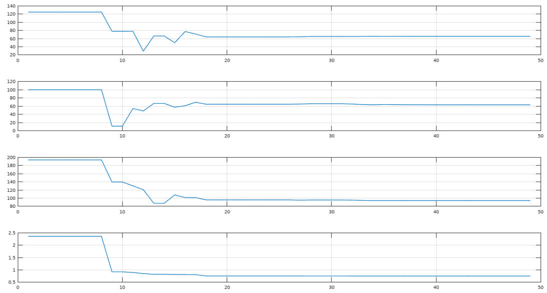


Fig. 10. Evolution of the best global particle of the population over time.

VI. Conclusion

In this paper, we have proposed a new automatic segmentation and tracking method of deformable structures in a minimally invasive surgery intervention such as laparoscopic cholecystectomy. The segmentation of anatomical structures is performed thanks to a modified PSO scheme to segment and track the deformable structure during the intervention, namely, the gallbladder in the case of laparoscopic cholecystectomy. The reconstructed 3D model is analyzed using a wavelet based method to perform the registration task. Therefore the system is able to track surgical instruments with possible interactive update of the color model guided by a priori anatomical knowledge. The only drawback of the proposed system is the need of the determination of a precise distance between the endoscope tip and the closest point of the corporal surface of the tracked anatomical structure. We are working on the development of such device to estimate precisely this distance. We intend to verify the effectiveness of the proposed method first on 3d printed deformable phantoms corresponding to real patients before testing the performance of the system intra-operatively.

VII. Conflict of Interest Disclosure

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

References

[1] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre, "Recent advances in augmented reality," *Computer Graphics and Applications, IEEE*, vol. 21, no. 6, pp. 34–47, 2001.

[2] R. Azuma *et al.*, "A survey of augmented reality," *Presence-Teleoperators and Virtual Environments*, vol. 6, no. 4, pp. 355–385, 1997.

[3] W. Reynolds, "The first laparoscopic cholecystectomy," *JSLS, Journal of the Society of Laparoendoscopic Surgeons*, vol. 5, no. 1, pp. 89–94, 2001.

[4] L. W. Traverso, "Carl langenbuch and the first cholecystectomy," *The American Journal of Surgery*, vol. 132, no. 1, pp. 81–82, 1976.

[5] L. Soler, S. Nicolau, J.-B. Fasquel, V. Agnus, A. Charnoz, A. Hostettler, J. Moreau, C. Forest, D. Mutter, and J. Marescaux, "Virtual reality and augmented reality applied to laparoscopic and notes procedures," in *ISBI*, 2008.

[6] K. G. Vosburgh and R. S. J. Estepar, "Natural orifice transluminal endoscopic surgery (notes): an opportunity for augmented reality guidance," *Studies in health technology and informatics*, vol. 125, p. 485, 2006.

[7] S. Connor and O. Garden, "Bile duct injury in the era of laparoscopic cholecystectomy," *British journal of surgery*, vol. 93, no. 2, pp. 158–168, 2006.

[8] J. Cullen, "Laparoscopic cholecystectomy: Avoiding complications," in *The SAGES Manual*. Springer, 2006, pp. 140–144.

[9] S. Cawich, D. Mitchell, M. Newnham, and M. Arthurs, "A comparison of open and laparoscopic cholecystectomy done by a surgeon in training," *West Indian Medical Journal*, vol. 55, no. 2, pp. 103–109, 2006.

[10] L. KOZUMPLÍK, "New classification of major bile duct injuries associated with laparoscopic cholecystectomy," *Scripta Medica (Brno)*, vol. 75, no. 6, pp. 283–290, 2002.

[11] G. S. Litynski, "Erich muhe and the rejection of laparoscopic cholecystectomy (1985): a surgeon ahead of his time," *JSLS, Journal of the Society of Laparoendoscopic Surgeons*, vol. 2, no. 4, pp. 341–346, 1998.

[12] R. Haluck, "Laparoscopic surgical instrument and method," 2003.

[13] T. Blum, H. Feußner, and N. Navab, "Modeling and segmentation of surgical workflow from laparoscopic video," *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, pp. 400–407, 2010.

[14] A. Jacquot, P. Sturm, and O. Ruch, "Adaptive tracking of non-rigid objects based on color histograms and automatic parameter selection," in *Application of Computer Vision, 2005. WACV/MOTIONS'05 Volume 1. Seventh IEEE Workshops on*, vol. 2. IEEE, 2005, pp. 103–109.

[15] S. Nicolau, X. Pennec, L. Soler, X. Buy, A. Gangi, N. Ayache, and J. Marescaux, "An augmented reality system for liver thermal ablation: Design and evaluation on clinical cases," *Medical Image Analysis*, vol. 13, no. 3, pp. 494–506, 2009.

[16] H. Kato, "Artoolkit: Library for vision-based augmented reality," *IEICE, PRMU*, pp. 79–86, 2002.

[17] D. Wagner and D. Schmalstieg, "Artoolkitplus for pose tracking on mobile devices," in *Proceedings of 12th Computer Vision Winter Workshop (CVWW'07)*, 2007, pp. 139–146.

[18] S. Nicolau, X. Pennec, L. Soler, and N. Ayache, "An accuracy certified augmented reality system for therapy guidance," *Computer Vision-ECCV 2004*, pp. 79–91, 2004.

[19] ——, "Evaluation of a new 3d/2d registration criterion for liver radio-frequencies guided by augmented reality," *Surgery Simulation and Soft Tissue Modeling*, pp. 1001–1001, 2003.

[20] D. E. Litwin and M. A. Cahan, "Laparoscopic cholecystectomy," *Surgical Clinics of North America*, vol. 88, no. 6, pp. 1295–1313, 2008.

[21] M. Feuerstein, "Augmented reality in laparoscopic surgery : New concepts for intraoperative multimodal imaging," Ph.D. dissertation, Technische Universitt Mnchen, 2007.

[22] H. Djaghloul, M. Batouche, and J.-P. Jessel, "Automatic pso-based deformable structures markerless tracking in laparoscopic cholecystectomy," in *Hybrid Artificial Intelligence Systems*. Springer, 2010, pp. 48–55.

[23] J. Kennedy, R. Eberhart *et al.*, "Particle swarm optimization," in *Proceedings of IEEE international conference on neural networks*, vol. 4, no. 2. Perth, Australia, 1995, pp. 1942–1948.

Fuzzy Data Mining for Autism Classification of Children

Mofleh Al-diabat
Department of Computer Science,
Al Albayt University,
Jordan

***Abstract*—Autism is a development condition linked with healthcare costs, therefore, early screening of autism symptoms can cut down on these costs. The autism screening process involves presenting a series of questions for parents, caregivers, and family members to answer on behalf of the child to determine the potential of autistic traits. Often existing autism screening tools, such as the Autism Quotient (AQ), involve many questions, in addition to careful design of the questions, which makes the autism screening process lengthy. One potential solution to improve the efficiency and accuracy of screening is the adaptation of fuzzy rule in data mining. Fuzzy rules can be extracted automatically from past controls and cases to form a screening classification system. This system can then be utilized to forecast whether individuals have any autistic traits instead of relying on the conventional domain expert rules. This paper evaluates fuzzy rule-based data mining for forecasting autistic symptoms of children to address the aforementioned problem. Empirical results demonstrate high performance of the fuzzy data mining model in regard to predictive accuracy and sensitivity rates and surprisingly lower than expected specificity rates when compared with other rule-based data mining models.**

Keywords—Autistic traits; data mining; fuzzy rules; statistical analysis

I. Introduction

Autism is a type of developmental condition initially listed under the umbrella of Diagnostic and Statistical Manual 4th edition text revised version (DMS-IV-TR) [1] as a type of Pervasive Developmental Disorder [PDD] [2]. Autism spectrum disorder [ASD] is defined as 'the challenges in social, communication, interaction, stereotyped movements, sensory and imagination skills, which significantly affect the behavioural performance of an individual'. According to the 2014 figures from the Disease Control and Prevention Centre [CDC], one child out of every 68 children is chronicled as a case of autism (1 per cent of the entire world population) [3]. By 2014, 3.5 million people in the USA had been diagnosed as cases of autism; the number of cases identified in the United Kingdom has risen 119.4 per cent from 2008 to 2014.

ASD screening is the process by which the autistic symptoms of an individual can be determined [4]. This is a crucial phase of ASD diagnosis as autism can't be identified by conventional clinical methods such as blood tests or body check-ups. There are various types of autism screening tools that involve direct observation, structured and semi-structured questionnaires and interviews [5]. Due to a lack of reliable measures in screening children for autism, in many situations the symptoms become visible only after they become adults. Therefore, the role of a viable screening instrument for identifying the risk of ASD at the preliminary stage is huge.

Existing ASD screening techniques rely on a simple domain expert, as well as a large number of questions that respondents have to answer, so these techniques have been criticized by scholars for being lengthy and subjective [5]-[9]. Therefore, developing detection systems that can be extracted using automated methods could be a promising direction. This approach of learning is called data mining and typically utilizes an historical dataset to discover effective hidden patterns for improving planning and the decision process [10], [11]. Recent initial studies in autism research, particularly ASD diagnosis, for example, [12]-[17] and others, indicated that data mining and machine learning techniques could enhance accuracy and efficiency of the diagnostic phase. However, there has been little headway in investigating data mining techniques within autism screening due to the unavailability of datasets. With the advancement of mobile technology, a recent dataset related to behavioural characteristics of autism has been proposed by [18].

This paper investigates fuzzy data mining models to detect autistic symptoms for cases and controls of children between the ages of 4-11 years. The proposed model learns If-Then rules based on different independent variables related to behaviour, i.e. AQ-10-Child [4], and other demographic features such as age, gender, and ethnicity. The dataset used in this research project consists of over 24 variables that have already been screened using a mobile application called ASDTests which was developed in 2017 [19]. A fuzzy rule based on data mining has been learnt using a Fuzzy Unordered Rule Induction algorithm (FURIA) [20]. The rules derived have been adopted to successfully distinguish individuals with ASD. In addition, these rules can be utilized to replace existing domain expert rules and possibly assist clinicians in referring individuals with ASD symptoms for further evaluation; additionally parents can now understand the relationship between autistic traits.

This paper is structured such that Section 2 discusses recent research developments related to the use of data mining in autism research, Section 3 presents data, features, the experimental setting, and results analysis. Finally, a conclusion is given in Section 4.

II. Literature Review

Investigated claims proposed by two other research studies regarding shortening the autism diagnosis related to the utilization of machine learning techniques in discriminating autism in the clinical context [21], i.e. [9], [22]. The researchers used 1949 instances [9], [22], that were obtained from the Autism Genetic Resource Exchange [AGRE] and Balance Independent [BID] datasets [23], [24]. Prior to experimentation, the dataset had been modified as [9], [22] eliminated instances that were not clear ASD cases. Then, the same machine learning techniques (tree-based algorithms) were used to classify individuals. The results of the [21] study revealed severe methodological and conceptual problems and, more importantly, no significant time reduction was found as claimed by the previous studies.

The authors in [17] explored the use of Twitter messages to feed into a data mining tool in order to obtain useful knowledge related to challenges, concerns and practices of autism, therefore raising awareness among people in the community. The ASD-related tweets and messages were collected by typing various keywords using the Twitter search engine to obtain the necessary data. The data was then analyzed in terms of Zipf's law criteria including message length, content, word frequency, hash tag frequencies, and parts of speech frequencies [6]. A further analysis was conducted to test whether the ASD tweets and non-ASD tweets could be automatically classified. The findings of the study concluded a number of common and differential characteristics related to ASD and non-ASD categories could be used to develop an automated mechanism to monitor the behaviours of the ASD community on social media.

The authors in [25] studied how data mining techniques can be used to enhance the impact of behavioural therapy on autistic individuals. Data was collected through videotaped sessions of approximately nine hours each from eight different autistic children who were receiving treatment therapy. During each session, the therapists recorded the four appropriate and inappropriate child behavioural types: their own playroom behaviour, behaviour with parents, behaviour with therapists, and behaviour with strangers. The findings of the research, based on data obtained through data mining techniques, indicated that behavioural therapy can increase appropriate behaviours and reduce any inappropriate behaviour of the autistic children. The rules discovered confirmed that the likelihood and frequency of appropriate and inappropriate behaviours can be predicted more accurately with more data.

The authors in [3] investigated nuroimaging patterns of autistic individuals to establish an effective mechanism to discriminate autism without the involvement of a long adminstration process that requires exclusive training and expertise. Functional Magnetic Resonance Imaging (fMRI) [26], [27] is used to capture the brain images of the subject when he is resting or idle. A total of 1035 fMRI instances were obtained from Autism Brain Imaging Data Exchange (ABIDE) [28] and then analysed to discover the pattern that could help to diagnosis autism. Deep learning techniques are used to classify and understand the unique features of neuro images of autistic individuals' brains and the functionality that can be used to diferentiate cases of autism from controls. The findings of the research suggested that autism can be differentiated 69 percent more accurately through neuroimaging patterns of the brain, than the conventional diagnosis methods, by using deep learning techniques like denoising autoencorders.

The authors in [29] examined the temporal variability of the functional connections (FC) using machine learning techniques and brain neuroimaging techniques for ASD classification. The node variability of the subject's brain is obtained to train different machine learning models on a large resting mind fMRI [26] data of ASD and non-ASD individuals obtained from ABIDE [28]. Machine learning classifiers such as Naive Bayes [30], Random Forest [31], Support Vector Machines [32] and the Multilayer Perceptron algorithm [33] were applied on 147 cases and 146 controls of autism obtained from ABIDE using Weka, open source marchine learning tool kit [34]. According to the results of the study, the machine learning models trained on different functional variabilitiy connections of the brain can achieve an accuracy of 62 percent in classifying and distingusing autism with a sensitivity of 60-65 per cent and specificity of 60+ percent.

The authors in [26] investigated whether machine learning can be an effective mechanism to diagnosis autism and Attention Deficit Hyperactive Disorder (ADHD). To ahieve the objective, the authors tested six different machine learning techniques on 2925 Social Responsive Scale [SRS] data obtained from Simons Simplex Collection version 15, Boston Autism Consortium and Autism Genetic Resource Exchange [35], [23]. The data relevant to 65 SRS items was stratified into 10 folders each comprising 10 percent of both ASD and ADHD data to perform cross validation. For each cross validation session, a minimal redundancy-maximal relevance [mRMR] feature selection method [36] was performed to rank all 65 items. The six machine learning algorithms including Support Vector Machines [32] linear discriminant analysis [21], Categorical lasso [16], tree-based algorithms (Decision Tree and Random Forest) [31], [35] and Logistics Regression Model [37] were tested on all the 65 rankings using the package Scikit-learn [38]. The results of the experiments showed that the majority of the machine learning techniques improve the accuracy of autism diagnosis. Particularly, a combination of Support Vector Machines, Logistics Regression, linear discriminant analysis and Categorical Lasso techniques produced the optimum level of performance in classifying autism and ADHD test instances.

The authors in [37] suggested a machine learning-based system to forecast ASD symptomology through the eye movement patterns of individuals. Initial experiments were carried out on two target groups of Chinese children. A total of 20 ASD children, 21 age-matched typically developing (TD) children, 20 IQ matched TD children (1st group), and 19 ASD, 22 IQ matched Intellectually Disabled (ID), and 28 age matched TD young adults and adolescents (2nd group). The eye movements and gazing patterns were captured through a Tobii T60 eye tracker. The images captured were analyzed using k-means [39] to identify the eye gaze coordinates on the spatial domains and to divide the face into different regions. ASD cases are anticipated to be distinguished based on the magnitude and directions of both the eye gaze coordinates and eye motions. A model similar to "bag of word" (BoW) is used to document the sequence of coordinates per image per person. The prediction models are developed using the Support Vector Machine (SVM) algorithm to avoid negative data and to identify linear decision boundaries. The subject level predictions with a global threshold are enabled as a scoring context to interpret functional boundaries and decision boundaries. The results of the experiments presented a greater potential and effectiveness in the proposed system for identifying symptoms of ASD.

The author in [38] evaluated the machine learning techniques used in prevailing ASD screening and diagnosis tools to identify their pitfalls to provide recommendations and guidance for future developments. Most of the previous research works on the similar topic of interest have addressed the quality, accuracy, technology usage and many other areas related to the computerised ASD diagnosis, but no study has yet addressed the different conceptual, implementation, and other data issues associated with various ASD tools. Most importantly many of the ASD tools have not integrated machine learning techniques into their screening and diagnosis process. Therefore [5], highlighted the machine learning techniques used in large prevailing ASD diagnostic instruments along with their conceptual issues and data and features issues like data imbalances, and provides a series of promising recommendations for future developers to overcome those issues.

III. Empirical Analysis

A. Data and Features

Controls and cases related to children (aged 4-11 years) have been collected using an ASD screening mobile application called ASDTests [40]. ASDTests was developed in 2017 to expedite ASD screening for different target groups including toddlers, children, adults and adolescents. In this paper, the focus is on instances related to the children category which have been collected based on the AQ-10-child ASD screening tool [4] using the ASDTests mobile application. Therefore, individual experiments were conducted on the children dataset only, which consists of 509 instances and 24 variables. The dataset has been obtained from its prospective author and covers the period between September 2017 and February 2018. Initially, the dataset was published in December 2017 with 292 instances at UCI data repository [41], but we were able to obtain from the dataset's owner the updated dataset with 227 child instances. The dataset contains 252 instances not on the spectrum (No ASD traits) and 257 instances with ASD traits; thus the dataset is somewhat balanced in regard to the target class variable. Initially, there were 24 independent variables including the target class. Most data instances relate to male participants with a ratio of 71.31 per cent (363 out of 509 instances). Moreover, 125 instances in the dataset were born with jaundice and 438 instances have been collected from parents. Table I depicts the primary variables that we have utilized prior to the data processing step. A number of variables have been discarded and not included in the table including: Country_of_Residence, Case_ID, Language, Screening_Type, Used_App_Before, since they have no added value and do not influence the classification of control and cases.

Independent variables A1-A10 shown in Table I correspond to the questions in the classic AQ-10-child screening tool and have been embedded within the ASDTests app. For simplicity, the authors of the dataset assigned these variables either "0" or "1" based on the answer given during the screening test by the participant. In particular, for questions 1, 5, 7, 10, "1" is assigned to the feature when the participant answers "Definitely" or "Slightly Agree" whereas, "1" will be given for "Definitely" or "Slightly Disagree" for questions 2, 3, 4, 6, 8 and 9. The dependent variable, which represents whether individuals have ASD traits, is associated with two possible values (Yes or No). This variable was assigned values based on the score obtained by individuals in the ASDTests app and was generated by the AQ-10-child tool. For a score larger than 6, "Yes" was assigned to the target variable for the instance, otherwise "No" was assigned. The process of assigning the values to the target variable was automated using the ASDTests app.

B. Settings

In this section, we investigate the performance of the fuzzy data mining algorithm called FURIA in detecting ASD traits for children and compare the performance with respect to different evaluation measures. To generalize the performance of FURIA, different data mining algorithms have been contrasted to reveal the upsides and the downsides of FURIA. In particular, we used JRIP, RIDOR and PRISM algorithms [25], [29] due to the fact they generate rules in the form of If-Then, as does FURIA, for fair comparison. In addition, these are rule-based data mining algorithms that have proved their merits in different classification applications, i.e. [42]-[44].

PRISM is a Covering algorithm that was developed to discover easy interpretable rules for decision-making by using a simple and effective metric called Expected Accuracy (EA). JRIP is a more advanced algorithm than PRISM that develops an optimization method and uses two subsets of data (growing, pruning) during the learning phase in order to reduce the number of rules generated. JRIP usually generates fewer rules than PRISM due to the pruning method implemented on the pruning set of data. RIDOR is a rule induction algorithm that generates exception in the format of rules. Lastly, FURIA is an extension of JRIP (RIPPER algorithm) which generates a fuzzy unordered set instead of classic ordered rules sets as JRIP. FURIA employs growing and pruning sets as JRIP in the process of rule learning and extraction. It learns rules sets per target class in a conventional strategy and then applies a stretch procedure to evaluate the rules sets derived. The outcome of FURIA is chunks of knowledge that can be used for decision-making especially in applications such as medical diagnosis. This is the primary reason for adopting FURIA to construct ASD classification models in order to detect ASD traits during the process of screening.

TABLE. I. FEATURES IN THE DATASETS

Variable No	Variable	Description
1.	A1	First question in AQ-10-Child screening tool
2.	A2	Second question in AQ-10-Child screening tool
3.	A3	Third question in AQ-10-Child screening tool
4.	A4	Fourth question in AQ-10-Child screening tool
5.	A5	Fifth question in AQ-10-Child screening tool
6.	A6	Sixth question in AQ-10-Child screening tool
7.	A7	Seventh question in AQ-10-Child screening tool
8.	A8	Eighth question in AQ-10-Child screening tool
9.	A9	Ninth question in AQ-10-Child screening tool
10.	A10	Tenth question in AQ-10-Child screening tool
11.	Age	Age of individual in numeric (years)
12.	Gender	Male or Female
13.	Ethnicity	Chosen from a list of predefined values
14.	Jaundice	Yes or No
15.	Family History	Whether any family members diagnosed with autism
16.	User	Who has taken the test (parent, self, relative, caregiver, etc)
17.	Target Class	The dependent variable (Yes/No). This variable was assigned based on the score obtained by individuals in the ASDTests app. If score larger than 6 “Yes” was assigned otherwise “No was assigned”.

All experiments of the data mining algorithms and FURIA have been conducted on WEKA, a machine learning platform that contains useful data mining, pre-processing and learning techniques [32]. In addition, a ten-fold cross validation procedure was adopted to conduct the data processing experiments. Lastly, all experimental runs have been conducted on a personal computing machine with 2.3 GHz processor and 8 RAM of memory.

C. Results and Discussions

Different evaluation methods, such as predictive accuracy, specificity and sensitivity among others, have been utilized to report the learning algorithms performance in classifying ASD test instances from the child dataset. Predictive accuracy is a common performance measure in classification that reveals the percentage of test data that was correctly detected from the total number of test instances. On the other hand, sensitivity represents the percentage of the test instances that is truly positive, and specificity represents the test instances that are truly negative. The accuracy of FURIA and the considered data mining algorithms on the child dataset are shown in Fig. 1. The figure pinpoints that classification models generated by FURIA are more accurate in detecting ASD traits than the remaining algorithm. In particular, the classification model of FURIA outperformed models produced by JRIP, PRISM and RIDOR by 3.14%, 7.66% and 0.98% on the child autism dataset. A principal reason for the superiority of FURIA is the rules fuzzification process and the stretching procedure that takes into account the order of the rule’s antecedent during the process of rule evaluation. This increases the rule’s purity and possibly data coverage making FURIA favours a more general rule than those that are specific. The sensitivity rate obtained by the considered data mining algorithms on the child dataset is shown in Fig. 2. The sensitivity rates derived are consistent with the predictive accuracy results in which FURIA outperformed the considered data mining algorithms. The sensitivity rate of FURIA is higher by 3.2%, 1.0% and 3.0% than JRIP, RIDOR and PRISM algorithms respectively. To evaluate the behaviour of FURIA we looked at the confusion matrix results obtained by its classification model. The confusion matrix results showed that only 14 instances with ASD traits have been incorrectly classified by FURIA as being without ASD traits, which is indeed a low number when compared with the remaining algorithms. To be specific, 42, 27, and 44 instances which are with ASD traits were misclassified by JRIP, RIDOR and PRISM algorithms. These numbers explain the higher predictive rate obtained by FURIA.

We investigated the false positives rates by deriving the specificity figures. Specificity (true negative rates) shows the percentage of participants who are without ASD and have been identified without ASD by the learning algorithm. Fig. 3 displays the specificity rates derived by the considered algorithms on the child dataset. Surprisingly, FURIA achieved lower specificity rates when compared with the remaining algorithms. We then investigated the false positive rates since they contribute largely in computing the specificity rate. From 252 instances, 33 which are actually without ASD have been misclassified by FURIA as being with ASD. In other words, there were 33 false positive instances generated by FURIA, compared with 18, 22 and 12 false positive instances generated by JRIP, RIDOR and PRISM algorithms respectively. These figures show that the specificity rate of PRISM is the highest, and the specificity rate of FURIA is the lowest, which is surprising. One possible reason for the higher false positive rates by FURIA and JRIP is the inability of this algorithm to differentiate among instances with limited ASD traits. These are instances that may show some autistic traits yet they are not classified to be on the spectrum by the screening tool. This shows a clear shortcoming of rule induction and fuzzy data mining algorithms, at least on the child data set considered in this paper.

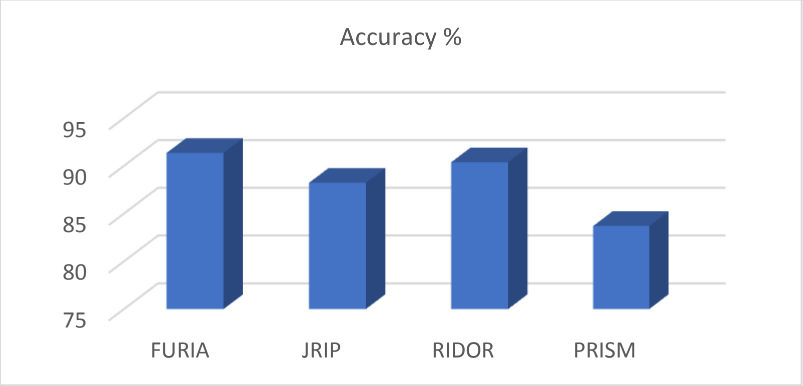


Fig. 1. Predictive accuracy derived by FURIA and the other Considered Data Mining Algorithms.

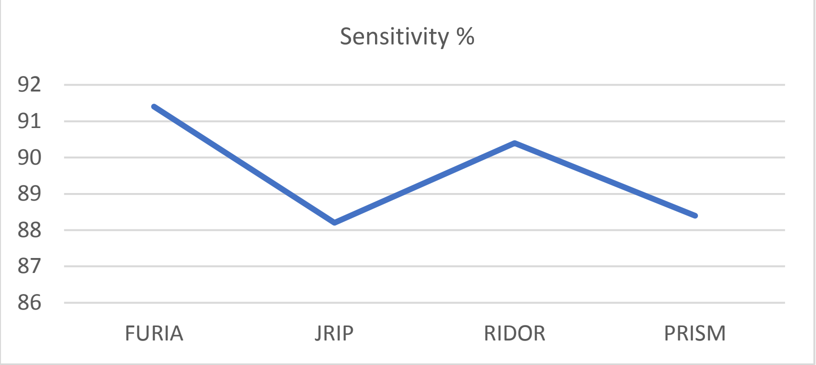


Fig. 2. Sensitivity rate derived by FURIA and the other Considered Data Mining Algorithms.

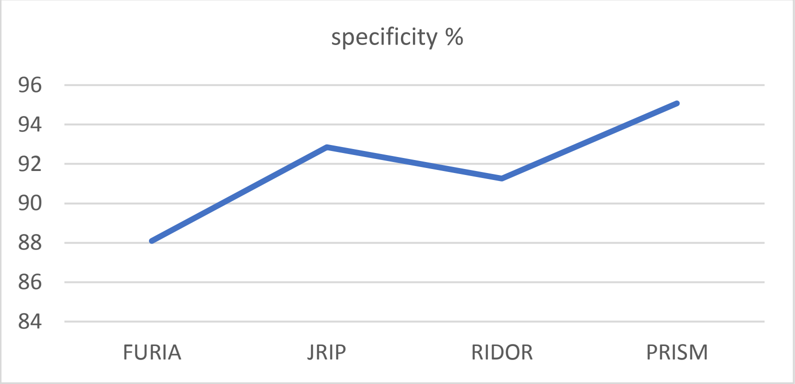


Fig. 3. Specificity rate derived by FURIA and the other Considered Data Mining Algorithms.

The fuzzy sets produced by FURIA are shown below: 29 fuzzy rules were derived by FURIA from the child autism dataset in which 11 rules are connected with target class “yes” and the remaining rules with class “no”. Based on the rules generated, the features related to AQ-10-child screening methods proved to be influential in detecting autistic traits particularly features such as A4, A7 and A9 appearing largely in the fuzzy rules sets. Specifically, features named A4, A7, A9, A2, A1, A10, A5, A3, A6 and A8 have appeared in the fuzzy rules sets 14, 11, 10, 10, 10, 12, 9, 9, 9, 9, respectively. This indicates that these features have high impact on detecting ASD traits and more important than demographic features in the child autism dataset.

Overall, FURIA produced useful chunks of knowledge that can be exploited by clinicians, parents, caregivers, and teachers among others, in understanding autism traits of children for better screening. When FURIA is integrated within screening tools of autism we expect that the automated fuzzy rules to be highly influential in detecting cases of autism for further referral and possibly to replace existing static domain expert rules.

```
FURIA rules:

===========

(A4 = 0) and (A8 = 0) and (A9 = 0) => Class=NO (CF = 0.99)
(A5 = 0) and (A10 = 0) and (A7 = 0) => Class=NO (CF = 0.98)
(A4 = 0) and (A1 = 0) and (A5 = 0) => Class=NO (CF = 0.98)
(A4 = 0) and (A10 = 0) and (Age in [-inf, -inf, 6, 7]) => Class=NO (CF = 0.98)
(A1 = 0) and (A2 = 0) and (A9 = 0) => Class=NO (CF = 0.98)
(A6 = 0) and (A5 = 0) and (A9 = 0) => Class=NO (CF = 0.98)
(A7 = 0) and (A3 = 0) and (A2 = 0) => Class=NO (CF = 0.96)
(A4 = 0) and (A7 = 0) and (A2 = 0) and (Family_ASD = no) => Class=NO (CF = 0.97)
(A10 = 0) and (A2 = 0) and (A1 = 0) => Class=NO (CF = 0.97)
(A8 = 0) and (A5 = 0) and (A3 = 0) => Class=NO (CF = 0.98)
(A6 = 0) and (A7 = 0) and (A10 = 0) => Class=NO (CF = 0.98)
(A4 = 0) and (A8 = 0) and (A1 = 0) => Class=NO (CF = 0.99)
(A9 = 0) and (A7 = 0) and (A8 = 0) and (A1 = 0) => Class=NO (CF = 0.98)
(A4 = 0) and (A6 = 0) and (A8 = 0) and (A2 = 0) => Class=NO (CF = 0.98)
(A9 = 0) and (A3 = 0) and (A4 = 0) => Class=NO (CF = 0.99)
(A9 = 0) and (A7 = 0) and (A5 = 0) and (Jaundice  = no) => Class=NO (CF = 0.98)
(A10 = 0) and (A2 = 0) and (Family_ASD = yes) => Class=NO (CF = 0.94)
(A6 = 0) and (A3 = 0) => Class=NO (CF = 0.99)
(A4 = 1) and (A5 = 1) and (A9 = 1) and (A10 = 1) => Class=YES (CF = 0.99)
(A8 = 1) and (A1 = 1) and (A3 = 1) and (A5 = 1) => Class=YES (CF = 0.96)
(A4 = 1) and (A7 = 1) and (A3 = 1) and (A6 = 1) and (A2 = 1) => Class=YES (CF = 0.99)
(A8 = 1) and (A10 = 1) and (A1 = 1) and (A6 = 1) and (A4 = 1) and (A3 = 1) => Class=YES (CF = 0.99)
(A7 = 1) and (A9 = 1) and (A1 = 1) and (A6 = 1) and (A10 = 1) => Class=YES (CF = 0.99)
(A4 = 1) and (A10 = 1) and (A7 = 1) and (A3 = 1) => Class=YES (CF = 0.99)
(A9 = 1) and (A10 = 1) and (Age in [-inf, -inf, 5, 6]) and (A8 = 1) and (A7 = 1) => Class=YES (CF = 0.98)
(A4 = 1) and (A10 = 1) and (A2 = 1) and (Ethnicity = asian) and (Age in [-inf, -inf, 7, 10]) => Class=YES (CF = 0.94)
(A9 = 1) and (A2 = 1) and (A1 = 1) and (A3 = 1) => Class=YES (CF = 0.97)
(A4 = 1) and (A10 = 1) and (A2 = 1) and (A5 = 1) and (A1 = 1) and (A6 = 1) => Class=YES (CF = 0.99)
(A8 = 1) and (Jaundice  = yes) and (A5 = 1) and (A7 = 1) and (A6 = 1) and (A10 = 1) => Class=YES (CF = 0.97)
```

IV. Conclusion and Future Works

Autism Spectrum Disorder (ASD) is one of the growing neurodevelopment conditions worldwide with many individuals undetected, making early screening crucial for individuals, family members and physicians. Most of the existing ASD methods consist of a large set of questions covering communication, social and repetitive behaviours and rely on domain expert rules with a basic scoring function to detect autistic traits. One promising approach that can automate the process of ASD screening and improve the accuracy and efficiency of the detection is the use of fuzzy data mining. In this paper, the Fuzzy Unordered Rule Induction algorithm (FURIA) has been evaluated for ASD traits detection. FURIA builds screening models in an automated way from historical controls and cases and then utilizes the models to detect the possibility of autistic traits in new individuals. The key strength of FURIA screening models is the fact that they contain useful chunks of knowledge (fuzzy rules) that not only clinicians and other medical staff can interpret but also family members, teachers and caregivers. These fuzzy rules are a source of information that can help different stakeholders understand the main influential factors for ASD and therefore proper individualized plans can be planned and developed to cater to the needs of people who fall within the spectrum. Empirical results based on real data collected recently from children between 4-11 years old using a mobile application called ASDTests, revealed that FURIA fuzzy rules were able to detect ASD traits with up to 91.35% classification accuracy and 91.40% sensitivity rate; these results were superior to other Greedy and Rule Induction techniques. Despite FURIA producing an acceptable specificity rate, i.e. 88.09%, other data mining techniques generated better specificity results.

One of the limitations of this study is not extensively considering feature assessment on the dataset and not considering other target datasets such as infants, adolescent and adults.

In near future, we are going to apply the fuzzy rules on datasets related to infants and adolescents and seek whether the performance will be sustained. In addition, we will investigate features that are similar among different age categories.

References

[1] American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders DSM-IV-TR (Text Revision). Washington: American Psychiatric Association.

[2] Biklen, D. (2005). *Autism and the myth of the person alone*. New York and London: New York University Press.

[3] Christensen, D., Baio, J., Braun, K., Bilder, D., Charles, J., Constantino, J., & Yeargin-Allsopp, M. (2014). *Prevalence and characteristics of autism spectrum disorder among children aged 8 years* — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012. Surveillance Summaries, 1–23.

[4] Allison, C., Auyeung, B., and Baron-Cohen, S. (2012). Toward brief "Red Flags" for autism screening: The short Autism Spectrum Quotient and the short quantitative checklist for autism in toddlers in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child Adolescent Psychiatry 51*(2), 202–12Y.

[5] Thabtah, F. (2018) Machine learning in autistic spectrum disorder behavioural research: A review and ways forward. *Informatics for Health and Social Care 43* (2), 1-20.

[6] Duda, M., Ma, R., Haber, N., & Wall, D. (2016). Use of machine learning for behavioural distinction of autism. *Translational Psychiatry*, 221.

[7] Liu, W., Yu, Z., Raj, B., Yi, L., Zou, X., & Li, M. (2015). Efficient Autism Spectrum Disorder Prediction with Eye Movement: A Machine Learning Framework. 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), 649-655.

[8] Kosmicki, J., Sochat, V., Duda, M., & Wall, D. (2015). Searching for a minimal set of behaviours for autism detection through feature selection-based machine learning. *Translational Psychiatry*, 1-7.

[9] Wall, D., Kosmicki, , J., DeLuca, T., Harstad, E., & Fusaro, V. (2012a). Use of machine learning to shorten observation-based screening and diagnosis of autism. *Traditional Psychiatry, 2*(4):e100.

[10] Mohammad, R., Thabtah, F., McCluskey, L. (2016) *An improved self-structuring neural network.* Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, 35-47. Auckland, New Zealand.

[11] Aburrous, M., Hossain, M. A., Dahal, K. & Thabtah, F. (2010) *Predicting phishing websites using classification mining techniques with experimental case studies.* Proceedings of the Information Technology: New Generations (ITNG). Las Vegas, Nevada, USA.

[12] Tejwani, R., Liska, A., You, H., Reinen, J., & Das, P. (2017). *Autism classification using brain functional connectivity dynamics and machine learning*. California, Santa Barbara: Cornwall University Library.

[13] Thabtah, F. (2017a). *Autism spectrum disorder screening: Machine learning adaptation and DSM-fulfilment.* Proceedings of the 1st International Conference on Medical and Health Informatics 2017, pp.1-6. Taichung City, Taiwan, ACM.

[14] Heinsfeld, A. S. (2017). *Identification of autism disorder through functional MRI and deep learning.* Rio Grande Do Sul: Pontifical Catholic University of Rio Grande Do Sul.

[15] Leroy, G., Irmscher, A., Charlop-Christy, M., Kuriakose, S., Pishori, A., Wurzman, L., . . . Stephanie, B. (2006). *Data mining techniques to study therapy success*. International Conference on Data Mining, 26 - 29 June 2006 (pp. 1-7). Las Vegas, USA: Institute of Electrical and Electronics Engineers (IEEE).

[16] Angelo, G., Rao, D., & Gu, C. C. (2009). Combining least absolute shrinkage and selection operator (LASSO) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies. *Genetic Analysis Workshop 16*, 25-63.

[17] Beykikhoshk, A., Arandjelovi´, O., Phung, D., Venkatesh, S., & Caelli, T. (2014). *Data-mining twitter and the autism spectrum disorder: A pilot study.* 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014) (pp. 349 - 356). IEEE Conferences.

[18] Thabtah, F. (2017c). *ASD dataset for children*. https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children++ [accessed January 15th, 2018].

[19] Biklen, D. (2005). *Autism and the myth of the person alone*. New York and London: New York University Press.

[20] Hhn J., Hllermeier E. (2009) FURIA: An algorithm for unordered fuzzy rule induction. *Data Mining Knowl. Discovery, 19,* (3), pp. 293-319, 2009.

[21] Bone, D., Goodwin, M., Black, M., Lee, C. C., Audhkhasi , K., & Narayanan, S. (2014). Applying machine learning to facilitate autism diagnostics: Pitfalls and promises. *Journal of Autism and Developmental Disorders, 1-48.*

[22] Wall, D., Dally, R., Luyster, R., Jund, J., & Deluca, T. (2012b). Use of artificial intelligence to shorten behavioural diagnosis of autism. Plos One, 7(8).

[23] Geschwind, P., Sowinski, J., Lord, C., Iversen, P., Shestack, J., Jones, P. Spence, S. (2001). The autism genetic resource exchange: A resource for the study of autism and related neuropsychiatric conditions. *American Journal of Human Genetics, 69*(2), 463.

[24] Gotham, K., Risi, S., Pickles, A., & Lord, C. (2007). The autism diagnosis observation schedule: Revised algorithm for improved diagnostic validity. *Journal of Autism and Developmental Disorders*, 613-627.

[25] Cohen W. W. (1995) *Fast effective rule induction*, Proceedings of the 12th International Conference on Machine Learning, pp. 115-123, 1995

[26] De la Iglesia-Vaya, M., Molina-Mateo, J., Jose, M., & Marti-Bonmati, L. (2013). Brain connections - resting state fMRI functional connectivity. *Novel Frontiers of Advanced Neuroimaging*, 236-286.

[27] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 3371–3408.

[28] Nielsen, J., Zielinski, B. A., Fletcher, P., Alexander, A. L, Lange, N, Bigler, E. D, . . . Anderson, J. S. (2013). Multisite functional connectivity MRI classification of autism: ABIDE results. *Frontiers in Human Neuroscience*, 1-12.

[29] Cendrowska J., PRISM: an algorithm for inducing modular rules. *Int. J. Man. Mach. Stud., 27* (1987) 349-370.

[30] Langseth, H., & Nielsen, T. (2006). Classification using hierarchical naïve Bayes models. *Machine Learning*, 135–159.

[31] Breiman, L. (2001). *Random forests.* Machine Learning, 5–32.

[32] Smola, A., & Scholkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing,* 199–222.

[33] Chaudhuri, B., & Bhattacharya, U. (2000). Efficient training and improved performance of multilayer perceptron in pattern classification. *Neurocomputing*, 11-27.

[34] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: An update. *SIGKDD Explorations 11*(1).

[35] Fischbach, G. D., Lord C. (2010) The Simons simplex collection: A resource for identification of autism genetic risk factors. *Neuron* (68), 192–195.

[36] Peng, H. C., Long, F., & Ding, C. (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(8, pp. 1226–1238, 2005.

[37] Barros, A., & Hirakata, V. (2003). Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. *BMC Medical Research Methodology,* 569-598.

[38] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., (2011). Scikit-learn: machine learning in python. *J Mach Learn Res 2011; 12*: 2825–2830

[39] MacQueen, J. B. (1967). *Some methods for classification and analysis of multivariate observations.* Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07.

[40] Thabtah, F. (2017b). *ASDTests. A mobile app for ASD screening*. www.asdtests.com [accessed March 10th, 2017].

[41] Olanow, C., & Koller, W. (1998). An algorithm (decision tree) for the management of Parkinson's disease. *Neurology,* 50-69. Barros

[42] Abdelhamid, N., Thabtah, F., & Abdel-jaber, H. (2017). *Phishing detection: A recent intelligent machine learning comparison based on models content and features.* 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), pp. 72-77. 2017/7/22, Beijing, China.

[43] Mohammad, R., Thabtah, F., McCluskey, L. (2014). Intelligent rule-based phishing websites classification. *IET Information Security, 8*(3): 153-160.

[44] Abdelhamid, N., and Thabtah, F. (2014). Associative classification approaches: Review and comparison. *Journal of Information and Knowledge Management (JIKM), 13*(3).

Training Difficulties in Deductive Methods of Verification and Synthesis of Program

Magdalina Todorova
Faculty of Mathematics and Informatics
Sofia University “St. Kl. Ohridski”
Sofia, Bulgaria

Daniela Orozova
Faculty of Computer Science and Engineering
Burgas Free University
Burgas, Bulgaria

***Abstract*—The article analyzes the difficulties which Bachelor Degree in Informatics and Computer Sciences students encounter in the process of being trained in applying deductive methods of verification and synthesis of procedural programs. Education in this field is an important step towards moving from classical software engineering to formal software engineering. The training in deductive methods is done in the introductory courses in programming in some Bulgarian universities. It includes: Floyd’s method for proving partial and total correctness of flowchart programs; Hoare’s method of verification of programs; and Djikstra’s method of transforming predicates for verification and synthesis of Algol−like programs. The difficulties which occurred during the defining of the specification of the program, which is subjected to verification or synthesis; choosing a loop invariant and loop termination function; finding the weakest precondition; proving the formulated verifying conditions, are discussed in the paper. Means of overcoming these difficulties is proposed. Conclusions are drawn in order to improve the training in the field. Special attention is dedicated to motivating the use of specific tools for software analysis, such as interactive theorem proving system HOL, the software analyzers Frama−C and its WP plug−in, as well as the formal language ACSL, which allows formal specification of properties of C/C++ programs.**

Keywords—Program verification; deductive verification methods; automated theorem provers; proof assistants; education

I. Introduction

Applying formal methods of program verification and synthesis is an important part of the training in the introductory courses of programming in some Bulgarian universities. The experience which is shared is gained as a result of delivering such training through courses in the disciplines Introduction to Programming, Object−Oriented Programming and Data Structures for Bachelor’s Degree students of specialties Informatics and Computer Sciences of Sofia University and Burgas Free University. Education is narrowed down to the following methods: Floyd’s method of inductive statements for verification of flowchart programs, Hoare’s method for verification of *while* programs, and Djikstra’s method for transforming predicates for verification and synthesis of Algol−like programs. The limitations in the choice of methods of program verification and synthesis are imposed by the consideration that the training under investigation is delivered during the first three semesters of the Bachelor’s Degree education, therefore the students lack sufficient knowledge in discrete mathematics and mathematical logic.

Some elements of the education are presented in the part, which is dedicated to the difficulties when applying deductive methods. Mere details on the training realization are provided in [1]-[4].

In [1], the following techniques are described, used in the education process in the field: axiomatic semantics, design by contract and generalized nets (GNs). Two main training approaches are considered. The first one combines the axiomatic semantics for proving total correctness of a procedural program with execution of the program [4]. The second one integrates axiomatic semantics in the GN models of the object−oriented programs under verification. The main stages of education and the process of education are considered. The results of the education are analyzed.

In [2], Floyd’s method and the method of transforming predicates for program verification are presented. The main stages and methodology of training in these methods of deductive verification are discussed.

In [3], Hoare’s method, the method of predicate transformer for synthesizing programs, runtime verification of programs and verification of object−oriented programs via developing their GN models are presented. The methodology of training in these methods of program verification and synthesis is considered. Examples of their application for the courses Introduction to Programming, Object−Oriented Programming and Data Structures are presented.

It is noted in the above cited articles that training in the field poses certain problems before the students; however, these problems has not been analyzed so far. Current article is dedicated to the analysis of the training difficulties; moreover, it presents means for overcoming these difficulties.

Bearing in mind the complexity of the field, the education is realized by: using simple examples; introducing the main terminology one at a time (e.g. precondition, postcondition, loop invariant, program specification); practicing the terminology not only during the lectures, but also during the seminars and lab sessions; applying deductive methods of verification is illustrated through automatized systems (such as HOL Interactive Theorem Prover [5] and the software analyzers Frama−C [6]); applying knowledge to realizing small projects.

Note that HOL and Frama−C systems are not introduced, due to the limited classes, only the results of their application are presented. The main goal is to demonstrate that knowledge about formal methods is useful and will support future software developers in designing, realizing and testing applications. Otherwise, part of the students may lose motivation for studying and applying formal methods in their practice.

The work is structured as follows. In Section II of the paper the difficulties, which students encounter when: defining the specification of the program under verification and synthesis; finding the weakest precondition; choosing an invariant and a termination function of the loop operator; proving the formulated verifying conditions, are analyzed. Dealing with difficulties during the training is presented in Section III. Some suggestions are given that assist: defining the specification; choosing an invariant and a termination function of the loop operator and proving the formulated verifying conditions. Recommendations for dealing with difficulties are described in Section IV.

II. Analysis of the Difficulties in Applying Deductive Methods of Program Verification and Synthesis

A. Defining the Specification

Program specification describes what has to be done as a result of program execution. At the beginning of the training students are introduced to Hoare's triple $\{Q\}\ S\ \{R\}$, where Q and R are predicates, and S is a program. This specification defines the total correctness of the program S with respect to Q and R. This can be interpreted as follows:

If executing S starts in a state which satisfies the predicate Q, what follows is that executing S terminates, after a limited time period, in a state which satisfies R.

The predicate Q is called *precondition*, and the predicate R – *postcondition.* Defining pre− and postconditions of some programs is a great challenge for some students. In most cases, these are programs solving tasks on each and existence. Proving the correctness of the specification $\{Q\}\ S\ \{R\}$ through applying Manna−Pnueli's rules [7], as well as proving partial correctness of S with respect to Q and R through applying Hoare's rules [8], is quite labor−intensive and demotivates even the best students to use the specification and the respective rules. That's why, for educational purposes, the more convenient specification, known as transforming predicate [9] is applied.

The transforming predicate $Wp(S, R)$ describes the set of all states, so that the start of the program execution from each of these states terminates, and the value of the output predicate R is *true*. $Wp(S, R)$ satisfies $\{Wp(S, R)\}\ S\ \{R\}$. Hoare's triple $\{Q\}\ S\ \{R\}$ is equivalent to $Q \Rightarrow Wp(S, R)$.

Finding $Wp(S, R)$, again, is far from a simple task, in most cases. For example, the definition of the transforming predicate of the loop operator is nearly unusable, but it helps to theoretically justify a methodology for verification and synthesis of code fragments containing loop operators [10].

That is why identification and check if the chosen specification holds is a complex activity. It requires of the students to have good knowledge of mathematical logic, as well as skills for defining and applying mathematical abstractions. This poses the main difficulty for training in the field: some of the students of specialty Informatics and Computer Sciences lack enough mathematical knowledge and skills. The complexity of the matter requires high motivation for applying these methods. However, first year students have little practical experience, they usually do not recognize the crucial importance of the formal methods for ensuring software quality. This demotivates them to apply any formal methods of software verification and synthesis. The lack of motivation is a great obstacle for the training in the field.

B. Choosing an Invariant and Termination Function for the Loop Operator

In order to verify a program formally, the following two tasks are to be solved:

- proving the partial correctness of the program with respect to given input/output specification;
- proving that the program terminates.

In order to solve these, the application of one of the above mentioned methods of deductive program verification and synthesis is discussed in the paper. The two tasks cannot be solved by any of these methods without finding and applying suitable invariants and termination functions of the loop operators of the program.

An invariant of a loop operator is a logical statement, which holds before the execution and after each execution of the loop operator.

The loop termination function is used to prove that the respective loop terminates. It gives the upper limit of the iterations to be completed by the end of the loop execution. The latter can be used to estimate the time left until the program ends.

In most cases, both loop invariant and loop termination function are not obvious. The task for correctly identifying them is of great importance for automatizing program verification. Solving this task is difficult, having in mind the volume and complexity of contemporary software, and the software realized for educational purposes, respectively. During the training process on finding the invariant of a loop operator, Gries's methodology [10] is applied. The loop invariant is seen as a weaker postcondition. The ways for finding a condition weaker than the postcondition, which are most commonly applied for educational purposes in finding the invariant of a loop are:

- deleting a conjunctive member;
- replacing a constant by a variable;
- enlarging the range of a variable.

Using only one of these three methods for generating a weaker postcondition sometimes does not lead to identifying a suitable invariant. In this case, a combination of these methods is performed, as well as a combination of the precondition and postcondition.

Finding an invariant these ways is a labor−intense work. In addition, some questions arise, such as: Which conjunctive member of the postcondition to be eliminated, so a suitable invariant to be found? In case of more constants, which one to be exchanged for a variable? What limits the variable area to be increased to, and whose variable area to be increased? Which methods to be combined? Which parts of the pre− and postconditions to be combined?

Finding the respective invariant is related to checking several logical conditions, which poses additional difficulties.

For example, in order for *P* to be an invariant of the *while* loop of the program for finding the factorial of the natural number *x* (the example traditionally used as a loop containing program fragment)

```
y = 1; z = 0;
while (z != x)
{ z++;
  y = y*z;
}
```

P must satisfy the following conditions:

$\{ P \wedge z \neq x \}$ z++; $y = y*z$ $\{ P \}$

$y = 1 \wedge z = 0 \;\Rightarrow P$

$P \wedge \neg (z \neq x) \;\Rightarrow y = x!$

where $y = x!$ is the postcondition.

In most cases, checking for meeting these conditions proves to be a complex task.

In order to find the loop termination function, an assumption can be applied that it gives the upper bound on how many iterations remain to be executed before loop termination occurs. The invariant of the loop operator suggests the definition of its termination function in almost all applications developed by the students. Therefore, the correct identification of an invariant is closely related to the choice of a loop termination function. Thus, the difficulties in finding the invariant also reflect on finding the loop termination function.

C. Proving the Formulated Verifying Conditions

Although proving the verifying conditions is narrowed down to proving the truthfulness of a system of implications, proving these implications manually, more often than not, is not an easy task. Even educational examples require the students to have basic knowledge in discrete mathematics and mathematical logic. They should also be able to work with propositions, in order to perform equivalence transformations (to know the laws of equivalence; the rules of substitution and transitivity). They have to have also at least basic knowledge on deductive proofs (inference rules; proofs and subproofs), and on predicates (extending the range of a state; quantification; free and bound identifiers; some theorems about textual substitution and states). In addition, they have to be aware of notations and conventions for arrays.

Supposedly, a great deal of this knowledge is to be acquired as a result of studying the following disciplines: Discrete Mathematics; Languages, Automata and Calculability; Introduction to Software Engineering; Algebra; Geometry and Mathematical Analysis, which are being taught in parallel with the courses in programming. However, a large number of the students have a different predisposition and are not motivated enough to study mathematics disciplines, as stated above.

Apart from the difficulties identified earlier, the following are of importance as well:

- Lack of environments for teaching in the field. There are not enough tools for automatizing the process of applying deductive methods, which to be applicable to training beginner programmers. Environments are needed to facilitate students in applying deductive methods of verification.
- There are no adequate didactic materials to support such education.
- The textbooks on the matter are not enough.

III. Dealing with Difficulties During the Training

Our experience in teaching formal methods at academic level, in addition to our observations on how graduates apply the knowledge in the field in their practice, made us believe that education in formal methods of verification and synthesis is useful and needed. It is only through training that these methods are applied in software industry. In order for this to be successful, measures for overcoming the difficulties described in Section II have to be taken.

Some suggestions for coping with these difficulties are proposed below.

A. Defining the Specification

In order to tackle the problems with defining the specification, the students need to get acquainted and are taught to apply some specification language. The experience with ANSI/ISO C Specification Language (ACSL) [11]-[13] provides ample evidence in favor of recommending it for applying formal methods of verification of C and some C++ programs. This language allows for relatively easily specifying properties of C and some C++ programs, after which these properties to be formally verified. ACSL is a Behavioral Interface Specification Language implemented in the Frama−C framework WP plug−in. Frama−C is an open code platform, analyzing source code written in the programming language C. It combines the following analysis techniques in a common framework: Frama−C's WP plug−in, Frama−C's value analysis plug−in, Frama−C's RTE plug−in and Frama−C's E−ACSL plug−in. Frama−C's WP plug−in is suitable for education purposes.

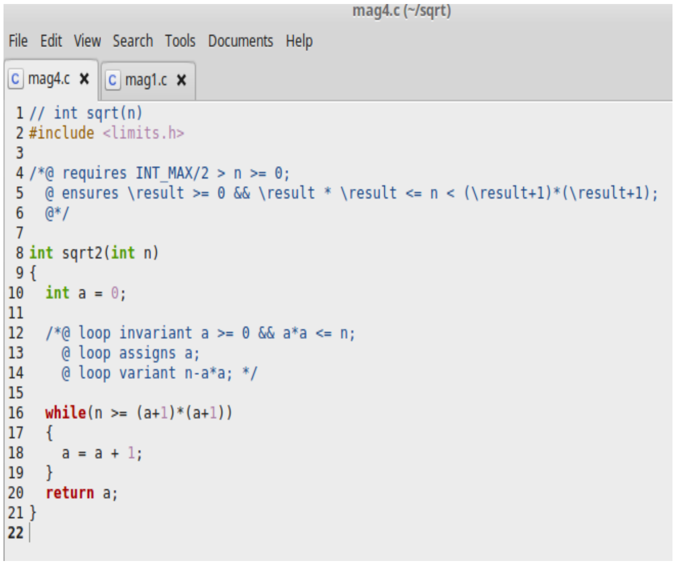
The Frama−C/WP plug−in enables deductive verification of C programs that have been annotated with ACSL. This plug−in uses Hoare−style weakest precondition computations to formally prove ACSL properties of a C code. Verification conditions are generated and submitted to external automatic theorem provers or interactive proof assistants [13].

By using the formal language ACSL, the software analyzers Frama−C and Frama−C's WP plug−in, defining of the program specification by the student will be narrowed down to: defining program precondition and postcondition, of the invariants and termination functions of the loop operators, used in the program. Thus, the learner will not have to find the transforming predicate, as well as to prove the formulated statements.

Fig. 1 shows the definition of the function *sqrt2*, which finds the biggest integer, whose square is not bigger that *n* (*n* is a given non−negative integer). The definition is annotated according to the specification, given in ACSL language.

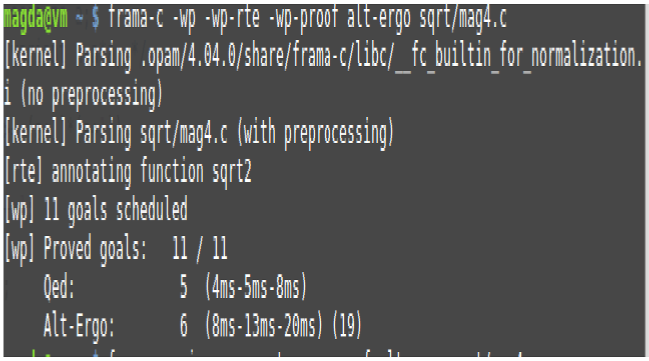
The precondition (*INT_MAX*/2 > *n* >= 0) is given via the requires clause, and the postcondition (*result* >= 0 && *result* * *result* <= *n* < (*result* + 1) * (*result* + 1)) – through the ensures clause. The invariant (*a* >= 0 && *a* * *a* <= *n*) and the loop termination function (*n* – *a* * *a*) are given by the clauses loop invariant and loop variant, respectively (see Fig. 1).

The result of realizing this specification through the WP plug−in of Frama−C (see Fig. 2) shows that 11 goals are achieved (5 goals are simplifications, done via the simplifier Qed that is integrated into Frama−C/WP, and 6 goals are proofs, done via the SMT solver Alt−Ergo).



```
// int sqrt(n)
#include <limits.h>

/*@ requires INT_MAX/2 > n >= 0;
  @ ensures \result >= 0 && \result * \result <= n < (\result+1)*(\result+1);
  @*/

int sqrt2(int n)
{
  int a = 0;

  /*@ loop invariant a >= 0 && a*a <= n;
    @ loop assigns a;
    @ loop variant n-a*a; */

  while(n >= (a+1)*(a+1))
  {
    a = a + 1;
  }
  return a;
}

```

Fig. 1. Function sqrt2, specified through ACSL.



```
magda@vm ~ $ frama-c -wp -wp-rte -wp-proof alt-ergo sqrt/mag4.c
[kernel] Parsing .opam/4.04.0/share/frama-c/libc/__fc_builtin_for_normalization.
i (no preprocessing)
[kernel] Parsing sqrt/mag4.c (with preprocessing)
[rte] annotating function sqrt2
[wp] 11 goals scheduled
[wp] Proved goals:   11 / 11
     Qed:             5  (4ms-5ms-8ms)
     Alt-Ergo:        6  (8ms-13ms-20ms) (19)
```

Fig. 2. Results of analysis of the function *sqrt2* via Frama−C.

B. *Choosing an Invariant and a Termination Function of the Loop Operator*

As the choice of an invariant of the loop operator is related to proving conditions, a means of facilitating the solving of this task is using automated theorem provers. Some of the most successful systems for theorem proving are: HOL Light, Mizar, ProofPower, Isabelle and Coq. An experience regarding formal verification of procedural and object−oriented programs using the theorem prover system HOL is shared in [14]. Through this theorem prover, it can be checked if the predicate *P*, chosen to be an invariant of the operator for the loop *while* (*B*) *S*, satisfies the conditions for an invariant:

$\{ P \wedge B \}\ S\ \{ P \}$

$Q \Rightarrow P$

$P \wedge \neg B \Rightarrow R$

Where *Q* is the precondition and *R* is the postcondition of the loop operator. It also can be used in proving if the loop termination function *t* satisfies the following conditions:

$P \wedge B \Rightarrow t > 0$ and

$P \wedge B \Rightarrow Wp(t1 = t;\ S,\ t < t1)$

Where *t1* is the value of *t* before executing the body *S* of the loop.

Another suitable module for finding a loop invariant is *Jessie*, complemented by *Apron* library. The module *Jessie* is included in Frama−C.

In order to find a loop termination function, the following strategy can be applied: wording down the functions of the loop termination function; formalizing the wording as a mathematical expression; checking if the mathematical expression satisfies the formal requirements for a loop termination function. It is recommended the checking to be performed both manually and via any of the automatic theorem provers.

C. *Proving the Formulated Verifying Conditions*

In order to prove the verifying conditions, the HOL Interactive Theorem Prover, as well as some of the theorem provers Qed, CVC4, Z3, Alt−Ergo, CVC3, E and Coq, which are supported by the Software Analyzers Frama−C can be applied.

IV. Recommendations for Dealing with Difficulties

The experience gained as a result of the training justifies the formulation of the following recommendations for dealing with difficulties in applying deductive methods of program verification and synthesis:

- Manual application with automated tools to be integrated. Thus, part of the problems will be avoided, and a better balance between simplicity, visualization and precision will be maintained.

- Programming environments, adequate to the educational goals, to be developed.

- Student motivation regarding studying the field to be increased. To this end, adequate samples to be designed, through which the advantages of deductive methods for ensuring the quality of industrial software and for decreasing the price of software project realization to be demonstrated.
- More efficient educational methods in the field to be introduced [15]. In addition to teaching through examples [16] and project−based learning approach [17], it is essential the training to be organized with respect to student knowledge level. For each level, appropriate methods to be chosen and applied.
- The teaching experience regarding the difficulties encountered during such education to be more widely discussed. The quality of education to be unified by employing cloud management [18].
- Cloud−based educational networks to be established in order to facilitate trans−institutional collaboration on creating and applying educational products and services in the field of programming, and formal methods of verification and synthesis, in particular [19].
- Appropriate formalization of programming language teaching to be made [20].
- Appropriate didactic materials on the matter to be designed to support the training.
- Studying the field to be given the status of a separate core discipline. Thus, each student will have to study and apply formal methods.

V. Conclusion

Education in the field of applying formal methods for developing correct software is the most efficient way of implementing these methods in software industry. The reason to state this is that a relatively large portion of the trained students continue using these methods in their further study at both Bachelor and Master degrees. This tendency is especially visible during courses such as Numerical Methods and Robotics [21]. Some of the graduate students who have completed this training also try to apply it in their practice as software specialists. Others continue their study at a PhD level in the field.

In order to overcome the challenges during the training in deductive methods of program verification and synthesis, an educational environment for verification of procedural and object−oriented programs is under development [22]. The environment is based on GNs and only provides tools for training in program verification so far. Future work includes expanding education framework with tools for supporting program synthesis, as well as integrating automatic systems for theorem proving in it.

References

[1] M. Todorova, “Applying program verification methods in software specialists education,” Proceedings of the 7th International Technology, Education and Development Conference, Valencia, Spain, pp. 6260−6270, 2013.

[2] M. Todorova and D. Orozova, “Applying deductive verification to bachelor degree courses in programming,” Proceedings of the the 10th Annual International Conference of Education, Research and Innovation, Seville, 16−18 November, 2017, pp. 5055−5065, 2017.

[3] M. Todorova and D. Orozova, “How to build up contemporary computer science specialists – formal methods of verification and synthsis of programs in introduction courses on programming,” Proceedings of the 9th annual International Conference of Education, Research and Innovation, Seville, 14th, 15−16 November, 2016, pp. 4249−4256, 2016.

[4] M. Todorova and P. Armyanov, “Runtime Verification of computer programs and its application in programming education,” Global Science and Technology Forum: International Journal of Mathematics, Statistics and Operations Research, vol. 1, No. 1, pp. 105−110, 2012.

[5] Realease Notes for HOL4, Kananaskis−11, https://hol−theorem−prover.org/kananaskis−11.release.html, 2012.

[6] Frama−C Software Analyzers, https://frama−c.com/index.html

[7] Z. Manna, Mathematical Theory of Informatics, Science and Art Publishing House, Sofia, 1983.

[8] C. A. R. Hoare, “An axiomatic basis for computer programming,” Communication of the ACM, vol. 12, No 10, 1969.

[9] E. Dijkstra, Discipline of Pprogramming, Prentice Hall, 1976.

[10] D. Gries, The Science of Programming, Springer−Verlag, New York, Heidelwerg, Berlin, 1981.

[11] V. Prevosto, ACSL Mini−Tutorial, CEA List, INRIA Centre de Recherche SACLAY – ILE de France.

[12] P. Baudin, J.Filliatre, C. Marche, B. Monate, Y. Moy, and V. Prevosto, ACSL: ANSI/ISO C Specification Language. Preliminary Design, version 1.4, http: //www.frama-c.cea.fr/ download/acsl_1.4.pdf, 2008.

[13] J. Burghardt, R. Clausecker, J. Gerlach, and H. Pohl, ACSL By Example. Towards a Verified C Standard Library, Version 14.1.1 for Frama−C (Silicon), 2017.

[14] M. Todorova, “Formal verification of procedural and object−oriented programs using the HOL theorem proof system,” Automatics and Informatics, John Atanasoff Union of Automatics and Informatics, ISSN 0861–7562, vol. XLII, No. 2, pp. 25−27, 2008.

[15] P. Armyanov, A. Semerdzhiev, K. Georgiev and T. Trifonov, “The effects of progressive evaluation and obligatory homeworks on student motivation and achievements,” Proceedings of the 12th International Technology, Education and Development Conference, Valencia, Spain, 2018, pp. 618−625, 2018.

[16] I. Donchev, “An approach to teaching object−oriented programming concepts by examples,” Thirty Seventh Spring Conference of the Union of Bulgarian Mathematicians, Borovets, April 2−6, 2008, pp. 335−341, 2008 (in Bulgarian).

[17] K. Kaloyanova, “An implementation of the project approach in teaching information systems courses,” 8th International Technology, Education and Development Conference, Valencia, Spain, pp. 7090−7096, 2014.

[18] S. Hadzhikoleva and E. Hadzhikolev, “QAHEaaS or quality assurance in higher education as a service,” Tem Journal, vol. 5, No.3, 2016, pp. 363−370, ISSN: 2217−8309.

[19] S. Hadzhikoleva, E. Hadzhikolev, S. Cheresharov, and L. Yovkov, “Towards building cloud education networks,” Tem Journal, vol.7, No.1, 2018, pp. 219−224, ISSN: 2217−8309.

[20] V. Dimitrov, Deriving semantics from WS−BPEL specifications of parallel business processes on an example, Computer Research and Modeling, vol. 7, No 3, pp. 445−454, 2015.

[21] I. Patias and V. Georgiev, Design of Robotic Systems, St Kliment Ohridski University Press, ISBN 978−954−07−4207−6, 2017.

[22] M. Todorova and K. Kanev, “Educational framework for verification of object−oriented programs,” The Joint International Conference on Human−Centered Computer Environments HCCE’2012, Hamamatsu, Japan, pp. 23−27, 2012.

Load Forecasting using Autoregressive Integrated Moving Average and Artificial Neural Network

Lemuel Clark P. Velasco, Daisy Lou L. Polestico, Gary Paolo O. Macasieb, Michael Bryan V. Reyes,
Felicisimo B. Vasquez Jr.
Mindanao State University-Iligan Institute of Technology
Premier Research Institute of Science and Mathematics
Iligan City, The Philippines

***Abstract*—Electric load forecasting is a challenging research problem due to the complicated nature of its dataset involving both linear and nonlinear properties. Various literatures attempted to develop forecasting models that utilized statistical in combination with machine learning approaches deal with the dataset's linear and nonlinear components to obtain close to accurate predictions. In this paper, autoregressive integrated moving average (ARIMA) and artificial neural networks (ANN) were implemented as forecasting models for a power utility's dataset in order to predict day-ahead electric load. Electric load data preparation, models implementation and forecasting evaluation was conducted to assess if the prediction of the models met the acceptable error tolerance for day-ahead electric load forecasting. A Java-based system made use of R Statistical Software implemented ARIMA(8,1,2) while Encog Library was used to implement the ANN model composing of Resilient Propagation as the training algorithm and Hyperbolic Tangent as the activation function. The ANN+ARIMA hybrid model was found out to deliver a Mean Absolute Percentage Error (MAPE) of 4.09% which proves to be a viable technique in electric load forecasting while showing better forecasting results than solely using ARIMA and ANN. Through this research, both statistical and machine learning approaches were implemented as a forecasting model combination to solve the linear and non-linear properties of electric load data.**

Keywords*—*Electric load forecasting; autoregressive integrated moving average; artificial neural network

I. INTRODUCTION

The fundamental characteristic that makes the electric power industry unique is the product: electricity. A single megawatt, like any other commodity, is frequently bought and resold a number of times before finally being consumed [1]-[3]. Load forecasting helps these power utilities make important decisions including decisions on purchasing electric power and load switching. There have been many tools and models used for electric load prediction. Commonly used models include Autoregressive Integrated Moving Average (ARIMA), Artificial Neural Network (ANN), time series and linear regression [2], [3]. In practice, hybrid models are being created by combining two models and have been proven to give a more accurate and more precise measure than using the individual models [1], [4]-[6]. But even these hybrid models would not always work for every electric load forecasting situation. Similar to non-hybrid models, they still depend on the type of data, the size of the data and the error handling mechanism [4], [5]. With this, power utilities would have to choose and ask for experts on recommendations regarding appropriate tools and models to be used in forecasting data. Electric load prediction conducted by these power utilities can be classified into long-term, medium-term, short-term and very short-term forecasting based on the forecasting horizon [3], [7], [8]. Short-term load forecasting is mainly used to forecast the day-ahead electric load that is why its accuracy directly affects the economic cost of operators in power utilities and markets [1]. Accurate load forecasting is helpful for security, stability, maintenance plans and economic operations in power grids. In order to obtain accurate load prediction, power utilities would need to use a forecasting tool that would work on their data and data structure.

ANN, a machine learning tool that is often used for day-ahead load forecasting exhibit certain performance characteristics similar to biological neural networks with elements capable of parallel processing like that of the human brain [2], [7], [9]. The major advantage of ANN is its flexible nonlinear modeling capability. With ANN, there is no need to specify a particular model form. Rather, the model is adaptively formed based on the features presented from the data. This data-driven approach is suitable for many empirical data sets like electric load where no theoretical guidance is available to suggest an appropriate data generation process [7], [10]-[12]. Consequently, ARIMA, popularly known as Box-Jenkins methodology is simple and yields accurate results, exhibiting its wide use by assuming that the future values of a time series have a clear and definite functional relationship with current values, past values and white noise. Although ARIMA models are quite flexible to the extent that they can represent several different types of time series, i.e. pure autoregressive (AR), pure moving average (MA) and combined AR and MA (ARMA) series, their major limitation is the pre-assumed linear form of the model which means that the ARIMA model has weakness in being able to read non-linear patterns [4], [12], [13]. Combining the two models, one which would handle the linearity and another for the non-linearity could give a better output than using just one of them.

A power utility company located in Mindanao, the Philippines has a short-term electric load forecasting system which utilizes linear regression in forecasting electric load. The current linear regression model employed by the existing load forecasting system of the power utility yields forecasted values above the international tolerance error standard of 5%. The technique used by the power utility company is valid, but there are still different techniques that could provide a better prediction using the electric load data composing of linear and non-linear properties. A hybrid model using ARIMA and ANN would be a viable solution because of its proven efficiency and affectivity in prediction [4], [13]. Through data preparation, hybrid model implementation and error measurement evaluation, this study aims to develop a day-ahead electric load forecasting model using ARIMA and ANN. This study hopes to contribute to researches in statistical and machine learning prediction technologies by implementing and evaluating a hybrid short-term electric load forecasting model that could aid power utilities in their decision-making, electric load planning and load power utilization.

II. Methodology

A. Electric Load Data Preparation

The dataset used for this study is from the three-year raw monthly electric load data that has been used by the existing system of the power utility from 2012 to 2014. However, only the electric load data coming from three metering points of 28,704 records from December 2013 to October 2014 were utilized since this range is best sufficient to fit an ARIMA and ANN model [1], [12]. As shown in Table I attributes of the historical electric load data were the following: metering point name, date, time, kilowatt delivered (KW_DEL), kilowatt per hour delivered (KWH_DEL) and kilo volt amps reactive hours delivered (KVARH_DEL).

TABLE I. Format of the Raw Electric Load Data

MP	DATE	TIME	KW_DEL	KWH_DEL	KVARH_DEL
NNN	NNN	NNN	NNN	NNN	NNN
NNN	NNN	NNN	NNN	NNN	NNN

These raw data from a .xls worksheets were then imported and stored in a PostgreSQL database. The data that was used as inputs in ARIMA modelling was identified as the kilowatt per hour (KW_DEL) column [10]. The KW_DEL is the energy delivered from the utility grid which is also the load to maintain and the basis for load prediction [5], [8]. Since the load consumption from the .xls worksheet is recorded per 15 minutes, the maximum load consumption among the four, 15-minutes recordings per hour will be set as the load for the hour [10]. The electric load data also contains scheduled and unscheduled power interruption records with zero values that could potentially cause the dataset to become out-of-range. To solve this, data correction was then done on the raw electric load data by removing empty or zero values and replacing them with electric load data reflective of the consumed electric load without the power interruption. There is currently no established standardized electric load data correction methodology for power interruptions, but researchers recommend a data correction method by replacing the outlying values with values from the electric load data of the preceding day with the same time frame as with the outlying value [6], [9], [10].

The residuals or the data that was generated from the ARIMA model also underwent a transformation process for the neural network to produce accurate forecasts. In neural networks, it is a best practice to transform input data before use since data transformation makes the training of the network faster and memory efficient resulting for the model to yield accurate forecast results [14]. In addition, neural networks only work with data usually between a specified range e.g. -1 to 1 or 0 to 1 [15]-[17]. Thus, transformation ensures that data is roughly uniformly distributed between the network inputs and the outputs [17]. The transformation technique will use a formula that is the same with the Min-Max normalization process in order for the values to be narrowed down into uniformed variation. Transformation process can be done by using the formula in (1) where *z* is the transformed value, *x* is the actual value, and *min(x)* and *max(x)* are respectively the minimum and maximum of the dataset. Hence, the transformation technique yielded a value between -1 and 1 and will be used in this study because the residuals have values which are less than 0 making the transformation technique suitable for fitting the data within the unity.

$$z = 2\,\frac{x - min(x)}{max(x) - min(x)} - 1 \qquad (1)$$

B. ARIMA and ANN Hybrid Model Implementation

In implementing the ARIMA+ANN hybrid model, the actual dataset which is in a database was initially processed in the ARIMA model. As shown in Fig. 1, there are two datasets that can be formulated with the ARIMA forecast: one of which is the linear forecast which by itself is also the ARIMA forecast, the other is the ARIMA residuals which is the difference of the actual dataset and the ARIMA forecast. The linear forecast is stored in a spreadsheet where it will be used later while the ARIMA residuals dataset is being processed in the ANN model. The process of combining the linear and nonlinear values is by adding the values of the same row of the two columns [12].

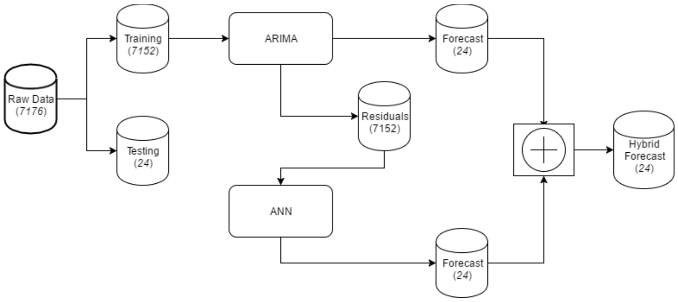


Fig. 1. Diagram of the ARIMA and ANN implementation.

In implementing the ARIMA and ANN models, the data from the database was read in a Java-based system. The ARIMA(8,1,2) model with 8 as the number of autoregressive terms, 1 as the number of non-seasonal differences needed for stationarity and 2 als the number of lagged forecast errors in the prediction equation was integrated into a Java-based system through the use of the R Statistical Software in order to simulate and calculate the linear results. An ANN model with a multilayer perceptron architecture that used Resilient

This work is supported by the Mindanao State University-Iligan Institute of Technology (MSU-IIT) as an internally funded research under the Premier Research Institute of Science and Mathematics (PRISM).

Propagation as the training algorithm and Hyperbolic Tangent as the activation function was also implemented in a Java-based system through the use of Encog Library in order to simulate and calculate the training and testing nonlinear results. Encog is a machine learning framework available for Java, .Net, and C++. Encog supports different learning algorithms such as Bayesian Networks, Hidden Markov Models and Support Vector Machines. However, its main strength lies in its neural network algorithms.

The residuals from the ARIMA model were then implemented in Encog with a maximum error of 0.0001 and a maximum iteration of 10000. As shown in Fig. 2, the ANN implemented in Encog used 1 input layer containing 24 input neurons, 1 hidden layer containing 17 hidden neurons and 1 output layer containing only 1 output neuron was used in making the network for the ANN architecture. The ANN model then delivered the output ANN forecast or the nonlinear forecast. These outputs are equivalent with the predictive size in the ANN process which corresponds to the number of forecast horizon, i.e. 24, covering the 24-hour day ahead forecasting with the 24 hourly values [2], [18]. The nonlinear forecast was then placed in the same spreadsheet with the linear forecast where the two values are then added to create an ARIMA+ANN forecast [12]. After the ARIMA+ANN hybrid model was implemented, the results from the hybrid model were assessed using error metrics to determine the accuracy. Mean Absolute Percentage Error (MAPE) and Mean Squared Error (MSE) were used as error metrics to quantify the difference between the ARIMA forecast, ANN forecast, ARIMA+ANN forecast and actual electric load.

The dataset used as validation set for the models was the actual consumed electric load data of October 21, 2014. After the predictions were generated by the models, the average MAPE and MSE were then calculated between the three models. Post-transformation of the data involves de-normalization or reversing the normalization process [14]. The de-transformation process was done by using the same formula used in the transformation process. A graphical representation of the computed results was then generated to illustrate the commonality of the actual and the predicted load values in a much better way.

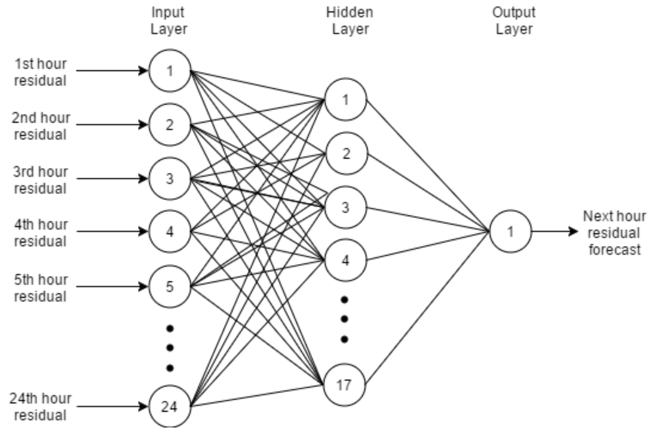


Fig. 2. ANN model architecture.

III. Results and Discussion

A. Electric Load Data Preparation Results

The granularity of the raw electric load data was originally per fifteen-minute containing the load consumption for every fifteen minutes that is why the maximum load consumption among the four, per 15-minutes recordings was chosen to reflect the hour's consumption. This is supported by studies which converted their fifteen-minute dataset to an hourly data for the reason that electric dataset are read hourly by spot markets [10], [19]. Furthermore, the fifteen-minute data was converted to hourly data by using the maximum load because it is also used by power utility in determining the hourly load. The process of choosing the maximum value from the four, 15-minutes records was performed using a Java code created by the researchers. After the maximum load was chosen, it was stored in another table in the database named the hourly table which has the following columns: time which is the combination of both the date and time column from the source table and the corresponding consumed electric load. The final result was a clean one-hour kilowatt delivered data which would serve as inputs to the models. The new number of observation would be 7176 with 7152 observations to be used for training the models and the last 24 observation for testing the models.

The number of observations to be used is just efficient for the models because a larger amount could lead to an over fitted ARIMA model while a smaller amount could lead to an under fitted. According to studies, an ARIMA model is only applicable to a definite and small amount of data [1], [4], [13]. The ARIMA model then generated residual data for ANN to process. After the raw date was fed into the ARIMA model, residuals were generated and data were plotted in the graph as shown in Fig. 3. The residual values that were plotted was a random distribution of values with -2680.893454503 as the minimum value and 3652.1309335015 as the maximum value. The values are within the range between -2680.893454503 and 3562.1309335015, which means that these values are the boundaries for the upcoming residual transformation. These values were used during the transformation process.

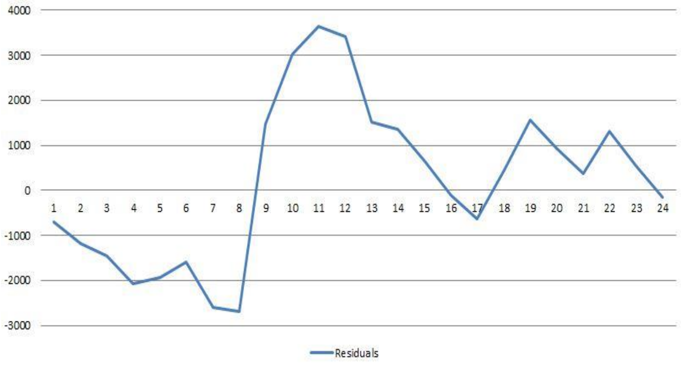


Fig. 3. Residuals dataset.

Shown in Table II are the actual residuals and normalized residual values which had values ranging from -1 to 1. The residuals had a value of -1, which is the minimum value after transformation for the minimum actual value -2680.893455 and 1, which is the maximum value after transformation for the maximum actual value 3652.130934 during the transformation. The researchers used Mix-Max transformation technique in transforming the residual values into uniformly distributed numbers between the network inputs and the output. This was supported by studies which used Min-Max transformation in order for the values to narrow down and would be used in ANN training [9], [10].

TABLE II. THE ACTUAL RESIDUAL AND NORMALIZED RESIDUALS

Actual	**Normalized**	**Actual**	**Normalized**	**Actual**	**Normalized**
-708.454	-0.377	1470.124	0.311	-642.885	-0.356
-1175.626	-0.525	3035.240	0.805	432.493	-0.017
-1459.631	-0.525	3652.131	1	1564.143	0.341
-2070.948	-0.614	3407.417	0.923	927.397	0.140
-1931.530	-0.763	1507.283	0.323	377.476	-0.034
-1601.018	-0.659	1345.452	0.272	1319.033	0.263
2596.552	-0.973	648.332	0.051	522.336	0.012
-2680.893	-1	-110.761	-0.095	-154.225	-0.807

B. *ARIMA and ANN Hybrid Model Implementation Results*

The output residuals of the ARIMA forecast was used as inputs to the ANN model. Shown in Fig. 4 are the ANN's forecasted values of the residuals for the next 24-hour data.

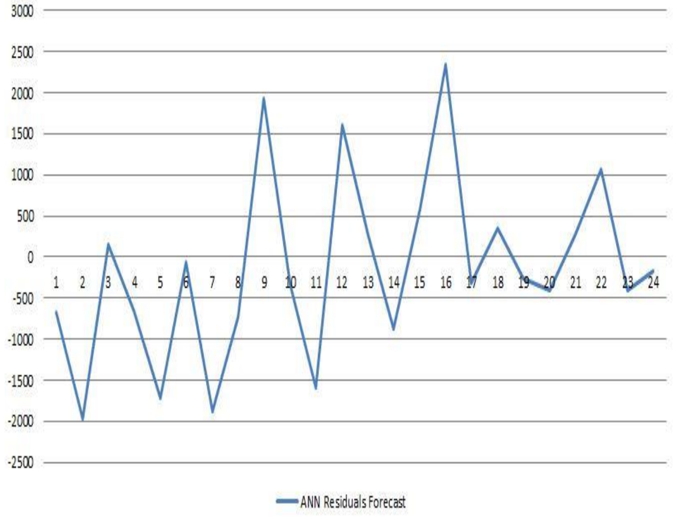


Fig. 4. ANN residuals forecast plot.

As observed, the data is in a nonlinear state which also came from the residuals which are initially nonlinear. The ANN forecast data-driven approach is suitable for this kind of empirical data sets where no theoretical guidance is available to suggest an appropriate data generating process [11]. In addition neural networks are flexible in terms of nonlinear modeling capability [9], [11], [12]. As shown in Table III, these values were then used in the next process as addends to be summed up with the ARIMA forecast.

TABLE III. ADDING THE LINEAR AND NONLINEAR FORECASTS

	ARIMA Forecast		**ANN Forecast**		**ARIMA+ANN Forecast**
1	26510.853	+	-183.272	=	26327.581
2	25565.625	+	-1245.971	=	24319.654
3	24966.430	+	-360.782	=	24605.648
…	…	+	…	=	…
24	28439.424	+	-463.978	=	27975.445

The ARIMA+ANN forecast or the hybrid forecast is the final result of the forecasting process. The results are based on a researcher's assumption that when added, the linear and nonlinear forecast taken from the ARIMA and ANN models would create the final output of the hybrid model [12]. To evaluate the forecasting performance, MSE and MAPE were computed for the ARIMA model output, ANN model output and ARIMA+ANN hybrid model output. Table IV shows the MSE of each model. The researchers observed that the ARIMA+ANN hybrid model output had the smallest MSE among the three models having a sum of 0.98, followed by the ARIMA model output with 1.23 and ANN model output of 16.90.

TABLE IV. MSE OF THE OUTPUT OF EACH MODEL

HOUR	ARIMA	ANN	ARIMA + ANN	HOUR	ARIMA	ANN	ARIMA +ANN
1	0.027	0.220	0.001	13	0.043	0.323	0.035
2	0.048	0.013	0.033	14	0.037	0.318	0.061
3	0.062	0.044	0.069	15	0.018	0.176	0.002
4	0.093	0.235	0.064	16	0.003	0.143	0.068
5	0.086	0.277	0.010	17	0.018	0.145	0.009
6	0.068	0.199	0.066	18	0.012	0.170	0.002
7	0.112	0.089	0.031	19	0.042	0.184	0.049
8	0.111	0.012	0.081	20	0.026	0.176	0.038
9	0.050	0.163	0.016	21	0.011	0.129	0.003
10	0.094	0.256	0.104	22	0.040	0.127	0.008
11	0.106	0.300	0.152	23	0.017	0.040	0.030
12	0.095	0.317	0.051	24	0.005	0.001	0.001

On all of the hours, the 24th hour gained the lowest MSE than the other hours while the 13th hour gained the highest MSE. For the ARIMA model output, the 16th hour gained the smallest MSE while the 7th hour has the highest MSE. For the ANN model output, the 24th hour gained the lowest MSE, while the 13th hour has the highest MSE. Lastly for the hybrid model output, the 24th hour gained the lowest MSE while the 11th hour has the highest MSE. The lowest MSE were found in the hours after 13 while the highest MSE are found on 13 and above it. From the result obtained, MSE from each model were below 0.5, and the MSE for both ARIMA and ARIMA+ANN hybrid models are below 0.3. According a study, a good predictive model has a MSE of below 0.5 [20]. Thus, the ARIMA model and the ARIMA+ANN model are suitable forecasting model for predicting electric load.

The MAPE of each model result was also calculated in order to evaluate the performance of each output parameter and assess whether the model was able to pass to the acceptable error for electric load forecasting. In calculating the MAPE, the forecasted outputs were subtracted with the actual values and then the difference was divided by the actual values. The results were then multiplied by 100 for percentage. Table V shows the MAPE of the three models. The ARIMA+ANN hybrid model has the lowest MAPE compared to the two other models, making it the best fitting model for the dataset. The MAPE of each model can also be used to check whether the models can be acceptable models for prediction. According to a research, the acceptable MAPE error for testing should be below 15% in order to say that the model is well-performing [21]. The ARIMA and ARIMA+ANN hybrid model was able to reach that goal. The MAPE of the ARIMA+ANN hybrid model was also below 5% which is the acceptable error of power utility making it a fitting model for use.

TABLE V. MAPE OF THE OUTPUT OF EACH MODEL

Model	**Mean Absolute Percentage Error**
ARIMA Model	5.11%
ANN Model	16.90%
ARIMA+ANN Hybrid Model	4.09%

The researchers also evaluated the ARIMA+ANN hybrid model if it can still maintain its forecasting accuracy by predicting the 2-days ahead, 3-days ahead, 4-days ahead and 5-days ahead and comparing the result to the actual days of October 22 - 25, 2014. RMSE and MAPE were used in evaluating the forecasting accuracy of the model. As shown in Table VI, there is an increase in the RMSE and MAPE of the 2-days ahead, 3-days ahead, 4-days ahead and 5-days ahead from the 1-day ahead forecast. The researchers observed that it is not reliable to use the same model for forecasting a number of days ahead. Instead, remodeling should occur every time the model is being used in forecasting the next day.

TABLE VI. FORECASTED RESULTS FROM OCTOBER 21 TO 25 OF 2014

	OCT. 21	OCT. 22	OCT. 23	OCT. 24	OCT. 25
RMSE	1741.632	3310.820	2500.884	2030.299	2767.667
MAPE	4.09%	8.71%	7.43%	4.99%	8.60%

A visualization of the difference between the actual consumed electric load values and the denormalized ARIMA forecasted outputs, ANN forecasted outputs, and ARIMA+ANN hybrid outputs was also generated for evaluation. As shown in Fig. 5, the ARIMA forecasted outputs were compared to the actual values of October 21.

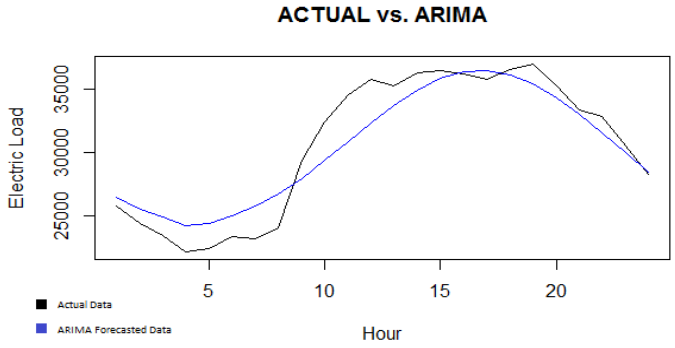


Fig. 5. Comparison of the actual load data and the ARIMA forecast.

It can be observed that in the hours between 1 and 10, the ARIMA model has higher forecasted values than the actual consumed electric load values. Meanwhile, in the hours between 10 and 15, the ARIMA model has lower forecasted values than the actual consumed electric load values. The values from 15 to 24 are very close to the actual values. Overall, the forecasted outputs from the ARIMA model were close to the actual values.

Shown in Fig. 6 is the ANN forecasted output as compared to the actual consumed electric load values. The forecasted outputs from ANN had a big difference to the actual data, compared to the ARIMA forecasted outputs. From the hours of 1 to 8, the ANN forecasted data was higher than the actual values, with the exception of hour 2 which is close to the actual value for that hour. The ANN forecast from the hours of 9 to 23, the forecasted outputs of ANN were very low that that of the actual values. Only on the hour of 24 that the forecasted data of the ANN has almost predicted the exact value of that hour. But as overall result, the forecasted outputs from the ANN model were very far from the actual values. The properties of the nonlinear estimators depend on the assumption that residual errors were independent and normally distributed with mean zero and correctly defined variance. Violations of this assumption can cause bias in parameter estimates, invalidate the likelihood ratio test and preclude simulation of real-life like data. The choice of error modelling is mostly done on a case-by-case basis from a limited set of commonly used models [7], [11], [16]. This is basically why ANN residuals forecast has a very far-off prediction as to the actual residuals.

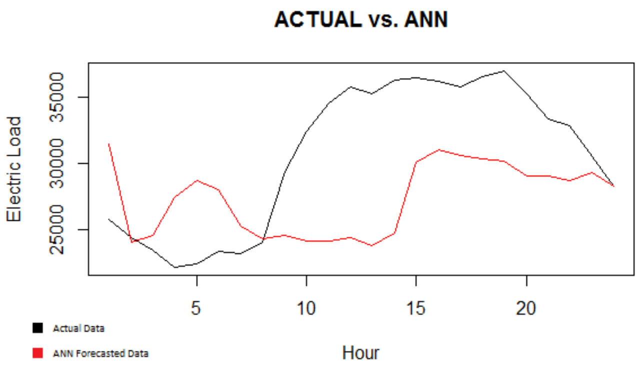


Fig. 6. Comparison of the Actual Load Data and the ANN Forecast.

The comparison between the actual values and the ARIMA+ANN hybrid forecasted values is shown in Fig. 7. On the hours of 3, 6, 7 and 16, the forecasted outputs of the hybrid model was higher than that of the actual values. On hour 16, the forecasted value was much higher compared to all the models having a value 2000 more than the actual value. While on the hours of 2, 10 to 13 and 20, the forecasted values of the ARIMA+ANN hybrid model was lower than that of the actual values. Moreover, the rest of the hours almost had the exact value as the actual values with difference ranging from an estimate of 250 to 10. The ARIMA+ANN hybrid model outputs have the most values closest to the actual values of the three models. This was supported by study which also used ARIMA+ANN in which the hybrid model is able to perform well in terms of accuracy for every component model used in isolation [12].

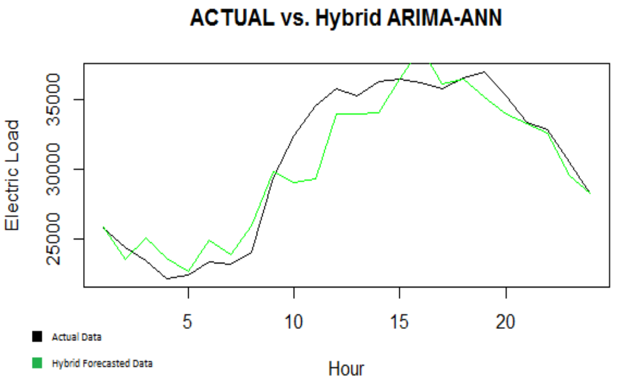


Fig. 7. Comparison of the Actual Load Data and the ARIMA+ANN Forecast.

A plot diagram shown in Fig. 8 shows the difference between all three models as to how each model are close to the actual consumed electric load values. From hours between 1 and 8, most of the forecasted values are higher than the actual values. ARIMA and the hybrid model are closest to the actual value while the ANN was farther away, except for hour 2. On hours of 9 to 22, most of the forecasted outputs were lower than the actual values, with the exception of hour 16 of the hybrid model and hour 17 of the ARIMA model which has values above the actual values. On hours 23 and 24, the values were near to the actual values with hour 24 of the three models almost captured the exact value.

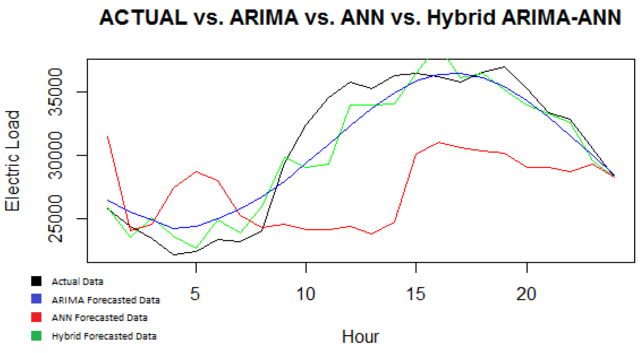


Fig. 8. Comparison of the Actual Load Data and Forecast of All Three Models.

As for the overall comparison of results, the forecasted outputs of the ARIMA+ANN hybrid model was closer than that of the ARIMA and ANN models. The forecast difference of the ANN model can be the result of the data input since only one column was fed to the ANN model and because of its usage as an error model. The ANN model being used acts like an error model since its inputs are the residuals of the ARIMA model, but the error model itself can still be used as an individual forecasting model [12]. Additionally, the researcher also stated in his study that the ANN error model may not give out good results if used individually. A separate study supports this assumption that the ARIMA+ANN hybrid model has better accuracy than the individual models because of how the ARIMA model caters to the linearity of the dataset and how the ANN model caters to the non-linearity of the same dataset [22].

IV. Conclusion and Recommendations

This study attempted to implement and evaluate the performance of ARIMA, ANN and ARIMA+ANN hybrid models in predicting day-ahead electric load. A Java-based system was created which calls R for the ARIMA model and integrates Engoc library for the ANN model. Compared to the ARIMA(8,1,2) and the ANN model which used Resilient Propagation as the training algorithm and Hyperbolic Tangent as the activation function, the ARIMA+ANN hybrid model yielded the best forecasting performance with a MAPE value of 4.09% and a RMSE value of 1959.41 ARIMA+ANN hybrid model also obtained an error rate which is below the acceptable tolerance error of 5%. Since the results of the ARIMA+ANN hybrid model has a lower MAPE than that of the ARIMA model and the ANN model, the hybrid model thus generate better result in prediction than solely using ARIMA and ANN.

This study only focuses on the ARIMA and ANN hybrid models; however, there are still other forecasting models that could also be viable in predicting day-ahead electric load. ARCH is one of the later model created that has the ability to read both linearity and non-linearity of data, however it is more into the linear side and could possibly be used in a hybrid model with ARIMA. Other hybrid models may also yield an even lower percentage error than that of this study and could be crucial in determining the better predicting model for an electric load dataset. R Statistical Studio and Encog library are two of the open source IDE and library, respectively, used in creating the hybrid model. R has no libraries that can allow it to be integrated in Java and is solely able to predict linear data while Encog has limited training algorithm and activation function. The possibility that using other development environments or libraries can lead to a better forecast or using a system that can handle both linear and nonlinear predicting models might gain a better result than using separate ones.

The results of this research clearly suggest that the use of a hybrid model that caters the linearity and non-linearity of a dataset proves to be a better technique for a day-ahead electric load forecasting rather than the use of an individual model.

Acknowledgment

The authors would like to thank Mr. Art Louie Bangcale for his technical assistance in the models evaluation. Additionally, the authors would also like to thank the support of the MSU-IIT Office of the Vice Chancellor for Research and Extension and PRISM-Premiere Research Institute in Sciences and Mathematics for their assistance in this study.

References

[1] H. Cui and X. Peng, “Short-Term City Electric Load Forecasting with Considering Temperature Effects: An Improved ARIMAX Model,” Mathematical Problems in Engineering, pp. 1-10, 2015, doi:10.1155/2015/589374.

[2] L. C. Velasco, N. Estoperez, R. J. Jayson, C. J. Sabijon, V. Sayles, “Day-ahead Base, Intermediate, and Peak Load Forecasting using K-Means and Artificial Neural Networks,” International Journal of Advanced Computer Science and Applications(IJACSA), 9(2), 2018, http://dx.doi.org/10.14569/IJACSA.2018.090210.

[3] I. Asenova, D. Georgiev, “Short-term Load Forecast in Electric Energy System in Bulgaria,” Power Engineering and Electrical Engineering, 8(4), 2010.

[4] M. Kumar and M. Thenmozhi, ”A Comparison of Different Hybrid ARIMA - Neural Network Models for Stock Index Return Forecasting and Trading Strategy,” International Journal of Financial Management, 1(1), 2012.

[5] Y. Yang, J. Wu, Y. Chen, C. Li, “A New Strategy for Short-Term Load Forecasting” Abstract and Applied Analysis, 1-9, 2013, doi:10.1155/2013/208964

[6] M. Meng, W. Shang, D. Niu, “Monthly Electric Energy Consumption Forecasting Using Multiwindow Moving Average and Hybrid Growth Models,” Journal of Applied Mathematics, 1-7, 2014, doi:10.1155/2014/243171.

[7] K. I. Ibraheem, and M. O. Ali, “Short Term Electric Load Forecasting based on Artificial Neural Networks for Weekends of Baghdad Power Grid,” International Journal of Computer Applications IJCA, 89(3), 30-37, 2014, doi:10.5120/15484-4263.

[8] L. C. Velasco, D. L. Polestico, D. M. Abella, G. Alegata, G. Luna, “Day-Ahead Load Forecasting using Support Vector Regression Machines, ” International Journal of Advanced Computer Science and Applications (IJACSA), 9(3), 2018, http://dx.doi.org/10.14569/IJACSA.2018.090305.

[9] P. Ramachandran, R. Senthil R., “An Approach in Artificial Neural Network in Predicting Power Load Forecasting For Short-Term of Indian Electrical Utility, ” I-manager’s Journal on Engineering and Technology, 2(2), 2007.

[10] L. C. Velasco, P. N. Palahang, J. A. Dagaang, "Next day electric load forecasting using Artificial Neural Networks", 2015 International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management HNICEM, 2015, IEEE, DOI: 10.1109/HNICEM.2015.7393166.

[11] S. K. Nanda, D. P. Tripathy, S. K. Nayak, S Mohapatra “Prediction of Rainfall in India using Artificial Neural Network (ANN) Models,” IJISA International Journal of Intelligent Systems and Applications, 5(12), 1-22, 2013, doi:10.5815/ijisa.2013.12.01.

[12] G. Zhang, “Time series forecasting using a hybrid ARIMA and neural network model,” Neurocomputing, 50, pp. 159-175, 2003, doi:10.1016/s0925-2312(01)00702-0.

[13] I. Khandelwal, R. Adhikari, G. Verma, “Time Series Forecasting Using Hybrid ARIMA and ANN Models Based on DWT Decomposition”, Procedia Computer Science, 48, pp. 173-179, 2015, doi:10.1016/j.procs.2015.04.167.

[14] N. Rotich, “Forecasting of wind speeds and directions with artificial neural networks,” LUT Energy, Lappeenranta University of Technology, 2014.

[15] R. K. Jain, “Normalizing tumor vasculature with anti-angiogenic therapy: A new paradigm for combination therapy,” Nat Med 7, pp. 987–989, 2001.

[16] F. Liu, Z. Wang, J. Wu, J. Wang, “A Hybrid Forecasting Model Based on Bivariate Division and a Backpropagation Artificial Neural Network Optimized by Chaos Particle Swarm Optimization for Day-Ahead Electricity Price,” Abstract and Applied Analysis, pp. 1-31, 2014, doi:10.1155/2014/249208.

[17] N. Vafaei, R. A. Ribeiro, Luis M. Camarinha-Matos, "Importance of Data Normalization in Decision Making: case study with TOPSIS method", proceedings of the 1st International Conference on Decision Support Systems Technologies, 2015 .

[18] J. Heaton, “Programming neural networks with Encog3 in Java,” MO: Heaton Research Inc., 2011.

[19] J. E. Gardner and R. L. Lehr, “Enabling the Widespread Adoption of Wind Energy in the Western United States: the Case for Transmission, Operations and Market Reforms,” Journal of Energy & Natural Resources Law, 31(3), pp. 237-285, 2013, doi:10.1080/02646811.2013.11435333.

[20] R. Veerasamy, H. Rajak, A. Jain, S. Sivadasan, C. P. Varghese, R. K. Agrawal, “Validation of QSAR Models - Strategies and Importance,” International Journal of Drug Design and Discovery, 2(3), 2011.

[21] H. Khosravani, M. Castilla, M. Berenguel, A. Ruano, P. Ferreira, “A Comparison of Energy Consumption Prediction Models Based on Neural Networks of a Bioclimatic Building,” Energies, 9(1), 57, 2016, doi:10.3390/en9010057.

[22] A. I. Elwasify, “A combined model between Artificial Neural Networks and ARIMA Models,” International Journal of Recent Research in Commerce Economics and Management, 2(20), pp. 134-140, 2015.

Interactive Visual Decision Tree for Developing Detection Rules of Attacks on Web Applications

Tran Tri Dang, Tran Khanh Dang
Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology, VNU-HCM
Ho Chi Minh City, Viet Nam

Truong-Giang Nguyen Le
IT Security Division
EVN Finance Joint Stock Company
Ho Chi Minh City, Viet Nam

***Abstract*—Creating detection rules of attacks on web applications is not a trivial task, especially when the attacks are launched by experienced hackers. In such a situation, human expertise is essential to produce effective results. However, human users are easily overloaded by the huge input data, which is meant to be analyzed, learned from, and used to develop appropriate detection rules. To support human users in dealing with the information overload problem while developing detection rules of web application attacks, we propose a novel technique and tool called Interactive Visual Decision Tree (IVDT). IVDT is a variant of the popular decision tree learning technique introduced in research fields such as machine learning and data mining, with two additionally important features: visually supported data analysis and user-guided tree growing. Visually supported data analysis helps human users cope with high volume of training data while analyzing each node in the tree being built. On the other hand, user-guided tree growing allows human users to apply their own expertise and experience to create custom split condition for each tree node. A prototype implementation of IVDT is built and experimented to evaluate its effectiveness in terms of detection accuracy achieved by its users as well as ease of working with. The experiment results prove some advantages of IVDT over traditional decision tree learning method, but also point out its problems that should be handled in future improvements.**

***Keywords*—Interactive analytics; security visualization; visual decision tree; web application security**

I. Introduction

Decision tree learning is a popular technique for solving object classification problem, which is a common task in research fields such as machine learning and data mining. One of the reasons for this popularity is its ability to handle multi-dimensional data well. Another notable feature of decision tree learning is its human friendly presentation of learning result. Compared with other classification techniques like artificial neural network or support vector machine, whose results are rather black boxes to external users, decision tree learning produces hierarchical sequences of classification rules that are easy to understand and follow. It is even better when these sequences of rules are displayed in visual forms in which the most popular one is node-link diagram. In these diagrams, nodes correspond to collection of data objects, and links (i.e. edges) correspond to conditions to partition similar objects into subsets. Human users can classify a data object manually by starting at the root node and subsequently following links whose conditions the object satisfies, until reaching a leave node, with a specific class label, i.e. the classification result.

The presentation of decision trees and the way those trees are constructed are more or less can be done manually by a human user. Indeed, there are several unique benefits by integrating human users into decision tree building process. Firstly, it is the direct application of users' specific domain knowledge to construct trees. Automatic tree building algorithms tend to serve as general solutions to classification problems, so it is difficult or not natural to inject expert knowledge to guide the tree building process. Furthermore, algorithms implementers are usually data scientists, not domain experts, hence they may not see the problems being solved the way actual users see. In contrast, when real users are involved into tree building process, they can guide the construction using their own knowledge and experience. As an example, although "first name" and "password" are both, textual data, their meaning and natural pattern are different. Considering them as text, as general decision tree algorithms do, will lost their important differences. These losses will not happen when domain knowledge is applied appropriately when building trees.

Secondly, when human users actively participate in tree construction phase, they will understand more clearly and use the resulted tree more effectively in classification phase. By joining the tree construction process, human users get an overview picture of the dataset and the distribution of each attribute. From this knowledge, they see the reasons behind node splitting rules and priorities among data object's attributes. This gained insight, in turn, helps them work more effectively with the resulted tree in classifying data objects. They can even start over the tree construction phase when seeing the final result not satisfying. Starting over with experience gained from previous works can create a total different tree. Again, this effect is not trivial to implement in automatic tree construction.

The last, but not least, benefit is the flexibility in dealing with different data types and forming node splitting conditions human users can achieve. In automatic tree construction, only a limited data types, such as numeric or nominal, can be used directly. More complicated types need to go through a reprocessing phase before they can be used. Because the reprocessing and construction phases are independent, it is not possible to determine in advance if a reprocessing method is suitable for a particular data attribute. This makes the reprocessing phase complex and time-consuming. On the other hand, in human-guided tree construction, reprocessing task is integrated into tree construction phase. As a result, using and changing reprocessing methods have instant feedbacks, thereby reducing time and complexity. Furthermore, unlike automatic algorithms which only produce simple predefined splitting rules, human users can enter any computable Boolean formulas for each tree node, making it quite flexible for classifying complex objects.

With the above benefits, it is natural to integrate human users into decision tree building technique to solve complex classification problems, one of which is the attack recognition of users' inputs on web applications. There are several reasons to integrate human security administrators into decision tree building process for the web application attack recognition problem. The first reason is computer attacks in general, and web application attacks in particular, are usually complicated. Because the web is one of the most popular platforms used today, its users have many differences in technical skills and background. This results in the determination whether a piece of user input is normal or not cannot be completely objective. For a particular instance of user input, one security administrator may see it as normal, yet another administrator may see it as abnormal. That is because besides the input data itself, it is necessary to consider its surrounding context, such as the user who creates it and the environment in which it resides, to make a fully informed decision. Without the direct involvement of a human administrator, it is not easy to integrate this context information into the classifier.

The second reason is related directly to the technique used to recognize web application attacks. As mentioned earlier, among machine learning methods, decision tree learning is one of the most human-friendly approaches, both for constructing classifiers as well as using learned results to classify data objects. As a result, integrating human users into decision tree learning is a natural choice to take advantages of the human reasoning capability and powerful computer data processing at the same time. This leads to the next question that we need to address: how to integrate human administrators into decision tree building effectively?

Because human users interact with computing systems via the user interface components, these are the places where most integration happens. The user interface should be designed to communicate effectively with its main users, in this case are the web application security administrators. Although there are many methods of communication, interactive visual interface is one of the most preferred methods. Unlike traditional text-based command line interface, interactive visual interface can provide more natural interaction and intuitive appearance to the users. Furthermore, when working with a high volume of input data, displaying it visually helps users overcome the information overload problem.

From the reasons mentioned above, in this paper we describe a method and tool called Interactive Visual Decision Tree (IVDT) whose purpose is to support web application security administrators in creating attack detection rules. IVDT is an interactive visual tool for building decision trees, specifically targeting at web application input classification. Given a collection of data input objects and their respective labels, security administrators can use IVDT to build visual decision trees which are used later to classify unknown inputs. But before proceeding further, it should be noted that although IVDT provide some unique advantages, it is not meant to replace automatic classification methods. Instead, we believe it should be used together with other techniques to utilize the strengths of all, especially in a complicated domain like security.

The rest of the paper is structured as follow: In Section II, we review the related works for this research; in Section III, the user interaction and data visualization of IVDT are described in detailed; the prototype implementation of IVDT is introduced in Section IV; Section V is used to describe our experiments and their results; and finally Section VI concludes the paper with lessons learned and a plan for future works.

II. Related Works

A. *Decision Tree Learning*

One of the most popular decision tree learning algorithms is the ID3 (Iterative Dichotomiser 3) proposed by Quinlan [1]. The input for this algorithm is a dataset containing data objects in many classes. Each data object in the dataset has the same number of data attributes and is in one particular class. At each step, ID3 selects a data attribute that has not been used and creates a formula on it to split the data objects into subsets. All data objects in a common subset have the same outcome when applying the split formula. The decision to select which unused attribute at a step to create a split formula is based on a value called information gain of that attribute.

The information gain (IG) is defined as

$$IG = H(parent) - \sum H(children) \tag{1}$$

In (1), *H(parent)* is the entropy of the parent node before splitting, and *ΣH(children)* is the weighted sum of the entropy of all child nodes that are the result of the split. Entropy of a node *N*, in turn, is defined as

$$H(N) = -\sum(p_i * logp_i) \tag{2}$$

In (2), p_i is the percentage of data objects in node *N* having *i* as the common class label. The goal of ID3, and other decision tree learning methods, in creating split conditions is to have the child nodes purer than the parent node. One of the disadvantages of ID3 is its inability to work on continuous domain data types without a preprocessing step.

Some of the limitations of ID3 are solved in C4.5, another decision tree learning method also proposed by Quinlan [2]. Among the improvements, C4.5 algorithm can handle continuous data attributes by generating a threshold for each of them, and use this value to split the parent node into two child nodes. Data objects having attribute values higher than the respected threshold are put into one child node, and the rest are put into the other child node. Another notable improvement of C4.5 compared with ID3 is that it can work with data objects missing values on some attributes. These missing values are simply not used in entropy and information gain calculations when deciding which attributes to use in split formulas.

CART (Classification and Regression Trees) is another popular technique for decision tree learning, proposed by Breiman et al. [3]. This technique is useful not only for classification, but also for regression. The output of CART is determined by the type of the dependent variable: if the dependent variable is categorical, the resulting tree is a classification tree; on the other hand, if the dependent variable is numeric, the resulting tree is a regression one. Like other decision tree learning methods, CART algorithm follows a greedy approach, i.e. at each step a split on a node is chosen as to maximizing the purity of the resulting child nodes. The purity metrics can be deviance, entropy, or gini index.

B. Interactive Decision Tree Learning

One of the earliest works on interactive decision tree learning was proposed by Ankerst et al. [4]. In that work, the authors used a multidimensional visualization method on the training data to support the users in selecting the split point optimally. The visualization technique is pixel-oriented, similar to the Circle Segments method [5], in which each attribute value is mapped to a particular color based on the class of the data object having that attribute value. Different attributes' values are positioned in separate areas. For data objects with D attributes, a circle with D segments, each segment for one attribute, is used to represent them. In each segment, the data values of the respective attribute are positioned from the center of the circle to the outer border line-by-line, each line is orthogonal to the segment halving line. The human users can select an unused attribute to create a split formula on it to grow the decision tree. They can also remove an attribute from the current decision tree, to backtrack to a previous state.

Another research on interactive decision tree learning is the PaintingClass by Teoh & Ma [6]. PaintingClass is a system for interactive construction, visualization and exploration of decision trees. Although the main objective of PaintingClass is to interactively construct decision trees, its other two objectives, visualization and exploration, are not less important. In fact, visualization and exploration features help PaintingClass's users have a better understanding of the underlying data, which in turn, make them be able to create small and accurate trees. Parallel coordinate visualization [7] is used in PaintingClass to display multi-dimensional objects in two-dimensional screen.

BaobabView is another work in this category [8]. Its user interface design has some differences compared with the previous works. The three most important visual areas of it are decision tree main view, attribute view, and confusion matrix view. The decision tree main view displays the decision tree being built in a node-link diagram. The nodes are containers of data objects, while the links display the flow of data objects from parent nodes to child nodes. Sizes of nodes and links correspond to the number of data objects they contain. Details about attributes' values of a selected node are shown in the attribute view. These values are sorted based on some impurity metrics such as information gain [1], gain ratio [2], and gini gain [3] to help users in seeing the overall data value distribution, and using that information to create appropriate split formula for the selected node. The confusion matrix view displays the correct and incorrect classified objects, to give users instant feedback for their choices.

C. Security Visualization

Security visualization is a multi-disciplinary research field studying the use of information visualization techniques to solve computer security problems. In security visualization systems, human user is an integral part. The reason behind this tight integration is to combine the powerfully graphical capability of computers with the efficiently visual analysis of human to solve complex security problems. This combination creates new advantages that are difficult to achieve when either human analysis or computer processing is used individually. One of the notable advantages is to help human users overcome the information overload issue when working with security data, thereby letting them make informed decisions.

In a research, Choi et al. used parallel coordinate method [7] to visualize network traffic for security administrators to detect large-scale internet attacks such as worms, DDoS, or network scanning activities [9]. Because each type of attacks has a particular visual pattern when visualized (the authors called them visual attack signatures), administrators can easily and quickly recognize them. After recognition, the administrators can further investigate the traffic data source in more detailed to see if the attacks are true or not. Although the final decision is made by human users, this visualization does enhance the decision making process by reducing the time and effort of administrators in analyzing high volume traffic data.

Security visualization is not only helpful for security experts, but also for average end users. End users need simple but effective software interface to accomplish their tasks. But if the tasks are security – related, the interface is rarely simple. One such task is sharing files between users. Because there are many rules governing the final permission of the shared files or folders, the interface to configure permission is rather complicated. To make the file sharing task simple and secure at the same time, Heitzmann et al. developed a visually secure interface for NTFS file system [10]. This interface uses Treemap [11] to display hierarchical folder structure with different colors for different permission levels. Because colors can be effortlessly differentiated by human, it is easy for a user to know if moving/copying files/folders to new locations violates their original permissions or not. Another task deserves looking at is web browsing. This is a popular task and involves many security decisions to be made by the users during a browsing session. As a result, browser software vendors invest much effort in designing visual interface components to communicate with their users about security information, such as sites protected by SSL/TLS, cookies, possible phishing sites, etc. A more comprehensive survey about browsers' security designs is given in [12].

One similar research of this work was proposed by Dang and Dang [13]. In that research, the authors described a security visualization technique to analyze user inputs on HTML web forms. Multi – levels zooming is used to provide administrators different levels of detailed views, depends on their needs. Unlike traditional text display method, with this visualization technique, the authors demonstrated that thousands of user input data objects can be displayed and analyzed at a time. Furthermore, built-in interactions support security administrators in selecting and viewing specific subset data in more detailed. By working directly with user input data, human administrators can have an overall understanding of the structure of the web application they are trying to protect as well as its environment (input forms, types of user, complexities of attacks, etc.). This understanding is not easily to obtain when they only work on the surface with automatic tools like IDS, IPS. The difference between that research and our work is that ours go a step further by providing a tool not only for analyzing user inputs, but also for developing attack detection rules in the form of decision trees.

III. Visualization and Interaction Design

A. *Problem and Solution Specification*

In this section, we describe the format of user input data used by administrators to develop attack detection rules and details about the decision trees being built. Although web users can input data at any accessible location, these data objects must follow some predefined structures. These structures are determined in advance by web developers. As an example, the data objects to log in to an e-commerce website may contain fields such as “username” and “password”, while the data objects to sign up an account may contain more fields like “first name”, “last name”, “email”, “address”, “telephone number”, etc. However, the data objects for a specific action will have the same structure. For each action, because of the similarity of its data objects, a decision tree can be built on it to classify data objects into either normal or attack.

More specifically, each user input record is considered as a separate data object. In general, if an action requires a data record with N input fields, the corresponding data object will have N attributes. In reality, administrators can choose to exclude some fields if they think these fields are not necessary for recognizing attacks. Each data object used in the tree building phase has a class label, which is either normal or abnormal. In the verification phase, the result tree is used to compute the class label for new data objects. In the beginning, all data objects are in the same node, root node, of the decision tree. The administrators then create a Boolean rule, with one unused attribute, for the root node. Data objects satisfy that rule are copied to a child node of the root, while the others are copied to the other child. This process continues until all data objects in a node having the same class, or when the percentage/number of data objects in one class is small/big enough. When a new object is put into the tree, it will follow the created rules until reaching a leave node whose label is assigned to the new node.

B. *Visualization and Interaction Design*

When building attack recognition decision tree, administrators selects a node which contains a collection of data objects, i.e. user input data records, enters a Boolean split expression for the selected node to spit its contained data objects into two child nodes. This process continues until the whole tree is built. To support the administrators in analyzing data and determining appropriate split expressions, we provide two supporting tasks, node analysis and tree analysis. The main tasks and their relationship are depicted in Fig. 1.

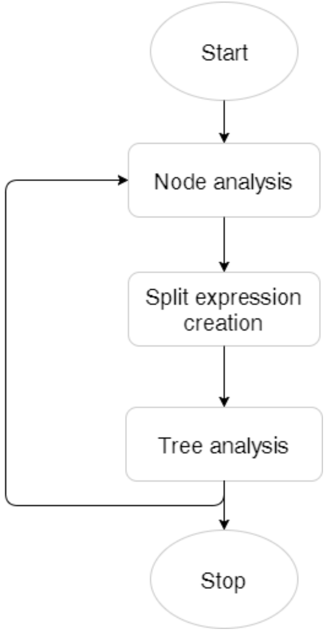


Fig. 1. The main tasks provided by ivdt for security administrators.

Node analysis: this task is used to inspect a selected node in details. The administrators can try different functions on any unused attributes to see how the data objects distribute. Usually, the pair of <*attribute*, *function*> that separates the data objects most should be chosen to create the split expression for the selected node. We provide a supported visual interface for this task, which is shown in Fig. 2. In Fig. 2, the pie chart displays the percentage of normal and abnormal data objects in the selected node. On the right of the pie chart are the distributions of data objects when custom functions are applied to unused attributes. The outputs of the custom functions are mapped to the X – coordinate, while the Y – coordinate is used to represent the number of data objects having the same X position, similar to the histogram presentation method. In Fig. 2, there are 2 distributions and it is easy to conclude that function 2 (segment (c) of Fig. 2) separates data objects better than function 1 (segment (b) of Fig. 2), and as a result, the administrators should use function 2 as the split expression.

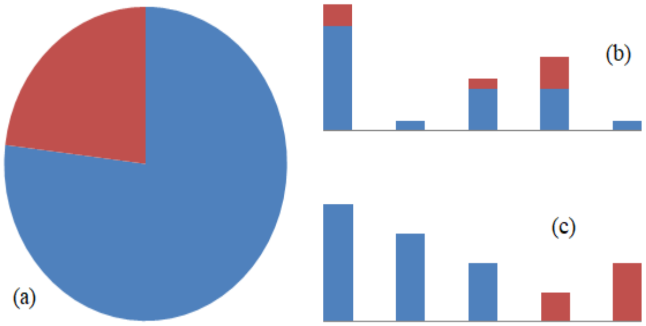


Fig. 2. Visualization of a node and custom functions: (a) Distribution of data objects over the whole node; (b) & (c) Distribution of data objects as histogram of custom functions on selected unused attributes.

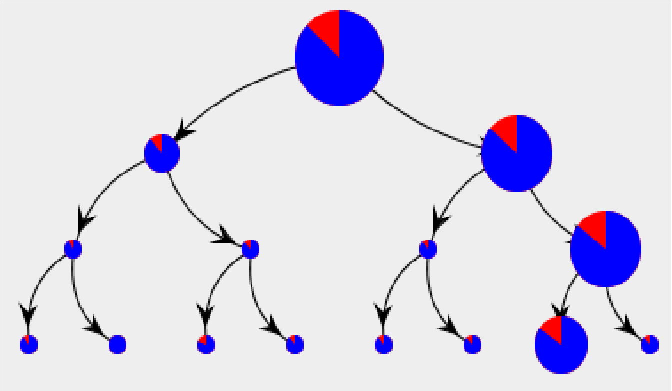


Fig. 3. Visualization of the attack detection decision tree.

Split expression creation: when the administrators finish analyzing a node, they can create a Boolean split expression for that node. The general form of the split expression is $S(a_i)$, in which S is any Boolean function and a_i is any unused attributes of the selected node. Because the split expression is a Boolean function, it maps the data objects into two groups. Data objects having the same output value are put into the same child node of the selected node. In this implementation, for no particular reason, the left child node is used to store data objects not satisfying the split expression and vice versa.

Tree analysis: after a split expression is created for a selected node, that node expands into two child nodes. This results in a new tree. The tree reflects the overall progress the administrators have made in developing attack detection rules. In contrast with node analysis, which provides a tool for local optimization, tree analysis focuses more on the global goal, i.e. it answers the question: how good is the attack detection decision tree being built? It does so by displaying the whole tree using pre-attentive visual attributes [14] to communicate crucial information with administrators quickly and naturally. More specifically, color is used for data objects' class label and node size is used for number of data objects in a node. These two graphical attributes together with tree's depth give administrators an approximate estimation of the goodness of the tree. Based on this subjective estimation, administrators can adjust their split expression strategy for new child nodes, or start over with a new tree. A sample visual decision tree is displayed in Fig. 3.

IV. Implementation

To experiment with the proposed technique, we've developed a prototype implementation of IVDT. The general architecture of the prototype is depicted in Fig. 4. The main components in the architecture are described below.

The User input Database: this is used to store the input records from external users of the target web application. Each input record contains a collection of <*name, value*> pairs. In each pair, *name* is a fixed value predefined in advance by the web application developer while *value* is an actual value entered by the web application user. All input records have the same structure, i.e. having the same number of <*name, value*> pair and the same *name* elements. If an input record has a different structure, there is a high chance that it is the direct result of form modification attack [15].

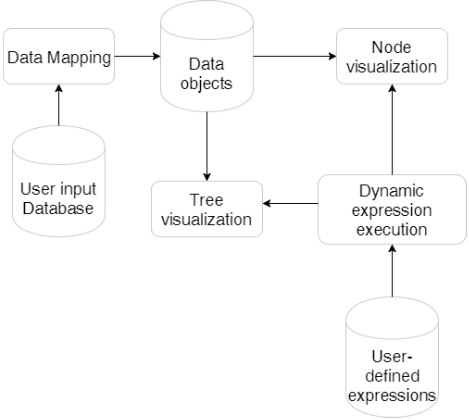


Fig. 4. Main components of the prototype implementation of IVDT.

The Data Mapping: the rules to retrieve data from the User input Database and map them into data objects suitable for processing are stored here. These rules specify the custom transformation for each input record field from its raw domain to another. For example, a textual input value of a field can be mapped into a numeric value for further processing. This component is also used to exclude fields that are not necessary in building decision tree. For example, fields like submit buttons always contain the same value for every input, so they do not provide any meaning in decision tree building and should be excluded at this step.

The Data objects: these objects are created from the user input records after applied the rules specified in the Data Mapping component. Besides a collection of <*name, value*> pair, each data object contains a class label, which is either normal or abnormal. The class label can be set by security administrators manually or by IDSs automatically.

The Dynamic expression execution: although the attributes contained in each data object can be of any type, e.g. Boolean, integer, real number, character string, etc. not all types are suitable for human analysis. In fact, we argue that choosing which data type/value to present an attribute is more or less a subjective decision. As an example, consider an attribute storing street addresses of buyers in an e-commerce web application, in our opinion, it will be more intuitive to analyze when these addresses are presented as distances to the target shop. But other people may prefer textual addresses to numeric distances and it is not unusual. To support different needs of different people, we develop the Dynamic expression execution component, which is responsible for converting attribute values from one domain to another at run time. The role of this component is somewhat similar to the role of Data Mapping component with one difference: the conversions here are done at run time. As a result, the transformation rules specified in Dynamic expression execution are not as complex as the ones specified in the Data Mapping component, but they can change data object attributes' values in real time. This real time support is important because it helps administrators in experimenting with different expressions to find the most suitable one to present an attribute. Once an expression is chosen, it is not difficult to turn it into a Boolean expression in order to create a split condition with that attribute. We use Java

Expression Library [16] to implement the Dynamic expression execution component.

The Node visualization: this component is responsible for displaying each selected node in detail according to the interface depicted previously in Fig. 2. As shown in Fig. 2, the selected node is visualized as a pie chart presenting the percentages of data objects with normal or abnormal class label. In addition to the pie chart, some bar charts presenting the distribution of the output values achieved by applying the Dynamic expression execution on an attribute are also displayed. The library JFreeChart [17] is used to implement the Node visualization.

The Tree visualization: this component is responsible for displaying the whole decision tree being built. The visualization method is depicted earlier in Fig. 3. As shown in Fig. 3, the relative size of each node corresponds to the number of data objects contained in that node. In turn, each node is visualized as a pie chart, similarly to the way the Node visualization does. Seeing the whole tree as it being built helps administrators in keeping track of their overall progress and also evaluating their current work result. We use JUNG (Java Universal Network/Graph Framework) [18] to organize tree elements, i.e. nodes, edges, layout, etc. as well as visualize it.

V. Experiments

There are two main objectives in our experiments. The first one is to measure the effectiveness of the proposed technique in recognizing attacks on web applications. Because attack detection rules are created by human users, in combination with the support of IVDT, the measured result does not only depend on our technique, but also depend on the skills and experience of the users. Despite this fact, the way users perform their analysis to create detection rules is also affected by the functionality of the provided tool, so the effectiveness we obtain is related to the proposed technique to some extent. In particular, the effectiveness is evaluated as the true positive rate and true negative rate of the resulted visual decision tree in classifying user inputs. The second objective of our experiment is to evaluate the ease or difficulty users have when using IVDT to create attack detection rules. For this objective, we follow a qualitative approach by interviewing the users directly.

A. Data Generation

Without loss of generality, the data being analyzed is supposed to be sent to a sign up action. This action is common on many types of web applications such as e-commerce, bulletin board, online social network, etc. On the input data, there are some fields which are easy to define which values are normal or not. Some examples are fields used to store email address, date of birth, and phone number. On the other hand, it is more difficult to define the patterns of normal values for fields like first name, last name, and password. Because both types of input fields are contained this example, we can see if there is a difference on the way users working on them.

The data objects in our experiments are generated automatically. They are labeled as either normal or abnormal. The abnormal data objects are malicious inputs on the sign up action. In particular, these malicious inputs are composed of SQL injection (SQLi) and Cross-site scripting (XSS) attacks. We choose SQLi and XSS because they are among the most popular attacks on web applications today [19], even though they existed long ago. The data generation processes are described below.

Normal data generation: We use the tool GenerateData [20] to generate normal data. It supports many different types of data such as human data (first name, last name, email, company, etc.), geo data (street address, city, region, latitude/longitude), credit card, text, numeric, etc. We also create a database table to store the artificial generated data. Each record of the database table corresponds to a data object and each table field corresponds to an attribute of the respective data object. There is an additional field used to store the label of the data object. For data objects generated by the GenerateData tool, their labels are assigned normal value.

Abnormal data generation: As described above, abnormal data are SQLi and XSS attacks, so they are generated by another tool. In the experiments, the attack values are created by the HackBar add-on for Mozilla Firefox browser [21]. The main purpose of this tool is to help web developers audit their code and look for security holes. The process we use to generate an abnormal data object is as follow: firstly, we use HackBar to generate an attack value; secondly, we select a random data object stored in the database table; and finally, we select a random field of the selected data object and replace its value with the generated attack value.

The total data objects generated is 1000, in which 800 are normal and 200 are abnormal. We further divide them into two sets: training set and testing set. Each set has the same number of normal and abnormal data objects (i.e. 400 normal and 100 abnormal objects in each set).

B. Experiment Settings

At this stage, our main objective is to get initial feedback about the proposed technique so that we can learn and improve them appropriately. Therefore, these experiments do not involve with many people. Instead, we invite three security practitioners, who are also members at our security lab, to try the prototype and give their opinion about it. These volunteer people all have adequate knowledge and skill about web application security. We record the action steps the security practitioners do during their assigned experiments, with their permission, to analyze further.

Before the security practitioners do their assigned tasks, we give them a brief tutorial about the main components of our prototype and the decision tree classification method. Because the security practitioners already know about SQLi and XSS attacks, these information are not covered in our tutorial. But we do stress the use of the Dynamic expression execution component as well as the visual display to analyze and develop SQLi and XSS attack detection rules. At the end of the tutorial, we tell the volunteers to write attack detection rules as precisely and compactly as possible.

C. Observed Action Steps

The common steps that the security practitioners do according to our observation are:

At first, when there is only the root node, they look through all attributes of the data objects. They do not apply any transformation on any field, but just sees the field's values in their originals. Maybe doing so can give them an overview of the distribution of values of each attribute.

After that, these people try some transformations on some attributes. We find that the most frequently used transformations they apply are length() and indexOf(), which return the length of an input string and the position of an input pattern in another input string respectively. Maybe these transformations are simple but effective enough to detect abnormal values, especially for human readable attributes like "first name" and "last name". These human readable attributes are also chosen before other attributes as split attributes.

Finally, for random attributes like password, it is more difficult for them to decide which value is normal and which one is not. As a result, the transformation rule they create on password field is rather complicated with many string functions combined together using Boolean functions such as AND and OR. Maybe foreseeing this complexity, split conditions for complex attributes are created near or at the leaves of the decision tree only.

D. Results and Discussion

After the security practitioners finish their assigned tasks, we ask them some questions to get their feedback on the usefulness or otherwise of the prototype. We also measure the detection performance of their result decision tree. Therefore, we divide our evaluation into two parts: quantitative result and qualitative result.

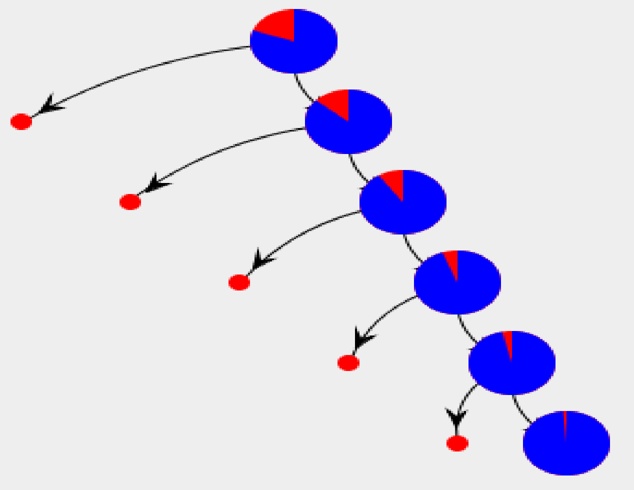


Fig. 5. All of the result trees have this similar form: they are skew toward the normal side.

Quantitative result: it takes around 15 minutes for the security practitioners to completely build attack detection rules in the form of a decision tree. Among the three people, one uses only 11 minutes to finish his task, while another uses 18 minutes. The action steps they do follow similarly to the ones described in the previous section. The height of all the resulted decision tree are 5 and they are rather unbalanced toward the normal class. This can be explained by the approach the security practitioners use to create the split condition for each node, i.e. these conditions focus on the recognition of attack patterns. So, as soon as a data object attribute's value matches such patterns, this data object is considered abnormal and the process can terminate immediately. On the other hand, when a data object attribute's value does not match any attack pattern, the remaining attributes should still be checked. When testing with our generated dataset described previously, the average true positive rate is 95% (92/97) and the average true negative rate is 100% (403/403). The visualization of one of the resulted trees is displayed in Fig. 5.

Qualitative result: we ask the security practitioners what they like or don't like about IVDT. Their responses are similar to what we expected. To summarize, all of them like the way we visualize each node (in Node analysis task) and the whole tree (in Tree analysis task) to support the tree building process. Two of them think the Dynamic expression execution feature is useful because it helps them in trying out different tests at run time, thereby being able to apply their relevant skills and knowledge directly. The volunteers also provide some useful feedbacks to improve IVDT. Two of them suggest that we should complement this tool with a feature to build trees automatically so that they can compare their results with the automatic result version, and/or they can use the automatic result as the base from which to make more refinements manually. One other important suggestion is that our tool should integrate with available web application IDSs or firewalls to reuse their large existing attack detection rules.

VI. Conclusions and Future Works

In this paper, we have proposed and developed a visual interaction technique to support security administrators in building detection rules of attacks on web applications. The technique is based on decision tree learning and is enhanced further by adding data visualization and user interaction supports into the tree building process. Unlike traditional decision tree learning, our technique is user – driven. In other words, human users manually create classification rules, with necessary supports from the IVDT tool. Our proposed technique possesses some advantages over traditional decision tree learning. Some of the most important advantages include:

Firstly, knowledge of users is applied directly. Because users directly analyze data objects and enter split conditions for each node, they can utilize their existing domain knowledge into the rules creation process. This can result in better classification trees.

Secondly, users gain not only result, but also insight about the target web application and its environment. When analyzing data to create split conditions, users do not only create decision trees but also have an overview picture of the data objects, attributes, and values' distribution. This insight is difficult to achieve when using an automatic tool.

Finally, classification rules can be made quite flexibly. Because classification rules are entered by human users, they can be quite flexible. For example, the data attributes used in split conditions can be of any type, not just numeric or nominal types; and custom functions used on data objects' attributes can produce many types of outcomes.

However, the proposed technique still has several issues that need to be addressed in the future.

Firstly, due to some external constraints, we can invite only three security researchers in our lab to join the experiments. Because of the small number of participants, the experiment result may not be representative for the general security administrators. In future experiments, we plan to invite more people with different backgrounds to have more feedbacks. In addition to security experts, we also invite students majored in computer science to join the experiments. The purpose of having computer science students in future experiments is to consider the appropriateness of using this tool for security teaching.

Secondly, because the number of data objects in the experiments is only average, we cannot evaluate the appropriateness of our using IVDT with very big data. For example, when there are many attributes in the training data objects, it is not trivial to select the first attribute to create the split condition. Similarly, when the decision tree being built is very big, displaying it on the computer screen at once in a meaningful way is also a difficult problem to which we will pay more attention in future works.

ACKNOWLEDGMENT

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number C2018-20-12.

REFERENCES

[1] J. R. Quinlan, “Induction of Decision Trees,” Mach. Learn., vol. 1, no. 1, pp. 81–106, 1986.

[2] J. R. Quinlan, “C4.5: Programs for Machine Learning,” Morgan Kaufmann San Mateo Calif., vol. 1, no. 3, p. 302, 1992.

[3] R. A. Breiman, Leo and Friedman, Jerome and Stone, Charles J and Olshen, “Classification and regression trees,” 1984.

[4] M. Ankerst, C. Elsen, M. Ester, and H.-P. Kriegel, “Visual classification: an interactive approach to decision tree construction,” in Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’99, 1999, pp. 392–396.

[5] M. Ankerst, D. Keim, and H. Kriegel, “‘Circle Segments’: A Technique for Visually Exploring Large Multidimensional Data Sets,” in Proc. IEEE Visualization ’96, Hot Topic Session, 1996, pp. 5–8.

[6] S. T. Teoh and K.-L. Ma, “PaintingClass: interactive construction, visualization and exploration of decision trees,” in Star, 2003, pp. 667–672.

[7] A. Inselberg, “The plane with parallel coordinates,” Vis. Comput., vol. 1, no. 4, pp. 69–91, 1985.

[8] S. van den Elzen and J. J. van Wijk, “BaobabView: Interactive construction and analysis of decision trees,” in IEEE S. Vis. Anal., 2011, pp. 151–160.

[9] H. Choi, H. Lee, and H. Kim, “Fast detection and visualization of network attacks on parallel coordinates,” Comput. Secur., 2009.

[10] A. Heitzmann, B. Palazzi, C. Papamanthou, and R. Tamassia, “Effective visualization of file system access-control,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2008, vol. 5210 LNCS, pp. 18–25.

[11] B. Johnson and B. Shneiderman, “Tree-Maps: S Space Filling Approach to the Visualization of Hierarchical Information Structures,” 1991, pp. 284–291.

[12] T. K. Dang and T. T. Dang, “A survey on security visualization techniques for web information systems,” Int. J. Web Inf. Syst., vol. 9, no. 1, 2013.

[13] T. T. Dang and T. K. Dang, “Visualization of web form submissions for security analysis,” Int. J. Web Inf. Syst., vol. 9, no. 2, pp. 165–180, 2013.

[14] A. Treisman, “Preattentive processing in vision,” Comput. Vision, Graph. Image Process., vol. 31, no. 2, pp. 156–177, 1985.

[15] D. Scott and R. Sharp, “Abstracting application-level web security,” in Proceedings of the eleventh international conference on World Wide Web - WWW ’02, 2002, p. 396.

[16] K. Metlov, “Java Expressions Library.” [Online]. Available: http://www.gnu.org/software/jel. [Accessed: 02-Oct-2017].

[17] D. Gilbert, “JFreeChart.” [Online]. Available: http://www.jfree.org/jfreechart/. [Accessed: 02-Oct-2017].

[18] J. Team, “Java Universal Network/Graph Framework.” [Online]. Available: http://jung.sourceforge.net/. [Accessed: 02-Oct-2017].

[19] OWASP, “OWASP Top 10 - 2017.” [Online]. Available: https://www.owasp.org/index.php/Top_10-2017_Top_10. [Accessed: 10-Jan-2018].

[20] GenerateData, “Generate Data.”. [Online]. Available: http://generatedata.com/. [Accessed: 10-Jan-2018].

[21] J. Adriaans and P. Laguna, “HackBar.” [Online]. Available: https://addons.mozilla.org/en-US/firefox/addon/hackbar/. [Accessed: 10-Jan-2018.

Developing a Candidate Registration System for Zambia School Examinations using the Cloud Model

Banji Milumbe, Jackson Phiri, Monica M Kalumbilo, Mayumbo Nyirenda
Department of Computer Science
The University of Zambia
Lusaka, Zambia

***Abstract*—Cloud computing has in the recent past gained a lot of ground in this digital age. The use of cloud technologies in business has broken barriers in sharing information making the world one big global village. Regardless of where one is, data or information can be received or sent instantly disregarding distance. In this research, we investigated the challenges in registering candidates for school examinations and availability of internet services in various parts of Zambia and then present a candidate registration process based on the cloud model which is aimed at resolving challenges of distances from examination centres to the examining body in order to register for examinations as well as improving the timelines and cutting down the back and forth movements in the whole process. The web based registration system was developed and tested and the testing ascertained connectivity, functionality and scalability of the system.**

Keywords—Cloud computing; candidate registration; online registration; Zambia; school examinations; bulk SMS; automation; information communication technology; ICT

I. Introduction

In this digital age, the advancements in technology have changed the way businesses operate. The proper, planned use of ICT can be highly beneficial but the fact of using ICT does not of itself automatically bring benefits [1]. The internet has made work easier in that one can access different facilities anywhere anytime, opening up numerous possibilities for doing business at a local and global level. Applications are being developed that enable information to be received in real-time regardless of the user's location. The availability of candidates information in good time would help examining bodies that conduct public examinations to adequately prepare accurate examination materials and reduce on errors.

The Examinations Council of Zambia (ECZ) created by an Act of Parliament Number 15 of 1983 whose main purpose was to set and conduct examinations and award certificates to successful candidates in Zambia. Its full launch and operationalisation was in 1987 as a semi-autonomous public institution [2]-[4]. Ten (10) years later, ECZ embarked on automating systems and localised the results processing as well as registration of candidates for school examinations. Different methods of getting candidates data for examinations were used starting from manual systems to some electronic systems which proved to have posed challenges in the accuracy, efficiency and effectiveness of such systems [5]. In the latest candidate registration method described by [6] was used by ECZ where examination centres were every year be provided with a CD containing the desktop candidate registration application. All Centres with prospective candidates for examinations would enter the candidate details and then create a text file of candidate details in a predetermined format which was later submitted to ECZ. This research investigated the candidate registration for school examinations in Zambia in order to develop a candidate registration system that utilises cloud services.

II. Literature Review

A. Introduction

The internet has become an invaluable and integral part of business and personal life in the modern world [1]. The rapid development of processing and storage technologies and success of the internet, have made computing resources become cheaper, more powerful and more available than ever before. This technological trend has enabled the realisation of a new computing model called cloud computing, in which resources are provided as general utilities that can be leased and released by users through the internet in an on demand style [7]. Internet has created a technology innovation, a new digital market place, rendering the need for centralised cloud service unavoidable [8]-[12]. While the enterprise begins to embrace the internet of things via the ability to communicate more digitally, the promises of business improvement at a reduced shared cost is leaking quietly.

Many sectors like governmental, non-governmental, profit and non-profit organisations have taken advantage of these technological developments to improve the way business is conducted and be up to speed with the rest of the world. In the education sector, many technologies have been adopted to enhance learning, teaching and also in the area of capturing and storing student information.

B. ICTs in the Education Sector

The prospects for the utilisation of new technologies in the field of education continue to be part of the transformations in the education sector with a strong bearing on the assessment and evaluation of the education system in public examinations. In a study conducted by Sibanda and Maposa, 2013 it was established that ICT can be utilised as an integral component to improve efficiency, effectiveness and excellence in learning, teaching and assessment [13]. Automated systems as alluded to by Obioma et al., 2013 offer some benefits in the education sector such as; lower long-term costs, instant feedback to students, creation of digital records of student growth and development, greater storage efficiency and increased productivity and low operational variability [14]. Employing new technology in any project implies certain inherent risks, so an adequate technology management is a precondition for a successful software development project [15].

C. Cloud Computing

Computing is being transformed to a model consisting of services that are commoditised and delivered in a manner similar to utilities such as water, electricity, gas, and telephony. In such a model, users access services based on their requirements regardless of where the services are hosted. Several computing paradigms have promised to deliver this utility computing vision [16]. Cloud computing has been defined by its characteristics by Buyya as follows, ‘Cloud computing is a parallel and distributed system consisting of a collection of inter-connected and virtualised computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements (SLA) established through negotiation between the service provider and consumers’ [7].

Cloud computing has the potential to dramatically change business models and the way people interact with each other because it provides access to large-scale remote resources in a very efficient and quick manner. It has the potential to level the playing field because it breaks barriers to entry [17]. Cloud computing is thought to be the solution to overcome the problem of processing large amounts of data. By using cloud computing the cost of implementing software solutions and storage of data is reduced significantly [18]. Using cloud based storage for large amounts of data is the key [19].

D. Web Applications Integrated with Bulk Short Message Service (SMS)

The wide use of mobile telecommunications has also brought about the integration of web based systems with mobile telecommunications especially with the GSM being the most successful digital mobile telecommunications used by millions of people in various countries in the world [20].

Use of mobile phone helps to have access to the system or receive alerts from the systems via mobile phone even when you are not connected to the internet. A useful service for very simple message transfer is the short message service (SMS), which can be used for “serious” applications as noted by [20].

A study conducted by [21] where use of SMS/USSD was proposed proved to be faster and more reliable in disseminating information on examination registration and results to candidates than the traditional computer.

III. Methodology and Scope

This study was conducted in all the ten (10) provinces of Zambia comprising provincial and district education offices and schools out of which 75% were secondary schools and 25% primary schools. The rationale behind having this proportion of schools was that secondary schools conduct up to three (3) different examinations per year while primary schools mainly conduct only one examination in a year. Non-probability sampling technique was used in this study for the 122 study participants by virtue of their specialised knowledge in the subject area.

The purpose of the baseline study was to establish the challenges with the registration of candidates and availability of internet services in the provinces. The questionnaires and interview guide were used to gather the data for this research. The results from the study were used as input or part of requirements gathering for the web based registration system. It also helped to ascertain the feasibility of implementing a web based registration solution in all examination registration centers in Zambia.

A. System Automation

Results of the baseline study and the regulations on registration for examinations in Zambia [22] were used to design a registration model based on cloud technologies.

The Web based candidate registration system that utilises the cloud model and integrates bulk SMS as well as barcode technology is shown in Fig. 1. The registration takes place at the examination centre after receiving notification through the integrated bulk SMS system that registration had begun. The Province and District users monitor the registration process by logging in the system, they are being able to view and generate statistical reports on the candidates registered. The integrated barcode technology in the system is to issue a card with a barcode containing candidate details which can be used during examinations and any other subsequent registrations for examinations.

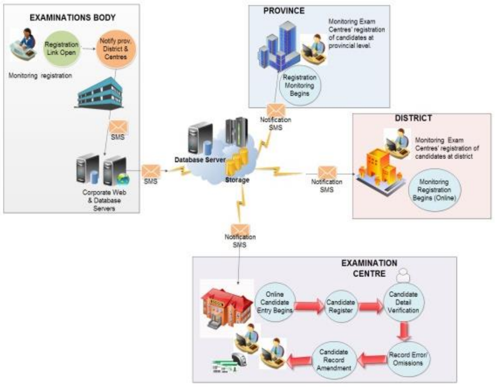

EXAMINATIONS BODY
Registration Link Open
Notify prov. District & Centres
Monitoring registration
SMS
Corporate Web & Database Servers
SMS
Database Server
Storage
PROVINCE
Monitoring Exam Centres' registration of candidates at provincial level.
Registration Monitoring Begins
Notification SMS
DISTRICT
Monitoring Exam Centres' registration of candidates at district
Notification SMS
Monitoring Registration Begins (Online)
Notification SMS
EXAMINATION CENTRE
Online Candidate Entry Begins
Candidate Register
Candidate Detail Verification
Record Error/ Omissions
Candidate Record Amendment

Fig. 1. Web based candidate registration model.

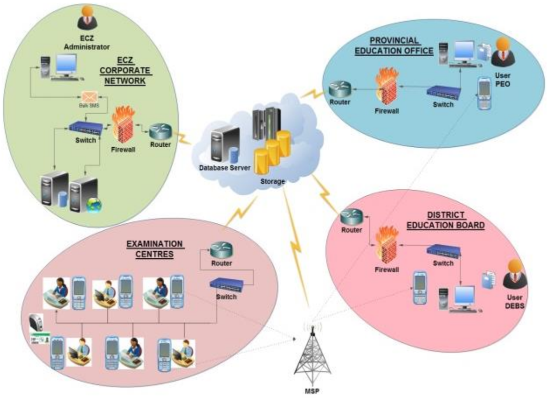


Fig. 2. Web based candidate registration system for ECZ.

B. System Architecture

The diagrammatic representation of the system architecture of the Web based Candidate Registration System (WCRS) for ECZ is shown in Fig. 2.

The architecture has the following components comprising the ECZ corporate network where the administrator opens the registration link and bulk SMS sent to all provinces, districts and Examination centre coordinators to inform them that the registration had opened. The other local networks are at the provincial education office, District Education board secretary's office and the examination centres, the cloud service constituting database storage and user application. The local service constituting database storage and user application as a backup measure is located at ECZ HQ. The Mobile Service Provider which is integrated with the web based application for sending and receiving the SMS is included.

As shown in Fig. 2, the system administrator at ECZ sends an SMS to all concerned parties in the registration process so that they could begin the registration. The web based candidate registration system is accessed using a web browser. The local backup server also exists as a backup measure in case of failures in the cloud. The local and cloud servers constantly synchronize to ensure data integrity and completeness

C. System Modelling and Design

The Unified Modelling Language (UML) was used to come up with Use case diagram. The use case diagram was developed based on the information gathered to incorporate the main actors in the candidate registration system like the system administrator and the user (Guidance Teacher). Fig. 3 shows the different actions that the System Administrator does with the system.

D. Sequence Diagram

The sequence diagram in Fig. 4 shows the business processes for the guidance teacher (user) in the system.

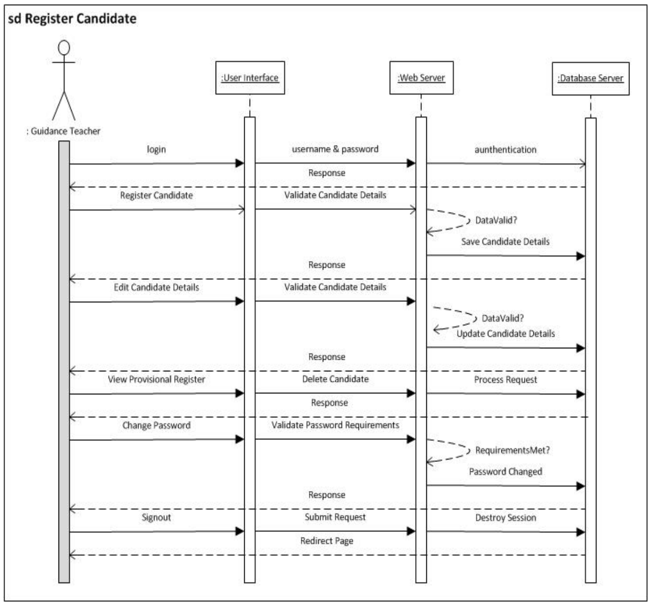


Fig. 3. System administrator activities use case diagrams.

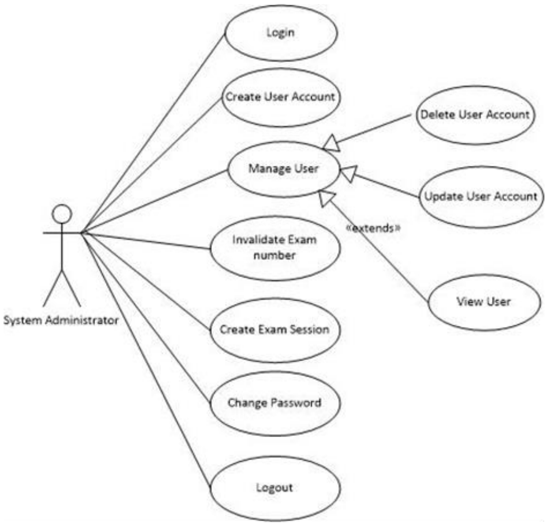


Fig. 4. Guidance teacher (user) sequence diagram.

IV. RESULTS

The focus of the study was to also establish that there was adequate infrastructure to deploy a web based candidate registration system in the various locations across the country. The results presented are derived from the responses by the respondents after data analysis.

1) Baseline study

The data collected was analysed and the results of the baseline study are presented in this section using charts.

a) Internet Accessibility

The research findings confirmed that internet was available and accessible through different means as indicated in Fig. 5. The following were the ways in which the 122 respondents access internet. Eighty (80) percent access internet through their personal mobile devices, like phones, tablets or dongles, 17 per cent said they accessed internet from their work place which was provided by their employers, while 1 per cent said that they access internet through public wifi like iconnect, i-Zone or wifi at shopping malls and 2 per cent said they accessed it through internet cafes. This means that people have access to internet services.

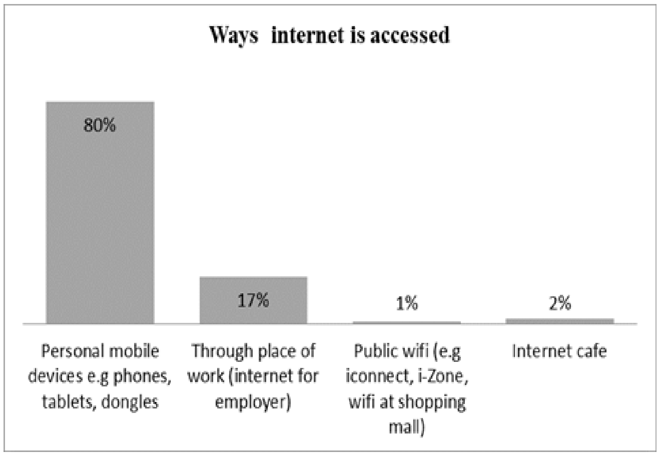


Fig. 5. Ways internet is accessed.

b) Availability of Mobile phone service

The participants also indicated the mobile network available in their area. The pie chart in Fig. 6 shows that there is at least one mobile service provider in the area making it possible to for them to have access to internet services via the mobile service providers. Information about the availability of mobile service providers was important because most of the access to internet is through these mobile service providers.

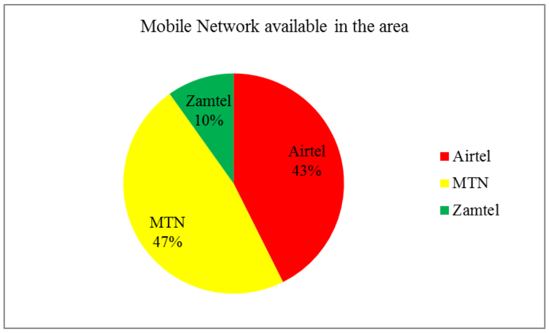


Fig. 6. Mobile service provider available.

c) Benefits of a web based candidate registration application

Benefits of a web based registration system were established and from the findings, it was clear that a web based candidate registration system was highly recommended. The reasons why respondents recommend a web-based system are indicated in Fig. 7 such as it being faster, efficient and effective, very convenient and many others.

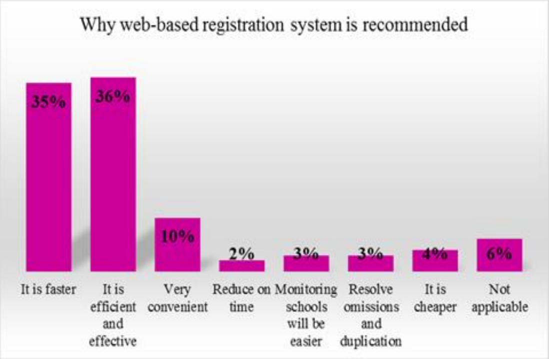


Fig. 7. Perceived benefits of a web application.

V. System Development and Testing

The system was successfully developed and tested.

A. System Development

A representation of some screen shots of the developed system is shown.

1) Administrator login and user creation: The system administrator is responsible for managing users in the system. For a user to be able to access the web based system modules, they must be registered in the system. All the data entered into the system should be traced to the user that entered. The administrator log in shown in Fig. 8 is to enable create a new user account in the system.

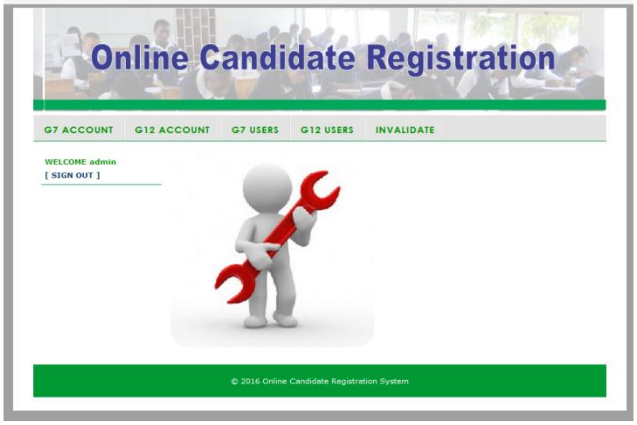
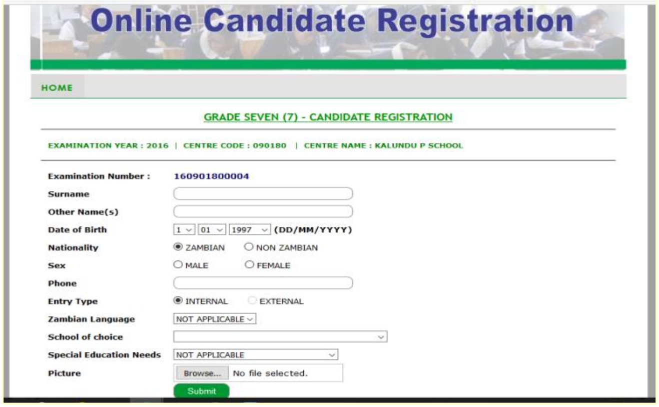


Fig. 8. Administrator sign in home page.

2) Registering a candidate: To register a candidate, you login to the appropriate level or grade and select the correct centre details. A candidate's examination number is automatically generated and the user can enter all the candidate details as shown in Fig. 9. Once the details have been entered, proceed to select subjects and then complete registration as shown in Fig. 10.

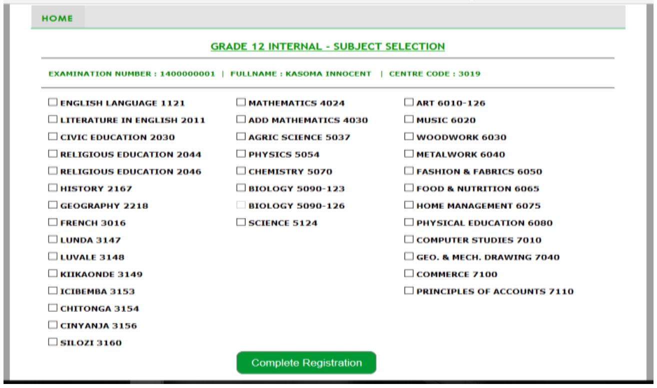


GRADE SEVEN (7) - CANDIDATE REGISTRATION

EXAMINATION YEAR : 2016 | CENTRE CODE : 090180 | CENTRE NAME : KALUNDU P SCHOOL

Examination Number :	160901800004
Surname	
Other Name(s)	
Date of Birth	1 / 01 / 1997 (DD/MM/YYYY)
Nationality	● ZAMBIAN ○ NON ZAMBIAN
Sex	○ MALE ○ FEMALE
Phone	
Entry Type	● INTERNAL ○ EXTERNAL
Zambian Language	NOT APPLICABLE
School of choice	
Special Education Needs	NOT APPLICABLE
Picture	Browse... No file selected.

Fig. 9. Candidate entry details screen.



GRADE 12 INTERNAL - SUBJECT SELECTION

EXAMINATION NUMBER : 1400000001 | FULLNAME : KASOMA INNOCENT | CENTRE CODE : 3019

□ ENGLISH LANGUAGE 1121
□ LITERATURE IN ENGLISH 2011
□ CIVIC EDUCATION 2030
□ RELIGIOUS EDUCATION 2044
□ RELIGIOUS EDUCATION 2046
□ HISTORY 2167
□ GEOGRAPHY 2218
□ FRENCH 3016
□ LUNDA 3147
□ LUVALE 3148
□ KIIKAONDE 3149
□ ICIBEMBA 3153
□ CHITONGA 3154
□ CINYANJA 3156
□ SILOZI 3160
□ MATHEMATICS 4024
□ ADD MATHEMATICS 4030
□ AGRIC SCIENCE 5037
□ PHYSICS 5054
□ CHEMISTRY 5070
□ BIOLOGY 5090-123
□ BIOLOGY 5090-126
□ SCIENCE 5124
□ ART 6010-126
□ MUSIC 6020
□ WOODWORK 6030
□ METALWORK 6040
□ FASHION & FABRICS 6050
□ FOOD & NUTRITION 6065
□ HOME MANAGEMENT 6075
□ PHYSICAL EDUCATION 6080
□ COMPUTER STUDIES 7010
□ GEO. & MECH. DRAWING 7040
□ COMMERCE 7100
□ PRINCIPLES OF ACCOUNTS 7110

Fig. 10. Subjects selection.

3) Reports menu: The web based candidate registration system has several reports that can be generated by the user as shown in Fig. 11.

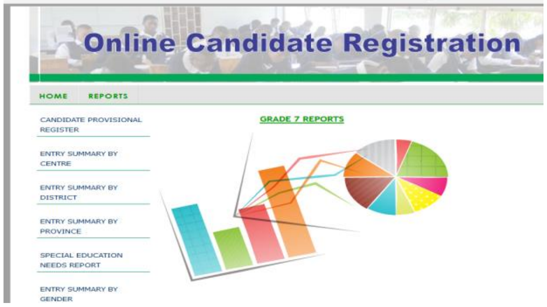

Online Candidate Registration
HOME
REPORTS
CANDIDATE PROVISIONAL REGISTER
ENTRY SUMMARY BY CENTRE
ENTRY SUMMARY BY DISTRICT
ENTRY SUMMARY BY PROVINCE
SPECIAL EDUCATION NEEDS REPORT
ENTRY SUMMARY BY GENDER
GRADE 7 REPORTS

Fig. 11. Reports selection.

B. *System Testing*

The testing of the web based candidate registration system was undertaken to verify connectivity to the application, functionality and scalability.

1) Test sample: The test sample comprised 40 schools, two (2) from each selected district from the ten (10) provinces. Of the schools and districts selected, there was a balance between rural and urban districts and schools. Fifty (50) candidates were selected in each of the schools to be registered using the web based candidate registration system.

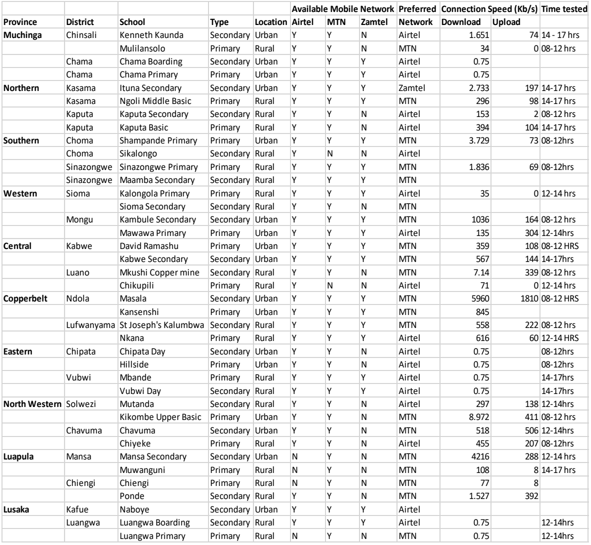
2) Testing results: Out of 40 schools selected for testing the system, 39 schools were used as test centres and all the 39 could access the web application, able to update candidate details and tested the various reports. Thirty six out of thirty-nine (92%) schools successfully registered all the 50 candidates sampled while three centres could not register all the 50 due to wrong examination numbers of the sampled candidates. Thus, the system could not retrieve candidate details to proceed with registration. This was part of testing the system as well to those conditions /rules as successfully applied.

3) Tools used for testing: During the system testing, laptops, mobile devices like tablets or mobile phones and mobile wireless 3G Routers were used in testing for the internet connectivity. The testing relied solely on the mobile phone service providers available in each area/school for internet connectivity. Even though the school had its own internet connection, it was not used because the testing wanted to establish availability of internet connectivity so that even schools that did not have internet of their own could easily purchase similar routers/modems and be able to access the web based registration system to register candidates.

4) Testing internet connectivity: The internet connection speed was tested at each of the schools where the pilot was conducted using the online website speed tester, www.speed.io. The download and upload speed was recorded as it showed under measured data of the online speed tester. The download speed ranged from 1.527 kb/s to 8972 kb/s while the upload speed ranged from 2kb/s to 506 kb/s. The time of testing the connection speed also had a bearing on the connection speed as at certain times some places were congested while in other places there was little congestion and the internet connection speed was very fast.

Fig. 12 shows that internet connectivity was available in all the schools that were used as sites for system testing both rural and urban schools. In each of these schools there was at least one mobile network provider available and internet access was possible using the 3G wireless modems for internet service. It should be noted though that in one of the centres, internet could only be accessed at a particular point which was under a guava tree. Despite that however, registration of candidates using the web based application system was successful.

5) System Performance and responsive: It was noted that the system response during the testing was good in most of the schools which accounted for 95 percent while 5 percent said it was fair. None of the test sites recorded poor system response which showed that the web based registration system generally performed very well.



					Available Mobile Network			Preferred	Connection Speed (Kb/s)		Time tested
Province	**District**	**School**	**Type**	**Location**	**Airtel**	**MTN**	**Zamtel**	**Network**	**Download**	**Upload**	
Muchinga	Chinsali	Kenneth Kaunda	Secondary	Urban	Y	Y	N	Airtel	1.651	74	14 - 17 hrs
		Mulilansolo	Primary	Rural	Y	Y	N	MTN	34	0	08-12 hrs
	Chama	Chama Boarding	Secondary	Urban	Y	Y	Y	Airtel	0.75		
	Chama	Chama Primary	Primary	Urban	Y	Y	Y	Airtel	0.75		
Northern	Kasama	Ituna Secondary	Secondary	Urban	Y	Y	Y	Zamtel	2.733	197	14-17 hrs
	Kasama	Ngoli Middle Basic	Primary	Rural	Y	Y	Y	MTN	296	98	14-17 hrs
	Kaputa	Kaputa Secondary	Secondary	Rural	Y	Y	N	Airtel	153	2	08-12 hrs
	Kaputa	Kaputa Basic	Primary	Rural	Y	Y	N	Airtel	394	104	14-17 hrs
Southern	Choma	Shampande Primary	Primary	Urban	Y	Y	Y	MTN	3.729	73	08-12hrs
	Choma	Sikalongo	Secondary	Rural	Y	N	N	Airtel			
	Sinazongwe	Sinazongwe Primary	Primary	Rural	Y	Y	Y	MTN	1.836	69	08-12hrs
	Sinazongwe	Maamba Secondary	Secondary	Rural	Y	Y	Y	MTN			
Western	Sioma	Kalongola Primary	Primary	Rural	Y	Y	Y	Airtel	35	0	12-14 hrs
		Sioma Secondary	Secondary	Rural	Y	Y	N	MTN			
	Mongu	Kambule Secondary	Secondary	Urban	Y	Y	Y	MTN	1036	164	08-12 hrs
		Mawawa Primary	Primary	Urban	Y	Y	Y	Airtel	135	304	12-14hrs
Central	Kabwe	David Ramashu	Primary	Urban	Y	Y	Y	MTN	359	108	08-12 HRS
		Kabwe Secondary	Secondary	Urban	Y	Y	Y	MTN	567	144	14-17hrs
	Luano	Mkushi Copper mine	Secondary	Rural	Y	Y	N	MTN	7.14	339	08-12 hrs
		Chikupili	Primary	Rural	Y	N	N	Airtel	71	0	12-14 hrs
Copperbelt	Ndola	Masala	Secondary	Urban	Y	Y	Y	MTN	5960	1810	08-12 HRS
		Kansenshi	Primary	Urban	Y	Y	Y	MTN	845		
	Lufwanyama	St Joseph's Kalumbwa	Secondary	Rural	Y	Y	Y	MTN	558	222	08-12 hrs
		Nkana	Primary	Rural	Y	Y	Y	Airtel	616	60	12-14 HRS
Eastern	Chipata	Chipata Day	Secondary	Urban	Y	Y	N	Airtel	0.75		08-12hrs
		Hillside	Primary	Urban	Y	Y	N	Airtel	0.75		08-12hrs
	Vubwi	Mbande	Primary	Rural	Y	Y	Y	Airtel	0.75		14-17hrs
		Vubwi Day	Secondary	Rural	Y	Y	Y	Airtel	0.75		14-17hrs
North Western	Solwezi	Mutanda	Secondary	Rural	Y	Y	N	Airtel	297	138	12-14hrs
		Kikombe Upper Basic	Primary	Urban	Y	Y	N	MTN	8.972	411	08-12 hrs
	Chavuma	Chavuma	Secondary	Urban	Y	Y	N	MTN	518	506	12-14hrs
		Chiyeke	Primary	Rural	Y	Y	N	Airtel	455	207	08-12hrs
Luapula	Mansa	Mansa Secondary	Secondary	Urban	N	Y	N	MTN	4216	288	12-14 hrs
		Muwanguni	Primary	Rural	N	Y	N	MTN	108	8	14-17 hrs
	Chiengi	Chiengi	Primary	Rural	N	Y	N	MTN	77	8	
		Ponde	Secondary	Rural	Y	Y	N	MTN	1.527	392	
Lusaka	Kafue	Naboye	Secondary	Urban	Y	Y	Y	Airtel			
	Luangwa	Luangwa Boarding	Secondary	Rural	Y	Y	Y	Airtel	0.75		12-14hrs
		Luangwa Primary	Primary	Rural	N	Y	N	MTN	0.75		12-14hrs

Fig. 12. Summary results of connectivity test.

VI. Conclusion and Future Scope

The study brought out important points on the availability and use of internet in various parts of the country, both rural and urban areas. This helped to ascertain that the web based candidate registration would be used without much difficulty in all schools to register candidates for school examinations.

The system test results validated the web based registration system and the next stage was to fully develop the system and use it at full scale for registration of candidates for school examinations in Zambia. The proposed web system would cut down on some unnecessary processes thereby reducing the time it takes to complete the candidate registration process.

The system only has one level of user authentication. The second level user authentication for using biometrics should be for future inclusion in addition to the use of CCTV in the registration rooms and GPS/GIS Location to enhance the security of the system in the cloud. An alternative fully fledged mobile application for popular and affordable mobile devices such as those that use the Android operating system should be provided in future.

ACKNOWLEDGMENT

We are grateful to the Director, Examinations Council of Zambia for allowing us to carry out this research at the Institution, ECZ employees and the Provincial Education Offices for assisting us in collecting data from Schools and Districts.

REFERENCES

[1] T. Lucey, Management Information Systems, 9th ed., London: BookPower, 2005.

[2] ECZ, “Examinations Council of Zambia Strategic Plan for 2009 to 2015,” Examinations Council of Zambia, Lusaka, 2009.

[3] Zambia, Laws of Zambia. Examinations Council of Zambia Act Cap 137 of 1983. Lusaka, Lusaka: Zambian Parliament, 1983.

[4] M. J. Kelly, “Education in a Declining Economy,” World Bank, Washington DC, 1991.

[5] Banji M. Shakubanza and Joe Kanyika, “Innovative Technologies in Enhancing Performance in the Conduct of Public Examinations: A Reflection on the Examinations Council of Zambia,” Journal of the Association for Educational Assessment in Africa, vol. 5, pp. 163-175, 2011.

[6] Banji Milumbe, “Automation of the Candidate Registration for School Examinations in Zambia using the Cloud Model,” in Proceedings of the IEEE International Conference in Information and Communication Technologies (ICICT)”, Lusaka, pp. 108-115, 2017.

[7] M.G. Avram Olaru, “Advantages and challenges of adopting cloud computing from an enterprise perspective,” Procedia Technology, vol. 12, pp. 529-534, 2014.

[8] T. U. Daim, “Exploring technology acceptance for online food services,” International Journal of Business Information Systems, vol. 12, no. 4, pp. 383-403, 2013.

[9] A. Antonov, “New business-oriented global/regional information network,” International Journal of Business Information Systems, vol. 12, no. 3, pp. 321-334, 2013.

[10] B. Gannon, “Outsiders: an exploratory history of IS in corporation,” Journal of Information Technology, vol. 28, no. 1, pp. 50-62, 2013.

[11] W. Venters and E. A. Whitley, “A critical review of cloud computing: researching desires and realities,” Journal of Information Technology, vol. 20, no. 1, pp. 113-126, 2012.

[12] A. Abareshi, W. Martin and A. Molla, “The role of information and communication technologies in moving toward new forms of organising,” International Journal of Business Information Systems, vol. 9, no. 2, pp. 169-188, 2012.

[13] F. Sibanda and R. S. Maposa, “The Ethics of ICT Assessment in Public Examinations: Reflections on the Zimbabwean Experience,” International Journal of Academic Research in Progressive Education and Development, vol. 2, no. 1, pp. 94-99, January 2013.

[14] G. Obioma, I. Junaidu and G. Ajagun, “The Automation of Educational Assessment in Nigeria: Challenges and implications for pre-service Teacher Education,” in Paper presented at the Annual Conference of the International Association for Educational Assessment, Tel-Aviv, Israel, 2013.

[15] P. Mangudo, M. Arroqui, C. Marcos and C. Machado, “Rescue of a whole-farm system: crystal clear in action,” International Journal of Agile and Extreme Sofware Development, vol. 1, no. 1, pp. 6-22, 2012.

[16] R. Buyya, “Introduction to the IEEE Transactions on Cloud Computing,” IEEE Transactions on Cloud Computing, vol. 1, no. 1, pp. 3-21, January-June 2013.

[17] S. Greengard, “Cloud Computing and Developing Nations,” Communications of the ACM, vol. 53, no. 5, pp. 18-20, May 2010.

[18] A. A. Tole, “Cloud Computing and Business Intelligence,” Database Systems Journal, vol. 5, no. 4, pp. 49-58, 2014.

[19] Mulima Chibuye and Jackson Phiri, “A Remote Sensor Network using Android Things and Cloud Computing for the Food Reserve Agency in Zambia,” International Journal of Advanced Computer Science and Applications, vol. 8, no.11, pp. 411-418, 2017.

[20] J. Schiller, Mobile Communications, 2nd ed., London: Addison Wesley, 2003.

[21] Lovemore Solomon and Jackson Phiri, “Enhancing the Administration of National Examinations using Mobile Cloud Technologies: A Case of Malawi National Examinations Board,” International Journal of Advanced Computer Science and Applications, vol. 8, no.9, pp. 294-305, 2017.

[22] ECZ, ‘Guidelines on the Administration and Management of Examinations in Zambia’, Examinations Council of Zambia, Lusaka, 2015.

Deep Learning Features Fusion with Classical Image Features for Image Access

Rehan Ullah Khan
Information Technology Department
Qassim University,
Al-Qassim, KSA

***Abstract*—Depending on the society, the access to the adult content can create social problems. This paper thus proposes a fusion approach for image based adult content filtering. The proposed approach merges the Deep Learning (DL) architecture and classical hand crafted feature extraction approaches. From the DL, we fuse the rich feature extraction capabilities of the Convolutional Neural Networks (CNNs) with the Correlograms features. We optimize the classification by integrating and modifying the Correlograms into skin Correlograms. The results show an increased performance by combining the DL learnt features with the classical hand crafted features. From an evaluation, the proposed approach achieves an Accuracy of 0.93. This work thus motivates the usage of classical hand crafted features to be exploited in the DL architectures for segmentation and detection scenarios.**

***Keywords*—Deep learning; content based filtering; content analysis; machine learning; support vector machines**

I. INTRODUCTION

The availability of the explicit adult content on the Internet is increasing rapidly. One of the main problems is that the accessibility of these media resources is becoming easy due to number of available solutions and the internet availability. This opens risks in terms of many social factors, and for many societies, accessibility of such content is crime. The most feared element, however, nowadays is the availability of these media resources to the children. This article thus targets such concerns in the form proposing a solution to media filtering using the fusion of DL and classical image features.

Convolutional Neural Networks (CNNs) have demonstrated its usefulness in the field of Computer Vision (CV) tasks of object detection and classification [1]. A shift has been observed from the hand crafted feature extraction to automated model learning [2], [3] from images. In this article, we investigate media filtering using the DL and classical feature extraction. Classical approaches require hand crafted feature extraction from given set of images. From the proposed approach of media filtering, we find that though DL provides good overall performance, a well-crafted feature set can augment DL approaches and increase detection performance.

We propose an approach for content filtering that flags the image as adult nature or safe image. Our approach exploit two things: Firstly, it uses the DL architecture to learn the rich set of features; secondly, it merges the innovative Deep Learning (DL) feature set and the classical feature extraction methods. We fuse the CNN based features with the Correlograms features. The classical feature set represents the skin color based Correlograms. We optimize the classification by integrating and modifying the correlation grams into skin Correlograms. The results show that it is useful to combine deep learnt features with the classical hand crafted features and achieve good overall performance. We get an increase of 5% detection performance by combining the DL features with the manual feature set. With this setup, we expect an already rich set of features to be exploited in the DL architectures for segmentation and detections.

The efforts to detect and possibly block explicit adult content are not new and there is interesting work regarding adult content detection in the state-of-the-art [4]-[7]. The work in [4] fuses the two DL approaches of AlexNet [3] and GoogLeNet [8] and reports that performance can be increased by fusing the two networks for adult content classification. The work in [6] targets skin locus detection for content filtering using the 24 colors transformations in widely available images and videos. The article [9] presents an evidence combination that includes video sequences, key-shots, and key-frames. The work in [10] is based on the adaptive sampled based analysis showing the usefulness of proposed method in the detection of minor pornographic sequences with 87% detection rate. The approach in [6] employs color based skin filtering and content based filtering based on the skin detection. Lopes et al. [5] uses the text retrieval approach for content filtering and is analyzed and evaluated using datasets, achieving a 93.2% detection rate. The authors in [11] demonstrate a framework to analyze the websites. The framework produces an augmented classification that is independent of the access scenarios. In [12], the authors combine key-frame based approaches with a MPEG-4 statistical analysis of the flow vectors. The work of [13] develops filtering for a Web-based P2P. Authors in [14] use two visual features for media access and filtering. First is the single frame, and the other visual based feature is the decision variable of multiple frames with the Discriminant analysis for optimization. Image filtering is actually similar to content based retrieval. The articles [15]-[18] discuss content image retrieval in detail. The work in [19] uses the Hue-SIFT for nude and explicit content detection achieving better results compared to the SIFT. The approach in [20] takes advantage of the motion flow vectors combining them with the audio features for filtering. Lee et al. [21] propose multi-modal approach comprising of three phases; hashing, real-time detection, and finally group of frames decision. The proposed approach of [22] employs optical flow for filtering and detection, achieving an acceptable 80% detection performance. Authors in [23] have similar approach of the motion estimation to the approach of [22] for media filtering. In [24], the authors present a fusion of audio features and video features based on the SVM classifier. The work in [25] uses spatial features and time-based features for content filtering with an accuracy of 94.4%.

II. Convolution Neural Networks (CNNs)

CNNs have proved to be a very useful and innovative tool of DL to learn image feature set and inherent relationship in low level features to higher level objects in images. The generic architecture of CNN contains interconnected layers and consists of repeated convolutional blocks, Rectified Linear Units (ReLU) and Pooling layers [3]. Convolutional layers perform convolution of the input with set of filters. The filters are then learnt during training. The non-linear behavior in data is modelled by the ReLU layer [3]. The pooling layers samples the input and consolidate image classes.

III. Proposed Deep Architecture

Fig. 1 shows the proposed fusion DL architecture. If fusion is not integrated, then the architecture is similar to the one proposed in [3]. However, we augment this architecture with fusion. The structure of the proposed fusion DL for feature extraction and classification of images is composed of different layers. Our DL architecture contains five Convolution layers. The Convolution layers are followed by three fully connected layers. Each layer uses kernel to filter its two dimension inputs. The coefficients for the kernel are calculated incrementally from the training process. The dot product operation is performed by the fully connected layers between the input and weights vectors. In this connected setup, each neuron is connected to all outputs. For learning, and expediting the learning process, each layer uses the ReLU. The Softmax layer gets its input from the last fully connected layer and thus produces probabilistic distribution for a bi-class problem of "adult" and "non-adult" image. Fig. 1 shows the proposed architecture to enhance and fuse the features from the color based Correlograms. SVM is used for classification after feature from DL and Correlograms are calculated.

IV. Experimental Setup and Results

L-norm distance between the two pixels and corresponding histogram is calculated as:

$$\text{L-norm-hist}= \mathbf{n}^2\boldsymbol{\pi}\mathbf{P}_{(p\ \epsilon\ \tilde{\imath})}[\mathbf{p}\ \epsilon\ \tilde{\imath}_{ci}] \qquad - \qquad (1)$$

$\tilde{\imath}$ is an image with color quantized as $c_1 \ldots c_m$. The π represents the product operation. We define the skin Correlograms as follows:

$$\mathrm{Cor}_{\_Skin}=\mathbf{P}(p_1\ \epsilon\ \tilde{\imath}p_{\ c},p_2\ \epsilon\ \tilde{\imath}p)\ [p_2\ \epsilon\ \tilde{\imath}p_{\ c}\,|\ |\mathbf{p_r}(p_2>(p_1\epsilon\text{Æ}(\tilde{\imath}))),\ |\ p_1-p_2|=k] \qquad - \qquad (2)$$

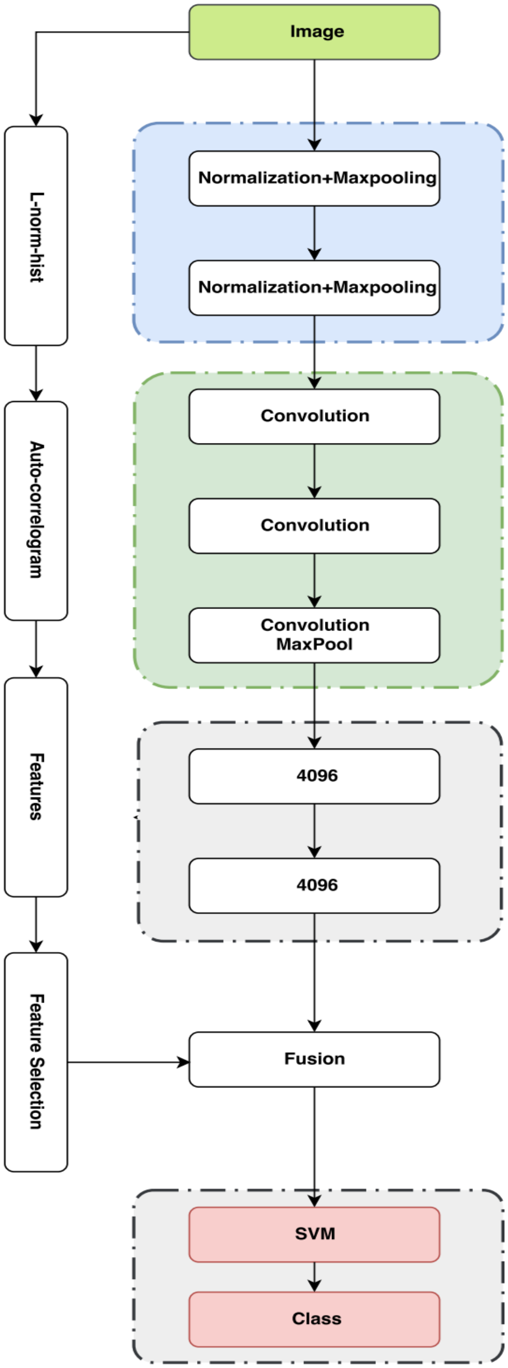


Fig. 1. Proposed Alex Correlogram Network (ACNet) adult image classification. The convolution layers (first five) are followed by two fully connected layers. The output of these layers is represented as DL feature set and merged with the correlograms features. The SVM learns and decides the nature of the test images.

Where *ĩp* is an image after classifier operation represented by Æ. In (2), the *ĩp* is a probabilistic output of the image representation from the classifier. This setup of feature extraction based on the Correlograms is integrated in the proposed DL architecture in Fig. 1. With the integration of the Correlograms, we represent the proposed DL network of Fig. 1 as the Alex Correlogram Network (ACNet). The convolution layers (first five) are followed by two fully connected layers. The output of these two layers is represented as DL feature set and merged with the Correlograms features. The SVM learns and decides the nature of the test images.

The layered architecture we use contains more than 55 million parameters trained for more than 10000 classes. Rather than training from the scratch, we use the weights of the ConvNet [14] obtained from 1.2 million images. Our input image is fed into the first layer and feature vector is obtained from the seventh layer. The weights for the classification layer are modified based on the labelled data. In the last layer, we are using the SVM. For a test image, the SVM outputs a binary decision “adult” or “non-adult”.

For an evaluation of the proposed architecture, we use the key-frames from the NDPI videos. Further details are available in [26]. Fig. 2 shows some samples.

Fig. 2. Sample images from NDPI [26].

We use the accuracy as an evaluation parameter as it is mostly used in the state of the art for similar problems and applications and is favorable for this evaluation as well. For comparison, we calculate the similar architectures of AlexNet (ANet) of [3], AGNets of [4]. For performance evaluation and comparison, we train 10 ANets, 10 AGNets and 10 ACNets using four of the five folds of NDPI videos. We set the fifth fold for testing. In the evaluation, our objectives are firstly to check the DL architectures with and without external feature fusion. Secondly, we compare the proposed approach to two networks fusion approach of [4]. The setup in [4] combines the AlexaNet and GoogleNet and represent it as the AGNet for performance enhancements. Interestingly, in the evaluation using Accuracy, our proposed approach of external feature fusion in DL architecture outperforms the AGNet. Fig. 3 shows the comparison of the three approaches using the Accuracy. ANet alone achieves an Accuracy of 0.88; the AGNet achieves an Accuracy of 0.90 and the proposed ACNet achieved an Accuracy of 0.93. The proposed approach thus gets 5% increase in detection performance compared to the ANet. The proposed approach achieves 3% increase compared to the AGNet.

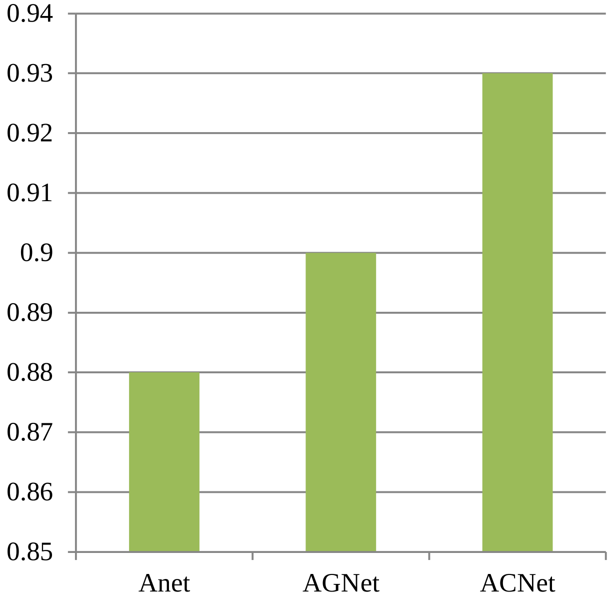


Fig. 3. Comparison based on accuracy between the proposed ACNet, ANet [3], AGNet [4].

The increased performance of our ACNet shows that it is possible to combine Deep learnt features with classical hand crafted features and achieve good overall performance. With this research direction, we expect an already rich set of features extraction paradigm to be exploited and used in conjunction with the DL architectures.

V. Conclusion

We proposed the fusion of the DL feature set with the classical hand crafted feature extraction paradigm. From the DL, we use the CNN feature set with the Correlograms feature set. We achieve 5% enhancement over CNN features alone (ANet) and 3% enhancement with the approach of (AGNet) which combines the fusion of two DL architectures. With this research direction, we expect an already rich set of feature extraction paradigm from last many years of research to be exploited and used in conjunction with the DL architectures achieving combined good overall performance. We hope that the new DL approaches will flourish that will combine multiple cues in supervised and unsupervised methods. Also, the execution of DL will be favored to the available hardware resources. Future work is directed towards semantic integration of skin learning in DL feature generation steps.

Acknowledgments

The work in this article is funded by the Dean of Scientific Research (SRD), Project number: 1335-coc-2016-1-12-S at the Qassim University, Kingdom of Saudi Arabia. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] D. Kotzias, M. Denil, N. de Freitas, and P. Smyth, "From Group to Individual Labels Using Deep Features," Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '15, pp. 597–606, 2015.

[2] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning Hierarchical Features for Scene Labeling," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. Curran Associates Inc., pp. 1097–1105, 2012.

[4] M. Moustafa, "Applying deep learning to classify pornographic images and videos," Nov. 2015.

[5] A. P. B. Lopes, S. E. F. de Avila, A. N. A. Peixoto, R. S. Oliveira, M. de M. Coelho, and A. de A. Araújo, "Nude Detection in Video Using Bag-of-Visual-Features," in 2009 XXII Brazilian Symposium on Computer Graphics and Image Processing, 2009, pp. 224–231.

[6] A. Abadpour and S. Kasaei, "Pixel-Based Skin Detection for Pornography Filtering," Iran. J. Electr. Electron. Eng., vol. 1, no. 3, pp. 21–41, 2005.

[7] R. Ullah and A. Alkhalifah, "Media Content Access: Image-based Filtering," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 3, 2018.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," Sep. 2014.

[9] E. Valle, S. Avila, F. Souza, M. Coelho, and A. de A. Araujo, "Content-Based Filtering for Video Sharing Social Networks," in XII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais—SBSeg 2012, 2011, p. 28.

[10] P. Monteiro, S. Eleuterio, M. De, and C. Polastro, "An adaptive sampling strategy for automatic detection of child pornographic videos."

[11] N. Agarwal, H. Liu, and J. Zhang, "Blocking objectionable web content by leveraging multiple information sources," ACM SIGKDD Explor. Newsl., vol. 8, no. 1, pp. 17–26, Jun. 2006.

[12] C. Jansohn, A. Ulges, and T. M. Breuel, "Detecting pornographic video content by combining image features with motion information," in Proceedings of the seventeen ACM international conference on Multimedia - MM '09, 2009, p. 601.

[13] J.-H. Wang, H.-C. Chang, M.-J. Lee, and Y.-M. Shaw, "Classifying Peer-to-Peer File Transfers for Objectionable Content Filtering Using a Web-based Approach."

[14] Hogyun Lee, Seungmin Lee, and Taekyong Nam, "Implementation of high performance objectionable video classification system," in 2006 8th International Conference Advanced Communication Technology, 2006, p. 4 pp.-pp.962.

[15] D. Liu, X.-S. Hua, M. Wang, and H. Zhang, "Boost search relevance for tag-based social image retrieval," in 2009 IEEE International Conference on Multimedia and Expo, 2009, pp. 1636–1639.

[16] J. A. Da, S. Júnior, R. E. Marçal, and M. A. Batista, "Image Retrieval: Importance and Applications."

[17] S. Badghaiya and A. Bharve, "Image Classification using Tag and Segmentation based Retrieval," Int. J. Comput. Appl., vol. 103, no. 15, pp. 20–23, Oct. 2014.

[18] A. N. Bhute and B. B. Meshram, "Text Based Approach For Indexing And Retrieval Of Image And Video: A Review," Apr. 2014.

[19] A. P. B. Lopes, A. P. B. Lopes, R. E. F. De Avila, A. N. A. Peixoto, R. S. Oliveira, and A. De A. Araújo, "A Bag-of-Features Approach based on Hue-Sift Descriptor for Nude Detection."

[20] N. Rea, G. Lacey, R. Dahyot, and C. Lambe, "Multimodal periodicity analysis for illicit content detection in videos," in 3rd European Conference on Visual Media Production (CVMP 2006). Part of the 2nd Multimedia Conference 2006, 2006, pp. 106–114.

[21] S. Lee, W. Shim, and S. Kim, "Hierarchical system for objectionable video detection," IEEE Trans. Consum. Electron., vol. 55, no. 2, pp. 677–684, May 2009.

[22] Y. S. L Li, "Objectionable videos detection algorithm based on optical flow," Comput. Eng., vol. 12, p. 77, 2007.

[23] Y. Qu, ZY; Liu, YM; Liu, Y; Jiu, K; Chen, "A Method for Reciprocating Motion Detection in Porn Video based on Motion Features," in IEEE International Conference on Broadband Network and Multimedia Technology, 2009, pp. 183–187.

[24] Z. ZHAO, "Combining SVM and CHMM classifiers for porno video recognition," J. China Univ. Posts Telecommun., vol. 19, no. 3, pp. 100–106, Jun. 2012.

[25] V. M. T. Ochoa, S. Y. Yayilgan, and F. A. Cheikh, "Adult Video Content Detection Using Machine Learning Techniques," in 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems, 2012, pp. 967–974.

[26] "Pornography Database." [Online]. Available: https://sites.google.com/site/pornographydatabase/. [Accessed: 09-Nov-2017].

Probabilistic Neural Network and Word Embedding for Sentiment Analysis

Saqib Alam, Nianmin Yao
School of Electrical Information and Electrical Engineering
Dalian University of Technology
Liaoning, Dalian 116024

***Abstract*—In the present days, Artificial Intelligence (AI) is an attractive area of research along with numerous practicable purposes and vigorous subject matters and tasks, such as, understand speech, natural language, diagnose medicine and support basic research. In this study deep learning (DL) techniques, i.e. Probabilistic Neural Network (PNN) and Word Embedding (WE) will be used for sentiment analysis. The entire proposed framework will be divided into three phases: (a) normalization, (b) word vectorization, and (c) execution of proposed model.**

Keywords—Deep learning; probabilistic neural network; word embedding; sentiment analysis

I. Introduction

Artificial Intelligence (AI) become a new trend, leaving behind other areas of research such as Big Data (BD), Internet of Things (IoT), Virtual Reality and many more in the past. In this new revolution of AI, DL turned into a hot zone for analysts, as a subset of Machine Learning (ML). It is a technique that uses many layers of non-leaner information processing for supervised or unsupervised classification, pattern analysis, transformation and feature extraction [1]. DL and its techniques have been used broadly in different areas of research and have an impressive impact on Natural Language Processing (NLP) [2]. In areas such as artificial vision and natural language processing, DL have impressive results because it is hard to use a strict mathematical model to characterize real-world images or languages. For example, it is almost near to impossible to write a powerful algorithm to detect handwriting or objects in an image, it is now a simple implementation of a DL algorithm that learns to perform tasks that exceed human accuracy levels [3]. Previously for sentiment analysis, some DL techniques were used which were enormously effective, such as word2vector [4]. In our this work we will use WE [4] with Probabilistic Neural Network (PNN) to increase the accuracy of sentiment analysis [5]. PNN is extensively adopted by researchers for pattern recognition and classification. We adopted PNN for some of its advantages, such as the training is easy and instantaneous in PNN [6]. PNN has Bayes optimal classification, much faster and accurate than multilayer perception networks [7]. PNN networks generate accurate predicted target probability scores.

This study includes five sections: (I) introduction and background of this study, (II) related work, (III) material and works which elaborate the three phases of this study, (IV) results and discussion of this research work, and (V) the conclusion.

II. Literature Review

Previously PNN was used mostly in image processing, such as [8] proposed in his work a modified PNN novel, the improvements were made by replacing the exponential activation function in the pattern layer of the PNN to the *complex* exponential function. The properties of the classical PNN such as fast training procedure and the convergence to the optimal Bayesian decision were same, but theoretically the recognition performance and space complexity should be approximately (R / C)2 / 3–times lower. During the experiment, they described a protocol for comparing image recognition methods using well-known data sets, in the situation of small sample problems. The experiments of modern deep neural network model show that due to its high accuracy and low computational complexity, the implementation of the maximum a posteriori (MAP) program is considered very promising in various tasks related to image recognition. Of course, the proposed algorithm is not the most accurate classifier in all cases, especially if the number of occurrences of each class is quite large. However, it has been shown that its method allows the treatment of PNN defects caused by brute force processing of all instances. In addition, unlike the original PNN, its modification allows you to choose a compromise solution between accuracy and computational complexity. Author in [6] reviewed two methods: PNN and the polynomial Adaline for classification based on Bayes strategy and nonparametric estimators for probability density functions. He found the most significant advantage of the PNN network is that the training is straightforward and can be used in real time, as the network can begin to summarize the new model as long as the patterns representing each category are observed. PNN has other advantages. (1) By selecting the appropriate smoothing parameter values, the shape of the decision surface can be made as complex as possible or as simple as possible. (2) The area where the decision area can approach the optimal minimum risk decision (Bayesian criteria). (3) The wrong sample can be tolerated. (4) Rare samples are sufficient to meet the performance of the network. (5) For the statistics that change over time, the old model may be overwritten by the new model. The PNN paradigm has been compared with the popular back propagation neural networks to classify and obtained data as actual measurements of electron emitters. In this particular experiment, PNN trained 200,000 times faster than back propagation. The disadvantage of PNN is the need to store and use the entire training database to test unknown patterns. In the case of very large databases and mature applications where test time is more important than training time, the Adalin polynomial paradigm has been developed, which does not have these limitations.

In this study [9] proposed a hybrid model was proposed Which includes a PNN and two layer restricted Boltzmann (RBM). The objective of this hybrid model of deep learning is designed to achieve better classification accuracy of the SA, i.e. negative and positive polarity, according to different situation, the model works very well. The experiments were conducted with Panga and Li, and Blitzer et al. datasets, binary classification was implemented in each data set. Accuracy been improved as compared to existing and advanced technology of [10].

In the review study [2] described plenty of studies related to sentiment analysis by using DL models. By analyzing all these studies, it was found that by using the DL method, SA can be analyzed in a more efficient and accurate manner. Since SA is used to predict user views, and the DL model is based on the prediction or imitation of the human mind, the DL model provides higher accuracy than the shallow models. DL networks are superior to SVMs and normal neural networks because DL networks have more hidden layers than normal neural networks with one or two hidden layers. The DL network can provide training in a supervisory/unattended manner. The DL network performs automatic extraction of functions and does not require manual intervention, so they can save time because no functional engineering is required.

In this work [11], proposed a supervised PNN structure determination algorithm. An important feature of this supervised learning algorithm is to directly consider the requirements of network size and classification error rate in the process of determining the network structure. Therefore, the proposed algorithm often leads to a rather small network structure with satisfactory classification accuracy.

In their study [12] proposed an adaptive system based on the learning algorithm Q(0), which selects and calculates the smoothing parameters of the PNN method. It includes all possible PNN models. These models differ in the way they represent smooth parameters. The basis of the new method is a selection algorithm based on Q(0) to IRC adjustment parameters. The proposed method has been tested in six data sets and compared with CRF in conjugate gradient method training, the algorithm SVM classification gene expression programming (GEP), the method k-means, perceptronie neural network and learning vector quantization. In these three classification problems, at least one of the NPSC, PNNV, or PNNVC patterns formed in the proposed process can ensure the highest average accuracy. In four out of six, the PNNS was the second since the last data classification. This means that the representation of the smoothing parameters in terms of vectors and matrices contributes to higher CRF predictions. As can be seen, she trained with the conjugate gradient method CRF gave the best accuracy of data classification for six cases. Therefore, the suggestion of any alternative probabilistic training method in neural networks is justified.

The author in [13] explained in their work how to learn more about what is being used on the tweets to show in the polarity of messages and words. They provide detailed information about their third-party online training program, the key to their success. The result creates a new state-of-the-art on the phrases and is second in part of the messages. All kinds of tests, of their system were the first one on both subtasks. Their network guide includes the use of distance supervised data that focuses on (clearing noisy texts from tweets) to enhance the weight of the network passed from the unsupervised neural network. Therefore, their solution combines two basic aspects of the IR components: unsupervised learning of text representation (WEfrom neural language model) and study of well-managed data (Fig. 1).

Previous work [5] suggested the impact of preprocessing technique on the accuracy of sentiment analysis of different ML algorithms. In that work it was suggested that removing of emoticons and stopwords alongwith stemming and word vectorization can improve the efficiency and accuracy of ML algorithms which are Naive Bayes, Support Vector Machine and Maximum Entropy. Due to the sarcastic nature of tweets removing of the emoticons can affect the accuracy, while stopwords such as *'a'*, *'is'*, *'the'*, *'it'* are the highest in the frequency due to which the desired results cannot be obtained till the removal of it. In some tweets users using a single English character many times for example *@stellargirl: looooooovvvvvveee my Kindle2*. In this work we will use PNN on the same dataset to improve the accuracy.

III. Material and Methods

Previous study [5] used some ML algorithms to enhance the accuracy of sentiment analysis, and in this study we will use some DL techniques for further incurring the accuracy.

A. Dataset

Twittratr[1] is an internet platform for Twitter's sentiment data, which using a series of negative and positive sentiments. For this work, we used the previous dataset of our research work which contains the Twittrat keyword list. This list includes 174 positive and 185 negative words [5]. For each tweet, we will count the number of negative and positive keywords that appears. The classifier returns a huge amount of polarities.

B. Preprocessing

Text can come in many forms, from single word lists to sentences and many paragraphs with special characters (such as tweets). As with any data science problem, a question can be raised that which steps should be taken to convert words into numerical values which can be understandable by the ML algorithms. Although this is not an exhaustive list, the preparation of the text is a complex art that requires the selection of the best tools, including data and questions. Many libraries and ready services are ready to help, but some may need to manually map terms and vocabulary. After preparing the dataset, supervised or unsupervised machine learning techniques can be used.

[1] http://twitrratr.com/

1) Cleaning data: Twitter is distinguished as a short messages; enclosure of URIs, usernames; special characters and topic markers. It often containing abbreviations and errors, some of these occurrence consist of linguistic noise, which makes make microblog part-of-speech tagging extremely difficult [14]. To avoid such a difficulty we removed emoticons, special characters and hashtages from our dataset.

2) Removal of stopwords: In today's world most of the data are available in textual form. Mostly this textual data congaing some words such as *"a", "the", "it", "as"* which are higher in frequency in every document and effecting the nature and accuracy of that document, if these high frequency words are not remove then they could interrupt the comparison calculation [15].

3) Stemming: In microbloging users often using some words with many alphabets, such as *@I loveoooooooo kindle.* Such kind of words can affect the efficiency of accuracy. To reduce a word to its proper stem is call stemming. The purpose of stemming is to find out the representative indexing form of a word by the purpose of truncation of affixes [16].

C. Word Embedding

WE is a natural language processing (NLP) technique which allows words or phrases to be mapped as vectors of real numbers. This process is important because many ML algorithms as well as deep neural networks require the input that should be vectorized and continues values vectors, as it cannot be done by strings of plain text. In [17], the author presented GloVe, a competitive set of pre-trained embeddings, suggesting that word embeddings was suddenly among the mainstream.

Logically, each feed-forward neural network which acquires words from a term as an input and embeds them as vectors into a lower dimensional space, and it then refine all through back propagation, essentially crop word embeddings as the weights of the first layer, referred as Embedding Layer.

The main dissimilarity between such networks and word2vec is complexity of its computational approach. The modernization and speedy growth in computational approaches improve its importance GloVe.

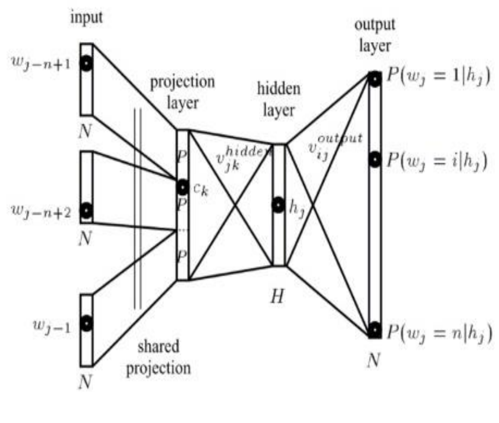


Fig. 1. Neural Language Model [18].

For illustration of words as vector an unsupervised algorithm was introduced by [17]. The training can be performed on combined global word-word co-occurrence statistics from a document. More specifically, the [17] stated that the relationship between the probabilities of the coexistence of two words (rather than their coexistence probabilities) is a factor that contains information, and therefore depends on the encoding of this information as a vector difference.

D. GloVe Algorithm

There was an enormous flow of articles regarding word vector representation after the publishing of Tomas Mikolov [4] work. Following that work, Stanford's Global Vector for Word Representation [17] was one of the best research work, which elucidated that why such algorithms and reformulated word2vec escalate as a particular nature of factorization for word co-occurrence matrices.

Below are the steps of the GloVe algorithm:

1) Collect word co-occurrence statistics in a form of word co-occurrence matrix X . Each element X_{ij} of such matrix represents how often word *i* appears in context of word *j*. Usually we scan our corpus in the following manner: for each term we look for context terms within some area defined by a *window_size* before the term and a *window_size* after the term. Also we give less weight for more distant words, usually using this formula:

$$\left(decay = \frac{1}{offset}\right) \tag{1}$$

2) Define soft constraints for each word pair: $\left(w_i^T w_j + b_i + b_j = \log(X_{ij})\right)$ where w_i - vector for the main word, w_j - vector for the context word, b_i, b_j are scalar biases for the main and context words.

$$J = \sum_{i=1}^{V} \sum_{j=i}^{V} f(X_{ij}) \left(w_i^T + b_i + b_j - (\log X_{ij})^2\right) \tag{2}$$

3) Define a cost function: Here f is a weighting function which helps us to prevent learning only from extremely common word pairs. The GloVe authors choose the following function:

$$f(X_{ij}) = \begin{cases} (\frac{X_{ij}}{x_{max}})^2 & if X_{ij} < x_{max} \\ 1 & otherwise \end{cases} \tag{3}$$

E. Probabilistic Neural Network

A probabilistic neural network (PNN) is a supervised network, which can be commonly used in decision making and classification problems [19]. PNN was firstly introduced by [20]. The immediate and easy training makes PNN's main advantage, and can be used for real-time as well [6].

F. Architecture of PNN

A PNN is an completion of a statistical algorithm, called kernel discriminate analysis in which the procedures are structured into a multi-layered feed-forward network with four layers, i.e. input layer, pattern layer, summation layer, and output layer.

- Input layer

This layer distributes the N number of input nodes to the neurons and every neuron symbolizes a predictive variable in this layer. According to the categorical variables, the N-1 neuron can be applicable on N number of categories. It normalizes the series of the values by deducting the medium and dividing by the inter-quartile range. After that the input neurons provides the values to every neurons in the hidden layer.

- Pattern layer

Pattern layer containing the Gaussian functions and for every case of training dataset the layer hold one neuron. Along-with the target values, it also stores the predictive variables values.

- Summation layer

The summation layer performs a sum operation of the outputs from the second layer for each class.

- Output layer

The output layer performs a vote, selecting the largest value. The associated class label is then determined.

G. PNN Algorithm

In Fig. 2, we exemplify the architecture of PNN with hidden layers. The sum of pattern nodes is the same of total of training sample. The synaptic weight $w_{ij}^{(P)}$ in the pattern to input is:

$$w_{ij}^{(P)}=x_i^{(j)} \tag{4}$$

Where, $x_i^{(j)}$ represents the i^{th} input node of the j^{th}sample at the input layer. And for the weight between pattern and summation layers $w_{jk}^{(S)}$ can be represent as:

$$w_{jk}^{(S)} = \begin{cases} 1 & if T_k^{(j)}=1 \\ 0 & otherwise \end{cases} \tag{5}$$

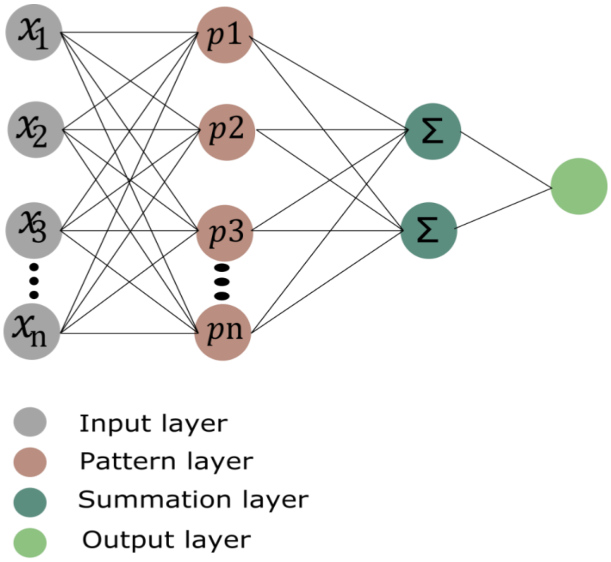


Fig. 2. Architecture of PNN.

Here the T_k^j value is 1 since the association of sample j with class k and otherwise 0s.

After the training procedure as shown in (4) and (5), the input classification pattern can be commenced as under:

$$n_j^P = \sqrt{\sum_i \left(w_{ij}^{(P)} - x_i\right)^2} \tag{6}$$

The pattern out can be calculated as:

$$P_j=\exp\left(-\frac{n_j^{(P)}}{2\sigma^2}\right) \tag{7}$$

Here σ is stander deviation of Gaussian distribution, which is a smoothing parameter corresponding representation.

The summation layer every single node symbolizes an individual class and can be expressed as:

$$S_k=\frac{1}{\sum_j w_{jk}^{(s)}} \sum_j w_{jk}^{(s)} . P_j \tag{8}$$

The input vector classifies into a precise single class by the output layer, if the output value is maximum from the input node at the summation layer:

$$y = \arg max_x S_k \tag{9}$$

H. Our Proposed Model

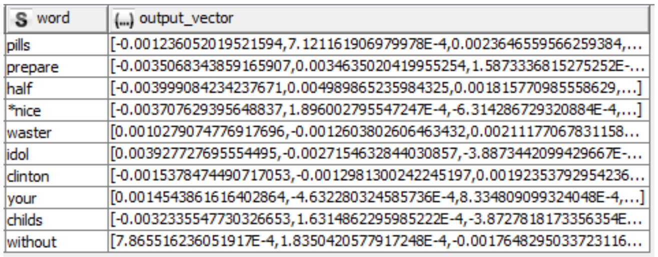
In our proposed model we used the dataset of 359 documents. The twitter data often containing *urls*, special characters and emoticons, besides this it contains unwanted words such as, *'i'*, *'the'*, *it*, *'a'* which mostly higher in the frequency and can affect the accuracy. As well as it includes words like *'looooooooovvvveeee'*, *'sooooooo'*. Table I shows the raw twitter data. Since we go further, we applied the preprocessing i.e. cleaning data, removal of stopwords and stemming to clean the dataset, which can be seen in Table II.

TABLE I. RAW TWITTER DATA

Stellargirl I loooooooovvvvvveee my Kindle2. Not that the DX is cool, but the 2 is fantastic in its own right.
I hate aig and their non loan given asses. :(
Jquery is my new best friend. $$$
i srsly hate the stupid twitter API timeout thing, soooo annoying!!!!!
How can you not love Obama? He makes jokes about himself.
@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D

TABLE II. TWITTER DATA AFTER CLEANING

Stellargirl love Kindle2. Not DX cool, but 2 fantastic right
hate aig their non loan given asses
Jquery new best friend
srsly hate stupid twitter API timeout thing, so annoying
how not love Obama makes jokes about himself.
A, that's bummer. should got David Carr of Third Day



S word	(...) output_vector
pills	[-0.00123605201952159 4,7.12116190697997 8E-4,0.00236465595662593 84,...
prepare	[-0.0035068343859165907,0.003463502041995525 4,1.5873336815275252E-...
half	[-0.003999084234237671,0.004989865235984325,0.0018157709855586 29,...]
*nice	[-0.0037076293956488 37,1.896002795547247E-4,-6.314286729320884E-4,...]
waster	[0.0010279074776917696,-0.001260380260646343 2,0.0021117706783115 8...
idol	[0.0039277276955544 95,-0.0027154632844030857,-3.887344209942966 7E-...
clinton	[-0.0015378474490717053,-0.0012981300242245197,0.0019235379295423 6...
your	[0.0014543861616402864,-4.632280324585736E-4,8.334809099324048E-4,...]
childs	[-0.0032335547730326653,1.6314862295985222E-4,-3.8727818173356354E...
without	[7.865516236051917E-4,1.8350420577917248E-4,-0.0017648295033723116...

Fig. 3. Word vectors.

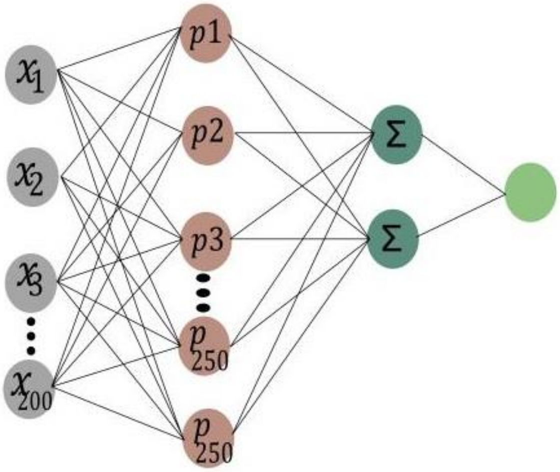


Fig. 4. The input and output nodes of our proposed model.

In the next step, we used WE [4] to convert the strings data into word vector implementing the batch size 1000 and layer size 200. In the following Fig. 3 words and their vectors can be seen.

Later on in the next step, we split the dataset into 70% train and 30% of test datasets, which divided it into 251 and 108 documents simultaneously.

In the last step, we applied PNN [7] on our dataset. As the layers size were 200 in our third step, the input nodes will also be 200 and pattern nodes will be 250 as well as showmen in Fig. 4. We kept the input and pattern layers sizes 200 and 250 simultaneously since the accuracy were at highest at this level.

IV. Results and Discussion

In our previous work [5], we applied preprocessing techniques along-with removing of emoticons on SVM, NB and MaxE algorithms and we got the accuracy results 81.63%, 91.81% and 88.27%, respectively on a dataset of 250 documents which can be seen in Table III.

TABLE III. Classifiers Accuracy after Preprocessing

Algorithms	**Accuracy after Preprocessing**
SVM	81.63%
NB	91.81%
MaxE	88.27%
PNN	98.00%

TABLE IV. PNN Accuracy Table

Algorithms	**Accuracy after Preprocessing**
SVM	81.63%
NB	91.81%
MaxE	88.27%
PNN	98.00%

TABLE V. Positive and Negative Predictions

	True positive	**False Positive**	**True Negative**	**False Negative**	**F-Measures**	**Accuracy**	**Cohen's Kappa**
Negative	120	0	127	3	0.987	-	-
Positive	127	3	120	0	0.988	-	-
Overall	-	-	-	-	-	0.988	0.975

To improve the accuracy we used PNN and WE which enhanced the results and can be seen in Table IV.

In the above Table IV, we can see the accuracy improved to 98% as compared to our previous work [5]. We can see in Table V that after applying Word Embedding and PNN on the dataset of 250 twitter documents we get 120 negative and 127 positive prediction document class and have only 3 wrong classified class as well, shown in Fig. 5.

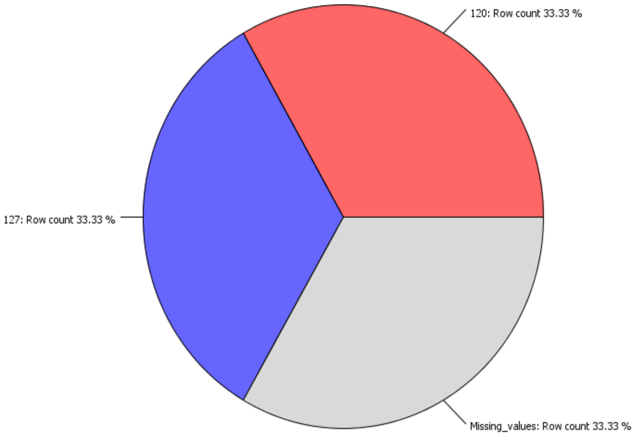


Fig. 5. Shows the row counts of positive, negative and missing values.

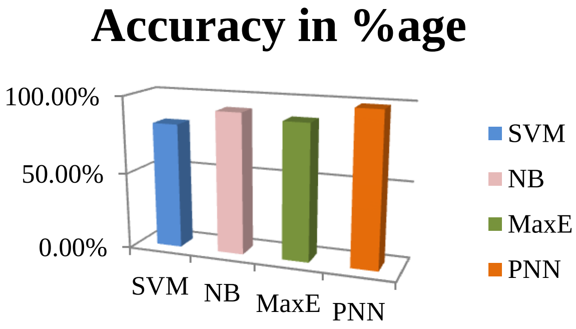


Fig. 6. Accuracy of SVM, NB, MaxE and PNN in per cents.

In Fig. 6, it can be seen that the accuracy of PNN is higher than other 3 algorithms, namely, SVM, NB and MaxE.

V. Conclusion

As in our previous work [5], we applied the preprocessing steps to improve the results. We observed in this work that as compared to Naive Bays, SVM and MaxE, the WE have tremendous effects on PNN. As compared to traditional techniques, PNN has higher accuracy and fast training time. Our investigational results on the basis of hybrid combination of WE and PNN could be a probable solution for enhancing the performance and accuracy of classification and as well decreasing the training time.

References

[1] L. Deng, G. E. Hinton, and B. Kingsbury, “New types of deep neural network learning for speech recognition and related applications: An overview,” 2013 IEEE Int Conf Acoust Speech Signal Process, pp. 8599–8603, 2013.

[2] Q. T. Ain et al., “Sentiment Analysis Using Deep Learning Techniques: A Review,” Int J Adv Comput Sci Appl, vol. 8, no. 6, pp. 424–433, 2017.

[3] T. J. O’Shea and J. Hoydis, “An Introduction to Deep Learning for the Physical Layer,” vol. 7731, no. c, pp. 1–13, 2017.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” Arxiv, pp. 1–12, 2013.

[5] S. Alam and N. Yao, “The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis.”

[6] D. F. Specht, “Probabilistic neural networks,” Neural Netw., vol. 3, no. 1, pp. 109–118, 1990.

[7] S. G. Wu, F. S. Bao, E. Y. Xu, Y.-X. Wang, Y.-F. Chang, and Q.-L.Xiang, “A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network,” pp. 1–6, 2007.

[8] A. Savchenko, “Probabilistic Neural Network with Complex Exponential Activation Functions in Image Recognition using Deep Learning Framework,” no. 14, pp. 1–14, 2017.

[9] R. Ghosh, K. Ravi, and V. Ravi, “A novel deep learning architecture for sentiment classification,” 3rd IEEE Int Conf Recent Adv Inf Technol, pp. 511–516, 2016.

[10] Y. Dang, Y. Zhang, and H. Chen, “A lexicon-enhanced method for sentiment classification: An experiment on online product reviews,” IEEE Intell Syst, vol. 25, no. 4, pp. 46–53, 2010.

[11] K. Z. Mao, K. C. Tan, and W. Ser, “Probabilistic neural-network structure determination for pattern classification,” IEEE Trans Neural Netw., vol. 11, no. 4, pp. 1009–1016, 2000.

[12] M. Kusy and R. Zajdel, “Probabilistic neural network training procedure based on Q(0)-learning algorithm in medical data classification,” Appl Intell, vol. 41, no. 3, pp. 837–854, 2014.

[13] A. Severyn and A. Moschitti, “Twitter Sentiment Analysis with Deep Convolutional Neural Networks,” Proc 38th Int ACM SIGIR Conf Res Dev Inf Retr - SIGIR 15, pp. 959–962, 2015.

[14] L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva, “Twitter part-of-speech tagging for all: Overcoming sparse and noisy data,” Proc Recent Adv Nat Lang Process, no. September, pp. 198–206, 2013.

[15] P. Runeson, M. Alexandersson, and O. Nyholm, “Detection of duplicate defect reports using natural language processing,” Proc - Int Conf Softw Eng, pp. 499–508, 2007.

[16] Y. Kadri and J. Y. Nie, “Effective stemming for Arabic information retrieval,” Proc Chall Arab NLPmt Int Conf Br Comput Soc, pp. 68–74, 2006.

[17] J. Pennington, R. Socher, and C. D. Manning, “GloVe : Global Vectors for Word Representation.”

[18] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, “Greedy Layer-Wise Training of Deep Networks,” Adv Neural Inf Process Syst, vol. 19, no. 1, p. 153, 2007.

[19] F. Vicino, “The Probabilistic Neural Network,” vol. 33, no. 2, pp. 335–352, 1998.

[20] D. F. Specht, “Probabilistic neural networks for classification, mapping, or associative memory,” IEEE Int Conf Neural Netw., pp. 525–532 vol.1, 1988.

Ranking Attribution: A Novel Method for Stylometric Authorship Identification

Marwa Taha Jamil, Dr. Tareef kamil Mustafa
College of Science, University of Baghdad
Baghdad, Iraq

***Abstract*—Stylometric Authorship attribution is one of the essential approaches in the text mining. The present research endorses a Stylometric method called Stylometric Authorship Ranking Attribution (SARA) overcomes the usual problems which are processing time and accurate prediction results, without any human opinion that relays on the domain expert. This new method also uses the most effective attributes used in the Stylometric authorship prediction frequent word bag counts, whether it was frequent single, pair or trio words attributes, which are the most successful attributes in Stylometric prediction, having more alibi for author artistic writing style for our authorship recognition and prediction proposed technique. The experiments show that the proposed method produces superior prediction accuracy and even provides a completely correct result at the final stage of our experimental tests regarding the dataset scope.**

Keywords*—*Data mining; text mining; Stylometric Authorship Attribution; SARA

I. Introduction

Data mining is the evaluation of observational data units to find authorized relationships and the evaluation of statistics in novel methods that are each obvious and beneficial to the statistics owner [1]. Text mining (TM) [2], additionally recognized as understanding discovery in textual database(KDT) [3] or textual content data mining [2], of which new fascinating expertise is created, many defined it also as the process of extracting previously unknown, understandable, achievable and practical patterns or understanding from the series of large and unstructured textual content information or corpus. Text mining uses the same evaluation approach and techniques as statistics mining. However, information mining requires structured data, whilst textual content mining aims to discover patterns in unstructured statistics [4]. The problem of text mining has gained growing attention in current years because of the big quantities of textual content data, which created a variety of social network, web, and other information-centric applications. Unstructured statistics is the most natural form of information which can be produced in any application scenario. As a result, there has been an extraordinary need for graph techniques and algorithms which can successfully manner a broad range of textual content purposes [1]. Another foremost issue is a multilingual text refinement dependency that creates problems. Only a few tools are available that aid multiple languages [5]. Text mining is generally composed of three steps: text preprocessing, text mining operations, post-processing. Text preprocessing tasks inclusive of information selection, classification and characteristic extraction normally convert the documents into intermediate forms, which have to be appropriate for distinct mining purpose. Text mining operations are the central phase of a text mining system and encompass clustering, association rule discovery, trend analysis, sample discovery and different know-how discovery algorithms. Post-processing tasks manipulate facts or understanding coming from text mining operations, such as comparison and resolution of knowledge, interpretation, and data visualization representation [6]. The upcoming sections in this research will illustrate the latest methods and approaches of a certain subfield in the text mining area that is concerned about the text corpus in literature and the writing style of its authors before stepping into the proposed method details.

II. Literature Review

An essential trouble in authorship attribution is the choice of stylometric aspects that are linguistic expressions of unique authors. Sets of proposed facets may vary, depending on accessible data, the supposed generality of their extraction approach and applicability to precise languages.

The easiest elements describe statistical residences of documents: word length, sentence length, and vocabulary richness. Function phrases are points primarily based on word frequencies. In contrast to text categorization problems, where the most established words are considered useless or even unsafe for classification, in authorship attribution problems they are frequently used as non-public fashion markers. However, not all the most universal phrases are exact candidates to be blanketed to that set of features: an important characteristic is an instability [7], i.e. the possibility to be replaced with the aid of every other word from the dictionary.

Other word-based elements are phrase sequences (n-grams). An instance of this approach can be observed in [8], the place classification using word sequences used to be examined on 350 poems in Spanish through five authors giving about 83% accuracy.

Features, which normally supply very excessive accuracy measures are personality n-grams, i.e. sequences of n characters extracted from phrases performing in documents. They are considered language independent, i.e. they can be extracted from texts in a variety of languages regardless of persona units used. See, for example, [9] for reviews on authorship attribution of English, Greek, and Chinese texts. In our opinion very accurate effects of their utility need to be handled with caution: there is an apparent useful dependence between report content and personality n-grams, so they may additionally represent and alternative representation of feature phrases (what is probable good) or they may also simply render document content material (what appears to be worse).

Tareef proposed a new Stylometric approach recognised as the Stylometric Authorship Balanced Attribution (SABA) which in a position to analyze texts in text mining, e.g., novels and performs by means of famous authors, attempting to measure the author's style, by way of deciding on some attributes that exhibit author's style of writing, assuming that these writers have a one of a kind way of writing that no different creator has, with greater accuracy prediction and impartial from human judgments, which ability that the technique does not count on the domain experts. This method is implemented by using merging three methods, which are called the computational approach, the Winnow algorithm, and the Burrows-delta method. The algorithm regarded an unguided mannequin and it tested in the English language correctly with noticeable prediction [10].

III. Stylometric Authorship Attribution

Stylometry is the study of writing style based totally on linguistic elements and is typically applied to authorship attribution troubles [11].

SAA was once begun as a "Content analysis" and was described as "understanding data now not as a series of bodily activities but as symbolic phenomena and to strategy their evaluation unobtrusively. Methods in the natural sciences do now not want to be worried about meanings, references, consequences, and intentions. Methods in social research that derive from these tough disciplines manipulate to omit these phenomena for convenience". The time period content material evaluation is about 50 years old. Webster's English Dictionary has listed it solely considering 1961 [12].

IV. Stylometric Authorship Balanced Attribution (SABA)

The SABA method is compared towards three different strategies the use of the computational approach, the Winnow algorithm method, and the Burrows-delta method. The results showed that the SABA method produces most useful prediction accuracy and even presents a completely right end result during the closing stage of the test [10].

The SABA method way is by neglecting the maximum values for the attribute frequencies and replacing it with "balanced" frequency. The idea that the right attributes are the "stabilized" or "balanced" attributes rather than attribute with the maximum frequencies. This means that in a written paragraph from a novel with assuming 10000 words, if a specific writer had used a specific word between 200-250 times in all of his books, then consider the attribute "word" has a "stabled" frequency percentage, hence is not a maximum frequency count[10].

V. Burrows Delta Method

While many methods have been utilized to the hassle of computerized authorship attribution, John F. Burrows's "Delta Method" [13] is an especially simple, yet effective. The purpose is to robotically determine, based on a set of known education archives labeled by using their authors, who the most probably creator is for an unlabeled check document. The Delta technique makes use of the most usual words in the education corpus as the facets that it makes use of to make these judgments. The Delta measure is described as: The suggestion of the absolute differences between the z-scores for a set of phrase variables in a given text-group and the rankings for the same set of word-variables in a target text [14].

VI. Methodology

Data is taken from the web site www.Gutenberg.org. The dataset is an incredible cross segment of nineteenth century English writing as appropriately as various work. Utilizing this accumulation; we assembled books from 5 of the best 100 most downloaded writers; collected 10 books from every one of the 5 writers and they are Charles Dickens, Jack London, William Shakespeare, Mark twain and Oscar Wilde.

Both algorithms (Burrow-Delta and SABA methods) sharing same first steps, starting by uploading the chosen novels in text mode (with .txt extension), steps of cleaning and chunking are performed (removing double spaces, punctuation marks, special characters, symbol and others) before the implementation of the process of transforming text into Microsoft Access 2010 database files; taking into account that every single record contains frequent or a pair or trio words.

All tests implemented in this experiment by using Microsoft Access 2010 database and Visual C#, and choose ten books for the famous author(Charles Dickens, Jack London, William Shakespeare, Mark twain and Oscar Wilde) (nine for Learn, one for test).

A. Burrow Delta Method

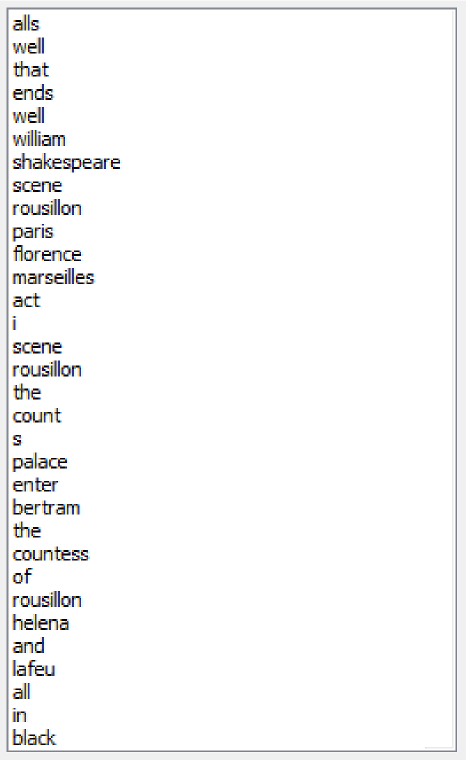
Burrow Delta represents the mean of the outright contrasts between the z-scores for an arrangement of word factors in a given text-gathering and the z-scores for a similar arrangement of word-factors in an objective text. The working steps will be implemented in detail in the case of frequent, pair, trio words.

The first step is to transform the book to be tested in text mode (.txt) into a separated list of book words. The final result of this is shown in Fig. 1. This operation will be executed for all learning and testing books.

Next, group the similar records, and calculate to the redundancy of these records, finally store the result in a separate table, the final result of this is shown in Fig. 2.

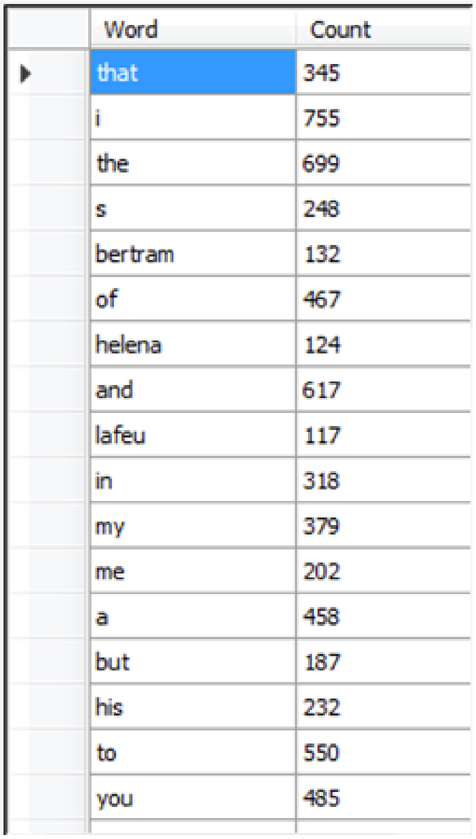
The next step is to cancel the differences between the size of books, by taking the percentage that speaks to the number of frequencies for each property separated by the entirety of frequencies for every one of the qualities multiplied by1000 in order to get a frequency that equal in weight for all used books and give true indication about the style of the author, the final result of this is shown in Fig. 3.

The following are making a stylometric map, by Merging and assembling all of the nine books (learning data) of the author which is being tested in a single table and make a relationship between their fields, calculate the arithmetic average of the redundancies.



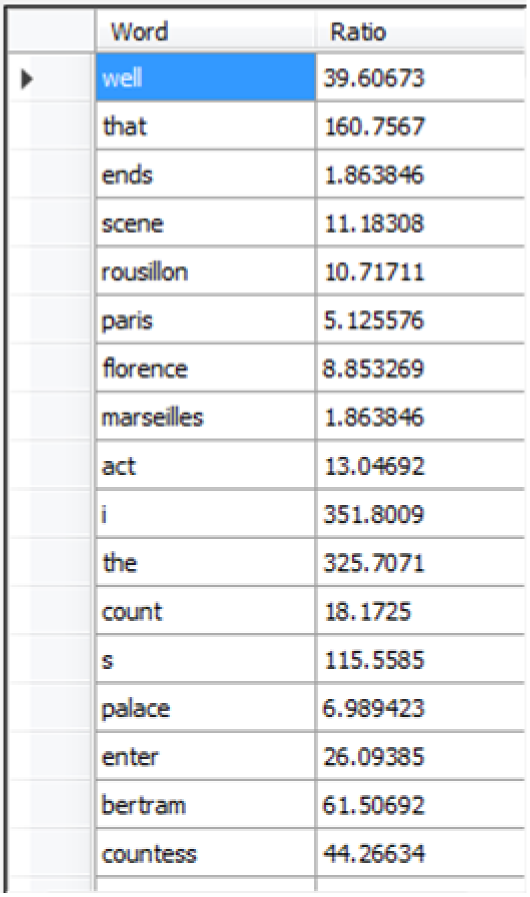
alls
well
that
ends
well
william
shakespeare
scene
rousillon
paris
florence
marseilles
act
i
scene
rousillon
the
count
s
palace
enter
bertram
the
countess
of
rousillon
helena
and
lafeu
all
in
black

Fig. 1. List of book words.



Word	Count
that	345
i	755
the	699
s	248
bertram	132
of	467
helena	124
and	617
lafeu	117
in	318
my	379
me	202
a	458
but	187
his	232
to	550
you	485

Fig. 2. Records redundancy.



Word	Ratio
well	39.60673
that	160.7567
ends	1.863846
scene	11.18308
rousillon	10.71711
paris	5.125576
florence	8.853269
marseilles	1.863846
act	13.04692
i	351.8009
the	325.7071
count	18.1725
s	115.5585
palace	6.989423
enter	26.09385
bertram	61.50692
countess	44.26634

Fig. 3. Book word ratio.

Index the total arithmetic average descending as shown in the following steps:

1) Merge all the learning data and save the result in a single table.

2) Assemble the result of merging data from the previous table and save the result in a new single table.

3) Make a relationship between their fields, and calculate the arithmetic average of the redundancies.

By calculating the average for all fields of the learning data and sorting it in descending, the stylometric map is ready now for the purpose of testing with other authors' books.

The Stylometric map is prepared for the purpose of examination and testing it, by building connections between the stylometric outline the five test books for all writers to get a new distribution of attributes based on the stylometric map that has been extracted.

In addition, this operation isolates the features that do not participate in any redundancy, that means if there are no common attributes between the learning books and testing books the main attributes will be isolated it by this operation, this step is important in order to make the stylometric map more stronger and reflecting a true style of the author.

After sorting the stylometric database map in the descending order based on the average percentage value for each attribute member in attribution set.

For Pearson, during the last step, select top 300 attributes that have the highest average percentage value in the stylometric map. Extract the Pearson correlation for the particular author's stylometric map from each of the five test books, hence giving five Pearson values. By having the weights for every parameter, increase each Pearson esteem by - 1 on the off chance that it is the wrong creator for the already known outcome or by +1 on the off chance that it is the correct writer.

For Spearman, a new table is configured that consist of 5 maps and 1 test. Each word corresponds to the ratio and the Rank (this rank is based on rank). Then works on it a word search function of the test, search on each map if found, take the rank for that word (only in this map), if not, they are compensated by zero. The result of this procedure is a table consisting of the test words only correspond to the word rank value and the rank of the word that was found at the specific map. The next step is applying spearman equation which also has a range between 1 to -1.

The Spearman connection between two factors is equivalent to the Pearson relationship between the rank estimations of those two factors; while Pearson's connection surveys straight connections, and Spearman's relationship evaluates the monotonic relationship.

B. *SABA Method*

The stylometric authorship balanced attribution (SABA) technique thought about an advancement of the calculation of Burrow-Delta strategy, this strategy relies upon the coefficient of difference (CV), which is spoken to as a factual estimation that isn't influenced by the perception of mean. Then will be analyzed and tried this calculation in English dialect in the regular, match and trio words.

In SABA technique, the trial of successive, match and trio words is like the Burrow Delta strategy in application, however there is basic contrast between them, precisely while choosing the highest point of 300 characteristics, these determinations rely upon the estimations of coefficient of variety (C.V), the accompanying case visit words can outline the real strides of removing the (C.V) And the strategy for choosing the required properties.

To apply SABA technique, rehash all the past strides as their request in the Burrow Delta strategy, at that point change the last stylometric guide to remove the estimations of the normal, the standard deviation (S.D) and the coefficient of variety (C.V) for every trait in the learning of the data, the (C.V) can be found by isolating the standard deviation by the mean itself, Finally, record the data in rising request in light of the estimations of the coefficient of variety (C.V) and select the main 300 qualities. In the wake of building connections between the last stylometric delineate the test books for all writers as we did on the Burrow Delta test, get the last successive test in SABA technique.

For Pearson, by having the weights for every parameter, duplicate each Pearson esteem by - 1 on the off chance that it is the wrong creator for the beforehand known outcome or by +1 in the event that it is the correct creator.

For Spearman, if there are no rehashed data esteems, a flawless Spearman relationship of +1 or −1 happens when every one of the factors is an ideal monotone capacity of the other. It merits saying that the utilization of Spearman is it requires less investment to contrast and Pearson and utilize basic numbers and less unpredictable in light of the utilization of the Rank rather than copies.

VII. Results

A. *Burrow Delta Method and Pearson*

The first step in this test is done on three authors only was the expectations of true and 0% error rate whether for frequent or pair or trio.

After applying it to five authors, it was found that there was an error of 20%.

- Frequent word

The following tables represent the final results for each author showing the prediction accuracy in the frequent word. The coefficient values in the highlighted cells are the highest value in each row, which indicates a fully correct prediction, as shown in Table I.

- Frequent pair

The following tables represent the final results for each author showing the prediction accuracy in pair word. The coefficient values in the highlighted cells are the highest value in each row, which indicates a fully correct prediction, as shown in Table II.

- Trio word

The following tables represent the final results for each author showing the prediction accuracy in trio word. The coefficient values in the highlighted cells are the highest value in each row, which not indicates a fully correct prediction, as shown in Table III.

- Summary

The results of the prediction for the frequent word and word pair were better than the trio. Although the results of trio words are less accurate than pair and frequent word, because the frequent word results and word pair don't contain any percentage of error prediction.

$$prediction\ error = \frac{Number\ of\ Mistakes}{Total\ Experiment\ Number} \times 100\%$$

$$Frequent\ word\ prediction\ error = \frac{0}{5} \times 100\% = 0\%$$

$$pair\ words\ prediction\ error = \frac{0}{5} \times 100\% = 0\%$$

$$trio\ words\ prediction\ error = \frac{1}{5} \times 100\% = 20\%$$

TABLE. I. PEARSON CORRELATION COEFFICIENT RESULTS IN THE FREQUENT WORD FOR EACH STYLOMETRIC MAP AGAINST FIVE OTHER AUTHORS TEST BOOKS

	Pearson in Dickens test	**Pearson in Shakespeare test**	**Pearson in Wilde test**	**Pearson in London test**	**Pearson in Twain test**
Dickens Stylometric Map	0.852674	0.586294	0.639235	0.655396	0.835108
Shakespeare Stylometric Map	0.66921	0.768426	0.615545	0.592736	0.718333
Wilde Stylometric Map	0.782839	0.622714	0.701601	0.638623	0.802786
London Stylometric Map	0.775219	0.554586	0.575797	0.738936	0.827077
Twain Stylometric Map	0.760399	0.597347	0.587423	0.70423	0.890519

TABLE. II. PEARSON CORRELATION COEFFICIENT RESULTS IN A FREQUENT PAIR FOR EACH STYLOMETRIC MAP AGAINST FIVE OTHER AUTHORS TEST BOOKS

	Pearson in Dickens test	**Pearson in Shakespeare test**	**Pearson in Wilde test**	**Pearson in London test**	**Pearson in Twain test**
Dickens Stylometric Map	0.657954	0.351077	0.304583	0.449358	0.610181
Shakespeare Stylometric Map	0.385408	0.607154	0.372539	0.340183	0.411975
Wilde Stylometric Map	0.484741	0.383454	0.515655	0.428261	0.560584
London Stylometric Map	0.492758	0.321433	0.251817	0.539384	0.491636
Twain Stylometric Map	0.532386	0.409412	0.326204	0.482538	0.684761

TABLE. III. PEARSON CORRELATION COEFFICIENT RESULTS IN TRIO WORD FOR EACH STYLOMETRIC MAP AGAINST FIVE OTHER AUTHORS TEST BOOKS

	Pearson in Dickens test	**Pearson in Shakespeare test**	**Pearson in Wilde test**	**Pearson in London test**	**Pearson in Twain test**
Dickens Stylometric Map	0.299347	0.073595	0.262487	0.293538	0.243407
Shakespeare Stylometric Map	-0.07354	0.220364	0.187214	0.092378	0.066574
Wilde Stylometric Map	0.215399	0.102512	0.339212	0.259741	0.264372
London Stylometric Map	0.259402	-0.02263	0.226403	0.349504	0.388979
Twain Stylometric Map	0.269146	0.108761	0.237624	0.371188	0.509085

However the experiment showed that the frequent word and word pair is the higher predicted values, and represents the best attribute according to the true prediction values for all results. This test use complex equations and numbers and take more time compared with the use of Spearman and Rank algorithm.

B. *Burrow Delta Method and Spearman*

The first step in this test is done on three authors only was the expectations of true and 0% error rate whether for frequent or pair or trio.

- Frequent word

The following tables represent the final results for each author showing the prediction accuracy in the frequent word. The coefficient values in the highlighted cells are the highest value in each row, which indicates a fully correct prediction, as shown in Table IV.

- Frequent pair

The following tables represent the final results for each author showing the prediction accuracy in pair word. The coefficient values in the highlighted cells are the highest value in each row, which indicates a fully correct prediction, as shown in Table V.

- Trio word

The following tables represent the final results for each author showing the prediction accuracy in trio word. The coefficient values in the highlighted cells are the highest value in each row, which indicates a fully correct prediction, as shown in Table VI.

- Summary

The results of the prediction for the frequent word, pair and trio were best possible, because of all results don't contain any percentage of error prediction.

$$prediction\ error = \frac{Number\ of\ Mistakes}{Total\ Experiment\ Number} \times 100\%$$

$$Frequent\ word\ prediction\ error = \frac{0}{5} \times 100\% = 0\%$$

$$pair\ words\ prediction\ error = \frac{0}{5} \times 100\% = 0\%$$

$$trio\ words\ prediction\ error = \frac{0}{5} \times 100\% = 0\%$$

TABLE. IV. SPEARMAN CORRELATION COEFFICIENT RESULTS IN THE FREQUENT WORD FOR EACH STYLOMETRIC MAP AGAINST FIVE OTHER AUTHORS TEST BOOKS

	Spearman in Dickens test	**Spearman in Shakespeare test**	**Spearman in Wilde test**	**Spearman in London test**	**Spearman in Twain test**
Dickens Stylometric Map	0.819973	0.330999	0.464541	0.480149	0.758905
Shakespeare Stylometric Map	0.406229	0.666872	0.305217	0.217681	0.490131
Wilde Stylometric Map	0.663092	0.389021	0.544767	0.411393	0.688602
London Stylometric Map	0.67824	0.234515	0.302373	0.648795	0.74929
Twain Stylometric Map	0.673973	0.329172	0.383627	0.549095	0.847885

TABLE. V. SPEARMAN CORRELATION COEFFICIENT RESULTS IN THE FREQUENT PAIR FOR EACH STYLOMETRIC MAP AGAINST FIVE OTHER AUTHORS TEST BOOKS

	Spearman in Dickens test	**Spearman in Shakespeare test**	**Spearman in Wilde test**	**Spearman in London test**	**Spearman in Twain test**
Dickens Stylometric Map	0.514772	-0.18968	-0.23314	0.062879	0.47578
Shakespeare Stylometric Map	-0.09823	0.404939	-0.14157	-0.33312	0.035592
Wilde Stylometric Map	0.158636	-0.0908	0.158825	-0.06006	0.386241
London Stylometric Map	0.218483	-0.30366	-0.41439	0.257186	0.336782
Twain Stylometric Map	0.227489	-0.15582	-0.24058	0.063646	0.577259

TABLE. VI. SPEARMAN CORRELATION COEFFICIENT RESULTS IN TRIO WORD FOR EACH STYLOMETRIC MAP AGAINST FIVE OTHER AUTHORS TEST BOOKS

	Spearman in Dickens test	**Spearman in Shakespeare test**	**Spearman in Wilde test**	**Spearman in London test**	**Spearman in Twain test**
Dickens Stylometric Map	-0.38487	-0.88527	-0.65054	-0.559	-0.52368
Shakespeare Stylometric Map	-0.99799	-0.67918	-0.80241	-0.93396	-0.89839
Wilde Stylometric Map	-0.62677	-0.86367	-0.445	-0.70271	-0.64397
London Stylometric Map	-0.51367	-0.9769	-0.74875	-0.30116	-0.26849
Twain Stylometric Map	-0.47065	-0.8657	-0.63585	-0.37992	-0.02225

However, the experiment showed that all test have perfect predicted values and represents the best attribute according to the true prediction values for all results. In this experiment the Speed and accuracy at a high rate, using the Spearman equation, which is less complex than Pearson's equation, it takes less time to compare with Pearson, work faster because taking from the test only 300 attributes means we did not adopt all the attributes values. Cancellation of CV and adoption of Ratio, use simple and less complex numbers because of the use of the Rank algorithm instead of the frequencies. Change the experience from 5 test 1 map To 5 map 1 test. It is worth mentioning that in this experiment was obtained perfect results.

C. SABA method and Pearson

- Frequent word

The following tables represent the final results for each author showing the prediction accuracy in the frequent word. The coefficient values in the highlighted cells are the highest value in each row, which not indicates a fully correct prediction, as shown in Table VII.

- Frequent pair

The following tables represent the final results for each author showing the prediction accuracy in pair word. The coefficient values in the highlighted cells are the highest value in each row, which not indicates a fully correct prediction, as shown in Table VIII.

- Trio word

The following tables represent the final results for each author showing the prediction accuracy in trio word. The coefficient values in the highlighted cells are the highest value in each row, which indicates a fully correct prediction, as shown in Table IX.

- **S**ummary

The results of the prediction for the frequent word and word pair were worse than the trio. Although the results of trio words are better accurate than pair and frequent word, because the trio word results don't contain any percentage of error prediction.

TABLE. VII. PEARSON CORRELATION COEFFICIENT RESULTS IN THE FREQUENT WORD FOR EACH STYLOMETRIC MAP AGAINST THREE OTHER AUTHORS TEST BOOKS

	Pearson in Dickens test	**Pearson in Shakespeare**	**Pearson in Wilde**
Dickens Stylometric Map	0.538038	0.428293	0.478812
Shakespeare Stylometric Map	0.555541	0.546308	0.500479
Wilde Stylometric Map	0.566413	0.451176	0.422471

TABLE. VIII. PEARSON CORRELATION COEFFICIENT RESULTS IN THE FREQUENT PAIR FOR EACH STYLOMETRIC MAP AGAINST THREE OTHER AUTHORS TEST BOOKS

	Pearson in Dickens test	**Pearson in Shakespeare**	**Pearson in Wilde**
Dickens Stylometric Map	0.490773	0.32738	0.300934
Shakespeare Stylometric Map	0.382706	0.44047	0.372926
Wilde Stylometric Map	0.405736	0.257335	0.294741

TABLE. IX. PEARSON CORRELATION COEFFICIENT RESULTS IN TRIO WORD FOR EACH STYLOMETRIC MAP AGAINST THREE OTHER AUTHORS TEST BOOKS

	Pearson in Dickens test	**Pearson in Shakespeare**	**Pearson in Wilde**
Dickens Stylometric Map	0.232979	0.06033	0.203679
Shakespeare Stylometric Map	-0.09015	0.18197	0.180051
Wilde Stylometric Map	0.146199	0.049749	0.336961

$$prediction\ error = \frac{Number\ of\ Mistakes}{Total\ Experiment\ Number} \times 100\%$$

$$Frequent\ word\ prediction\ error = \frac{2}{3} \times 100\% = 66\%$$

$$pair\ words\ prediction\ error = \frac{1}{3} \times 100\% = 33\%$$

$$trio\ words\ prediction\ error = \frac{0}{3} \times 100\% = 00\%$$

However the experiment showed that the frequent word and word pair is the less predicted values, and represents the worse attribute according to the true prediction values for all results. Use CV, this cause the consumption time to be longer than the ratio used. It also has long equations and complex numbers. Because there is a false expectation in the frequent (Table VII) and Pair (Table VIII), this test was not applied to all authors because the error rate will increase.

D. SABA Method and Spearman

- Frequent word

The following tables represent the final results for each author showing the prediction accuracy in the frequent word. The coefficient values in the highlighted cells are the highest value in each row, which not indicates a fully correct prediction, as shown in Table X.

- Frequent pair

The following tables represent the final results for each author showing the prediction accuracy in pair word. The coefficient values in the highlighted cells are the highest value in each row, which indicates a fully correct prediction, as shown in Table XI.

- Trio word

The following tables represent the final results for each author showing the prediction accuracy in trio word. The coefficient values in the highlighted cells are the highest value in each row, which not indicates a fully correct prediction, as shown in Table XII.

- Summary

Results of the prediction for the trio word was best possible, because of other results contain percentage of error prediction.

$$prediction\ error = \frac{Number\ of\ Mistakes}{Total\ Experiment\ Number} \times 100\%$$

$$Frequent\ word\ prediction\ error = \frac{1}{3} \times 100\% = 33\%$$

$$pair\ words\ prediction\ error = \frac{0}{3} \times 100\% = 0\%$$

$$trio\ words\ prediction\ error = \frac{1}{3} \times 100\% = 33\%$$

However the experiment showed that the pair word is the higher predicted values, and represents the best attribute according to the true prediction values for all results.

Use CV, this cause the consumption time to be longer than the ratio used. It also has long equations and complex numbers. Because there is a false expectation in the frequent (Table X) and Pair (Table XI), this test was not applied to all authors because the error rate will increase.

TABLE. X. SPEARMAN CORRELATION COEFFICIENT RESULTS IN THE FREQUENT WORD FOR EACH STYLOMETRIC MAP AGAINST THREE OTHER AUTHORS TEST BOOKS

	Spearman in Dickens test	**Spearman in Shakespeare**	**Spearman in Wilde**
Dickens Stylometric Map	0.47352	0.094352	0.223405
Shakespeare Stylometric Map	0.32946	0.359226	0.17037
Wilde Stylometric Map	0.402448	0.102134	0.153217

TABLE. XI. SPEARMAN CORRELATION COEFFICIENT RESULTS IN THE FREQUENT PAIR FOR EACH STYLOMETRIC MAP AGAINST THREE OTHER AUTHORS TEST BOOKS

	Spearman in Dickens test	**Spearman in Shakespeare**	**Spearman in Wilde**
Dickens Stylometric Map	0.28902	-0.19774	-0.19657
Shakespeare Stylometric Map	-0.09847	0.171194	-0.14929
Wilde Stylometric Map	0.073531	-0.25251	-0.10344

TABLE. XII. SPEARMAN CORRELATION COEFFICIENT RESULTS IN TRIO WORD FOR EACH STYLOMETRIC MAP AGAINST THREE OTHER AUTHORS TEST BOOKS

	Spearman in Dickens test	**Spearman in Shakespeare**	**Spearman in Wilde**
Dickens Stylometric Map	-0.37887	-0.86592	-0.61241
Shakespeare Stylometric Map	-1.01489	-0.71803	-0.80471
Wilde Stylometric Map	-0.62162	-0.90985	-0.36788

VIII. CONCLUSIONS

The first contribution is gain, a better prediction accuracy by involving the statistical Pearson correlation and Spearman correlation as a main weighting factor in the SABA and burrows method. And do not overlook that using the Spearman algorithm which is less complex compared to Pearson with the burrows algorithm led to optimal prediction results. The next contribution is improving the feature extraction process by introducing a new set of more dependable attributes, such as the word pair and the trio, in addition to the use of classical frequent words. The results showed that using Spearman correlation coefficients measure leads to, zero error prediction, Speed, and accuracy at a high rate, the Spearman Equation which is less complex than the Pearson Equation and it takes less time to compare with Pearson. The main consideration in this treatise is that the results are best when used ratio rather than CV, use simple numbers and less complicated because of the use of the Rank algorithm instead of frequencies matches. Conducting optimal predictors result in SARA compared with SABA and burrows. Replace ratio value with attribute ranks make the calculations more easy and speedy.

REFERENCES

[1] Kotu, V., & Deshpande, B. (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann.

[2] Stańczyk, U. (2016). The class imbalance problem in construction of training datasets for authorship attribution. In *Man-Machine Interactions 4* (pp. 535-547). Springer, Cham.

[3] Korasidi Andriana Maria (2016) "Authorship Attribution Forensics: Feature selection methods in authorship identification using a small e-mail dataset." Master thesis, University of Athens.

[4] Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

[5] White, D. R., & Joy, M. S. (2004). Sentence-based natural language plagiarism detection. *Journal on Educational Resources in Computing (JERIC)*, *4*(4), 2.

[6] Zhang, Y., Chen, M., & Liu, L. (2015, September). A review on text mining. In *Software Engineering and Service Science (ICSESS), 2015 6th IEEE International Conference on* (pp. 681-685). IEEE.

[7] Cohen, J. (1988). Statistical power analysis for the behavioral sciences 2nd edn.

[8] Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *42*(1), 59-66.

[9] Sullivan M. "Fundamentals of Statistics". Pearson Education. Canada, 2010.

[10] Mustafa, Tareef Kamil (2011), "Stylometric authorship balanced attribution prediction method". PhD thesis, Universiti Putra Malaysia.

[11] Dauber, E., Overdorf, R., & Greenstadt, R. (2017, June). Stylometric Authorship Attribution of Collaborative Documents. In *International Conference on Cyber Security Cryptography and Machine Learning* (pp. 115-135). Springer, Cham.

[12] Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Sage.

[13] Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, *17*(3), 267-287.

[14] Burrows, J. (2002). The Englishing of Juvenal: computational stylistics and translated texts. *Style*, *36*(4), 677-698.

Dynamic Data Aggregation Approach for Sensor-Based Big Data

Mohammed S. Al-kahtani
Dept. of Computer Engineering
Prince Sattam bin Abdulaziz University, Saudi Arabia

Lutful Karim
School of Information and Communications Technology
Seneca College of Applied Arts & Technology, Canada

***Abstract*—Sensors are being used in thousands of applications such as agriculture, health monitoring, air and water pollution monitoring, traffic monitoring and control. As these applications collect zettabytes of data everyday sensors play an integral role into big data. However, most of these data are redundant, and useless. Thus, efficient data aggregation and processing are significantly important in reducing redundant and useless data in sensor-based big data frameworks. Current studies on big data analytics do not focus on aggregating and filtering data at multiple layers of big data frameworks especially at the lower level at data collecting nodes (sensors) that reduce the processing overhead at the upper layer, i.e., big data server. Thus, this paper introduces a multi-tier data aggregation technique for sensor-based big data frameworks. While this work focuses more on data aggregation at sensor networks. To achieve energy efficiency it also demonstrates that efficient data processing at lower layers (sensor) significantly reduces overall energy consumption of the network and data transmission latency.**

***Keywords*—Data aggregation; big data; sensor networks; energy efficiency; clustering**

I. Introduction

The time of spreadsheet is over. A Google search, a barcode scan, a voice message, a picture of a car, a tweet among others all contains data that can be collected, analyzed and monetized. Indeed in today’s time, we manage and store our life online. Data are gathered from smart phones, laptops and tablets that collect and transfer information on what people do. However, this is just the beginning. Most devices including our TVs, watches and even washing machines will collect and transmit messages. With the growing amount of information that exceed quintillion of bytes, new machines and techniques more powerful than the normal computer had to be created to allow us to make sense of the zeros and ones. Super computers and various algorithms have helped one so far in the real time analysis of those increasingly larger amounts of information. Nevertheless, for more efficient data mining, one always has to be on the chase for new methods.

The term Big Data refers to large volume of data sets. In the last few years, with the increase in the amount of digital information around us, the term has gained in popularity. As we speak, many professional in the field are working on finding better data mining ways to cope for the future. Sensors, mobile phones and other devices all generate big data. One can simply question what is the advantage of collecting so much information and how can it be useful for any company? The simplest example to answer such a question is the grocery stores/supermarkets. These stores offer various promotions and discounts upon using their cards such as Air Miles, Optimum card etc. These cards generate big data in the form of collected information in regards to demand and supply among various parameters stated in the contract signed by the customer. All the information are gathered and once processed, they help companies improve their businesses in various ways. Indeed, the primary goal of collecting these huge datasets is to look for meaningful patterns by using optimal processing.

Emergence of sensor networks also play a major role in the rise of big data as thousands of sensor network applications collect huge amount of data that require processing. Hence, sensors data processing can be considered as a part of big processing. As sensors produce redundant data we can aggregate data to reduce and represent them in a meaningful way in big data framework. However, works on big data presented in [9]-[13] do not talk on sensor-based big data aggregation, they mostly talk about architecture and network theory of big data, data mining, and application of big data.

As sensors-based big data aggregation is an important area of research to reduce computational cost as well as energy consumption this paper introduces a sensor data aggregation approach for a multi-tier big data framework. The proposed aggregation approach is designed in three layers to ensure that sensors data aggregation is facilitated at the lowest layer. As the proposed communication framework only consists three layers of communication and processing devices (i.e., sensors, gateway node that connects to Internet, and big data server) this data aggregation approach has three layers.

The proposed data aggregation allows both cluster-based and tree-based network topologies and thus, considered as a hybrid data aggregation approach. Clustering is used in most sensor network applications especially, they are greatly required for emergency or real-time applications such as rescue operations, health, and traffic monitoring to reduce data transmission latency (results in reduced data processing delay and overhead at big data server). On the other hand, tree-based approach achieves efficiency in non-real time applications where achieving energy efficiency is more important than data transmission delay. The proposed approach works by selecting a few nodes that work as active nodes [19] to collect and aggregate data for a certain period of time unless the residual energy of these nodes become critical. While most clustering algorithms [1], [4]-[8], [18]-[20] allow all member nodes of a cluster to actively work at any time instant the proposed approach selects only a few nodes as active to work at any time instant that cover the whole network area. The proposed approach allows other nodes to work as alternative nodes that take the responsibility of active nodes only when any active node fails. This results in fault tolerance and energy efficiency. The rest of the paper is organized as follows.

Section II briefly presents literature on sensor data aggregation approaches. Section III briefly presents the working principle of the proposed data aggregation approach. Section IV analyzes the performance of the proposed data aggregation approach and compares it with tree and cluster-based approaches in terms of energy consumption and data transmission latency. Experimental (simulation) setup and results are presented in Section V. Finally, the summary of the paper and future works are presented in Section VI.

II. Related Work

Current research on big data analytics include distributed algorithms to process big data, network architecture and application of big data, MapReduce paradigm that works on big data [9]-[15]. The existing distributing algorithms to process and aggregation big data are mostly done at high performance big data server. These studies [9]-[15] do not consider data aggregation at multiple layers especially sensor data aggregation at the data collecting side as a way to reduce computational cost. Hence, we studied and presented a few literatures on sensor data aggregation as follows as a plan to integrate an improved sensor data aggregation approach in our proposed sensor-based big data framework.

Directed diffusion (DD) is a flat data aggregation approach where a node *A* broadcasts its interest and the node *B* that senses data related to the interest message transmits to *A* though multiple paths. Later, the node *A* selects the shortest path for further data transmission through a reinforcement packet. However, DD requires a large number of data transmissions. Hence, Cluster diffusion with Dynamic Data Aggregation Approach (CLUDDA) [3], [16] is introduced to only propagate event of interest and interest event between cluster head and cluster members. In case, the cluster head resides far from the cluster members, it consumes huge energy.

Tree-based approaches are good for small networks with fewer nodes. However, these algorithms suffer from a single point of failure where the failure of a single node disconnects the data transmission path from leaf node to the root. Among many tree-based approaches, energy aware distributed heuristic (EADAT) [17], Power efficient data gathering and aggregation protocol (PEDAP) [18] based on a spanning tree to maximize the lifetime of the network and Power-Aware PEDAP (PEDAP-PA) [18] are more popular. Chain-based data aggregation techniques, such as power efficient data gathering protocol for sensor information systems (PEGASIS), have been proposed [20] where each sensor transmits only to its closest neighbor. As this approach does not guarantee the shortest data transmission path from the furthest nodes of the chain to the sink a multiple-chain scheme is introduced in [20]. Again, this approach does not provide the shortest data transmission distance. Hence, the greedy chain construction algorithm, which constructs the chain by starting at the furthest node from the sink and considers it as a chain head, was proposed in [5]. Every time a non-chain node is added to the chain, this new node is considered as a new chain head until all nodes are added to the chain.

A multiple chain scheme has also been proposed in [22]. In this approach, the network is divided into four zones and each zone is centered at the node that is closest to the center of the sensing region. A linear that ends at the centre node is created for each zone. The multiple chain schemes aim to decrease the total distance of transmitting data as nodes broadcasts. In the greedy chain construction scheme proposed in [12], the process starts by selecting the chain head. The farthest node from the sink is selected as the chain head. At each step, a non-chain node, *A* is added to the chain head if *A* is closest to the chain head. The procedure stops whenever all nodes are added to the chain. This approach is further improved by including the non-chain node to the chain as a chain leader that provides the shorted distance as compared to other nodes if included into the chain as a leader.

In the grid-based data aggregation method [18], each grid has a data aggregator and all sensors in a grid transmit data to the grid aggregator while in the in-network data aggregation, data are aggregated at parent nodes as they are being transmitted towards sink at the root of the tree. The work in [5] presents a hybrid data aggregation scheme that combines the best features of grid-based and In-network aggregation schemes. The network topology is initially constructed based on in-network data aggregation approach. Once an event is detected by a sensor, the sensor follows in-network data aggregation scheme if the data is received from a static sensor application. If data is from a mobile sensor application, grid-based approach is used for data aggregation. Among other approaches, the work done in [26] introduces a cluster-based data aggregation approach where cluster head uses three different approaches to reduce redundant data collected from neighboring nodes (i.e., huge processing burden on cluster head), [27] introduces identity-based aggregate signature (IBAS) scheme for sensor-based secure data aggregation that provides data integrity as well as reduce bandwidth usage.

In sensor network, nodes receive data only when they are in active state that introduces the idea of properly utilizing the limited number of active time slots of sensor nodes with the goal of reducing data aggregation latency. The minimum latency aggregation schedule (MLAS) in most duty cycle WSN allows low latency and collision free aggregation schedule. However, this approach uses fixed structure aggregation methods and requires all sensor nodes are always awake. The work done in [28] introduces a distributed aggregation algorithm for duty-cycle WSNs, in which the aggregation tree and a conflict free schedule are generated simultaneously without using any fixed aggregation structure. The work done in [29] introduces an approximation algorithm to construct a maximum lifetime data aggregation tree that uses an adjustable transmission power level to achieve higher network lifetime while most work consider fixed transmission power. In [30], authors introduce a cluster-based approach for in-network aggregation. This approach uses an energy efficient routing strategy that uses multi-path routing tree and performs data fusion and data aggregation at intermediate nodes. While most data aggregation approached do not consider data security and privacy issues, Vakilinia et al. [31] presents data privacy preserving data aggregation/fusion approach for crowdsensing that uses linear transformation and homomorphic encryption scheme to obtain secured aggregated data. However, these approaches are complex and computationally expensive.

The work done in [32] presents several data fusion techniques such as approaches based on neural network, genetic algorithm, fuzzy logic, particle swarm optimization, steiner tree-based approach and data selection-based summation fusion. In [33] Yan M. introduces Forecast Algorithm of Data Aggregation (FTDA) data fusion algorithm based on the time prediction model, which predicts a time when data may differentiate from the data at current time. This model has the ability to proactively identify data redundancy and reduce energy consumption. However, approaches presented in [32], [33] work for small scale sensor networks, require more computational power and hence, have space to make them more energy efficient.

Most approaches that we have presented in this section do not consider selecting a fewer number of nodes as active nodes and allowing all other nodes to remain in sleep state (or idle) that reduce the network energy consumptions. Also they do not consider the type and priority of data packets for data aggregation. Hence, we introduce a multi-tier data aggregation approach that (1) uses both cluster and tree-based approaches, (2) selects only a few nodes as active node while keep all other nodes in sleep state, (3) assigns type and priority to each data packet.

III. Proposed Architecture and Approach

This section presents the high level architecture of the proposed data aggregation framework of big data along with the low level data aggregation and filtering scheme at sensor networks.

A. High Level Architecture

The proposed big data aggregation and filtering framework works in three layers, (1) Lower Layer: aggregates data at sensors (2) Middle Layer: aggregates data at base station (3) Upper Layer: aggregates data aggregation at big data server in distributed manner.

Fig. 1 illustrates such as a big data framework that only has three data communication layers. For instance, sensors at lower layers sense data and transmit those data to sink node or base station (BS). Then, the BS processes or aggregates data and transmit the aggregated data to the central big data sever through Internet. Finally, the big data server aggregates data by distributing it to commodity computers. Hence, the proposed hybrid data aggregation scheme has three data aggregation layers. The computational efficiency of big data sever at upper layer depends on data aggregation at data at middle and lower layers as low power nodes at these layers can aggregate and filter data to some extent even though nodes at upper the layer have higher computational power. However, existing big data aggregation approaches in literature are mostly only designed for upper layer at big data server. Hence, the computational cost or time at the server is not reduced as these approached do not consider any lower layers preprocessing of data (such as preprocessed at lower layers at sensors).

By designing efficient data aggregation approach at the lower level sensor nodes the overall computational costs at the upper layer big data server can be reduced, which is the objective of this paper as the data aggregation scheme reduces the volume of sensor's data that will be transmitted to the upper layer. Thus, this approach reduces data aggregation and processing overhead at the upper layer in NoSQL or other non-relational database systems for big data. The upper layer also consists of emergency response centre. The sink or base station at middle layer transmits emergency or time critical data to the emergency response centre before sending it to NoSQL database servers for processing/filtering and future storage.

Sensor networks are being used for many applications. These applications can be classified as (1) real-time and (2) non-real-time. Real-time applications such as health monitoring have more priority than non-real-time applications (i.e., real-time emergency data should have more priority than non-real-time data). Hence, data aggregation approaches should be designed considering the priority of sensor applications or data types. Most existing approaches [1], [4]-[8], [18], [23] do not consider this criteria to design a data aggregation approach.

Moreover, data processing at upper layer (i.e., at big data servers) should also consider the type of data so that data can be stored based on their categories for future use. Data aggregations at the lower and middle layers based on data types and sensor applications will ease the data processing at the upper layer. Thus, this paper introduces an energy efficient application dependent data aggregation approach for sensor-based big data frameworks. Sensors are programmed to have a data type field in their packets so that other sensors or devices that receive the data packet can identify the type of applications and perform data aggregation based on the data type [21]. This field also helps to store data at the appropriate locations in big data server for further processing and use.

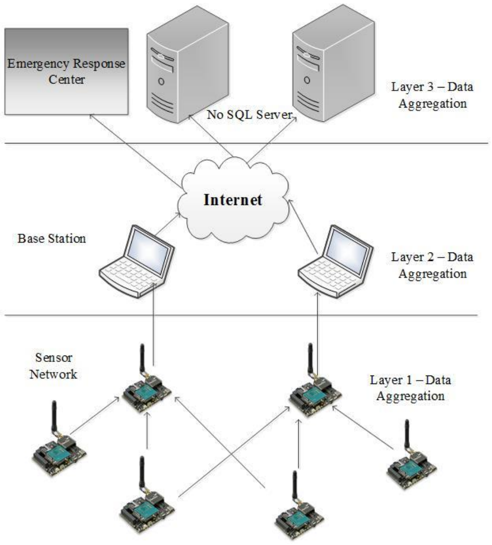

Emergency Response Center
No SQL Server
Layer 3 – Data Aggregation
Internet
Base Station
Layer 2 – Data Aggregation
Sensor Network
Layer 1 – Data Aggregation

Fig. 1. 3-tier sensor-based big data aggregation framework.

Routing protocols can be proactive (periodic) and reactive (event-based). For periodic routing protocols, data are sensed and transmitted periodically – at a certain time interval. In reactive routing protocols, data are transmitted only when a certain event is triggered. Sensors will also be programmed to contain a field (i.e., routing type) in their data packet that data transmission mode. For instance, if the routing field is set to 1 it will represent the periodic data transmission of emergency/real-time applications. Otherwise, data transmission will be event-based. Data aggregation at sensors also depends on this field. In the proposed approach, emergency real-time data will be only aggregated or filtered at sensors to avoid transmitting redundant data (i.e., data with the same information that has already been transmitted) that will reduce network energy consumption and also allow the sink to transmit data faster to the emergency response centre. Moreover, more data aggregation and processing takes place at the middle layer (at base station or sink node) compared to that at the lower layer (i.e., at the sensor) since sensors have limited power and processing capabilities. Thus, big data servers at the upper layer are expected to receive partially structured data to reduce the overall processing overhead of big data framework.

B. Proposed Hybrid Sensor Data Aggregation Scheme

The proposed hybrid data aggregation scheme classifies sensor-based applications into the following categories.

1) Real-time, emergency, time critical applications – such as traffic monitoring, battlefield surveillance and health monitoring.

2) Non-real-time applications – agriculture, air pollution monitoring.

The lower layer sensors transmit data to the upper layers through gateway nodes. Fig. 2 illustrates such a scenario. However, data aggregation approaches may achieve energy and computation efficiency using dynamic network topologies based on the requirement of sensors applications. For example, sensors are programmed to form cluster-based topology for emergency real-time applications and tree topology for non-real-time applications (details of cluster formation, tree formation and CH selections are presented in [24]). In cluster-based topology, sensors collect and transmit data at their allocated timeslot to the cluster head (CH). Then the CH transmits to the gateway and end station. As this type of topology ensures the minimum number of hops to transmit to the end node data aggregation using cluster-based approach is expected to achieve computational, data latency as well as energy efficiency. In cluster-based data aggregation, once a cluster is formed and CH is selected the CH selects a minimum number of nodes as active node for any time instant while other nodes remain in sleep state (or idle). We use the work done in [19] to select active nodes. Active nodes of a cluster sense and transmit data to CHs while aggregates and filters data to discard redundant data. On the other hand, idle nodes (in sleep state) do not perform data sensing, transmission and aggregation. By discarding a large number of redundant and useless data in emergency applications this approach ensures faster transmission of data to the central server [1].

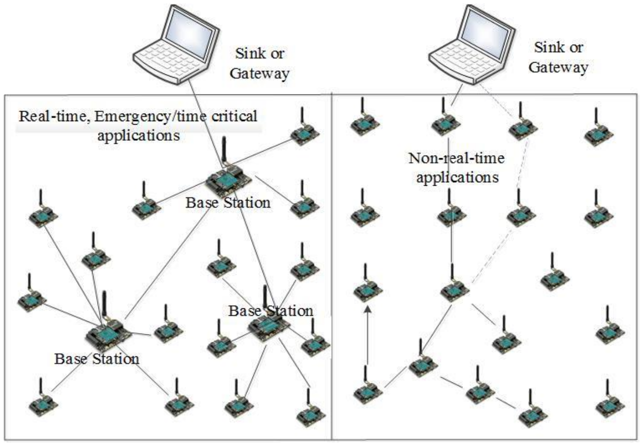

Sink or Gateway
Sink or Gateway
Real-time, Emergency/time critical applications
Non-real-time applications
Base Station
Base Station
Base Station

Fig. 2. Layer 1 data aggregation.

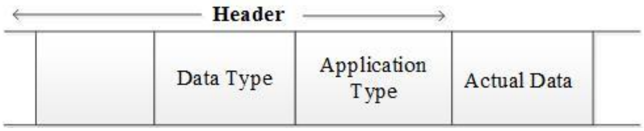


Fig. 3. Format of sensor data packet.

On the other hand, achieving energy efficiency is more important than achieving reduced end-to-end delay in non-real-time applications, such as agriculture, farming, pollution monitoring. Tree-based hierarchical topology may create the shorter path that uses more hops as compared to cluster-based topology. As distance is less the energy consumption will be less (energy consumption is directly proportional to the distance between two nodes [2], [3]). Thus, in tree-based data aggregation approaches, a sensor transmits data through the shortest path from itself to the sensor gateway.

In tree-based approach, nodes are identified to locate at different levels of the hierarchy considering the gateway node is the root of the hierarchy. Nodes residing one-hop away from the gateway can be considered to locate at the level 1 and so on. Then, the shortest path from the sensor gateway node to the active leaf nodes will be created using the method presented in [19]. Data transmission starts at sensors of the lowest level. For instance, active sensor nodes (or leaf nodes) sense and transmit the event of interest to the active nodes at the upper level. Parent nodes in this tree structure always perform data aggregation using different aggregation functions such as MAX, MIN, MEAN, MEDIAN and SUM and transmit again to the active nodes at the upper level until data reaches at the sensor gateway at root. Thus, this energy consumption of the active nodes in this approach is well distributed and the total network energy consumption is expected to be lower even though the number of hops from the sensor to the gateway is more as compared to the cluster-based counterpart of this proposed approach. However, the tree-based data aggregation may result in increased end-to-end data transmission delay as data from a node passes through several number of hops and is processed at each node for a certain time period. Thus, the proposed hybrid, dynamic and application-based data aggregation scheme offers a trade-off between energy efficiency and data transmission delay.

TABLE I. REPRESENTING TERMINOLOGIES BY SYMBOLS

Name of the terminology	**Symbol**
Sensor Network	$G(V, E)$
Non-real-time applications	nr
Real-time applications	r
Data type	Dt
Application type	AP
Tree-based topology	Tr
Cluster-based topology	Cl
Cluster head	CH
Level in a hierarchy	L
Active nodes	AN
Alternative nodes	Al
Gateway Node	G

Normally, sensor networks are used for a specific application by forming a specific network topology. Using the proposed data aggregation scheme, the sensors in a network can be reused to other applications and are able to change their topology if the application changes. Data packets have a number of fields and one field is used to set the application type. Once sensors receive a data packet from the gateway with the changed application field, it reconstructs the topology. Fig. 3 illustrates a sensor data packet that contains fields to identify data type and application type for the proposed data aggregation framework.

Sensor networks are mostly designed for a specific application and hence, a data aggregation scheme (cluster or tree-based) can be pre-established. However, the data aggregation scheme can also be constructed on-demand based on the types of packets that sensors transmit. This dynamism allows sensor networks to be used or re-used in multiple applications. Algorithm 1 presents the pseudo-code for the proposed sensor data aggregation approach. Table I lists the symbol used for different terms in Algorithm 1.

Algorithm I: Proposed Hybrid Data Aggregation Scheme

```
Randomly pick a node $i$
Set $node_i \leftarrow active$
$activenodeset \leftarrow \{i\}$
while $WholeNetCovered \neq TRUE$
    pick node j randomly
    if $j \neq i$ & not in activenodeset & $NetCoverage(node_i) \cap NetCoverage(node_j) = Null\ or\ Minimal$
then $node_j \leftarrow active$
        $activenodeset \leftarrow \{i, j\}$
    else
        $node_j \leftarrow alternative$
        $alternativeenodeset \leftarrow \{j\}$
    end if
end while
Remaining nodes $\in$ sleep-mode
If $AP = nr$ then
    $G(V, E) \leftarrow Tr$
    form shortest path with $AN$ in leaf nodes to $g$
    $AN$ at different $L$ transmit multi-hop
else if a $AP = r$ then
    $G(V, E) \leftarrow Cl$
    select CHs from $AN$s
    $AN$ in each cluster transmit towards $CH$s
end if
while $G(V, E)$ in work do
    if $AP = r$ & $Dt = r$ then
        for each $AN_i$ in cluster $j$ do
            transmit data to $CH$
            $CH$ filters redundant data & transmit to $g$
        end for
    else if $AP = r$ & $Dt = nr$ then
        reconstruct $G(V, E) \leftarrow Tr$
        aggregation level $\leftarrow L$i
        CH aggregates using MAX, MIN, SUM, REDUCE & other functions based on $AP$
        CH transmits aggregated data to $g$ directly or through other $CH$s
    else if $AP = nr$ & $Dt = nr$ then
        aggregation level $\leftarrow L$i
        $CH$ aggregates using MAX, MIN, SUM, REDUCE & other functions based on $AP$
        $CH$ transmits aggregated data to $g$ directly or through other $CH$s
    else if $AP = nr$ & $Dt = r$ then
        reconstruct $G(V, E) \leftarrow Cl$
        for each $AN_i$ in cluster $j$ do
            transmit data to $CH$
            $CH$ filters redundant data & transmit to $g$
        end for
    end if
end while
```

IV. Performance Analysis

In this section, the performance of the proposed data aggregation scheme will be analyzed in terms of networks energy consumption and data transmission delay. Then we will set up the network simulator based on some assumptions and measure the performance of the proposed hybrid data aggregation scheme as compared to the tree and cluster-based approaches.

A. Energy Model

The energy model in [2], [3] is used to evaluate the performance of the proposed data aggregation approach as we only consider data transmission and reception energy consumption in this evaluation. This model considers that energy consumption is proportional to data transmission distance. The energy consumption of a node for transmitting data of n_{data} bytes to another node, which are at distance d apart is

$$E_{TX} = n_{data} \times \varepsilon_{data} + n_{data} \times d^2 \times \varepsilon_{air} \tag{1}$$

However, the energy consumption of a node for receiving a data packet is independent of distance and is denoted as follow.

$$E_{RX} = n_{data} \varepsilon_{data} \tag{2}$$

Where ε_{data} is the energy consumption of a sensor node in its electronic circuitry and ε_{air} represent the energy consumptions in RF amplifiers for propagation loss.

B. Estimation of Energy Dissipation

Let us assume that the number of sensor applications = n_{app} and the number of non-real-time applications that use tree topology = n_{nr}

The number of real-time applications that use cluster-based topology = n_r .

$$\therefore n_{nr} + n_r = n_{app} \tag{3}$$

Let us assume that each network has the same number of nodes, n_{node}.

Therefore, the total number of nodes = $n_{node} \times n_{app}$.

1) Existing cluster-based method

Let us assume that the number of clusters in each network is n_{cl} .

Therefore, the number of nodes in each cluster,

$$n_{nodecl} = \frac{n_{node}}{n_{cl}} . \tag{4}$$

Let us assume that each network has 2 level or hierarchy. We denote these levels as L_1 and L_2 . Also, we consider that the level that is closer to the gateway is L_2 . So, the number of clusters in each level is $\frac{n_{cl}}{2}$.

Let us assume that the distance between an active member node and CH = d_{avg}

The average size of a data packet that is transmitted from a member node to CH is n_{data}.

Therefore, the total network energy consumption for transmitting a data packet to a cluster is

$$E_{TX1CL} = \left(\frac{n_{node}}{n_{cl}} - 1\right) \times (n_{data} \times \varepsilon_{data} + n_{data} \times d_{avg}{}^2 \times \varepsilon_{air}) \tag{5}$$

The energy consumption of a CH for receiving data from an active member node is

$$E_{RXCL} = \left(\frac{n_{node}}{n_{cl}} - 1\right) \times (n_{data} \times \varepsilon_{data}) \tag{6}$$

Similarly, the energy consumption of a CH to transmit data packet to the sensor gateway is given as

$$E_{TX2CL} = n_{agdatacl} \times \varepsilon_{data} + n_{agdatacl} \times d_{CH}{}^2 \times \varepsilon_{air} \tag{7}$$

Where the aggregated data size at CH is $n_{agdatacl}$ and the average distance between CH and sensor gateway is d_{CH}

Thus, the total transmission energy consumption in a cluster-based data aggregation scheme is

$$\begin{aligned} E_{TXCL} &= n_{cl} \times (E_{TX1CL} + E_{TX2CL}) = n_{cl} \times (n_{data} \times \varepsilon_{data} + n_{data} \times d_{avg}{}^2 \times \varepsilon_{air}) \\ &+ n_{cl} \times (n_{agdatacl} \times \varepsilon_{data} + n_{agdatacl} \times d_{CH}{}^2 \times \varepsilon_{air}) \end{aligned} \tag{8}$$

2) Proposed hybrid approach

This section presents the proposed data aggregation scheme both for when (1) modifications are done based on cluster-based topology for real-time applications, and (2) modifications are done based on tree-based topology for non-real-time applications.

Proposed approach is based on cluster-based topology for real-time applications

Let us assume that the number of nodes that reside in sleep mode = n_{idle}.

Therefore, the number of active nodes in a cluster including CH is

$$n_{activeprcl} = n_{nodeprcl} - n_{idle} . \tag{9}$$

If we substitute (9) into (5) we find the energy consumption of active nodes in a cluster for transmitting data to CH, which is given as follows:

$$E_{TX1PRCL} = (n_{nodeprcl} - n_{idle} - 1) \times (n_{data} \times \varepsilon_{data} + n_{data} \times d_{avg}{}^2 \times \varepsilon_{air}) \tag{10}$$

Similarly, if we substitute (9) into (6) we obtain the energy consumption of a CH for receiving data from an active member node of a cluster, which is given as follows:

$$E_{RXPRCL} = (n_{nodeprcl} - n_{idle} - 1) \times (n_{data} \times \varepsilon_{data}) \tag{11}$$

The energy consumption of a CH for transmitting a data packet to the sensor gateway is given as

$$E_{TX2PRCL} = n_{agdataprcl} \times \varepsilon_{data} + n_{agdataprcl} \times d_{CH}{}^2 \times \varepsilon_{air} \tag{12}$$

In (12) the aggregated data size at a CH is $n_{agdata-prcl} \le n_{agdata-cl}$ and the average distance between a CH and sensor gateway is d_{CH} .

Thus, the total transmission energy consumption in the cluster-based proposed data aggregation approach is:

$$\begin{aligned}E_{TXPRCL} &= E_{TX1PRCL} + E_{TX2PRCL} = (n_{nodeprcl} - n_{idle} - 1)\\ &\times (n_{data} \times \varepsilon_{data} + n_{data} \times d_{avg}{}^{2} \times \varepsilon_{air}) + n_{agdataprcl} \times \varepsilon_{data}\\ &+ n_{agdataprcl} \times d_{CH}{}^{2} \times \varepsilon_{air}\\ &= n_{activeprcl} \times (n_{data} \times \varepsilon_{data} + n_{data} \times d_{avg}{}^{2} \times \varepsilon_{air}) +\\ &n_{agdataprcl} \times \varepsilon_{data} + n_{agdataprcl} \times d_{CH}{}^{2} \times \varepsilon_{air}\end{aligned} \tag{13}$$

Proposed approach is based on tree-based topology for non-real-time applications

Again let us assume that the number of levels from leaf nodes to the sensor gateway is 2.

The number of active nodes in each level is $n_{activeprtr}$.

The proposed data aggregation approach that uses tree topology creates the shortest path from a leaf node to the sensor gateway. We assume that the size of a data packet that is sensed at a leaf node is $n_{dataprtr}$ and the size of aggregated data packets at the upper level nodes is $n_{agdataprtr}$. The average distance between the nodes at level 1 and level 2 is d_{L1prtr} and between the nodes at level 2 and the sensor-gateway is d_{L2prtr}

Therefore, the average distance (shortest) between the leaf node and the sensor-gateway node is given as

$$d_{L1prtr} + d_{L2prtr} \tag{14}$$

Thus, the energy consumption of active nodes at $L1$ for transmitting data to the nodes at $L2$ is given as

$$E_{TX1PRTR} = n_{activeprtr} \times (n_{data} \times \varepsilon_{data} + n_{data} \times d_{L1-prtr}{}^{2} \times \varepsilon_{air}) \tag{15}$$

The energy consumption of active nodes at $L2$ for receiving data from nodes at $L1$ is given as follows

$$E_{RXPRTR} = n_{activeprtr} \times (n_{data} \times \varepsilon_{data}) \tag{16}$$

Similarly, the energy consumption of all active nodes at $L2$ for transmitting data packets to the sensor-gateway is given as

$$\begin{aligned}E_{TX2PRTR} &= n_{activeprtr} \times (n_{agdataprtr} \times \varepsilon_{data}\\ &+ n_{agdataprtr} \times d_{L2prtr}{}^{2} \times \varepsilon_{air})\end{aligned} \tag{17}$$

Thus, the total energy consumption for transmitting data in the tree-based proposed data aggregation approach is given as:

$$\begin{aligned}E_{TXPRTR} &= E_{TX1PRTR} + E_{TX2PRTR} = n_{activeprtr} \times\\ &(n_{data} \times \varepsilon_{data} + n_{data} \times d_{L1prtr}{}^{2} \times \varepsilon_{air}) + n_{activeprtr} \times\\ &(n_{agdataprtr} \times \varepsilon_{data} + n_{agdataprtr} \times d_{L2prtr}{}^{2} \times \varepsilon_{air})\\ &= n_{activeprtr} \times \varepsilon_{data}(n_{data} + n_{agdataprtr}) + n_{activeprtr} \times \varepsilon_{air}\\ &(n_{data} \times d_{L1prtr}{}^{2} + n_{agdataprtr} \times d_{L2prtr}{}^{2})\end{aligned} \tag{18}$$

3) Existing tree-based method

Let us assume that the number of nodes at level of the tree = n_{tr} and the number of hops to transmit data from a leaf node to the sensor-gateway = 2

Let us assume that the average distance from L1 nodes to L2 nodes = d_{L1tr}

The average distance from L2 nodes to sensor-gateway= d_{L2tr}

The size of data sensed at the lowest level leaf nodes= n_{datatr}.

Then, the size of aggregated data at L2 nodes = $n_{agdatatr} \geq n_{agdataprtr}$ (19)

In this approach, all nodes are kept in the inactive mode. Transmission energy consumption of L1 nodes as given in (20).

$$E_{TX1TR} = n_{tr} \times (n_{data} \times \varepsilon_{data} + n_{data} \times d_{L1tr}{}^{2} \times \varepsilon_{air}) \tag{20}$$

Similarly, reception energy consumption of nodes at L2 is given as:

$$E_{RX-TR} = n_{tr} \times (n_{data} \times \varepsilon_{data}) \tag{21}$$

And energy consumption for transmitting data from nodes at L2 to the sensor-gateways deduced using (22).

$$E_{TX2TR} = n_{tr} \times (n_{agdatatr} \times \varepsilon_{data} + n_{agdatatr} \times d_{L2tr}{}^{2} \times \varepsilon_{air}) \tag{22}$$

Thus, the total transmission energy consumption is

$$\begin{aligned}E_{TXTR} &= E_{TX1TR} + E_{TX2TR} = n_{tr} \times (n_{data} \times \varepsilon_{data}\\ &+ n_{data} \times d_{L1tr}{}^{2} \times \varepsilon_{air}) + n_{tr} \times (n_{agdatatr} \times \varepsilon_{data}\\ &+ n_{agdatatr} \times d_{L2tr}{}^{2} \times \varepsilon_{air})\\ &= n_{tr} \times \varepsilon_{data}(n_{data} + n_{agdatatr}) +\\ &n_{tr} \times \varepsilon_{air}(n_{data} \times d_{L1tr}{}^{2} + n_{agdatatr} \times d_{L2tr}{}^{2})\end{aligned} \tag{23}$$

4) Comparison of energy consumption among cluster-based, tree-based and hybrid approach

Case 1: Non-real-time sensor applications using tree-based topology.

Since it is known that $n_{activeprtr} < n_{tr}$ we can conclude from (18) and (23) that

$$E_{TXPRTR} < E_{TXTR} \tag{24}$$

Similarly, we can conclude from equations (16) and (21) that

$$E_{RXPRTR} < E_{RXTR} \tag{25}$$

Case 2: Real-time sensor applications that use cluster-based topology.

It has been shown that $n_{active-prcl} < n_{cl}$, so, we can conclude from (8) and (12) that

$$E_{TXPRCL} < E_{TXCL} \tag{26}$$

Similarly, $E_{RXPRCL} < E_{RXCL}$ (27)

Case 3: Comprises of both real-time and non-real-time applications. Let us assume that the number of non-real-time and real-time applications are n_1 and $n_{2,}$ respectively. Then, the transmission energy consumption for the proposed data aggregation approach will be given as

$$n_1 \times E_{TXPRTR} + n_2 \times E_{TXPRCL} \tag{28}$$

Where the transmission energy consumption for the cluster-based approach will be denoted as

$$n_1 \times E_{TXCL} + n_2 \times E_{TXCL} \tag{29}$$

Similarly, the transmission energy consumption for the tree-based approaches will be given as

$$n_1 \times E_{TXTR} + n_2 \times E_{TXTR} \tag{30}$$

Since $E_{TXPRCL} < E_{TXCL}$ comparing (28) and (29) we find that transmission energy consumption of the proposed approach will be less than the transmission energy consumption of the cluster-based approach. Similarly, as $E_{TXPRTR} < E_{TXTR}$ comparing (28) and (30), we find that transmission energy consumption will be less than that of tree-based approach.

We will find the similar result for data reception energy consumption (i.e., reception energy consumption of the proposed approach will be less than that of the cluster and tree-based approaches)

C. Analysis on Data Transmission Latency

In the cluster-based method, the active member nodes of a cluster transmit data packets to the CH. Then the CH aggregates and transmits the processed data to the sensor-gateway. If the time allocated to the active member node and CH are T_c and T_{ch}, $T_{ch} > T_c$ as the CH performs data sensing, data transmission, reception and aggregation.

The data transmission latency for the cluster-based method will be as presented in (31).

$$D_{cl} = n_{cl} \times \left(\left(\frac{n_{node}}{n_{cl}} - 1 \right) \times T_c + T_{ch} \right) \tag{31}$$

1) Proposed hybrid approach

The data transmission latency for the proposed cluster-based approach

$$D_{prcl} = n_{cl} \times \left(n_{activeprcl} \times T_c + T_{ch} \right) \tag{32}$$

The data transmission latency for the proposed tree-based method is presented in (33).

$$D_{prtr} = n_{activeprtr} \times T_{L1prtr} + n_{activeprtr} \times T_{L2prtr} \tag{33}$$

The number of active nodes in each level of the proposed tree-based method is presented in (34).

$$n_{activeprtr} < n_{no\det r} \tag{34}$$

2) Existing tree-based method

The number of nodes in each level is assumed to be same $= n_{no\det r}$ and duration of timeslot allocated to each node at the lowest level is T_{L1tr}.

The duration of timeslot allocated to each node at the upper level is $T_{L2tr} > T_{L1tr}$.

This is because the upper level nodes perform data aggregation and transmit aggregated data to the sensor-gateway.

Thus, the data transmission latency for tree-based approach will be

$$D_{tr} = n_{no\det r} \times T_{L1tr} + n_{no\det r} \times T_{L2tr} \tag{35}$$

3) Comparison of data transmission latency

Case 1: If all sensor applications of the proposed approach are non-real-time and use tree-based topology

By comparing (41), (42) and (43) we can conclude that $D_{prtr} < D_{tr}$

Case 2: If n_1 sensor applications of the proposed approach are non-real-time and n_2 applications are real-time, the data transmission latency will be

$$\begin{aligned} n_1 \times D_{prtr} + n_2 \times D_{prcl} &= n_1 \times (n_{activeprtr} \times T_{L1prtr} \\ &+ n_{activeprtr} \times T_{L2prtr}) + n_2 \times n_{cl} \times \left(n_{activeprcl} \times T_c + T_{ch} \right) \end{aligned} \tag{36}$$

For tree-only approach the data transmission latency will be

$$\begin{aligned} (n_1 + n_2) \times D_{prtr} &= n_1 \times (n_{no\det r} \times T_{L1tr} + n_{no\det r} \times T_{L2tr}) + \\ & n_2 \times (n_{no\det r} \times T_{L1tr} + n_{no\det r} \times T_{L2tr}) \end{aligned} \tag{37}$$

As $n_{activeprtr} < n_{no\det r}$ and $T_{L1prtr} < T_{L1tr}$ we can conclude from (36) and (37).

$D_{pr} < D_{tr}$ (i.e., data transmission latency of proposed approach is lower than tree-based approach).

$D_{pr} < D_{cl}$ (i.e., data transmission latency of proposed approach is lower than cluster-based approach).

From the above analysis, we conclude that the energy consumption and data transmission delay of the proposed sensor-based data aggregation approach at layer 1 is less than that of traditional cluster and tree-based schemes.

D. Computational Complexity

If the number of active nodes at each level *l* in the proposed tree-based approach is $n_{l(activeprtr)}$ and the number of levels in the network is L_{prtr} the total number of active nodes will be $\sum_{l=1}^{L_{prtr}} n_{l(activeprtr)}$.

Thus, the number of packets transmitted by each active node of the network at their predefined timeslot is $\sum_{l=1}^{L_{prtr}} n_{l(activeprtr)}$.

If we define the complexity of the algorithm based on the number of message transmission, which is a function of the number of nodes from each level at the predefined timeslot then the processing complexity of the proposed approach based on tree topology is O (*n*) where *n* is the number of nodes transmitting data packets.

Similarly, we can show that the processing complexity of proposed approach based on clustering will O (*n*).

V. Validation of the Proposed Approach

To validate our proposed hybrid data aggregation and filtering technique for sensor-based big data frameworks we considered the scenarios presented in the section.

A. Simulation Setup

We designed and implemented a simulator to implement the proposed data aggregation approach using C programming language rather than using the existing simulators, NS-2, OPNET, NS-3 many sensor network and big data functionalities are not available in these simulators. Moreover, we have more control on implementing the new concept of sensor-based big data.

Real experiments or testbed always give accurate result as compared to simulation. However, real experiments are not always possible due to the unavailability of sensors and other components. Hence, simulation is being used to replace experimental work in sensor networks and other fields to a great extent. Hence, we decided to perform simulation to evaluate the performance of the proposed data aggregation scheme that works at layer 1 of the big data architecture and compared with the traditional cluster and tree-based approach as presented before. We use network energy consumption, network lifetime and data transmission latency as the performance metrics. Each time the simulator was run for a certain number of rounds and we run the simulator a certain number of times. The outputs are calculated as an average of these results. We define the performance metrics and related terms as follows:

Round – is a period of time comprises a number of network setup and operation phases.

Data transmission latency – is considered as the end-to-end data transmission delay, i.e., the time required to transmit data from an active node to the sensor gateway or base station.

Energy consumption – is the total energy consumed by a sensor to transmit, receive and aggregate data.

We simulated an area of size 100 meters x 100 meters as the network size. As this network area is considered as small, the network is divided into only 4 clusters and 20-30 nodes are randomly deployed on an average into each cluster (100 nodes in total into the network). For this small network area deploying 100 sensors can be considered as a large number of sensors that collect huge amount of data, i.e., big data. The proposed data aggregation approach still works even if we increase the size of the network and the number of sensors in this ratio (large scale). Simulation parameter and their respective values of our paper [25] are also used in this paper.

The simulator was run for rounds between 5000 and 30000 for different experiments to compare energy consumption between low (5000 rounds) and high (30,000 rounds) number of network setup phases. The sensor gateway is placed at the outside of the network area which is located at the co-ordinate (55, 101). During the network operation phase, cluster head allocates a number of timeslots to each node. However, each nodes receives different number of timeslots based on their distance or level from the sensor gateway. For instance, nodes which are closer to the sensor gateway require more time to sense data, receive data from lower levels of nodes, aggregate and transmit data. Hence, these nodes require more time (i.e., timeslots) as compared to the nodes that reside far from the sensor gateway at the lower levels. Table I lists the parameters and their values that are used in the simulation.

B. Simulation Results

Fig. 4 shows that the energy consumption of the proposed data aggregation approach is much lower than that in traditional tree and cluster-based approaches because the proposed approach selects only a few active nodes and most other nodes remain in idle state whereas the traditional approaches consider all nodes as active. Moreover, the proposed approach uses both cluster and tree-based approaches based on the type of data it senses and balances the energy consumption. Fig. 5 demonstrates that the data transmission latency of the proposed data aggregation scheme are less than that of the tree and cluster-based data aggregation approach because the CH receives data from a few active cluster member nodes in cluster-based approach and the parent node receives data from a few active child node, which require less time for the CH and a parent node to process and further transmit data to the next level.

From the result presented in Fig. 4 about the network energy consumption we can deduce that the network lifetime of the proposed scheme is expected to be more than those of cluster and tree-based approaches. Figure 6 demonstrates our claim that the network lifetime of the proposed hybrid data aggregation approach is much more than that in the traditional tree-based and cluster-based data aggregation approaches. We can further justify the presented results as follows:

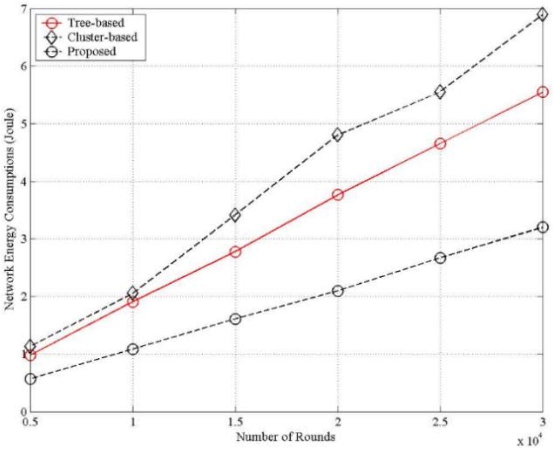


Fig. 4. Comparison of network energy consumption.

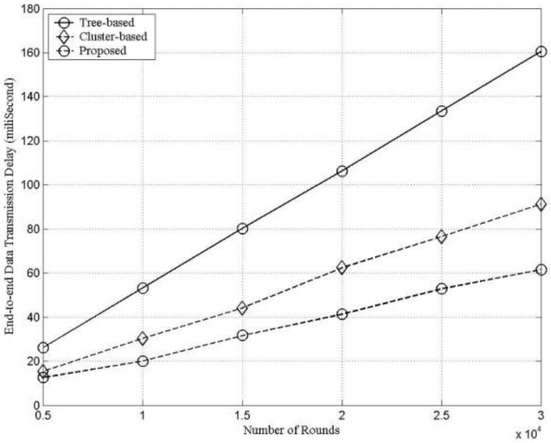


Fig. 5. Comparison of data transmission delay.

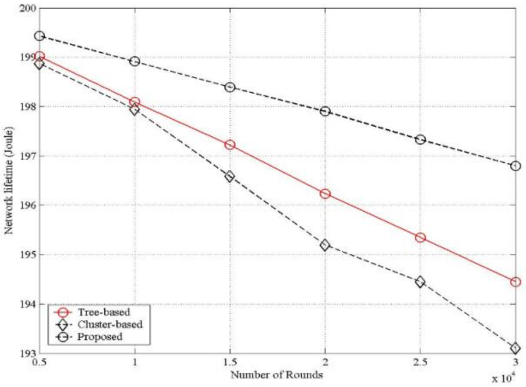


Fig. 6. Comparison of network lifetime.

In tree-based data aggregation schemes, upper level nodes wait until nodes at the lower levels transmit data to the upper levels. This results in higher data transmission latency. Moreover, a large number of active nodes at each level results in data redundancy, and data processing overhead. Cluster-based approach allows all cluster members to transmit data to the cluster head (CH). Thus, the CH requires much energy to process the received data. As some of the CHs might be far away from the sensor gateway, it consumes much energy of the CH to transmit the large aggregated data. In its own case, the proposed data aggregation scheme selects only a few active nodes that cover the whole network, this provides lower processing overhead and reduce the total network energy consumption (i.e., higher network lifetime). Processing and transmitting data from a fewer active nodes will also result in less data transmission latency. In summary, Table II compares the existing tree and cluster-based data aggregation approaches with the proposed hybrid approach based on different features.

VI. Conclusion and Future Works

We introduced a sensor-based big data aggregation approach in this paper. This approach works in multiple layers. However, we focus on aggregating redundant and unstructured sensors data at the lowest level of this framework at sensor nodes. The proposed hybrid data aggregation scheme uses either an efficient cluster-based data aggregation when data are transmitted from real-time or emergency sensor applications or a tree-based approach for non-real-time sensor applications. Experimental results demonstrate that the proposed hybrid and dynamic data aggregation scheme is better than traditional cluster and tree-based schemes in terms of network energy consumption, network lifetime and data transmission latency. This results in less amount of (unprocessed) data by big data server at upper layers to further faster data aggregation and filtering. In future, we plan to design and implement and efficient (computational) data aggregation scheme for upper layers at big data server. Also, we plan to implement the proposed approach in testbed (real experiments) and compare with more existing approaches to justify its effectiveness. Securing sensor data aggregation approaches against attacks, i.e., Sybil, wormhole, blackhole, bogus information, modification of sequence number through the use of public and private key cryptography and encryption mechanisms is significantly important even though those approaches require more computations. Hence, we plan to implement computation-efficient secure data aggregation approaches as part of our future research in this direction.

TABLE II. Comparison of Different Data Aggregation Methods

Features	Tree – based	Cluster-based	Proposed
Flooding interest propagation	√	X	X
Initially, sink receives data through multiple paths	√	X	X
All nodes in the network are active (i.e., they sense, send and transmit data)	√	√	X
A few active nodes cover the whole network area	X	X	√
Form clusters and send event of	X	√	√

interest to CH			
Dynamic data aggregation	X	X	√
Static data aggregation	√	√	X
Support fault tolerance	X	X	X
Tree structure with a single point of failure	√	X	X
Name of existing approaches	DD, FEDA, DABDR, TAG	CLUDDA SUMAC, OCABTR	PROPOSED HYBRID

References

[1] Karim L., Nasser N. and Salti T., "Routing on Mini-Gabriel Graphs in Wireless Sensor Networks", *IEEE WiMob*, pp. 105-110, China, Oct 2011.

[2] Heinzelman W.B., "Application Specific Protocol Architectures for Wireless Networks", *PhD thesis*, Massachusetts Institute of Technology, June 2000.

[3] W. R. Heinzelman, A. Chandrakasan and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, 2000, pp. 10 pp. vol.2. doi: 10.1109/HICSS.2000.926982

[4] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann and F. Silva, "Directed diffusion for wireless sensor networking," in *IEEE/ACM Transactions on Networking*, vol. 11, no. 1, pp. 2-16, Feb 2003.

[5] R. Rajagopalan and P. K. Varshney, "Data-aggregation techniques in sensor networks: A survey," in *IEEE Communications Surveys & Tutorials*, vol. 8, no. 4, pp. 48-63, Fourth Quarter 2006.

[6] Takaishi, D., Nishiyama, H., Kato, N., Miura, R., "Toward Energy Efficient Big Data Gathering in Densely Distributed Sensor Networks," in *Emerging Topics in Computing, IEEE Transactions on* , vol.2, no.3, pp.388-397, Sept. 2014.

[7] Paulo Jesus, Carlos Baquero and Paulo Sergio Almeida, "A Survey of Distributed Data Aggregation Algorithms", *IEEE Communication Surveys and Tutorials*, Vol. 17, No. 1, First Quarter 2015.

[8] Yu Du, Fengye Hu, Lu Wang and Feng Wang, "Framework and challenges for Wireless body area networks based on big data," in *Digital Signal Processing* (*DSP*), *2015 IEEE International Conference on*, pp.497-501, 21-24 July 2015.

[9] Andreu-Perez, J.; Poon, C.C.Y.; Merrifield, R.D.; Wong, S.T.C.; Yang, G.-Z., "Big Data for Health," in *Biomedical and Health Informatics, IEEE Journal of*, vol.19, no.4, pp.1193-1208, July 2015.

[10] Chao Wu, Birch, D., Silva, D. Chun-Hsiang Lee, Tsinalis, O. and Guo, Y., "Concinnity: A Generic Platform for Big Sensor Data Applications," in *Cloud Computing, IEEE*, vol.1, no.2, pp.42-50, July 2014.

[11] Fu Xiao, Chongshen Zhang, and Zhijie Han, "Big Data in Ubiquitous Wireless Sensor Networks", *International Journal of Distributed Sensor Networks*, Volume 2014, Article ID 781729.

[12] Michael, K., Miller, K.W., "Big Data: New Opportunities and New Challenges", in *Computer*, vol.46, no.6, pp.22-24, June 2013.

[13] Hunter, P., "Journey to the centre of big data," in *Engineering & Technology*, vol.8, no.3, pp.56-59, April 2013.

[14] Fan Ye, Alvin Chen, Songwu Lu, Lixia Zhang, "A Scalable Solution to Minimum Cost Forwarding in Large Sensor Networks", *10th International Conference on Computer Communications and Networks* (*ICCCN2001*), Scottsdale, Arizona USA. October 15-17, 2001.

[15] Fan Ye, Haiyun Luo, Jerry Cheng, Songwu Lu, and Lixia Zhang, "A two-tier data dissemination model for large-scale wireless sensor networks", In *Proceedings of the 8th annual intl conference on Mobile computing and networking* (MobiCom '02). New York, USA, pp. 148-159, 2002

[16] S. Chatterjea and P. Havinga, "A Dynamic Data Aggregation Scheme for Wireless Sensor Networks". *In ProRISC 2003, 14th Workshop on Circuits, Systems and Signal Processing*, 26-27 November 2003, Netherlands.

[17] Ding, M.; Xiuzhen Cheng; Guoliang Xue, "Aggregation tree construction in sensor networks," in *Vehicular Technology Conference, 2003. VTC 2003-Fall. 2003 IEEE 58th*, vol. 4, pp. 2168-2172 6-9 Oct. 2003.

[18] Hüseyin Özgür Tan and Ibrahim Körpeoğlu, "Power efficient data gathering and aggregation in wireless sensor networks" *SIGMOD Rec.* 32, vol. 4, pp. 66-71, December 2003

[19] L. Karim, N. Nasser, T. Sheltami. A fault-tolerant energy efficient clustering protocol of a wireless sensor network. *Wireless Communications & Mobile Computing*, vol. 14, Issue 2, 2012, pp. 175-185.

[20] Ahmed A. Ahmed, Hongchi Shi, and Yi Shang, "Survey on Network Protocols for Wireless Sensor Networks," in *Proc. Intl. Conf. Information Technology: Research and Education*, 11-13 Aug. 2003.

[21] Mohammed S. Al-kahtani, "Efficient Cluster-Based Sleep Scheduling for M2M Communication Network", *Arabian Journal for Science and Engineering*, August 2015, Vol 40, Issue 8, pp. 2361-2373.

[22] Harichandan, P., Jaiswal, A., and Kumar, S., "Multiple Aggregator Multiple Chain routing protocol for heterogeneous wireless sensor networks," in *Signal Processing and Communication (ICSC), 2013 International Conference on*, vol., no., pp.127-131, 12-14 Dec. 2013.

[23] Karim L., Nasser N., Abdulsalam H. and Moukadem I., "An Efficient Data Aggregation Approach for Large Scale Wireless Sensor Networks", *GLOBECOM 2010*, pp. 1-6, Miami, USA.

[24] Karim, L., Mahmoud, Q. H., Nasser, N., and Khan, N, "An integrated framework for wireless sensor network management", *Wireless Communications and Mobile Computing*, 2014, *14*(12), 1143-1159.

[25] Lutful Karim and Mohammed S. Al-kahtani, "PDDA: Priority-based, Dynamic Data Aggregation Approach for Sensor-based Big Data Framework", *The 7th IEEE Annual Information Technology, Electronics and Mobile Communication Conference* (IEEE IEMCON), University of British Columbia, Vancouver, Canada, 13 - 15 October 2016, pp. 1-7.

[26] H. Harb, A. Makhoul, S. Tawbi and R. Couturier, "Comparison of Different Data Aggregation Techniques in Distributed Sensor Networks," in *IEEE Access*, vol. 5, pp. 4250-4263, 2017.

[27] L. Shen, J. Ma, X. Liu, F. Wei and M. Miao, "A Secure and Efficient ID-Based Aggregate Signature Scheme for Wireless Sensor Networks," in *IEEE Internet of Things Journal*, vol. 4, no. 2, pp. 546-554, April 2017. doi: 10.1109/JIOT.2016.2557487

[28] Q. Chen, H. Gao, S. Cheng, J. Li and Z. Cai, "Distributed non-structure based data aggregation for duty-cycle wireless sensor networks," *IEEE INFOCOM 2017*, Atlanta, GA, 2017, pp. 1-9.

[29] H. C. Lin and W. Y. Chen, "An Approximation Algorithm for the Maximum-Lifetime Data Aggregation Tree Problem in Wireless Sensor Networks," in *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3787-3798, June 2017. doi: 10.1109/TWC.2017.2688442

[30] R. Vinodha and S. Durairaj, "Data gathering cluster-based approach for in-network aggregation," *2016 Intl Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*, Pudukkottai, 2016, pp. 1-4.

[31] I. Vakilinia, J. Xin, M. Li and L. Guo, "Privacy-Preserving Data Aggregation over Incomplete Data for Crowdsensing," *2016 IEEE Global Communications Conference (GLOBECOM)*, Washington, DC, 2016, pp. 1-6.

[32] Jayashri B. S and G. R. Rao, "Reviewing the research paradigm of techniques used in data fusion in WSN," *2015 International Conference on Computing and Communications Technologies (ICCCT)*, Chennai, 2015, pp. 83-88. doi: 10.1109/ICCCT2.2015.7292724

[33] M. Yang, "Data aggregation algorithm for wireless sensor network based on time prediction," *2017 IEEE 3rd Information Tech and Mechatronics Engineering Conference (ITOEC)*, Chongqing, 2017, pp. 863-867.

Measuring the Effect of Use Web 2.0 Technology on Saudi Students' Motivation to Learn in a Blended Learning Environment

Sarah M.Bin-jomman, Mona Al-Khattabi
Department of Information Systems
College of computer and information sciences
Al-Imam Muhammad Ibn Saud Islamic University
Riyadh, Saudi Arabia

***Abstract*—Students' motivation to learn is the goal of the educational process around the world. There is a close link between learning outcomes and students' motivation to learn. Thus, the success of blended learning in Saudi higher education depends on not only using different teaching methods and massive expenditures on technology but also on students' motivation to learn. The main objective of this study is to measure the effect of using the Web 2.0 technology on students' motivation to learn in a blended learning environment through their attention, relevance, confidence, satisfaction inside in this environment. This study used a randomized experimental research design to examine differences in student's motivation based on their use of Web 2.0 tools in a blended environment in the Computer Science at Al-Imam Muhammad Ibn Saud Islamic University (IMSIU). This study adopted Keller's ARCS model of motivation to develop a comprehensive framework of factors that affect the use of Web 2.0 tools in blended learning environment. A questionnaire was conducted to collect data from students. Throughout our investigation, we found that there was a statistically significant difference at the level of 0.05 in overall student motivation between the experimental and control groups resulting from the using Web 2.0 tools technology. Moreover, students using Web 2.0 tools were found to exhibit a statistically significant higher degree of motivation. The results of this study can help decision makers readjust the learning strategy by realizing the importance of using Web 2.0 tools as the main platform in Saudi higher education.**

***Keywords*—Web2.0 tools; blended learning; motivation; ARCS model**

I. Introduction

The use of technology has become a necessity and a trend in all countries of the world in various sectors. In education sector technology and the internet have changed the concept of the traditional classroom, which in the past depended on the students and the teachers being within the boundaries of the university classrooms. E-learning (distance learning) made educational process effective and unique by giving students the opportunity to create a new style of learning environment [1]. Despite the benefits of e-learning, it lacks some matters such as a student's sense of isolation outside the framework of the traditional learning community [2]. Then generate the concept of blended learning, which fills the weakness of online learning because it serves as a bridge constituting a balance between the use of technology (such as computer, learning management system, e-mail) and traditional learning, which based on face-to-face meetings, to form integrated educational environments for students to create a meaningful educational community. The main objective of blended learning is to improve the ways and methods of instruction, increase flexibility and reduce time restrictions so a student can choose when and where to learn. According to [3], blended learning created motivation to learn within that environment since it fits the student's needs and circumstances.

Learners may face challenges in building technical skills and self-control in blended learning environment. According to the theory of learning by [3] "*learning is socially situated within community of practice*". Thus, learning requires social interaction through the integration of the Web 2.0 technologies as the main platform in a blended learning environment. Dealing with learning methods will change completely and education will become not only a way to get a certificate awarded to a student for gaining employment but also a motivation for learning to build an effective strategy that based on a learner-centered approach.

The goal of education is enhanced quality of education for everyone all over the world. Currently, students do not use wiki, blogs, social networks and other tools effectively in education. Web 2.0 is not just technology but considered as a dynamic social platform that changed the concept of people when using the web. It solves the problems by allowing people to share information on the Internet. [4] investigated the benefits of social networks such as Facebook and Twitter and his found many benefits such as facilitating the participation of content, improving discussions among students, enhancing intimacy between students, and meeting people who have the same interests. The variety of Web 2.0 technologies are also a benefit because they allow for a degree of user choice when deciding on the best method of learning [18]. There is lack of integration of Web 2.0 technology with universities in Saudi Arabia. This paper aims to examine the effect of the Web 2.0 technology on students' motivation to learn in a blended learning environment. To help decision makers realize the relationship between utilizing the Web 2.0

technology and students' motivation to learn to readjust an educational process strategy in Saudi Arabia.

The remainder of this paper is structured as follows: Section II explores the basic concepts of blended learning. Section III describe categories of blended learning environment. Section IV investigates the importance of use web 2.0 in education sector while Section V explains motivation to learn as the critical factor in education process. Finally, research method and model to measure the effect of using the Web 2.0 technology on students' motivation to learn in a blended learning environment are presented in Section VI. Research results are discussed in Section VII. Finally, Section VIII concludes the paper.

II. Basic Concepts of Blended Learning

In the past, traditional learning and online learning were separate from each other because of the difference in teaching methods. Innovations in technology and investments in education facilitated human interactions synchronously and asynchronously to integrate traditional learning into an online learning environment to constitute a blended learning concept [5]. There was no agreed standard definition for blended learning among academics and practitioners. Some researchers referred to blended learning a "buzzword", but many agree that a blended learning system refers to a combination of traditional face-to-face classroom instruction and any of a wide variety of computer-mediated instruction, including Web 2.0 [6]-[8]. The author in [9] defined blended learning as "*a system, which focuses on optimizing achievement of learning objectives, by applying the 'right' learning technologies, to match the 'right' personal learning style, to transfer the 'right' skills, to the 'right' person, at the 'right' time*". Furthermore, blended learning represents as "*the organic integration of thoughtfully selected and complementary face-to-face and online approaches and technologies*" [10].

III. Categories of Blended Learning Environment

Blended learning represents the convergence of digitally distributed enabled by ubiquitous broadband internet connectivity using a combination of synchronous and asynchronous programs and applications and the traditional classroom-learning environment. Blended environments have the potential for four dimensional integrations along the following continuums: (a) space (physical/face-to-face vs. distributed), (b) time (live/synchronous vs. asynchronous), (c) fidelity (rich/all senses vs. text only), and (d) humanness (high human/no machine vs. no human/high machine [11]. For example, an online course can add synchronous distributed interactions to the distributed space using live chats or webinars. Fidelity can be managed using multimedia presentation, videos, or guest speakers while humanness may be enhanced using virtual communities or group messaging technologies. As these four dimensions are used to create new solutions, blended instruction will evolve. Moreover, [6] argues that the learner has the opportunity to gain learning in an environment that combines face-to-face learning and online learning, and he suggests that it is not sufficient for the institution to have fully online learning that is largely separate from traditional learning. He classified blend into three categories. Table I summarizes the main differences between these categories.

Furthermore, Studies have shown that blended learning increased attention, relevance and satisfaction with online classes by developing their capacity for reflection. Multiple modes of delivery for course content, using a wiki, blog, social media and webinar, improved the learning process and improved participants' grades [12], [13]. In [14], the author conducted a comparison of a traditional and blended course in undergraduate Management. Blended class participants more easily formed social bonds with classmates, felt safe to communicate thoughts and ideas freely, and expressed a sense coherence with group goals. According to [15], blended courses create a sense of significant social learning compared with courses through online learning or traditional learning alone. They mentioned that blended learning should have the following three characteristics:

- The learner becomes the center of the learning process rather than the instructor.
- There is improved interaction between students, between student and content, and between student and instructor.
- There is an integration of formative (monitor student learning) and summative assessment (evaluate student learning) mechanics of the student and the teacher.

TABLE I. Blended Learning Main Categories

Main categories	**Description**	
	Definition	***Example***
Enabling blend	Provide the same opportunities or learning experience to learners by selecting their courses in different modes.	Face-to-face and online.
Enhancing blend	Allow adding of changes to the pedagogy but not fundamental change to the method of teaching.	In traditional face-to-face, including supplementary online resources for courses.
Transformation blend	Provide a fundamental change in the method of teaching by using new modern technological approaches in teaching.	Transform the model that depended receives information to model that construct knowledge.

Learners may face challenges in building technical skills and self-control in blended learning environment. Thus, learning requires social interaction through the integration of the Web 2.0 technologies as the main platform in a blended learning environment [3]. Dealing with learning methods will change completely and education will become not only a way to get a certificate awarded to a student for gaining employment but also a motivation for learning to build an effective strategy that based on a learner-centered approach.

They offer the best opportunity to use learning settings based on student-centered strategies [15].

IV. Web2.0 Technology as Platform in Education

The goal of education is enhanced quality of education for everyone all over the world. Web 2.0 refer to group of web-based technology that promote user-generated content, sharing information and users' capabilities communication [16]. It has many tools and services, such as blogs, wikis, social networks, Ajax, RSS, tagging, etc. Dale Dougherty, a vice president of O'Reilly Media Inc. in a conference, which discussed the future of the web, first appeared the term web 2.0 in 2004 [17].

The authors in [18] indicated four-dimensional of web 2.0 that integration along the following continuums: interactivity, real-time user control, social participation (sharing), and user–generated content. These four-dimensional promote student participation and engagement, which are essential to the instructional learning dynamic. For example, Twitter are environment depends on the generation of content and collaboration user between user which mean sharing opinions, posting, comment, assessment, discussions, and exchange of experiences. Web2.0 technology considers personalization to the learner by matching learning preferences or needs and tracking behavior to the specific interests of different learners it is not just technology but is considered a dynamic social platform that changed the concept of people when using the web [19].

Faculty and instructional designers are increasingly expected to incorporate Web 2.0 technologies and applications into their teaching. Educators sense a sort of "*moral panic*" in higher education to change teaching and learning practices to meet the demands of the online generation [18]. The trend toward technology convergence was consistent with the core educational principles that academic outcomes are improved by increasing student engagement and improved social interactions, both learner-learner and learner-instructor interactions. The author in [20] reported that learner-content interactions are more important indicators of learner outcomes than learner-instructor or learner-learner interactions. Moreover, it solves the problems by allowing people to share or understand information via the Internet. Web 2.0 applications for education were designed to improve student engagement using all three types of interactions. That confirmed the importance of Web 2.0 by provides tools that support the teaching and learning process, which helps students improve their performance not only in boundary universities but also from these technologies.

V. Motivation to Learn

Students' motivation to learn is the goal of the educational process in new methods of learning, such as blended learning. There is a close link between learning outcomes and students' motivation to learn. Motivation considered as "*a process that requires students to perform physical or mental activities for achieving their goals*" [21]. It is a critical element for learning and accounts for between 16% and 38% of the variance in studies on university student learning variance [22], [23].

Motivation remains the critical factor in learning despite developments in pedagogy and drastically improved educational technologies. As such, studies to improve teaching and learning based on technology adoption should consider the principles of motivation and their application to technology adoption. There are two mainly type of motivation: intrinsic and extrinsic. Intrinsic motivations come from inside the individual. Thus, inner satisfaction can drive a student to learn and achieve success. Extrinsic motivations come from outside the individual, such as a student's success in learning by obtaining external rewards [24]. There are many studies showed that the students' intrinsic and extrinsic motivations have a significant impact on the educational process. Especially, collaborative-based learning can be an effective way for the learner to be the center of the learning process. Learning environments should foster intrinsic learning motivation.

A. Keller's Model of Motivational Design ARCS Model

Enhancing student learning motivation and participation is crucial for the teaching and learning of new knowledge or skills since motivation would affect how instructors and students interact with learning materials. In the era of Web 2.0 could be a potentially novel method to engage instructors and students in meaningful teaching and learning activities. The author [25] developed the ARCS model which stands of (attention, relevance, confidence and satisfaction) model to understand, predict, and develop strategies for improving motivation to learn.

ACRS model is based on expectancy – value theory of motivation which develops motivational strategy [26]. This theory has two components namely the student can expect to succeed in the learning and the value of learning to the learner. Table II summarizes the characteristics of the ARCS model. The ARCS model synthesizes behavioral, cognitive, and effective learning theories into a single unified framework for examining motivation and academic success. Keller's ARCS Model of Motivational Design suggests that student motivation may be affected by to improving the motivational appeal of instruction [27].

TABLE II. Characteristics of the ARCS Model

Characteristics of the ARCS Model	**Definition**
Attention	Educators must have the student's fairly constant attention.
Relevance	How closely the course content connects each student's personal experiences, hopes, dreams, or desires.
Confidence	Function of establishing student's positive expectation for success.
Satisfaction	Positive feelings associated with the relationship between the amount of effort expended and one's accomplishments and learning experiences.

VI. Methodology and Hypotheses

The aim of this study is to investigate the effect of using Web 2.0 technology to motivate students to learn within a blended learning environment. This study adopted ARCS model in order to develop a comprehensive framework of factors that effect of use Web 2.0 in blended learning environment. The study uses an experimental research design with a control group and an experimental group. This research design is used to fulfill the objective of gauging the variation in a phenomenon (Web 2.0 technology), as well as measure changes in outcomes (levels of learning motivation). The control group is taught using traditional learning without integrating Web 2.0 tools. The experiment group is taught using traditional learning with integrated Web 2.0 tools.

A. Hypotheses of the Study

Following hypotheses were constructed to examine the effect the Web 2.0 technology in blended learning on students' motivation to learn in Saudi higher education:

H_0: The student that using Web 2.0 tools in the blended environment will no perceive positive effect on their motivation.

H_1: The student that using Web 2.0 tools in the blended environment will perceive positive effect on their motivation.

B. Sampling Procedure

Sampling refers to the process of selecting a sample as a small portion or subset of a defined population [28]. The purpose of this study was to make use of a sample to generalize the findings in a particular population about how Web 2.0 technology tools enhance the learning motivation of students in higher education. Thus, the sample for this study is mainly based on simple random sampling. The control group and experiment group were selected randomly. This type of sample offers high generalizability of findings [28]. The sample population for this study was primarily on female students in the Computer Science at Imam Mohammad Ibn Saud Islamic University (IMSIU) during the second semester 2018. The study was used a convenience sample of 60 students assigned to two groups. Group 1 consisted of 30 students was attended a class that adopted Web 2.0 techniques. Group 2 consisted of 30 students was attended a class without using Web 2.0 techniques. This study focused on the three Web 2.0 tools listed in Table III since they are among the most widely used and recognized by students for education [29].

TABLE III. Web2.0 Tools Usage and Participants Number in Experimental Group

Web 2.0 Tools	**Participants Number**
Social Networks	10
Blog	10
Wiki	10
Total	30

TABLE IV. Measurment Scale Items for Students Motivation

Factors	**Measurement Items**	**References**
Attention	1. There was something interesting at the beginning of this course that got my attention.	[30]
	2. These materials are eye-catching.	
	3. This course is so abstract that it was hard to keep my attention.	
	4. This course has things that stimulated my curiosity.	
	5. The amount of repetition in this course caused me to get bored sometimes.	
	6. I learned some things that were surprising or unexpected	
	7. The variety of exercises, illustrations, etc., helped keep my attention on the course.	
Factors	**Measurement Items**	**References**
Relevance	1. It is clear to me how the content of this material is related to things I already know.	[30]
	2. There were examples that showed me how this material could be important to some people.	
	3. Completing this course successfully was important to me.	
	4. The content of this material is relevant to my interests.	
	5. There are explanations or examples of how people use the knowledge in this course.	
	6. This course was not relevant to my needs because I already knew most of it.	
	7. I could relate the content of this course to things I have seen, done, or thought about in my own life.	
	8. The content of this course will be useful to me.	
Factors	**Measurement Items**	**References**
Confidence	1. When I first looked at this course, I had the impression that it would be easy for me	[30]
	2. This material was more difficult to understand than I would like for it to be.	
	3. After reading the introductory information, I felt confident that I knew what I was supposed to learn from this course.	
	4. As I worked on this course, I was confident that I could learn the content.	
	5. The exercises in this course were too difficult.	
	6. After working on this course for a while, I was confident that I would be able to pass a test on it	
	7. I could not really understand quite a bit of the material in this course.	
	8. The good organization of the content helped me be confident that I would learn this material.	
Factors	**Measurement Items**	**References**
Satisfaction	1. Completing the exercises in this course gave me a satisfying feeling of accomplishment.	[30]
	2. I enjoyed this course so much that I would like to know more about this topic.	
	3. I really enjoyed studying this course.	
	4. The wording of feedback after the exercises, or of other comments in this course, helped me feel rewarded for my effort	
	5. I felt good to successfully complete this course.	
	6. It was a pleasure to work on such a well-designed course.	

C. Data Collection and Instrumentation

The authors used a questionnaire to collect the participants' perspectives about motivation to learn with and without using Web 2.0 tools in a blended course in Saudi higher education. This questionnaire designed according to Keller's ARCS model. This model focuses on measuring the effect of use web 2.0 technologies on students' motivation to learn in a blended learning environment. Five-point Likert scale used to determine the participants' perspectives for the level of agreement\disagreement expressed by them on each item, in which, 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree. This study was used the Instructional Materials Motivation Survey (IMMS), a 36-item situational measure of motivation based on the ARCS model [30]. The IMMS measures all four ARCS Factors: attention, relevance, confidence, and satisfaction. The IMMS was selected to evaluate whether the blended-learning experience induces attention, relevance, confidence and satisfaction, and measures students' motivation levels. Some items were slightly modified and other items were dropped for a total of 29 items in this study. Table IV illustrates the measurement scale items for student motivation.

VII. Data Analysis and Results

Cronbach's coefficient alpha (CA) is examined to measure coefficient of stability and Pearson correlation coefficient to measure internal consistency. The value of alpha coefficient should be greater than the threshold value of 0.70 to be accepted [28]. The results of the reliability test for the Factors are likely to be accepted because greater than 0.70. As shown in the Table V the alpha coefficients for the Factors ranged from 0.70 to 0.88. Thus, it confirmed that all items of respondents' answers in this study were consistency and stability.

TABLE V. Reliability Statistics

Factors	No. of items	Cronbach's alpha
Attention	7	0.70
Relevance	8	0.71
Confidence	8	0.78
Satisfaction	6	0.76
Overall	29	0.88

The data values of this study were normally distribution. The independent samples t-test was made to determine the difference between the students in the use Web 2.0 and those without using Web 2.0 technology. The groups were compared with respect to overall student motivation factors: attention, relevance, confidence and satisfaction. This analysis is appropriate to the aim of this study, thus we need to compare the means of two groups, and especially appropriate as the analysis for the posttest-only two-group randomized experimental designs.

A. Attention Analysis

The result of analyzing the data by using t-test formula shows that there is a significance increase in students' attention after they use web 2.0 technologies. Table VI indicates that the mean of the control group, the group do not use web2.0 tools (24.10), and the mean of experimental group, the group use web 2.0 tools (score is 25.87), the standard devotion of the control group, the group do not use web 2.0 tools is 2.67 and the standard devotion of experimental group, the group use web 2.0 is 2.83. This mean that there is increase in mean of experimental group, the group use web 2.0 tools in attention factor. The result has shown the t- value at a degree of significance is 0.016.

B. Relevance Analysis

The result of analyzing the data by using t-test formula shows that there is a significance increase in students' relevance after they use web 2.0 technologies. Table VI indicates that the mean of the control group, the group do not use web 2.0 tools is (28.63), and the mean score of experimental group, the group use web 2.0 tools score is (31.50) the standard devotion of the control group, the group do not use web2.0 tools is (4.33) and the standard devotion of experimental group, the group use web 2.0 tools is (3.99). This mean that there is increasing in mean of experimental group, the group use web 2.0 tools in relevance factor. The result has shown the t- value at a degree of significance is (0.010).

C. Confidence Analysis

The result of analyzing the data by using t-test formula shows that there is a significance increase in students' confidence after they use web 2.0 technologies. Table VI indicates that the mean of the control group, the group do not use web2.0 tools is (27.17), and the mean score of experimental group, the group use web 2.0 score is (29.23) the standard devotion of the control group, the group do not use web2.0 tools is (2.84) and the standard devotion of experimental group, the group use web 2.0 tools is (4.05). This mean that there is increasing in mean of experimental group, the group use web 2.0 tools in confidence factor. The result has shown the t- value at a degree of significance is (0.026).

D. Satisfaction Analysis

The result of analyzing the data by using t-test formula shows that there is a significance increase in students' Satisfaction after they use web 2.0 technologies. Table VI indicates that the mean of the control group, the group do not use web2.0 is (22.50), and the mean score of experimental group, the group use web 2.0 tools score is (26.90) the standard devotion of the control group, the group do not use web.20. is (3.89) and the standard devotion of experimental group, the group use web 2.0 tools is (3.56). This mean that there is increasing in mean of experimental group, the group use web 2.0 tools in confidence factor. The result has shown the t- value at a degree of significance is (0.000).

According to the results of four factors attention, relevance, confidence and satisfaction, the values of t-test were significant at the 0.05, 0.01 levels in each dimension attention, relevance, confidence and satisfaction. This indicates that there are statistically significant differences concerning these motivations of students that using Web 2.0 tools accordance with the differences in the groups in favor with Web 2.0 technology (experimental group). This result shows that there was a greater increase in motivation scores for the experimental group than for the control group. From this, we can conclude that using Web 2.0 technology affect students' motivation to learn in a blended learning environment.

E. Hypotheses Testing Result

There is significant difference in motivation, as measured by mean IMMS score, between a sample of students using Web 2.0 tools in a blended environment (experimental group) and a control group. To test the hypotheses, t-tests and p-value analysis tests were used to explore the motivation of students using Web 2.0 tools in a blended environment between the control and experimental groups. We found that there was a statistically significant difference at the level of 0.05 in overall student motivation between the experimental and control groups resulting from the using Web 2.0 tools technology. Moreover, students who using Web 2.0 tools technology were found to exhibit a statistically significant higher degree of motivation. Thus, the alternative hypothesis (H_1) was acceptable.

VIII. CONCLUSION AND IMPLICATIONS OF THE STUDY

This study tries to provide the best ways to deliver instruction methods and learning in Saudi higher education for the use of a blended learning environment. It aims to contribute to improving the educational process that measures the effect of using Web 2.0 technologies, including blogs, wikis, and other social networks, on motivation to learn inside blended learning. Essentially, the success of blended learning in Saudi higher education depends not only on using different teaching methods and massive expenditures on technology but also on students' motivation to learn, which reflects their creativity, exploration, and performance improvement and satisfaction. Our study found that there are statistically significant differences concerning these motivations of students using web 2.0 tools accordance with the differences in favor with web 2.0.

TABLE VI. T- TEST ANALYSIS

2	**Groups**	**Mean**	**N**	**Std. Deviation**	**t- value**	**df**	**P= Sig.**
Attention	Without Web 2.0	24.10	30	2.67			
	With Web 2.0	25.87	30	2.83	2.89	58	0.016*
Relevance	Without Web 2.0	28.63	30	4.33			
	With Web 2.0	31.50	30	3.99	2.665	58	0.010*
Confidence	Without Web 2.0	27.17	30	2.84			
	With Web 2.0	29.23	30	4.05	2.288	58	0.026*
Satisfaction	Without Web 2.0	22.50	30	3.89			
	With Web 2.0	26.90	30	3.56	4.571	58	0.000**
Overall	Without Web 2.0	102.40	30	11.03			
	With Web 2.0	113.50	30	10.53	3.987	58	0.000**

Students' motivation to learn is the goal of the educational process in new methods of learning, such as blended learning. There is a close link between learning outcomes and students' motivation to learn. Thus, the Web 2.0 technologies should be integrated into educational process. Web 2.0 technology to enrich teaching environments and they give creative and practical ideas to teachers on the use of these tools in Teaching. Thus, Web 2.0 technology encourages students to not only view and experience information on the Internet, but to also create and share their knowledge and opinions. In this sense, the overall purpose of this research was to investigate potential of using different Web 2.0 tools in blended learning as well as their advantage.

This study provides the following significant implications:

- The ARCS model is used in this study in order to measure the effect of using the Web2.0 technology on students' motivation to learn. After applying the experiment to two samples as an independent samples t-test, the result shows a positive effect of using tools on students to learn. Thus, Academics or educator may inspire to acquaint on other models that used in the blended learning to enhance motivation to learn.
- The finding can help study can help decision makers readjust the learning strategy by realizing the importance of using Web 2.0 as the main platform in Saudi higher education.

Further research could focus not only on whether blended coursework and Web 2.0 technology in general impacted student motivation, but on what types of this technology specifically had the strongest impact on motivation. The limitation of this study was conducted Keller's model of motivational design ARCS Model on small sample sizes. Thus, the results might change when applying ARCS model on bigger sample sizes.

References

[1] U.S. Department of Education, Office of Innovation and Improvement, "Evaluating Online Learning: Challenges and Strategies for Success", Washington, D.C.: Education Publications Center, 2008.

[2] C. Clement, and K. Jones, ''Blended Learning vs. Traditional Classroom Settings: Assessing Effectiveness and Student Perceptions in an MBA Accounting Course", Journal of Educators Online, vol. 4, no. 1, pp. 1-15, 2007.

[3] J. Lave, and E. Wenger, Situated learning. Cambridge [England]: Cambridge University Press, 1991.

[4] S. Abram, "Web 2.0, Library 2.0, and Librarian 2.0: Preparing for the 2.0 World". Library and Information Services in Astronomy, vol. 377, pp.161-167, 2007.

[5] J. Confrey, "How Compatible are Radical Constructivism, Sociocultural Approaches, and Social Constructivism?" in Constructivism in Education, Steffe LP and Gale,J Ed.,),Hover: UK, 1995.

[6] R. Graham, "Blended learning systems: Definition, current trends, and future directions," In The Handbook of Blended Learning: Global Perspectives, Local Designs. J.Curtis ,R. Graham, ,Ed., San Francisco, CA: Pfeiffer Publishing,2006, pp. 1-32.

[7] E.Rooney, "Blending learning opportunities to enhance educational programming and meetings", Association Management, vol.55, no,5 ,pp.26–32, 2003.

[8] J. Young, "Hybrid teaching seeks to end the divide between traditional and online instruction", Chronicle of Higher Education, vol. 35, no. 2, pp.33-34, 2002.

[9] C. Reed & H. Singh, "Getting the best of both world", Training Strategies for Tomorrow, vol.16, no .4, pp.12-14, 2002.

[10] R. Garrison, and N. Vaughan, Blended Learning in Higher Education: Framework, Principles, and Guidelines. San Francisco: Jossey-Bass, 2008.

[11] R. Graham, and D. Dziuban, C.D. Core research and issues related to blended learning environments, Handbook of research on educational communications and technology, 3rd ed., J. Spector, M.Merrill, J.Van Merrienboer, and M.Driscoll, Ed., Mahwah, NJ: Lawrence Earlbaum Associates, 2003.

[12] J. Lei, "Quantity versus quality: a new approach to examine the relationship between technology use and student outcomes", British Journal of Educational Technology, vol.41, no.3, pp.455–472, 2010.

[13] H. Lim, and L. Morris, "Learner and instructional factors influencing learning outcomes within a blended learning environment". Educational Technology & Society, vol.12, no.4, pp.282–293, 2009.

[14] R., Garrison, and N. Vaughan, Blended learning in higher education: Framework, principles, and guidelines, San Francisco: Jossey-Bass, 2007.

[15] C. Dziuban, J. Hartman, and P.Moskal, "Blended learning", EDUCAUSE Center for Applied Research Research Bulletin, 2004.

[16] E. Maloney, "What Web 2.0 can teach us about learning", Chronicle of Higher Education, vol.25, no.18, pp.26, 2007.

[17] T. O'Reilly, "What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software", International Journal of Digital Economics., no.65, pp.17-37, 2007.

[18] K. Laudon, and J.Laudon, "Telecommunications, the Internet, and Wireless Technology," in Management information systems: Managing the digital firm, 12th ed., Upper Saddle River, N.J.: Prentice Hall, 2012.

[19] R. Aucoin, "A Study of Students' Perceptions of the Use of Web 2.0 Applications in Higher Education," M.S. thesis, Univ. British Columbia, Vancouver, Canada, 2014.

[20] M. Bernard, C. Abrami, E. Borokhovski, A. Wade, M.Tamim, A. Surkes and C. Bethel, "A meta-analysis of three types of interaction treatments in distance education", Review of Educational Research, vol.79,no.3 pp.1243-1289,2009.

[21] D. Schunk, P. Pintrich, and J. Meece, "Motivation in education: Theory, research, and applications," 3rd ed., Upper Saddle River, NJ: Merrill Prentice Hall, 2008.

[22] L. Rodgers and J. Withrow-Thorton, "The effect of instructional media on learner motivation", International Journal of Instructional Media, vol.32, no.4 pp.333-340, 2005.

[23] A. Ocak, and M. Akçayır, "Do motivational tactics work in blended learning environments? The ARCS model approach", International Journal of Social Sciences & Education.vol.3, no.4, pp.1058-1070, 2013.

[24] C. Walker, A. Barbara and A. Robert, "Identification with academics, intrinsic/extrinsic motivation, and self-efficacy as predictors of cognitive engagement", Learning and Individual Differences, vol. 16, pp. 1-12, 2006.

[25] M. Keller, Use of the ARCS model of motivation in teacher training, ERIC, ED 288520, 1983.

[26] S. Green, "Using an expectancy-value approach to examine teachers' motivational strategies", Teaching and Teacher Education, vol.18, no.8, pp. 989-1005, 2002.

[27] B. Huett, K. Kalinowski, L. Moller, and C. Hutt, "Improving the motivation and Retention of Online Students Through the Use of ARCS-based E-mails", Journal of Distance Education, vol. 22, pp. 159-176, 2008.

[28] U. Sekaran, and R. Bougie, Research Methods for Business: A Skill Building Approach, Haddington: John Wiley & Sons, 2010.

[29] R. Michelle, "Web 2.0 Use In Higher Education", European Journal of Open, Distance and e-Learning, vol. 17, no. 2, pp.130-142, 2014.

[30] N. Loorbach, O. Peters, J. Karreman, and M. Steehouder, "Validation of the Instructional Materials Motivation Survey (IMMS) in a Self-Directed Instructional Setting Aimed at Working with Technology", Instructional Setting British Journal of Educational Technology, vol.46, no.1 ,pp.204-218 ,2015.

The Impact of Motivator and Demotivator Factors on Agile Software Development

The Case of Pakistan

Shahbaz Ahmed Khan Ghayyur, Salman Ahmed
Department of Computer Sciences and Software Engineering,
International Islamic University
Islamabad, Pakistan

Saeed Ullah
Department of Computer Science
Federal Urdu University of Arts, Science and Technology
Islamabad, Pakistan

Waqar Ahmed
Department of Statistics,
Comsats University Lahore,
Lahore, Pakistan

Abstract—**Since the last decade, Agile software development has emerged as a widely utilized software development method keeping in view the developing countries of South Asia. The literature reports significant challenges and barriers for agile in software industry and thus the area still has significant problems when considered with this domain. This study reports an industrial survey in Pakistani software industry practices and practitioners to elicit the indigenous motivator and demotivators of agile paradigm in Pakistan. This study provides a concrete ranking of motivator and demotivator factors which influence the agile paradigm. A lack of proper training and other identified issues indicate that the adoption of agile is in preliminary phases and serious effort is required to set the direction right for success of agile paradigm and its adopting institutions. The survey is conducted in 23 companies practicing agile organizations and involves 90 agile practitioners. Reports of 67 practitioners were finally selected after careful selection against selection criteria for this study. The results indicate various alarming factors which are different from reported literature on the subject. Tolerance to work is the most important motivating factor among Pakistan agile practitioners, likewise lack of resources is the highest demotivating factor. A detailed ranking list of motivators and demotivators and comprehensive data analysis has been provided in this paper which influences strongly the agile software development issues in Pakistan.**

Keywords—Agile software development; motivators; demotivators; success factors; barriers; agile methods; software development life cycle

I. Introduction

Agile software development is a repetitive method to produce acute and disciplined software development. Research study suggest that agile is the mostly used software development technique all over the world but in under developing countries due to their less domain knowledge and lack of experience there exist many barrier to proper implementation of agile methods [1]. As there is new shift of software industry from SDLC to Agile there exist many individuals and collective problem which caused hurdle in implementation of agile methodology [2]. These hurdles exist in individual and communal manner. To gauge these issues, a survey has been conducted in Pakistan to check the impact of motivators and demotivator factors in agile software development in Pakistani Software industry. This survey will help to enhance productivity of software and reduced number of demotivator factors [3].

In Pakistan, Agile is nourished as Emerging field. In past decade, due to the lack of interest and strategy of software practitioners, software industry face many economical issues but as the agenda of agile become popular there following increase immensely [4]. The formation of PSEB is also an initiative to ensure assistance of software industry. In recent years, Agile boom has become a latest trend in software industry. According to PSEB, about 70% of software organization has converted or thinking to convert their development method on Agile.

In ASD, due to their iterative nature has less failure ratio than SDLC. For this reason, many organizations local or international software industry moving towards agile due to their well-defined set of rules and well organization teams [5]. Motivators factors in agile plays a key role in development of agile industry. These motivating factors provide ability to work on self-determination and to made better product.

As the most common concern for an organization is to provide productive software to their end-users and this phenomena is only achieve by providing motivation to their employee and to avoid the demotivator factor as possible [6]. Motivators plays important role in enhancing people and technical skills. A lot of work has been done to motivator SDLC practitioners but agile has less contribution in this regard [7].

A. The Need of Empirical Analysis

As agile is mostly used method in software industry, but it requires a lot of work on employee's motivation for their full adaption. The main concern is to remove the practitioners' anxiety for adoption of agile method. In this regard, motivator and demotivator factors play a critical role [8]. These factors can be used to alter the software productivity and these factors can provide new power to agile industry.

In Pakistan, software industry is growing day by day but due to the higher failure ratio of projects is become a worry sign for software development organization. For this reason, they want to trade towards agile but due to the barriers exist in the form of demotivator factors they can't fully yield the concern results. Due to which a survey study is conducted to gauge the concern of agile practitioners. For this purpose, motivator and demotivator factors has been collected from literature and a survey is been conducted which rank these motivators and demotivators according to Pakistan software industry.

This study will also contribute to gather the motivator factors of agile, which is present in dispersed form and need to analyse. The literature is mainly covering the motivator factor of SDLC but agile is neglected. There exist a gap to empirically analysis of the motivator and demotivator factors [9]. By finding these motivator factors, list of demotivator factors can reduced which eventually result in quality software product.

The arrangement of this article is ordered as: Section II briefly describe the literature review regarding motivators and demotivators factors. Section III explains the research methodology used during research. Section IV describes the in-detail analysis and results regarding survey. Research contribution is discussed in Section V, Section VI is covering portion of results and discussion and finally Section VII describe conclusion and future work.

II. Related Work

Motivators and Demotivators has a vital role in software productivity. This portion will provide a brief literature review of work done in motivators and demotivators of Agile Software Development. De O et al. [10] provides a detailed list of motivators and demotivators in software development life cycle. Afterward they propose model of motivation of software Engineering(MOCC) in which they divide the motivator into different category .Highsmith and Cockburn [11] are the member of agile formation team, they provide the benefit of adopting agile software development. Akhtar et al. [12] has conducted a survey in Pakistani software industry about the barrier exist in Scrum method, there findings suggest that there exist many flaws in full fledge implementation of Scrum in Pakistani software industry. Hassan et al. [13] briefly describe the challenges exist in full adoption of scrum in Pakistan. There finding suggest that scrum is newly implemented in Pakistan that's why they require adequate training for their full implementation. Wagener [14] listed down detailed list of motivators and demotivators in SDLC afterward they categorize the motivator factor into three groups: organizational, people and technical. Chow and Cao [15] has conducted survey among 109 Agile Teams among different organization, on the basis of survey they find new motivator factor of agile software development. Baddoo and Hall [16] has done a detailed analysis among SDLC factors in which they have found that rewards and incentive to employees can increase their productivity. Asghar and Usman [17] has done a Systematic literature review of Motivators and Demotivators of Software development life cycle, they proposed a model of motivation for Pakistan industry in which they claim Hofstede's cultural issue is the biggest barrier in this region.

III. Research Method Detail

Due to the limitation in research and difference in survey method, mail method in questionnaire and personal method are selected. Research questions are shown in Table I.

A. Research Questions

TABLE I. Research Questions

Sr.no	Research Questions	Motivation
1	What are the motivator and de-motivator factors existed in Agile Software Development?	This question will provide a detailed list of motivators and demotivators of agile.
2	What are the impact of motivators and demotivators on software industry?	This question aims to provide a detailed discussion of impact of motivators and demotivators in software industry.

B. Questionaire Design

The Questionnaire is divided into three sections

Section 1: Include Respondents profile.

Section 2: Include company's profile.

Section 3: Include Motivator and Demotivator of Agile.

C. Data Collection Techniques and Methodologies

Questionnaire is floated via two methods, email and Personal contacts. In both these medium our target is those organization who are fully or partially practices agile. For this purpose, list of organizations has taken from PASHA and PSEB and try to target maximum population. Along with pass on strategy personal contacts also been made to target more organization following agile. A total of 25 software organizations were visited. A total of 25 companies were chosen to provide the research method with pre-requisite to following agile fully or partially.

D. Sample

The whole population (23 agile companies) employees were the sample of the study. As we have limited sample size that's why regular follow ups with respondents containing email and telephone calls and meetings are arranging to get the maximum number of respondents. Some appreciation cards and other incentive are also arranging to get the maximum number of accurate Reponses.

E. Identification of Agile Practitioners

In this survey, one thing is assured that all respondents must possess agile background. The background considers regarding agile will be fully and partially usage. For this purpose, companies are visited personally plus email and phone are used to convince agile practitioners to fill the survey. Cross questioning has been made to verify the respondent's record to double check the practitioner knowledge regarding agile.

Identification of agile practitioner is check using three steps:

1) Respondents are currently working / have already worked at organization which practices agile.

2) Respondents are currently working / have already worked in an organization where at least one agile method is used, e.g. Scrum, Kanban, etc.

3) Respondent must be willing to give interview in given time slot.

F. Compilation of Issues

Once the feedback of survey is received, compilation work has been started. For this purpose, two software Microsoft Excel and SPSS (Version 24) are used to get better view of respondent's behaviour towards the survey. After accessing all feedback responses, a list of issue was extracted based on respondents output.

G. Interview to Resolve Open Issues

Interview is conducted to address some open issues which can't be address in questionnaire. A session of two interviews with practitioners using agile is conducted in which open issue are briefly discussed. The opinion is included in conclusion.

H. Identification of Renowned Agile Practitioners

Selection of renowned agile practitioners has been collected on following criteria:

1) At least five years of agile experience.

2) Worked in an organization using agile more than two years.

3) Taken and conducted agile trainings in past two years.

4) Achieve agile certifications.

I. Compilation of Data (Interview and Survey Result)

A total number of 25 companies were visited. Participating companies were selected from given number of respondents give the information about the motivator and demotivator of Agile and different initiative to reduce the demotivator factors. The companies were chosen to provide the cross section of current profile, total working experience, experience usage of agile method, extend of usage of agile method and preference of most using agile practices.

J. Analysis Method Used

There are two major analysis one is qualitative and other one is quantitative. Both techniques are used to measure the more accuracy of respondent's feedback.

K. Quantitative Analysis

Quantitative analysis is best used analysis technique to measure the respondents result more accurately. In quotative analysis rather than question and their answer numeric data is prominent by which significant of research is prominent. Our focus is to target the quantitative analysis to get a more accurate result with respect to motivator and demotivator of agile. Table II shows the key aggregate on Surveys response.

TABLE II. KEY AGGREGATE ON RESPONSE

Total Number of Software Companies Surveyed	25
Total Number of Software companies using agile	23
Companies working on Offshore Development:	14
Companies working on In-house Development:	07
Companies working on Both:	02
Small-Medium Companies:	14
Large Companies:	09
Total Number of Software Practitioners Contacted	90
Total Number of Software Practitioners Responded	67
Respondent's Total Experience (3-5 Yrs):	20
Respondent's Total Experience (1 to >3 Yrs):	22
Respondent's Total Experience (5-10 Yrs):	10
Respondent's Total Experience (10> Yrs):	15
Business Analyst/ Professional services	02
Project Management	09
Team Lead	04
Junior Software Developer	07
Senior Software Developer	32
Software test Engineer	05
Quality Assurance	08
Total Number of Questions in Questionnaire	19
Mandatory Questions:	16
Optional Questions:	03
Total Number of Strongly Agreed Responses	427
Number of Agreed Responses:	1525
Number of Neutral Responses:	425
Number of Disagreed Responses:	167
Number of Strongly Disagree	23

IV. ANALYSIS AND RESULTS

Author has already study the motivator and demotivator factors and identified issue according to agile software development and categorize into three factors: People, technical and organization. The same motivator and demotivator factors are used in a survey conducted in Pakistani Agile Software industry. The aim of this survey is to find out the higher rank motivator and demotivator factors and then results shown below is used to find out the issues of agile practitioners and compare results with the literature to increase the motivator factors in ASD.

TABLE III. CRONBACH'S ALPHA FOR PILOT STUDY

Scales	**K**	**Cronbach's Alpha (α)**
Motivation factors	36 items	0.895
Demotivation factors	24 items	0.923

To check the reliability of survey, Cronbach alpha test is applied. Motivator contain 36 factors whereas demotivator contains 24 factors, Cronbach alpha test shows that both values are highly reliable.

Cronbach on survey

Table III shows reliability analyses of scales used by the motivation factor and demotivation factors, α = .895 and α = .923 respectively.

A. *Profile of Respondents*

1) Gender based respondents: Empirical analysis result shows that male respondents are more than female respondents. They have 59 and 8 frequencies respectively.

Following Fig. 1 shows the gender respondents of pilot study.

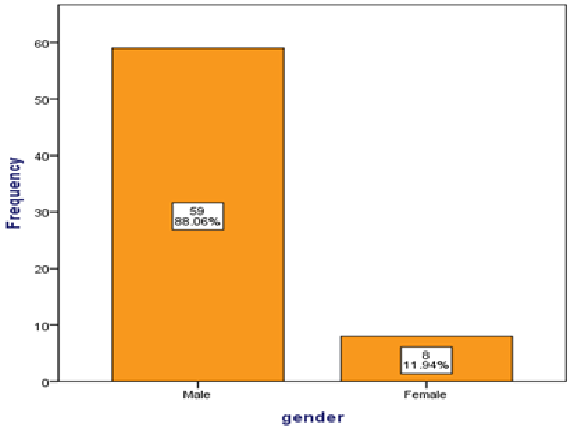


Fig. 1. Pilot study gender wise respondents.

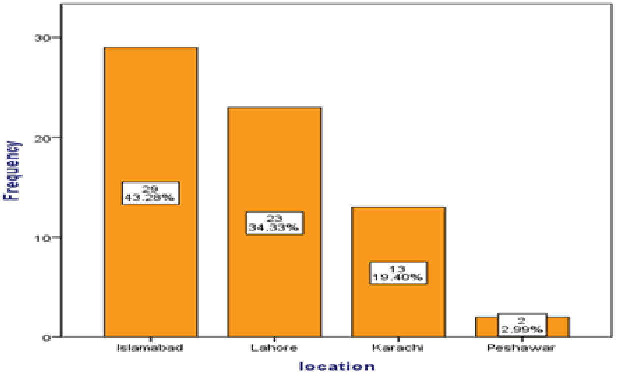


Fig. 2. Cities wise respondents.

2) Location based respondents: Islamabad has more frequency of respondents than other cities of Pakistan. Its frequency is 29 whereas Lahore has 23, Karachi has 13 and Peshawar has 2 respondent's frequencies.

Fig. 2 shows the cities by which responders fill the pilot study.

B. *Current Profile based Respondents*

According to our respondent's Senior software developers has more number of respondent's frequency which is 32 whereas project Managers has 9, Quality assurance engineer has 8, Junior software developer has 7, software test engineer has 5, team lead has 4 and Business analyst has 2 respondents.

The following Fig. 3 shows the total experience of responders

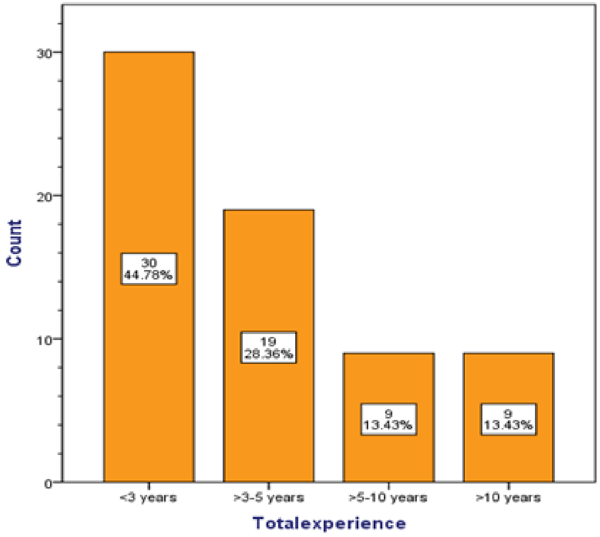


Fig. 3. Respondents total experience.

C. *Agile Experience of Respondent's*

The result indicates that the respondents having 1 to less than three years' experience are 43%, while less than one year has 30%, the experience from 3 to 5 years are 14%, the respondent's having experience from less than 5 to 10 years are 9% and the respondents having experience more than 10 years are 4%.

Following Fig. 4 is depicted the agile experience of respondents.

Following Table IV shows the demographic profile of respondents based on gender, location, current profile, total experience and agile experience. To make better understanding of results, the results are shown in frequency as well as percentage.

D. *Extend of Usage of Agile Methods*

The following Fig. 4 shows the responders usage of agile method to different type of project. The results indicate that agile is been used for majority of projects and is using large number in organizations ongoing projects.

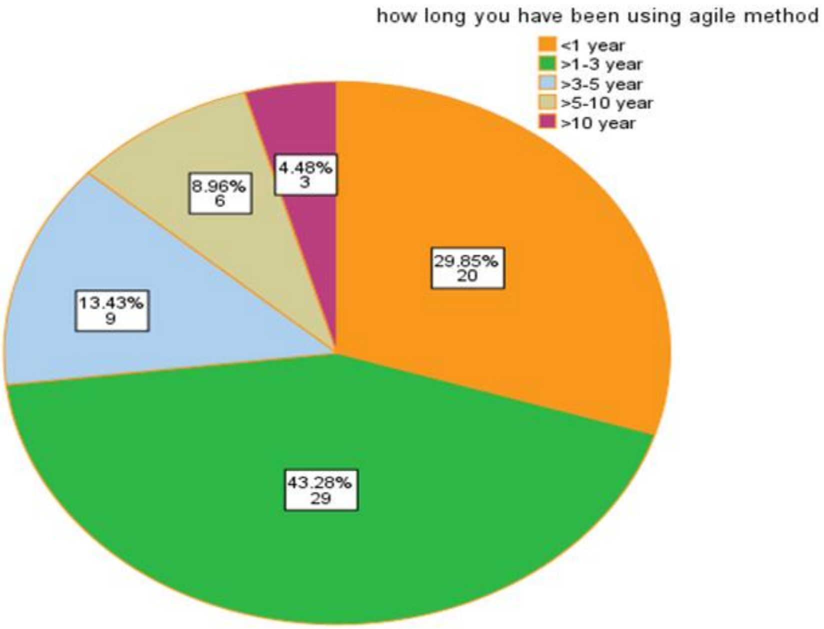


Fig. 4. Duration of agile usage.

TABLE IV. SOCIO-DEMOGRAPHIC PROFILE OF PILOT STUDY

Category		**Frequency (%)**
Gender	Male	59(88.1)
	Female	8(11.9)
Location	Islamabad	29(43.3)
	Lahore	23(34.3)
	Karachi	13(19.4)
	Peshawar	2(3.0)
Describe your current profile	Business Analyst/ Professional services	2(3.0)
	Project Management	9(8.9)
	Team Lead	4(6.0)
	Junior Software Developer	7(10.4)
	Senior Software Developer	32(47.8)
	Software test Engineer	5(7.5)
	Quality Assurance	8(11.9)
Total Work experience	<3 years	30(44.8)
	3 years to 5 years	19(28.4)
	>5 years to 10 years	9(13.4)
	>10 years	9(13.4)
How long you have been using Agile methods	Less than a Year	20(29.9)
	1-3 years	49(43.3)
	>3 years to 5 years	9(13.4)
	>5 years to 10 years	6(9.0)
	>10 years	3(4.5)

E. Usage of Agile Methods

Fig. 5 shows the respondents answer of using different agile methods. The result indicates that Scrum is mostly used method with usage of 20%, extreme programming has usage of 11% and crystal-clear method has least usage of 1%.

On a question of most using agile practice, Fig. 6 the respondents agree on planning iteration with 55%, daily stand ups have 22%, Iteration retrospective has 4.5% and Review meeting has 3% usage in respondent's organizations.

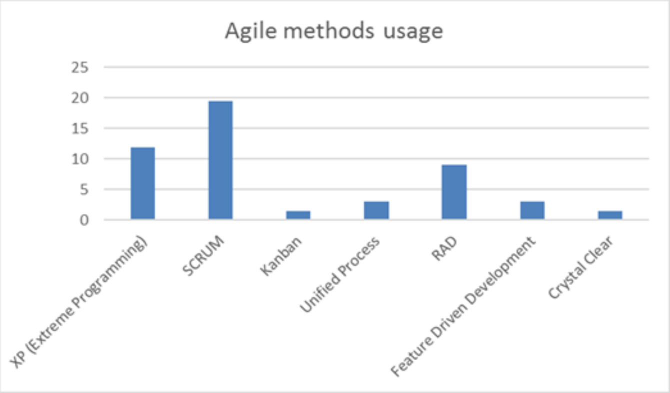


Fig. 5. Agile usage method usage of agile practices.

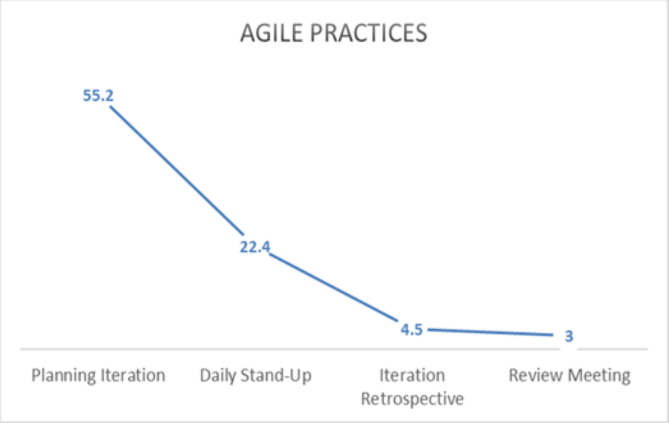


Fig. 6. Agile practices usage.

Following Table V shows the information about the agile usage in which first block is answering about extend of usage of agile methods in majority, large, small and other projects. In most fluent agile method scrum has 19%, extreme programming has 20%, and crystal clear has least 1% agile method. In preference of agile method, the core agile practices such as planning iteration, Daily Stand-ups, Iteration retrospective and review meeting.

TABLE V. AGILE METHOD YOU ARE FLUENT MOST (PILOT STUDY)

Category		**Frequency (%)**
Extend of Agile Methods	Majority of Projects	32(47.8)
	Large number	13(19.4)
	Small number	10(14.9)
	Just Started	2(3.0)
	In learning phase	9(13.4)
	Have never used	1(1.5)
In which Agile method you are fluent most	XP (Extreme Programming)	8(11.9)
	SCRUM	13(19.4)
	Kanban	1(1.5)
	Unified Process	2(3.0)
	RAD	6(9.0)
	Feature Driven Development	2(3.0)
	Crystal Clear	1(1.5)
	Team Software Process	2(3.0)
	Agile Modeling	7(10.4)
	N/A	3(4.5)
Preference of Agile Practice	Planning Iteration	37(55.2)
	Daily Stand-Up	15(22.4)
	Iteration Retrospective	3(4.5)
	Review Meeting	2(3.0)

F. Respondents Responses

1) Motivators responds: Following Table VI are the motivators ranking evaluated by the respondent's results. According to which tolerance to work has most 98% while the eliminated managerial politics is the least number of motivator with 25%.

TABLE VI. RESPONDENTS MOTIVATORS RESULT

Motivating Factors	Strongly Agree	Agree	Disagree	Strongly Disagree
Rewards and incentives	21	35	0	2
Management Supportive role	21	35	4	0
Well defined coding standard	15	30	2	0
Career path	11	43	0	0
Better working environment	20	42	2	0
Variety of work	14	35	5	0
Technically challenging work	10	36	1	2
Successful company experience	11	36	2	1
Trust	20	30	3	0
Identify with the task	11	36	1	1
Sufficient resources	8	32	2	0
Development needs addressed	13	40	2	0
Feedback	19	33	1	0
Recognition	13	36	0	0
Autonomy	7	42	3	1
Work balance	16	32	4	1
Management contribution	13	37	2	0
Sense of Responsibility	24	36	0	0
Sense of belonging	15	34	3	0
Equity	10	45	1	0
Job security	12	29	6	3
Self-organizing teams	15	39	2	1
Eliminate Politics	17	21	8	1
Right amount of documentation	7	33	4	2
Tolerance to work	62	4	0	0
Life Insurance	29	37	0	0
Annual Award System	56	10	0	1
Recreational tours	50	15	1	0
Staff Dinner	43	22	2	0
Leave on demand	47	16	1	1
Recording suggestions	40	26	1	0
Client Availability	47	19	0	1
Recreational facility	41	23	1	1
Follow standard Practices	48	18	0	1
Managing Self respect	57	9	1	0
Knowledgeable Team Leader	57	8	2	0

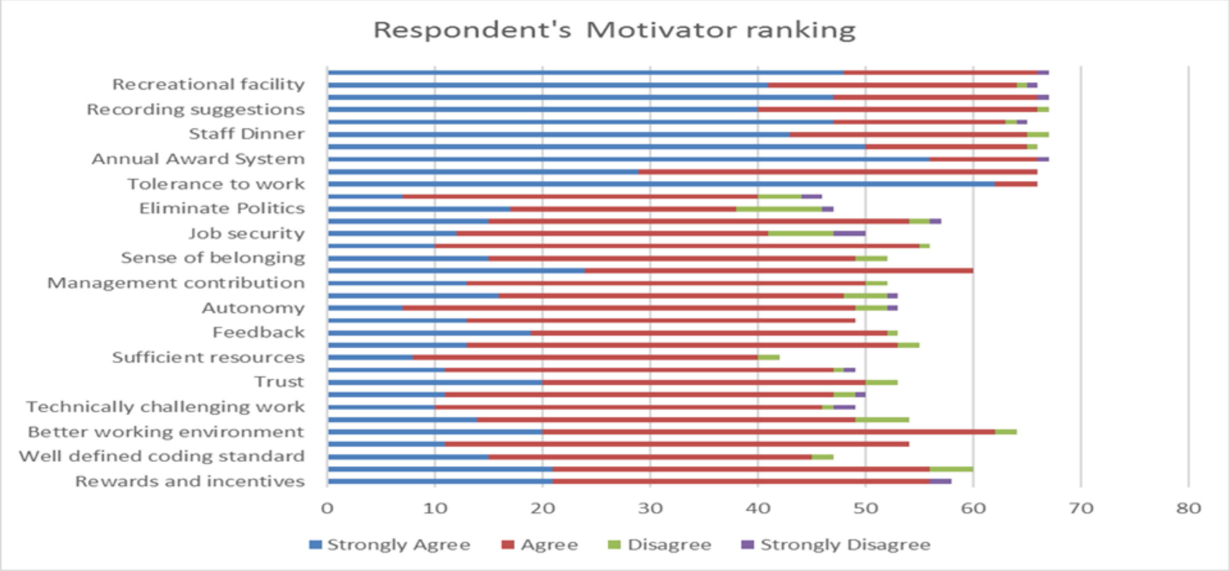


Fig. 7. Respondents motivators ranking.

Following Fig. 7 shows the respondents motivator ranking into more accurate graphical work in which tolerance to work has highest motivator value.

2) Demotivators respondents: Following are the respondent's ranking about the demotivator factors of agile software development (Table VII). The result indicates that lack of resources is the biggest demotivator factor among all factors while unrealistic goals are the least demotivator factor.

Following Fig. 8 shows the respondents demotivator ranking into more accurate graphical work in which lack of resources has highest demotivator value.

TABLE VII. RESPONDENTS DEMOTIVATOR RESULTS

Demotivating Factors	Strongly Agree	Agree	Disagree	Strongly Disagree
Communication Barrier	24	29	7	0
Lack of relationship opportunities	13	29	11	0
Unrealistic goals	20	21	6	0
Injustice in Promotions	15	29	8	2
Poor quality software	13	23	10	4
Political Environment	14	24	9	1
Uncompetitive pay	16	24	12	0
Unsupportive management	24	15	9	0
Lack of influence	10	28	11	3
Unfair reward system	19	27	8	1
Non-interesting work	16	19	8	2
Inequity/Personal preferences	12	29	7	0
Risk	3	39	7	0
Stress/Pressure	14	32	8	1
Less Documentation	37	28	0	0
Restricted Social Networking	36	31	0	1
Job threatening	45	18	2	1
Lack of Resources	47	20	0	0
Political Background	36	30	1	0
Late Hours	42	25	0	0
Sectarian Discrimination	36	29	1	1
Lack of Team work	37	28	0	0
Prohibition of change	34	31	0	0
Long Term Project	32	35	0	0

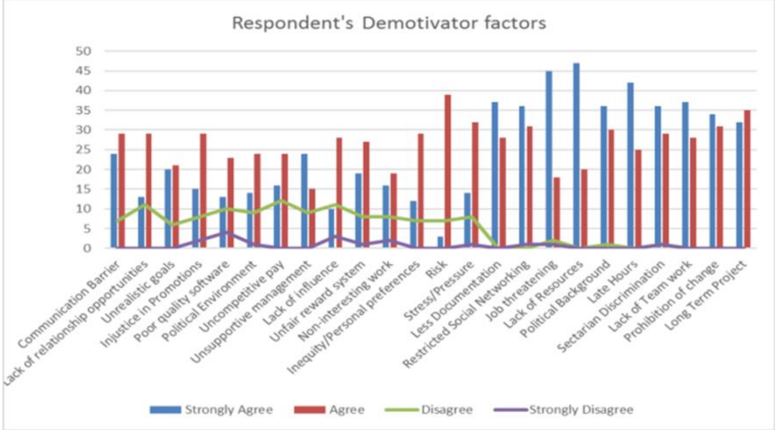


Fig. 8. Respondents demotivators ranking.

G. Correlation Factors (Answering RQ # 1)

Following Table VIII shows the correlation between the motivator factors. The factor which has 0 to 0.25 value has weak positive correlation. The factors having value from 0.25 to 0.75 has medium positive correlation and the factor has more than 0.75 value has strongest positive correlation. Likewise, if the factor has 0 to -0.25 value has weakest negative correlation, if a factor has -0.25 to -0.75 value has medium negative correlation and if a function has less than -0.75 has strongest negative correlation.

TABLE VIII. Correlation between Motivator Factors

Correlations																																				
	M1	M2	M3	M4	M5	M6	M7	M8	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M 21	M 22	M23	M 24	M25	M26	M27	M28	M29	M29	M29	M30	M31	M32	M33
M1	1.00	0.34	0.56	0.27	0.27	0.23	0.31	0.34	0.45	0.23	0.22	0.29	0.30	0.43	0.13	0.19	0.35	0.10	0.24	0.39	0.41	-0.11	0.12	0.40	0.49	0.11	0.10	0.24	-0.10	0.10	-0.19	-0.32	-0.07	0.16	0.00	-0.06
M2		1.00	0.35	0.29	0.54	0.13	0.21	0.40	0.27	0.39	0.46	0.35	0.39	0.45	0.08	0.41	0.55	0.37	0.23	0.46	0.19	0.17	0.05	0.18	0.11	-0.08	0.37	0.02	-0.06	0.17	-0.05	-0.13	-0.14	-0.05	0.16	0.00
M3			1.00	0.37	0.34	0.27	0.32	0.43	0.55	0.40	0.32	0.29	0.39	0.48	0.32	0.40	0.55	0.34	0.39	0.24	0.29	-0.13	0.24	0.31	0.12	0.21	0.15	0.01	0.09	0.13	0.23	0.11	0.04	0.04	-0.12	-0.06
M4				1.00	0.28	0.23	0.20	0.49	0.36	0.22	0.22	0.43	0.31	0.36	0.28	0.38	0.37	0.26	0.34	0.24	0.25	0.12	0.28	0.09	-0.01	0.07	0.28	0.25	0.36	0.14	0.24	-0.19	-0.07	0.04	0.29	0.14
M5					1.00	0.27	0.20	0.50	0.50	0.37	0.32	0.28	0.41	0.44	0.37	0.35	0.54	0.46	0.24	0.44	0.20	0.42	0.36	0.25	0.18	-0.06	0.21	0.07	-0.01	0.12	0.12	-0.10	0.06	0.15	0.06	0.27
M6						1.00	0.48	0.35	0.35	0.45	0.11	0.09	0.50	0.17	0.28	0.42	0.38	0.35	0.44	0.26	0.19	0.29	0.25	0.44	0.12	-0.17	0.14	-0.02	0.05	-0.12	-0.01	-0.01	0.03	0.28	0.21	0.37
M7							1.00	0.34	0.35	0.47	0.18	0.32	0.36	0.26	0.41	0.56	0.21	0.26	0.68	0.54	0.28	0.24	0.24	0.40	0.31	-0.09	0.19	0.24	0.26	-0.02	-0.03	-0.13	0.07	0.04	0.16	0.21
M8								1.00	0.59	0.50	0.44	0.49	0.32	0.44	0.40	0.49	0.38	0.37	0.45	0.41	0.24	0.16	0.40	0.32	0.18	0.09	0.12	-0.08	0.07	0.01	0.17	-0.23	-0.10	0.16	-0.04	0.19
M9									1.00	0.42	0.23	0.31	0.39	0.37	0.26	0.22	0.40	0.37	0.23	0.15	0.21	0.14	0.23	0.21	0.24	-0.02	0.11	0.17	0.09	-0.10	0.05	-0.15	-0.05	-0.03	-0.05	0.01
M10										1.00	0.49	0.34	0.43	0.20	0.34	0.33	0.50	0.31	0.34	0.43	0.23	0.33	0.17	0.38	0.11	0.03	0.35	0.01	0.01	0.03	0.11	-0.05	-0.06	0.13	0.14	0.37
M11											1.00	0.51	0.08	0.38	0.22	0.28	0.40	0.14	0.24	0.31	0.00	-0.02	-0.06	0.29	-0.02	-0.05	0.29	-0.08	0.13	0.06	0.09	-0.22	0.04	-0.21	0.02	0.08
M12												1.00	0.23	0.21	0.38	0.23	0.31	0.23	0.26	0.45	0.12	-0.03	-0.02	0.24	0.17	-0.14	0.23	-0.12	0.01	0.00	0.05	-0.32	0.06	-0.12	-0.02	0.05

M13	M14	M15	M16	M17	M18	M19	M20	M21	M22	M23	M24	M25	M26	M27	M28	M29	M30	M31	M32	M33	M34	M35
0.21	0.10	0.11	0.19	0.20	-0.04	0.07	0.20	0.20	0.50	0.39	0.32	0.08	0.17	0.20	0.17	0.27	0.22	-0.16	0.20	0.35	0.49	1.00
0.27	0.16	-0.12	0.19	0.32	-0.04	-0.10	0.20	0.28	0.33	0.18	-0.04	-0.10	0.17	0.45	0.27	0.06	-0.14	-0.25	0.10	0.25	1.00	
0.24	-0.01	-0.05	0.17	0.08	0.05	0.16	0.08	0.30	0.12	0.37	0.18	0.27	0.34	-0.07	-0.02	0.02	0.06	0.06	0.03	1.00		
0.08	0.21	0.01	-0.08	0.09	0.01	0.11	0.10	-0.02	-0.06	0.11	0.18	-0.20	0.28	0.16	0.26	0.26	0.16	0.07	1.00			
-0.08	-0.16	-0.01	-0.13	-0.21	0.01	-0.02	-0.37	-0.14	-0.10	-0.07	-0.30	-0.02	0.23	-0.08	-0.11	0.12	0.45	1.00				
0.02	0.16	-0.04	0.05	-0.08	-0.10	0.12	-0.09	0.04	-0.06	0.17	0.12	0.06	0.28	0.01	0.32	0.27	1.00					
0.07	0.10	-0.07	-0.08	0.11	0.13	0.01	0.18	-0.02	0.10	-0.16	-0.01	-0.01	0.12	0.23	0.29	1.00						
0.19	0.22	0.12	0.14	0.01	-0.04	0.25	0.12	-0.05	0.11	0.05	0.02	-0.19	0.26	0.13	1.00							
0.11	0.22	-0.26	0.15	-0.01	-0.05	0.09	0.20	0.29	0.21	0.21	0.13	0.16	0.40	0.17								
0.38	0.21	-0.02	0.22	0.48	0.26	0.18	0.26	0.13	0.26	0.07	0.19	-0.11	-0.05	1.00								
-0.03	0.26	-0.02	0.10	-0.08	-0.24	0.01	-0.06	0.33	-0.10	0.33	-0.04	-0.03	1.00									
0.19	0.07	0.09	-0.04	-0.03	-0.09	0.14	0.20	0.28	0.08	0.05	0.39	1.00										
0.45	0.28	0.24	0.37	0.37	0.22	0.45	0.33	0.41	0.24	0.40	1.00											
0.25	0.37	0.28	0.55	0.20	0.11	0.26	0.17	0.58	0.32	1.00												
0.46	-0.06	0.19	0.21	0.26	0.34	0.09	0.15	0.21	1.00													
0.32	0.24	0.13	0.48	0.23	0.20	0.14	0.33	1.00														
0.40	0.32	0.24	0.45	0.36	0.38	0.43	1.00															
0.49	0.44	0.34	0.53	0.22	0.50	1.00																
0.65	0.29	0.16	0.38	0.47	1.00																	
0.59	0.44	0.37	0.44	1.00																		
0.32	0.38	0.39	1.00																			
0.18	0.23	1.00																				
0.39	1.00																					
1.00																						

H. Comparison from Literature (Answering RQ # 2)

This section provides the concrete information about the literature comparison with our survey. Based on the solid result a participant agreement and disagreement has been detailed discussed. Following Table IX shows the comparison between the findings in literature with our survey, respectively.

TABLE IX. MOTIVATORS FACTORS IN LITERATURE AND COMPARED WITH SURVEY

Motivators from Literature	**Motivators in this Survey**	**Participants Agreement / Disagreement with Literature**
1. “Working in company that is successful (e.g. financially stable)” [18]	• Working in successful company (47 / 67=70%)	**Strongly Agree**
2. “*Good Management* is cited 3 times as motivator due to the open communication and workload balance in agile projects” [19]	• Supportive management (56 / 67 = 83 %)	**Strongly Agree**
3. “Factors unrelated to team interactions are not included, such as financial compensation and job security. The findings we present here are all based on how an individual’s behaviour within a team might motivate or de-motivate other developers.” [20]	• Job security (41 / 67) = 61 %)	**Agree**
4. “we could conduct our experiment in a company under real working conditions with employees of the company. Now, however, our internal validity is threatened, because we cannot control the influence of confounding variables like programming experience” [21]	• Working with others/teamwork (62 / 67) = 92 %)	**Strongly Agree**
5. “Consistent with prior research individuals on both teams were personally motivated by factors such as interesting and challenging work, responsibility and the opportunity for growth and development as part of a defined career path.” [22]	• Career path (54 / 67) = 80 %)	**Strongly Agree**
6. “The allocation of office space, putting developers in close with each other, the emphasis on face-toface communication, the availability of appropriate development tools, and close customer collaboration require a great deal of external support to be implemented” [23]	• Appropriate working conditions (60 / 67) = 89 %)	***Strongly Agree***
7. “*Variety of work,* the iteration planning meeting provides a forum in which team members can easily and openly verbalize their preference to work on specific task(s) in order to improve their knowledge and skills in a certain area, which is motivating when, “*people want areas of work where they would learn the most… to acquire certain skills”* [22]	• Variety of work (49 / 67) = 73 %)	**Agree**
8. “We ranked the motivators by their relative frequency in the results. The most frequent general motivator we found is *technically challenging work* (M1), in which work is not mundane and is technically challenging” [19]	• Technically challenging work (46 / 67= 68 %)	***Agree***
9. “Interestingly, based on responses to other questions, it does not seem to matter whether the manager is perceived as actually understanding the issues faced by practitioners, or whether rewards and incentives for successful SPI are established” [24].	• Rewards and incentives (56 / 67= 83 %)	***Strongly Agree***
10. “*Trust/respect:* All three agile practices were identified as an important component of building trust in an agile team due to the increase in verbal communication. In particular, the stand-up is a daily touch-point for all team members, which requires team members (co-located and distributed) to meet and communicate with each other on a daily basis and “*keeps the lines of communication open.”* [22]	• Trust/respect (50 / 67= 74 %)	***Agree***
11. “The allocation of work in many agile teams and also in this team makes it easy for developers to identify *12.* with tasks that have been fulfilled. The user story represents a task that produces a visible part of the software.” [25]	• Identify with the task (47 / 67 = 70 %)	***Agree***
13. “*Limited supply of software engineers.* Several sources2,3 have indicated that the current US shortage of software personnel is between 50,000 and 100,000 people and that the suppliers (primarily university computer science departments) do not have sufficient resources to meet the future demand”. [9]	• Sufficient resources (40 / 67=59 %)	a. ***Agree***
14. “For some, learning and development opportunities may have a higher motivational impact, while for others compensation or supportive superior may be more important”. [26]	• Development needs addressed (53 / 67= 79 %)	b. ***Strongly*** c. ***Agree***
15. “To make team-based performance evaluation more effective team members can act as both evaluators and those being evaluated. Six companies introduced *360-degree feedback*, in which all team members evaluate one other (as opposed to managers appraising subordinates), thus capturing voluntary contributions and mentorship”.[27]	• Feedback (52 / 67 = 77 %)	d. ***Strongly*** e. ***Agree***
16. “He concluded that recognition, security, and sense of belonging were more important to productivity and morale or motivation, and a friendly relationship with the supervisor was very important in securing the loyalty and cooperation of the team” [28]	• Recognition (49 / 67=73 %)	f. ***Agree***

17. "The motivating potential of a job is determined by the degree of richness of five core job dimensions: skill variety, task identity, task significance, autonomy and feedback from the job. The job's motivating potential score (MPS) is computed from the survey responses on the core job dimensions". [29]	• Autonomy (49 / 67= 73 %)	g. ***Agree***
18. "Project managers have to deal with peaking workloads, making it difficult to achieve a work-life balance. Particularly, the temporary nature of project work is a challenge for project managers. Often, there is an uncertainty about future assignments, including the nature of the assignment, its location, and future work colleagues" [26]	• Work balance (48 / 67= 71 %)	h. ***Agree***
19. "In addition, factors such as career development, a sense of belonging and making a contribution to the entire system, receiving positive feedback, and having autonomy were also identified as important motivational factors for project managers" [26]	• Management contribution (50 / 67 = 74 %)	i. ***Agree***
20. "Santana and Robey's (1995) model suggests that managerial, team member or self-control of tasks influences the level of job satisfaction felt by an employee. Two of these motivators are represented in the new model by 'good management', and 'empowerment/responsibility' but the notion of other team members controlling tasks is not explicitly mentioned". [18]	• Sense of Responsibility (60 / 67 = 89 %)	j. ***Strongly*** k. ***Agree***
21. "factors such as career development ,a sense of belonging and making a contribution to the entire system, receiving positive feedback" [26]	• Sense of belonging (47 / 67 =73 %)	Agree
22. "Equity Theory (Homans and Adams in (Couger and Zawacki, 1980)) explains motivation in terms of matching the inputs that practitioners bring to a job (experiences, qualifications, etc.) with appropriate outputs (pay, responsibility, authority, etc.)." [16]	• Equity (55 / 67 = 82 %)	Strongly Agree
23. "trade-offs across sensor, networking, fusion, command-control, software infrastructure elements of a SISOS and more, along with additional trade-offs between performance, security, usability, safety, and fault tolerance" [30]	• Tolerance to work (66 / 67 = 98 %)	Strongly agree
24. "Based on results of a survey with 1005 managers and technical employees in an insurance company" [26]	• Life Insurance (29 / 67 =43 %)	Disagree
25. "Risk mitigation practices include career path development, mentoring junior staff to provide replacements for key personnel, incremental completion bonuses, flowdown of contract award fees to project performers, and recognition initiatives for valued contributions". [30]	• Annual Award System (66 / 67 = 98 %)	Strongly agree
26. <new Add Motivator 1>[31]	• Recreational tours (65 / 67 = 97 %)	Strongly agree
27. <new Add Motivator 2>[32]	• Leave on demand (64 / 67 = 95 %)	Strongly agree
28. <new Add Motivator 3>[33]	• Recording suggestions (40 / 67 =59 %)	Agree
29. <new Add Motivator 4>[34]	• Client Availability (66 / 67 =98 %)	Strongly disagree
30. <new Add Motivator 5>[35]	• Follow standard Practices (66 / 67 =98 %)	Strongly Agree
31. <new Add Motivator 6>[36]	• Knowledgeable Team Leader (54 / 67 =80 %)	Strongly Agree
32. <new Add Motivator 7>[37]	• Managerial Politics (17 / 67) = 25%)	Strongly disagree
33. <new Add Motivator 8>[38]	• Right amount of documentation (33 / 67) = 49%)	Disagree
34. <new Add Motivator 9>[39]	• Staff dinner (23 / 67) = 34 %)	Strongly disagree
35. <new Add Motivator 10>[40]	• Recreational facility (41 /67) = 61%)	Agree

V. Research Contribution

Literature review predicts that there is less work done on motivators and demotivators of agile software development and need a strong analysis that can increase the software performance and productivity. This research aims to provide solid background to agile practitioners to increase their satisfaction level by prioritizing their motivators factors. For this purpose, survey data analysis method is selected.

VI. Results and Discussion

Prioritization of motivators and demotivators has been done by the help of software industrial survey. The main target of this research is to increase the motivation level of agile practitioners by increasing no of motivator and decreasing demotivator factors respectively. Our result indicates that, rewards and incentive and well-defined coding standard has strong correlation factors with value of 0.56, while recreational tours has weakest correlation factors with value of -0.19. In Management Supportive role, work load has highest correlation factors with 0.55 and staff dinner has weakest correlation factor. In well-defined coding standard, work load has highest correlation factor with 0.55 and job security is least correlation factor and vice versa. In comparison of our findings with literature, we have concluded that Knowledgeable team leader, leave on demands, tolerance to work, sense of responsibility and arranging recreational tours are the top motivators factors while staff dinner, life insurance and managerial politics are least motivating factors in agile software development.

Our result indicates that the participant responds tolerance to work as most strongly motivator factor, annual award system, manageable self-team and knowledgeable team leader are responds as other strong motivator factors. Besides the motivator factors, prioritization of demotivator factors has also been performed, lack of resources is the biggest demotivator factor while other strong demotivator factors include job threatening and late hours sittings. These findings lead to predict a guideline for agile practitioners that have strong impact on one's productivity.

VII. Conclusion and Future Work

This survey is conducted on 23 software companies of Pakistan who have implemented agile methods. There are total 67 agile practitioners who have participated in this research. The survey is the extended version of empirical research and case study of systematic mapping and literature review conducted on agile software developing. For this purpose, Pakistan a developing country is been chosen to evaluate our result. This research has revealed more motivator and demotivator factors than existing literature. The analysis has been done to find the top rank motivator and demotivator factors. Our result indicates that the tolerance to work is the highest motivator factor while managerial politics is the last. Likewise, lack of resources the most demotivator factor while the unrealistic goal is the least demotivator factor. These motivators and demotivator must be mitigated, in order to successful implementation of agile in their organizations.

The future work of this research us an implementation of model of motivator for agile practitioners. Another extension of this work is needed to find out motivator and demotivator factors according to core agile practices like planning iteration, iteration retrospective, daily stand ups and review meeting. By implementing motivator and demotivator factor on these agile practices we can attain more in-depth knowledge of this research.

References

[1] A. Law and R. Charron, "Effects of agile practices on social factors," ACM SIGSOFT Softw. Eng. Notes, vol. 30, no. 4, p. 1, 2005.

[2] Z. Masood, R. Hoda, and K. Blincoe, "Motivation for Self-Assignment: Factors Agile Software Developers Consider," in 2017 IEEE/ACM 10th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE), 2017, pp. 92–93.

[3] O. Dieste, E. R. Fonseca C., G. Raura, and P. Rodriguez, "Professionals Are Not Superman: Failures beyond Motivation in Software Experiments," in 2017 IEEE/ACM 5th International Workshop on Conducting Empirical Studies in Industry (CESI), 2017, pp. 27–32.

[4] A. C. C. França, T. B. Gouveia, P. C. F. Santos, C. A. Santana, and F. Q. B. da Silva, "Motivation in software engineering: A systematic review update," 15th Annu. Conf. Eval. Assess. Softw. Eng. (EASE 2011), pp. 154–163, 2011.

[5] P. C. Chen, C. C. Chern, and C. Y. Chen, "Software project team characteristics and team performance: Team motivation as a moderator," in Proceedings - Asia-Pacific Software Engineering Conference, APSEC, 2012, vol. 1, pp. 565–570.

[6] A. Cockburn and J. Highsmith, "Agile software development: The people factor," Computer (Long. Beach. Calif)., vol. 34, no. 11, pp. 131–133, 2001.

[7] P. E. McMahon, "Bridging agile and traditional development methods: A project management perspective," CrossTalk, no. 5, pp. 16–20, 2004.

[8] M. Lindvall et al., "Empirical Findings in Agile Methods," Proc. Extrem. Program. Agil. Methods, XP/Agile Universe 2002, pp. 197–207, 2002.

[9] B. Boehm and R. Turner, "Management challenges to implementing agile processes in traditional development organizations," IEEE Softw., vol. 22, no. 5, pp. 30–39, 2005.

[10] C. De O. Melo, C. Santana, and F. Kon, "Developers motivation in agile teams," Proc. - 38th EUROMICRO Conf. Softw. Eng. Adv. Appl. SEAA 2012, no. March 2015, pp. 376–383, 2012.

[11] J. Highsmith and A. Cockburn, "Agile Software Development : The Business of Innovation," Science (80-.)., vol. 34, no. 9, pp. 120–123, 2001.

[12] M. J. Akhtar, A. Ahsan, and W. Z. Sadiq, "Scrum adoption, acceptance and implementation (A case study of Barriers in Pakistan's IT Industry and Mandatory Improvements)," Proc. - 2010 IEEE 17th Int. Conf. Ind. Eng. Eng. Manag. IE EM2010, pp. 458–461, 2010.

[13] Colleen Frye, "Agile by the numbers: Survey finds more adoption, but age-old problems." [Online]. Available: http://searchsoftwarequality.techtarget.com/news/1372395/Agile-by-the-numbers-Survey-finds-more-adoption-but-age-old-problems. [Accessed: 24-Jul-2017].

[14] R. P. Wagener, "Investigating critical success factors in agile systems development projects/Ruhan Wagener.," no. November, 2012.

[15] T. Chow and D.-B. Cao, "A survey study of critical success factors in agile software projects," J. Syst. Softw., vol. 81, no. 6, pp. 961–971, 2008.

[16] N. Baddoo and T. Hall, "Motivators of Software Process Improvement: An analysis of practitioners' views," J. Syst. Softw., vol. 62, no. 2, pp. 85–96, 2002.

[17] I. Asghar and M. Usman, "Motivational and de-motivational factors for software engineers: An empirical investigation," Proc. - 11th Int. Conf. Front. Inf. Technol. FIT 2013, pp. 66–71, 2013.

[18] H. Sharp, N. Baddoo, S. Beecham, T. Hall, and H. Robinson, "Models of motivation in software engineering," Inf. Softw. Technol., vol. 51, no. 1, pp. 219–233, 2009.

[19] C. de O. Melo, C. Santana, and F. Kon, "Developers Motivation in Agile Teams," in 2012 38th Euromicro Conference on Software Engineering and Advanced Applications, 2012, pp. 376–383.

[20] S. Beecham, H. Sharp, N. Baddoo, T. Hall, and H. Robinson, "Does the XP environment meet the motivational needs of the software developer? An empirical study," in Proceedings - AGILE 2007, 2007, pp. 37–48.

[21] J. Feigenspan, C. Kästner, S. Apel, and T. Leich, "How to Compare Program Comprehension in FOSD Empirically – An Experience Report."

[22] O. Mchugh, K. Conoby, and M. Lang, "Motivating agile teams: A case study of teams in ireland and sweden," in 5th International Research Workshop on Information Technology Project Management (IRWITPM 2010), 2010, pp. 71–83.

[23] G. Asproni, "Motivation, Teamwork, and Agile Development," Agil. Times, vol. 4, no. 1, pp. 8–15, 2004.

[24] J. D. Herbsleb and D. R. Goldenson, "After the Apraissal: A systematic survey of CMM experience and results," pp. 323–330, 1996.

[25] B. Tessem and F. Maurer, "Job Satisfaction and Motivation in a Large Agile Team," Lncs, vol. 4536, no. 5020, pp. 54–61, 2007.

[26] S. Seiler, B. Lent, M. Pinkowska, and M. Pinazza, "An integrated model of factors influencing project managers' motivation - Findings from a Swiss Survey," Int. J. Proj. Manag., vol. 30, no. 1, pp. 60–72, 2012.

[27] K. Conboy, S. Coyle, X. Wang, and M. Pikkarainen, "People over process: Key challenges in agile development," IEEE Softw., vol. 28, no. 4, pp. 48–57, 2011.

[28] L. Šteinberga and D. Šmite, "Towards Understanding of Software Engineer Motivation in Globally Distributed Projects," in 2011 IEEE Sixth International Conference on Global Software Engineering Workshop, 2011, pp. 117–119.

[29] J. D. Couger, V. Halttunen, and K. Lyytinen, "Evaluating the motivating environment in Finland compared to the United States—a survey," Eur. J. Inf. Syst., vol. 1, no. 2, pp. 107–112, 1991.

[30] A. Cockburn et al., "Advanced Software Technologies for Protecting America."

[31] D. Hutchison and J. C. Mitchell, Agile Processes in Software Engineering and Extreme Programming. 1973.

[32] S. Ahmed, K. Ghayyur, S. Ahmed, and A. Razzaq, "Motivators and Demotivators of Agile Software Development : Elicitation and Analysis," vol. 8, no. 12, pp. 304–314, 2017.

[33] M. Kropp and A. Meier, "Agile Success Factors A qualitative study about what makes agile projects successful," no. May 2015, 2015.

[34] S. Ahmed, K. Ghayyur, S. Ahmed, M. Ali, A. Razzaq, and N. Ahmed, "A Systematic Literature Review of Success Factors and Barriers of Agile Software Development," vol. 9, no. 3, pp. 278–291, 2018.

[35] A. C. C. Franca, D. E. S. Carneiro, and F. Q. B. da Silva, "Towards an Explanatory Theory of Motivation in Software Engineering: A Qualitative Case Study of a Small Software Company," 2012 26th Brazilian Symp. Softw. Eng., pp. 61–70, 2012.

[36] D. V Nithyanandan, "Work value as motivation among software professionals," Manag. Prudence J., vol. 1, no. 1, pp. 23–27, 2010.

[37] O. McHugh, K. Conboy, and M. Lang, "Using Agile Practices to Influence Motivation within IT Project Teams," Scand. J. Inf. Syst. (Special Issue IT Proj. Manag., vol. 23, p. pp 85-110, 2011.

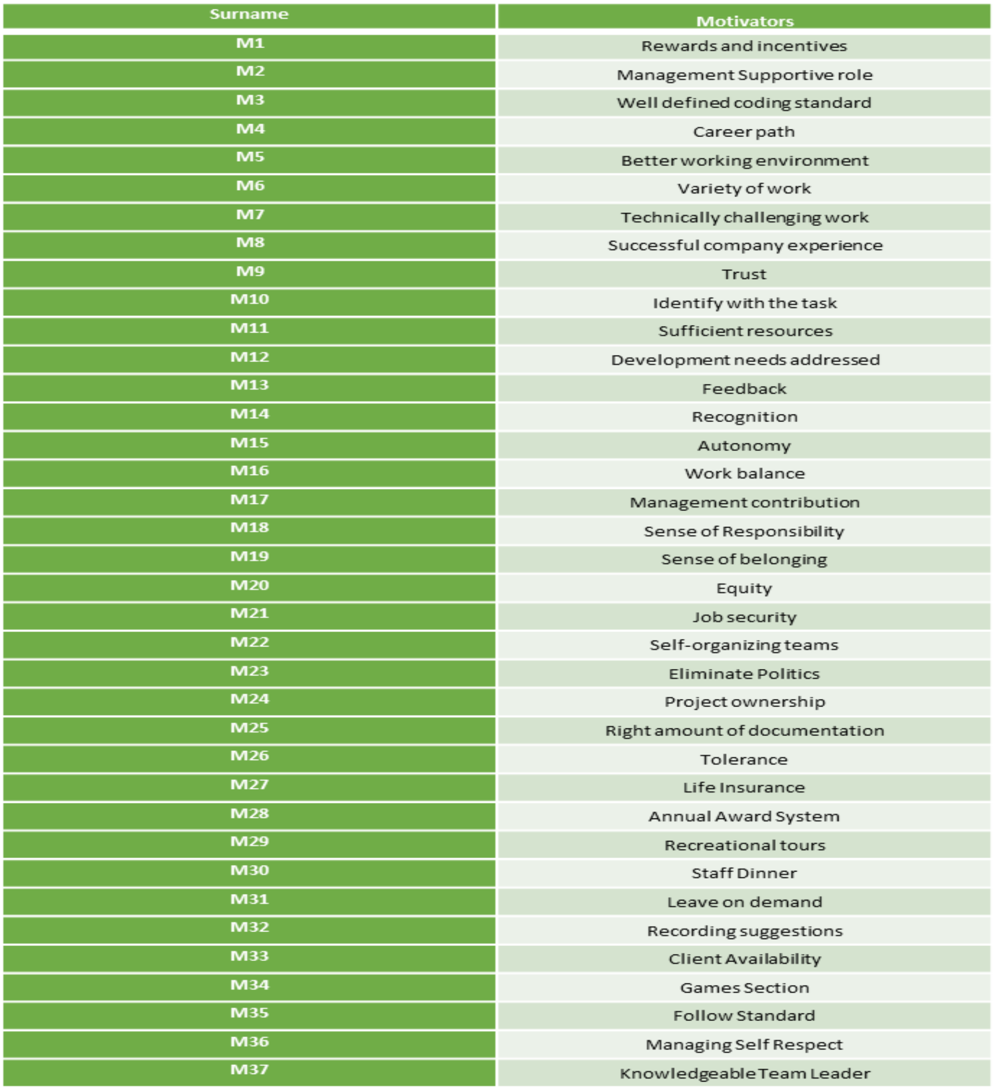
[38] S. Misra, V. Kumar, U. Kumar, K. Fantazy, and M. Akhter, "Agile software development practices: evolution, principles, and criticisms," Int. J. Qual. Reliab. Manag., vol. 29, no. 9, pp. 972–980, 2012.

[39] A. Baird and F. J. Riggins, "Planning and Sprinting: Use of a Hybrid Project Management Methodology within a CIS Capstone Course," J. Inf. Syst. Educ., vol. 23, no. 3, pp. 243–257, 2012.

[40] D. Woit and K. Bell, "Do XP customer-developer interactions impact motivation? findings from an industrial case study," Proc. 7th Int. Work. Coop. Hum. Asp. Softw. Eng. - CHASE 2014, pp. 79–86, 2014.

APPENDIX

Fig. 9 shows the detailed list of motivators acronyms used in articles.



Surname	Motivators
M1	Rewards and incentives
M2	Management Supportive role
M3	Well defined coding standard
M4	Career path
M5	Better working environment
M6	Variety of work
M7	Technically challenging work
M8	Successful company experience
M9	Trust
M10	Identify with the task
M11	Sufficient resources
M12	Development needs addressed
M13	Feedback
M14	Recognition
M15	Autonomy
M16	Work balance
M17	Management contribution
M18	Sense of Responsibility
M19	Sense of belonging
M20	Equity
M21	Job security
M22	Self-organizing teams
M23	Eliminate Politics
M24	Project ownership
M25	Right amount of documentation
M26	Tolerance
M27	Life Insurance
M28	Annual Award System
M29	Recreational tours
M30	Staff Dinner
M31	Leave on demand
M32	Recording suggestions
M33	Client Availability
M34	Games Section
M35	Follow Standard
M36	Managing Self Respect
M37	Knowledgeable Team Leader

Fig. 9. Appendix.

Social Success Factors Affecting Implementation of Agile Software Development Methodologies in Software Industry of Pakistan: An Empirical Study

Muhammad Noman Riaz
Department of Computer Science
Virtual University of Pakistan
Lahore, Pakistan

Athar Mahboob, Attaullah Buriro
Department of Electrical Engineering
KFUEIT
Rahim Yar Khan, Pakistan

Abstract—**During the past few years it has been observed that the implementation of Agile software development methodologies have become a part and parcel in software development projects not only in large and developed organizations but also in small organizations despite the existence of misapprehension that Agile methodologies are only valid for large scale projects and established organizations. Keeping in view the potential of Agile software methodologies and with the aim of eliminating this misconception, a mixed method methodology was adopted to conduct a study for determining the social factors that contribute or have influence in the successful implementation of Agile software development methodologies. In this study, face-to-face interview sessions were conducted with 271 software professionals that include Portfolio/Program/Project Managers, Scrum Masters and Product Owners representing 28 software development companies operating in Pakistan to gauge the influence of social factors on the success of Agile software projects. The study concluded that the size of the project has nothing to do with the success of a project or otherwise but there exist certain other factors like visionary leadership, degree or level of Agile software practices, congruence value, etc. contribute significantly in success of a project.**

Keywords—Agile methodologies; social factors; congruence value; visionary leadership; software developers

I. Introduction

It is a well-established fact that Agile software development methodologies are and applicable and beneficial for the small sized teams and small scaled projects but the implementation and scaling of these Agile methods are daunting and challenging for the software organizations [1]-[4]. However, contrary to this apprehension large organizations are suffering from the same problems [1]. As the main focus of Agile software development methodologies is the interaction and style of communication among the project team members as well as how the leadership is involved in the planning, execution and monitoring of the project. Furthermore, the main cause of failure of software. development project is not due lack in technological prowess but it is mainly due to social factors such as lack of effective communication among team members at all levels [5], lack of interest in software development, misaligned teams, etc. [6]. So, it is of utmost importance to gain understanding about how and at what extent these social factors contribute and influence in the success of Agile based projects. Furthermore, the more interest lies in the success or failure of large scale IT projects as these projects are more prone to failures [7]. From research [3], [5], [6] it is evident that social factors like communication barriers, misaligned teams, ineffective leadership must be scaled up to a desired level before implementation of Agile methodologies in both large and small scale projects.

During the study, the factors like communication among team members and factors pertaining to leadership were assessed with respect to their contribution in the success of Agile based software project. Also, the greater emphasis was paid on impact and influence of these social factors on complex l and large scale software projects. Finally a conceptual model was proposed based on qualitative interview sessions based on best proven practices and literature. Afterwards, each considered success factor in the proposed conceptual model was assessed and tested in order to determine the relative significance of success factor in the success of Agile based software project.

II. Problem Statement and Relevance

The objective of this is to determine the influence of social factors in the success of Agile software development methodologies in small and large scale projects. For this purpose we have two objectives: (1) to verify the previously identified social success factors; and (2) to propose and subsequently test and validate several success factors independently and examine their relationship and influence necessary for the success of Agile project. This study is theoretically related with the previous studies in terms of role of communication and leadership style and presents deep insight of how communication management and leadership styles are appropriate in ever changing environments like Agile Software Development Projects. Secondly, the research gap that exists between social success factors and Agile based software development is likely to be reduced. Also, the results of the study would assist the Agile practitioners in the improvement of communication and leadership practices in the organizations.

III. Theoretical Background

In this study the selection of factors responsible for success in Agile based software projects have been based on the preceding research conducted by different researchers in different parts of the world. The prior research works depict that the there are several 'people related factors' that significantly contribute in Agile software development. In this study initially we have considered social success factors like communication style, style of leadership, congruence value, Agile practices adaption degree, and size of the project. The success of the project has been related with the term *'effectiveness'* - to what extent and level the project team members have managed to meet or exceed the desired outcomes of quality as demanded by the customer [8] by evaluating multiple points or ratings of success.

The ever changing and at times unrealistic requirements of the organizational leadership are a huge challenge in successful implementation of Agile methodologies especially in large scale projects [3]. A style of leader can evaluated and assessed on the basis of Transactional and Transformational leadership styles [9]. The Transactional leadership style refers to social transactions in which the rewards to the member and expectations that the management has from the team member must be explicitly stated and communicated to the member, and the focus of the management is for a brief period of time. However, as far as the Transformational leadership is concerned it is related with motivation, inspiration, expression of vision and emotional engagement or involvement of employees in the project keeping in view the long term engagement and commitment. It is our expectations that the Transformational leadership is more beneficial than Transactional leadership in Agile software development as the latter put more emphasis on well-being of team members as well encourages healthy communication among the team members at all levels.

From [5] it is evident that the lack of communication among team members of the project creates misunderstandings within the project team that eventually leads to the failure of the project. This lack of communication problem can be catered by encouraging the team members to share their ideas, experiences and knowledge with other members of the project team that helps in trust building and strengthens the interpersonal relationship binding [10]-[12], and these factors are considered as critical success factors for Agile software development [11]. Furthermore, the informal way of communication helps in conveying information without any delay and with the first hand knowledge from the source person to the intended recipient and hence facilitates the decision makers to quickly react on the problem(s) and continuously changing requirements as in Agile software development projects. Therefore, we can say that the style of communication rather than frequency of communication is more critical and important for the success of project and the Agile based software projects are no exception.

In addition how the team members interact with each other and what communication protocol they are following the similarity in values and goals (personal and professional) they possess greatly affects their interpersonal relationships [12], [13]. In case the team members have different values, goals and objectives then diversity occurs in the team and that increases relationship conflicts, decreases level of satisfaction among team members and as a consequence to this the performance of the software development team affected adversely [13], [14]. Therefore, we can say that level of congruence is critical for the alignment of teams and subsequent success of Agile based software project.

Another variable that we have across is the degree of adoption of Agile practices in the organization. This variable indicates perception of agility among project team members. The size of the project refers total number of team members involved in each Agile software project.

IV. Research Strategy

A. Research Method

In this research work, the high profile professionals of software industry of Pakistan were identified who have been involved in successful implementation of Agile software development methodologies. The study has been divided in two phases namely: In phase 1, an exploitative phase in which based on the interview sessions with Agile software practitioners a thorough and comprehensive model was developed and phase 2, in which the developed or proposed model was quantitatively validated.

In Phase 1, an explorative study phase, qualitative interviews were conducted with high profile professionals working in highly reputable software companies of Pakistan and have proven record of successful Agile software development methodologies in large scale projects. During the interview sessions the topics like general aspects of the projects, leadership style and communication pattern practiced during course of the projects. Based on these interview sessions and previously conducted research a comprehensive conceptual model has been developed. The five candidate success factors are: (1) transformational style of leadership; (2) style of communications; (3) congruence value; (4) agility degree; (5) size of project, as shown in Fig. 1 with arrows depicting the interrelationship of each social success factors with one another. Based on the interview sessions we expected that the social success factors other than project size would positively affect the project success. Also, it is expected that style of leadership and style of communication would be meditated with congruence value.

In Phase 2, the aim was to validate the developed conceptual model. The Hypothesis pertaining to relationships between social success factors were tested based on the data collected from 271 software professional that include Professional Scrum Masters, Program and Project Mangers, Product Owners and Team Leads and team members working on 52 different projects and associated with 28 software companies registered with Pakistan Software Export Board (PSEB). To facilitate the study at least one team member, Project / Program Manager, etc. were asked to fill out a questionnaire that permits the comparison of roles in a specific project. An online questionnaire having five sections along with demographic information was distributed to each of the subject / respondent. The included five sections of the questionnaire are: (1) demographic information of the respondent / subject; (2) Agile degree with which organizations are practicing Agile methods; (3) style of leaderships, the style of leadership was evaluated with the help of Multifactor

Leadership Questionnaire (MLQ) [9]; style of communication; (4) congruence value; and (5) success of project. The respondents responded to questions in each of these sections through 5-Point Likert scale in order to depict the level success perception. The values from all the scales fall within the acceptable reliability limits (Cronbach Alpha), but transactional leadership, and that has been eliminated from the dataset.

Furthermore, in order to determine the contribution of each of the candidate success factors in the project success a Regression and Meditation analyses was carried out at both project and individual level. At project level, a project in which individuals are performing specific roles in cluster was considered as one unit of analysis, thus allows the projects to be compared. At individual level, each respondent was considered as a single unit of analysis and hence allows the comparison of different roles.

B. Ethical Considerations

This research work has used only the published data and the documents related interview sessions will be made available to public. Therefore, the ethical conflict is not expected to develop.

C. Research Limitations

The research in bounded to some limitations as the literature related to Agile software development methodologies as most of Pakistan software companies practicing some software development methodologies rather than using Agile methodologies.

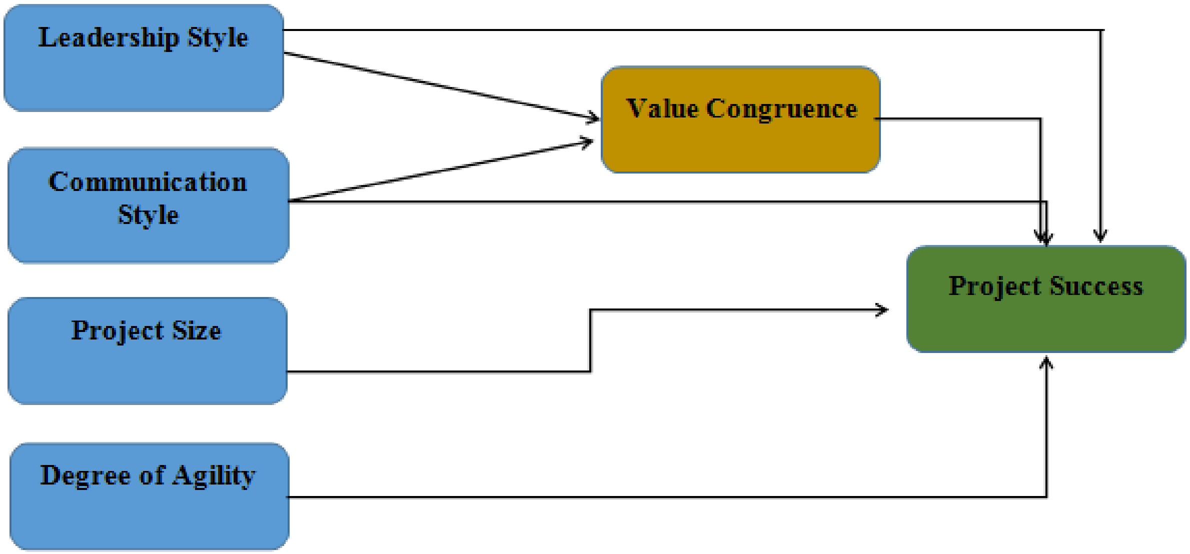


Fig. 1. A conceptual model.

V. Discussion and Results

After collecting the data from the respondents, we carried out T-tests measurement to confirm that there lies no significant difference in the replies of Program / Project Managers, Professional Scrum Masters, Product Owners, project team members, etc. in the interpreting the project success factors. Therefore, in this study the success of project has been interpreted in non - differentiated manner.

Before beginning the testing and validation of developed conceptual model, a Regression Analysis was carried out to determine the relationship between proposed success factors and the success of a project. The analysis results reveal that a significant and positive correlation exists between proposed candidate success factors with the success of a project. However, the size of a project does not reflect any positive correlation with the project success. Furthermore, the results of Regression Analysis depict there exist a predictive relationship between (1) project success and transformational leadership, (2) project success and value congruence, and (3) project success and degree of agility. Hence, we can say that based on Regression Analysis the most important and critical predictors are transformational leadership, congruence value and degree of Agility in the project.

On the basis of Regression Analysis a Meditation Analysis [14] was carried out when the predictors have a significant relationship with the proposed mediator (value congruence) and the success of project. The results depict that the value congruence was a meditating factor in the proposed conceptual model between value congruence and project success. Furthermore, the results depict that a full mediation was present between transformational leadership and success of project while a partial mediation was observed between degree of agility and project success. Such results clearly show that a high value congruence within the team members, keeping in view that a value congruence is both a mediating and predictor factor as far as the success of project is concerned.

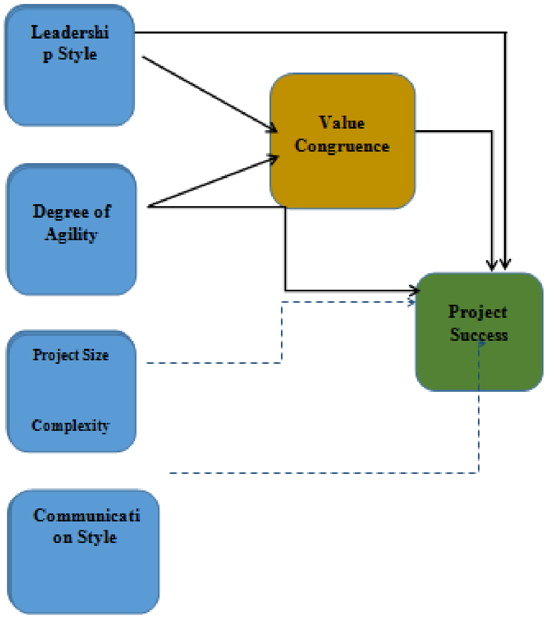


Fig. 2. Revised conceptual model.

The results of Regression Analysis also show that size of the project has nothing to do with the success of the project encouraging the application of Agile methodologies as far as there exist a high degree of transformational leadership, a high degree of agility and a high degree of value congruence.

The results obtained after carrying out such statistical analyses appreciate the refinement of initially developed conceptual model. The findings depict that it is not the style of communication but it is a degree of agility that predicts the success of Agile project, provided it is meditated by value congruence. The revised conceptual model is shown in Fig. 2 in which the significant relationships are depicted by bold arrows.

VI. Practical Implications

As from the conducted study we concluded that there exist three critical factors that can have a significant impact on the project success, namely: transformational leadership, degree of agility and value congruence. In order to determine the extent of impact of identified factors on the success of project, the considered projects have been divided into groups. One group is dedicated for the projects that have mediocre score on candidate success factors and other group which scored high on identified success factors. The value congruence and degree of agility have shown a large amount of impact on the project success, 0.50 and 0.45 respectively on Likert-Point scale while transformational leadership scores 0.06 as shown in Fig. 3.

Fig. 4 below depicts the extent to which the three identified success factors affect each other and / or act independently. For this purpose the projects are inspected based on the scores the group of projects scored on project success. The result shows that for project success the scores have been increased monotonously that leads to the fact that all three identified candidate success factors should be given maximum attention to optimize the project success.

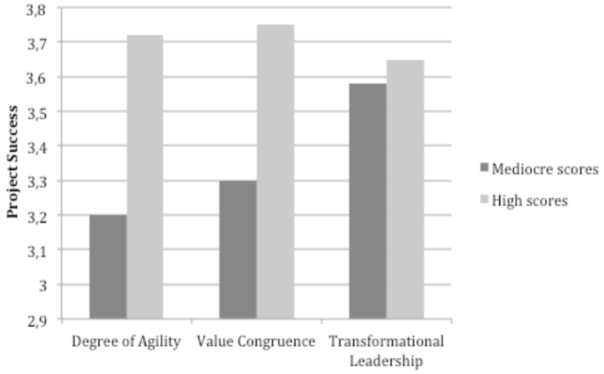


Fig. 3. Critical success factors.

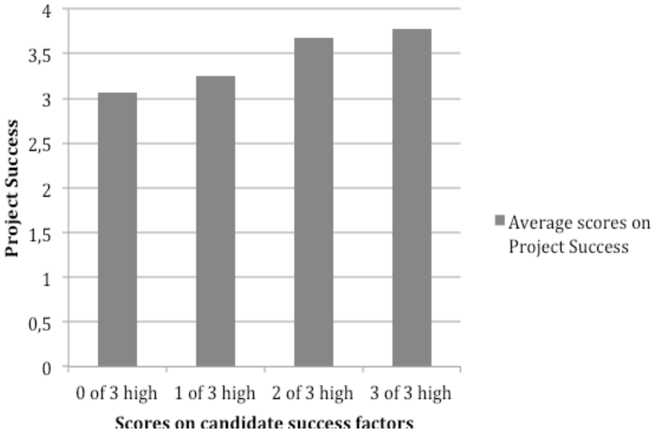


Fig. 4. A trinity of agile project success.

The findings encourage us to focus more on the alignment of values like vision, priorities and organizational goals in Agile projects. The alignment of such values can be achieved by ensuring informal communication among team members on regular basis and maintaining transformational style of leadership [10], [11].

VII. Conclusion and Future Research Work

The study identifies the social factors that may have a profound impact and influence on the success of the project and assessed the role of project size in project success. The study empirically identifies the social success factors based on communication in Agile software development projects by giving an insight of how actually these factors contribute in project success. Furthermore, it has also been revealed from the study that size of the project does not directly contribute in project success or failure. So, the project managers, Scrum masters etc. must focus on the practice of transformational leadership, agility and value congruence.

Furthermore, it is also interesting to note that the Agile methods can work seamlessly in large project however, the Agile puts more stress on short sprints and small teams. It is pertinent to clarify here that the project size does not necessarily have no impact on project success or failure but it is not an explaining factor here. Yes, the large size have more chances to fail as compare to small size projects but this cannot be compulsorily be elaborated by the size of the project. The study clearly reveals that the application of Agile software development methodologies can be beneficial for the project when the identified social factors like transformational leadership, value congruence and agile are of high value.

In future, our plan is to increase size and diversity of the sample and study be conducted for a longer period of time in order to further validate the proposed conceptual model.

References

[1] K.Beck, *Extreme programming explained: embrace change.* Addison Wesley Professional, 2000.

[2] D.J. Reifer, F. Mauer, and H. Erdogmus, "Scaling agile methods," *software*, IEEE, vol. 20, no. 4, pp. 12-14, 2003.

[3] B. Boehm, "Get ready for agile methods, with care," *Computer*, vol. 35, no. 1, pp. 64-69, 2002.

[4] J. Eckstein, *Agile software development in the large: Diving into the deep.* Addison-Wesley, 2013.

[5] M. Bloch, S. Blumberg, and J. Laartz, "Delivering large- scale IT projects on time, on budget, and on value," *Harvard Business Review*, 2011.

[6] K.A. Jehn, "A multimethod examination of the benefits and detriments of intragroup conflict. *Administrative Science Quarterly,* vol. 40, no. 2, pp. 256-83, 1995.

[7] J.R. Hackman, *The design of work teams. In J. Lorsch (eds.),* Prentice Hall, 1987.

[8] B.M. Bass, *Leadership and performance beyond expectations*, Free Press, 1985.

[9] J.R. Turner, and R. Müller, "Communication and co- operation on projects between the project owner as principal and the project manager as agent," *European Management Journal,* vol. 22, no. 3, pp. 327-336, 2004.

[10] S. Nerur, R. Mahapatra, and G. Mangalaraj, "Challenges of migrating to agile methodologies, *Communications of the ACM,* vol. 48, no. 5, pp. 72-78, 2005.

[11] J.R. Hackman, Groups that work (and those that don't), Jossey-Bass, 1990.

[12] K.A. Jehn, G.B. Northcraft, and M.A. Neale, "Why differences make a difference: A field study of diversity, conflict and performance in workgroups, *Administrative science quarterly,* vol. 44, no. 4, pp. 741-763, 1999.

[13] R.M. Baron and D.A. Kenny, "The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations," *Journal of Personality and Social Pshycology,* vol. 51, pp. 1173-1182, 1986.

[14] R.M. Baron and D.A. Kenny, "The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations," *Journal of Personality and Social Pshycology,* vol. 51, pp. 1173-1182, 1986.

A Multi-Criteria Decision Making to Rank Android based Mobile Applications for Mathematics

Seren Başaran, Oluwatobi John Aduradola
Computer Information Systems
Near East University
Lefkoşa 98010 via: Mersin 10 Turkey, Cyprus
European Centre for Research and Academic Affairs (ECRAA)
PO Box 1045, Lefkoşa via: Mersin 10, Turkey, Cyprus

***Abstract*—Exponential growth in the amount of mobile applications for Mathematics has led users to confusion and difficulty in selecting proper application manually which suits to their needs. Therefore, there exists an imperative need for automated and efficient selection of mobile applications for Mathematics where users still heavily trust either application store ratings or the content rated by the application developer. In this study, fuzzy scale weights together with ELECTRE I (ELimination and Choice Expressing REality) were used to solve a typical multi-criteria decision making problem on ranking selected mobile applications for Mathematics with respect to given set of criteria. The alternatives are mobile applications for Mathematics and were chosen from Google Play Store through considering top five highest user ratings and high usage frequencies. Ten sets of criteria on technical and pedagogical aspects specific to mobile applications and five alternatives were used in the ranking process. Findings suggest that ELECTRE I with fuzzy scale weights are remarkably practical for outranking and selection processes. Particularly in the case of unclear and imprecise ratings, this method could offer substantial solution.**

***Keywords*—ELECTRE; mobile applications for mathematics; multi-criteria decision making; pedagogical requirements; technical requirements**

I. Introduction

Mobile devices have become very dominant in our lives. The use of mobile devices has been extended merely from making calls and sending text messages to improved ability to execute various applications in demand. This took forward the usage and now smart phones have capability to support mobile learning [1]. The shifting and integration of mobile technologies in education milieu has caused users to use their own mobile devices for teaching and learning practices. Mobile devices are considered to be more affordable than PCs and laptops [2]. Authors in [3], reported that mobile phones are already be the part of the higher education for teaching and learning online courses. In addition, the researcher in [4] remarked upon benefiting from mobile phones for educational practices. He mentioned that it is possible to learn “anything, if developers designed it right”. As time goes by several mobile applications for learning certain subjects have been developed to make learning easy but prior to when these applications are been made available to end users some tests should be performed to ensure it is of satisfying quality, reliable and it meets the specific criteria or requirements.

The authors in [5] defined mobile applications for learning as mobile applications that make it possible for users to exercise learning in a changeable position. These mobile applications could establish anytime, anywhere learning environment [6], [7]. This technique of learning provides more flexibility and freedom to the learner which as a result fosters higher adoption rates by many individuals and educational institutions.

Numerous mobile applications for learning were introduced which assist in learning Mathematics at various sub disciplines of Mathematics and other fields as well [8]. Particularly mobile applications for Mathematics allow users to evaluate mathematical functions, giving graphical abilities and provide some sorts of mobile calculators. Mobile technologies that provide support to learning Mathematics via using mobile devices have likewise been expanding in the course of the most recent decade [9].

However, the researcher in [10] reported that there are over 4000 mobile applications specific to Mathematics to select from which have paved the ways for myriads of options to make selection on which mobile learning application to adopt. This scenario has led many individuals making a premature selection of mobile applications for mathematics because making an efficient selection from more than 4,000 applications seems tedious and time consuming, thus making the proper selection is crucial to enhance applications’ continuity in usage and enhancements in development. However, with the help of automated decision making techniques such as multi-criteria decision making, the burden on the decision makers will be minimized considerably.

Hence, to address this problem an evaluation framework with automated selection process model was proposed to provide a roadmap for making a reliable selection of mobile applications for Mathematics. So far, there is only one study to focus on evaluating and selecting suitable applications for mathematics by using multi-criteria decision making approach through considering technical and non-technical aspects such as user satisfaction [11].

Numerous frameworks to evaluate a software in general from technical point of view exist where some of the well-known models are; ISO/IEC 9126, ISO/IEC 25010, FURPS, [12]-[14], etc. It was stated that ISO 9126 software quality characteristics could be beneficial for evaluating mobile applications in general [15]. But till now, no such evaluation is available involving technical and pedagogical aspects together particularly for the evaluation of mobile applications. This study aims to adopt a framework from two viewpoints; technical and pedagogical aspects.

There are numerous multi-criteria decision making (MCDM) techniques which are utilized for the purpose of decision-making such as ANP, AHP, FAHP, SMART, ELECTRE, PROMETHEE, TOPSIS, etc. The researcher in [16] summarized some of the most frequently used MCDM methods in evaluating digital learning objects as; ELECTRE, Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), multiplicative exponential weighting (MEW), simple additive weighting (SAW) and Fuzzy Analytic Hierarchy Process (FAHP). These MCDM methods can be compared according to trustworthiness, perceived simplicity, quality and robustness. Among these, ELECTRE method is used for choosing the most suitable action from a given set of actions. ELECTRE is among most widely used MCDM methods that can be applied to many practical activities. ELECTRE method works with an input of criteria based ratings of alternatives by decision maker(s) which is named as decision matrix and preference information of the criteria expressed as weights and thresholds [17]. The ELECTRE I technique for picking the most suitable activity from a given arrangement of activities was contrived in 1965. The ELECTRE is for 'Disposal ET ELimination Et Choix Traduisant la REalité (Elimination and Choice Expressing the Reality). ELECTRE is an outstanding MCDM strategy that has a background marked by fruitful genuine applications. ELECTRE I requires the contribution of criteria assessments for the options, called choice network and inclination data, which are communicated as weights, limits, and different parameters [17].

This study aims to apply fuzzy scale weights with ELECTRE I to obtain the outranking of alternatives through adopted technical and pedagogical criteria.

II. Related Work

It was predicted in an internet report that in 2017, 268,692 millions of total free and paid-for downloads of mobile applications was available[1]. In 2018, it is normal that this number will increment to 254 billion downloads with 48% increase in download rate in 2017 as compared to 2013[1]. These statistics reveal that mobile devices are heavily taking part in our lives day by day: at home, at work, in the public and in teaching and learning as well. Particularly speaking, there exist several mobile applications to deal with numerous endeavors and instructive applications for practicing Mathematics. The consistent utilization of multipurpose innovations empowers the variability in mobile applications for learning. The researcher in [10] remarked upon that there are more than 4000 mobile applications to choose from. The same study also identified that “Despite the rapid expansion of the use of mobile applications in the educational domain, there is a lack of empirical studies as to their effectiveness in supporting learning, particularly in relation to Mathematics”. This absence of available research likewise reaches out to the employments of applications by instructors. This scenario has led many users into making a premature selection of mobile applications for mathematics because making an efficient selection from over 4000 applications seems tedious and time consuming. Thus, making right selection is crucial to enhance its continuous usage and developments. Hence, to address this problem a software quality model to provide a roadmap for making a reliable selection of mobile learning application for mathematics from different but conflicting options are inevitable. Particularly the evaluation of mobile applications for Mathematics by employing multi criteria approach have been neglected by the literature where Mathematics is the fundamental field which constitutes the basis for science and engineering. Therefore evaluating any mathematics learning related mobile applications is indispensable and will be quite beneficial for users.

Regardless of the abundance and expanding usage of mobile applications only one study was located in the extant literature to evaluate mobile applications particularly for Mathematics using MCDM methods [11]. There exist no other studies either for selecting or evaluating mobile applications in general by applying any of the MCDM methods. In only located relevant study, the authors in [11] proposed an adopted model defining both quality and user satisfaction used to evaluate mobile applications for Mathematics by utilizing hybrid Fuzzy AHP and TOPSIS approaches together. 11 criteria used were based on the technical and non-technical aspects specific to the mobile applications for Mathematics. The technical aspects were adopted from the ISO 9126 model of while non-technical aspects were considered as user satisfaction. The weight of each criterion derived was determined through using Fuzzy AHP approach while the alternatives as mobile applications for mathematics were ranked by applying TOPSIS.

The rest of the studies mentioned below have applied ELECTRE method into different disciplines from supplier selection to personnel, network, environmental impact, m-commerce candidate partner and project selection etc. Most of these studies employ FAHP and ELECTRE together. The researchers in [18] applied Fuzzy AHP and ELECTRE to cover the issue of a network selection where the network alternatives were ranked, utilized fuzzy numbers since the importance of criteria cannot be exactly defined to integrate subjective judgment in decision-making. The authors in [19] proposed a methodology that was carried out on a hybrid approach; fuzzy AHP–ELECTRE approach. The criteria weights were computed with the FAHP method, eight criteria were used in this study. In addition, fuzzy ELECTRE I was utilized to the alternatives. The study ended with an aggregate matrix to rank the alternatives finally. The researcher in [20] proposed an M-commerce partner selection method that uses a hybrid MCDM approach, AHP and ELECTRE I with a set of 13 criteria and 5 m-commerce candidate partners where AHP determined the weight of the 13 criteria and ELECTRE I ranks the candidate partners. The researchers in [21] applied ELECTRE I to select proper supplier with 4 different suppliers for computer

[1] Gartner, Inc., 2014. Gartner Says Mobile App Stores Will See Annual Downloads Reach 102 Billion in 2013. [online] Available at: http://www.gartner.com/newsroom/id/2592315 [Accessed 25 May 2018].

hardware and 13 criteria. The researchers in [22] used ELECTRE I to select most suitable personnel. 7 criteria and 5 decision makers involved in this study for selecting most suitable from five personnel using ELECTRE method and were again ranked by using AHP by considering 5 decision makers to rank 5 projects with respect to 5 criteria as financial, solution delivery, strategic contribution, risk management and environmental factors. Researchers in an earlier study employed ELECTRE I to rank 5 projects through using ranking tool [23]. The same study also highlighted the dominance of ELECTRE method over other MCDM methods through the inclusion of thresholds and outranking [23].

In general, evaluation by using MCDM methods is based on some set of criteria. Few studies were located in the literature for the evaluation of the quality of mobile applications despite their exponentially growing usage rates [24], [25]. The authors in [25] stated the difficulty of finding quality evaluation models specific to mobile learning applications. This study encompasses adopted technical and pedagogical aspects together as the selection criteria from located studies. Therefore, the main technical requirements subject to this study are; user interface (usability, navigation and orientation), reliability and maintainability (error free, easiness of installation, easiness of upgrade), efficiency and performance (energy consumption, responsiveness) and the pedagogical requirements are; content quality, content presentation and content organization. The selected requirements are crucial in expanding users' engagement, inspiration, learning, capability, and capacities.

Despite growing usage rates in mobile applications in general and abundantly available mobile applications particularly for Mathematics learning, there should be less time demanding and easier automated ways for users to select proper application to their use. To remedy this problem multi-criteria decision making methods can be applied to rank or evaluate the quality of the mobile applications for Mathematics which is a fundamental field of study. So far, technical and non-technical aspects were considered but pedagogical aspects were understated by the current literature. Also studies on mobile learning applications that are specific to quality evaluation frameworks are seldom. ELECTRE I method which is quite practical in addressing particularly ranking problems are frequently seen in the studies used along with another method namely FAHP in existing studies. Therefore, in the light of above, this study adopts 10 technical and pedagogical aspects as criteria to rank 5 top rated mobile applications for Mathematics by using fuzzy scale weights with ELECTRE I method.

III. Methodology

A. Alternatives: Mobile Applications for Mathematics

There are several mobile applications for Mathematics and they come in different forms depending on features like design, functionalities, purpose, limitations and target audience. This study targeted Android applications only because they are open source and have most populous mobile application store. These applications are distributed digitally via official Google Play store on the Android OS platform, which is either available freely or at some price. The Google Play store host millions of Android applications of different categories, such as social, games, education, security, etc. Users based on their experience rate these applications. Five mobile applications of Mathematics for adults were selected as alternatives for the evaluation based on their respective Google App Store user rating of at least 4.0 out of 5 and download rates greater than 1000 users. In addition, similarity in the features of the applications was also considered. Table I shows selected applications and user ratings with download numbers. As mobile applications continue to grow rapidly and gain popularity, different platforms have been developed to create and allow users to download these applications. 3.4 million mobile applications in October 2017 are available for download and the application store gives users the ability to express opinions through reviews and ratings[2]. By looking at the extant literature the most crucial criteria involving technical and pedagogical aspects were involved in the adopted evaluation framework.

TABLE I. ALTERNATIVES

Alternatives	Rating (0-5)	Downloads(in 2017)
Mathematics(A_1)	4.1	45182
Cymath(A_2)	4.5	2174
MalMath(A_3)	4.6	75216
MathPapa(A_4)	4.7	5098
Math 42(A_5)	4.7	2174

B. ELECTRE I Method

ELECTRE is frequently employed to find most suitable alternative with several set of criteria. Experts can select the most suitable choice through outranking alternatives via pairwise comparisons using concordance and discordance matrices. This MCDM approach has a special capacity to point out the exact motives of a decision-maker suggest an appropriate result through its ranking.

The ELECTRE I method steps were described as follows: If a problem has a number of alternatives $E1$,$E2$,$E3$,…,Ea and b number of criteria $F1$,$F2$,$F3$,…,Fb . Each alternative is rated to b criteria.

Step 1: K number of decision makers (DM) denoted as $D1$, $D2$, D3… DK. DMs rate weights verbally. Verbal variables are transformed into (l, m, u) which is a fuzzy number. k=1, 2… K and j=1, 2… b and the aggregated fuzzy significance weights can be;

$$\alpha_j^l = \min\{\gamma_{jk}\}\alpha_j^m = \frac{1}{K}\sum_{k=1}^{K}\gamma_{jk}\ \alpha_j^u = \max_k\{\gamma_{jk}\} \qquad (1)$$

Calculating weights, then normalization of aggregated fuzzy significance weights are:

$\widetilde{w} = (w_{\mathrm{j}}^{l}, w_j^{\mathrm{m}}, w_j^{u})$

Where

[2] AppBrain: Free versus paid Android apps. http://www.appbrain.com/stats/ free-and-paid-android-applications. (Last accessed: May 2018)

$$w_j^l = \frac{1/\alpha_j^l}{\Sigma_{j=1}^n 1/\alpha_j^l} w_j^m = \frac{1/\alpha_j^m}{\Sigma_{j=1}^n 1/\alpha_j^m} w_j^u = \frac{1/\alpha_j^u}{\Sigma_{j=1}^n 1/\alpha_j^u} \qquad (2)$$

Finally, the normalized aggregated fuzzy significance weight matrix is

$$\overline{W} = [\widetilde{w}_1, \widetilde{w}_2, \dots, \widetilde{w}_b] \qquad (3)$$

Step 2: A selection matrix with the aid of X= (xij) axb is fashioned for each criterion:

$$x = \begin{vmatrix} \chi_{11} & \chi_{1b} \\ \vdots & \vdots \\ x_{a1} & x_{ab} \end{vmatrix} \qquad (4)$$

Step 3: The normalized decision matrix R= (rij) axb by calculating rij, which shows the normalized criteria.

$$r_{ij} = \frac{\chi_{ij}}{\sqrt{\Sigma_{i--1}^a x_{ij}^2}} \qquad (5)$$

$$R = \begin{vmatrix} \gamma_{11} & \gamma_{1n} \\ \vdots & \vdots \\ r_{m1} & r_{mn} \end{vmatrix} \qquad (6)$$

Step 4: Seeing that each criterion has an exclusive weight, the weighted normalized decision matrix is constructed by taking product of significance weights of criteria and the values inside the normalized fuzzy selection matrix. . V = (v_{ij}) $_{axb}$ for i=1,2,…,a and j=1,2,…,b where $v_{ij} = r_{ij}$ x $\widetilde{w}_j$

$$V^l = \begin{vmatrix} v_{11}^l & v_{1b}^l \\ \vdots & \vdots \\ v_{a1}^l & v_{ab}^l \end{vmatrix} \; V^m = \begin{vmatrix} v_{11}^m & v_{1b}^m \\ \vdots & \vdots \\ v_{a1}^m & v_{ab}^m \end{vmatrix} V^u = \begin{vmatrix} v_{11}^u & v_{1b}^u \\ \vdots & \vdots \\ v_{a1}^u & v_{ab}^u \end{vmatrix} \qquad (7)$$

Step 5: weighted normalized fuzzy choice matrix was used to calculate concordance indices and sets are calculated with the use of the and pairwise assessment most of the options, respectively. If p and q are two options, the concordance index Cpq represents the pairwise contrast between p and $_q$ $(A_p \longrightarrow A_q)$. C_{pq} is the gathering of attributes where Ap is higher than or equal to Aq.

$$C_{pq}^l = \sum_{j^+}^{-} w_j^l \; C_{pq}^m = \sum_{j^+}^{-} w_j^m \; C_{pq}^u = \sum_{j^+}^{-} w_j^u \qquad (8)$$

Where j$^+$ are attributes covered inside the concordance set Cpq.

Step 6: The discordance indices mean the variances in judgment among alternatives p and q $(Ap \longrightarrow Aq)$. Dpq represents that Ap is worse than or equal to Aq. The discordance indices are calculated as;

$$D_{pq}^l = \frac{\Sigma_{j^+}\left|v_{pj^+}^l - v_{qj^+}^l\right|}{\Sigma_j\left|v_{pj}^l - v_{qj}^l\right|} D_{pq}^m = \frac{\Sigma_{j^+}\left|v_{pj^+}^m - v_{qj^+}^m\right|}{\Sigma_j\left|v_{pj}^m - v_{qj}^m\right|}$$

$$D_{pq}^u = \frac{\Sigma_{j^+}|v_{pj^+}^u - v_{qj^+}^u|}{\Sigma_j|v_{pj}^u - v_{qj}^u|} \qquad (9)$$

Where, j$^+$ are attributes contained inside the concordance set D$_{pq}$. Later, 2 threshold values were calculated by taking the mean of all the indices in concordance and discordance matrices.

Step 7: The very last concordance and discordance indices are computed as follows:

$$C_{pq}^* = \sqrt{\prod_{z=1}^Z C_{pq}^Z} \;\; , D_{pq}^* = \sqrt{\prod_{z=1}^Z D_{pq}^Z} \; where \; Z = 3 \qquad (10)$$

Step 8: In the end, Boolean concordance and discordance indices are calculated to decide high-quality alternative. Alternative with the minimal net concordance index and most discordance index is the satisfactory alternative among every alternative.

$$\tilde{C}_i = \sum_{i=1}^a C_{pq} - \sum_{i=1}^a C_{qp} \; \widetilde{D}_i = \sum_{i=1}^a D_{pq} - \sum_{i=1}^a D_{qp} \qquad (11)$$

C. Triangular Fuzzy Number(TFN)

Fuzzy is set of numbers where the quantity is not specific, it can also be addressed as an extension of the popular Boolean logic whose sets is not just 0 and 1 but a connection of different values where each and every value is assigned a weight. It can be defined as a set of values ranging from one interval to another. Fuzzy attaches more dynamism to expressions. So far, many types of fuzzy numbers exist such as triangular, trapezoidal, octagonal, pyramid, pentagonal, diamond and hexagonal fuzzy numbers. Among them, triangular and trapezoidal fuzzy numbers are the most frequently applied due to the ease of use and simplicity. The researcher in [16] gave detailed description on the use of triangular and trapezoidal fuzzy numbers. A triangular fuzzy number consists of the set of three real numbers ranging from minimum, most expected and maximum weights. Fig. 1 below depicts the triangular fuzzy number with its three values; $a_{1,}a_2$ and a_3 . Fig. 2 represents the membership function used in this study that converts linguistic variables into triangular fuzzy numbers and into crisp values by calculating mean value of each TFN known as defuzzified crisp values at the interval [0, 1]. The linguistic variable and crisp values were adopted from [26].

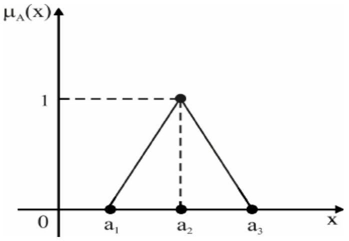


Fig. 1. A triangular fuzzy number.

TABLE II. LINGUISTIC SCALE AND DEFUZZIFIED CRISP VALUES

Linguistic Scale	**TFN**	**Fuzzy scale weights**
Excessively Low significance (EXL)	(0,0.002,0.004)	0.002
Extremely Low significance (EL)	(0.002,0.004,0.144)	0.050
Very Low significance (VL),	(0.004,0.144,0.362)	0.170
Low significance (L)	(0.144,0.362,0.364)	0.290
Semi-Low significance (SL)	(0.362,0.364,0.504)	0.410
Neither Low Neither High significance (NL,NH)	(0.364,0.504,0.722)	0.530
Semi-High significance (SH)	(0.504,0.722,0.724)	0.650
High significance (H)	(0.722,0.724,0.864)	0.770
Very high(VH)	(0.724,0.864,0.887)	0.825
Extremely High significance (EH)	(0.864,0.887,0.889)	0.880
Excessively High significance (EXH)	(0.887,0.963,1)	0.950

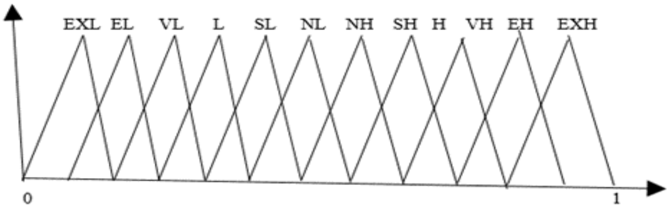


Fig. 2. TFN representation of linguistic variables.

D. Framework for Ranking

Fig. 3 shows the proposed evaluation criteria based upon technical and pedagogical aspects of the mobile applications that were adopted from [24] and [25].

The main adopted technical requirements are; user interface (usability(C_1), navigation and orientation(C_2), reliability and maintainability (error free(C_3), easiness of installation(C_4), easiness of upgrade(C_5), efficiency and performance (energy consumption(C_6), responsiveness(C_7) and pedagogical requirements are; content quality(C_8), content presentation(C_9) and content organization(C_{10}).

E. Ranking Procedure

Fig. 4 represents the steps followed during ranking of 5 alternatives using fuzzy scale weights and ELECTRE I method. First phase involves about deciding about the number of decision makers. In this case, expert who evaluated the alternatives with respect to given set of criteria has a PhD on Mathematics Education which could be considered as qualified for this kind of task. Later alternatives were decided for ranking process. Then pairwise comparison of each criteria was done in order to determine their significance over each other. Afterwards, chosen alternatives were evaluated by the expert with respect to each criteria. Finally, the ranking by using ELECTRE I method was performed to obtain the most suitable mobile application for Mathematics.

IV. RESULTS

The ranking process was implemented in five mobile applications for mathematics using fuzzy scale weights obtained from triangular fuzzy numbers (TFN). For ELECTRE I method, a pairwise matrix for correlation was made utilizing a relating fuzzy scale.

The 11-point linguistic scale of an earlier study was adopted to evaluate pairwise comparison for the each pair of criteria by the decision maker. Later, defuzzified crisp values were used in this study as fuzzy scale weights. The corresponding eleven linguistic scale are; excessively low significance (EXL), extremely low significance (EL), very low significance (VL), low significance (L), semi-low significance (SL), neither low neither high significance (NL NH), semi-high significance (SH), high significance (H), very high significance (VH), extremely high significance (EH) and excessively high significance (EXH) independently, which were used to depict the importance of weights of every criteria. Table II represents the linguistic scale and the corresponding crisp values that were adopted from [26]. In addition, Table III shows Step 1 involving pairwise comparison of criteria with each other that were evaluated by the decision maker using linguistic scales stated in Table II. Using Step 2, the value of pairwise comparison of each criterion was converted into fuzzy scale weights and their corresponding reciprocal value in Table IV. The sum column represents the sum of the rows for each criterion and weight column values were calculated by dividing sum to the number of criteria. Using Steps 3 and 4, normalized weight values were obtained dividing weight value for each criterion to the sum of the weight values. Sum of the normalized weights given in Table IV are 1.

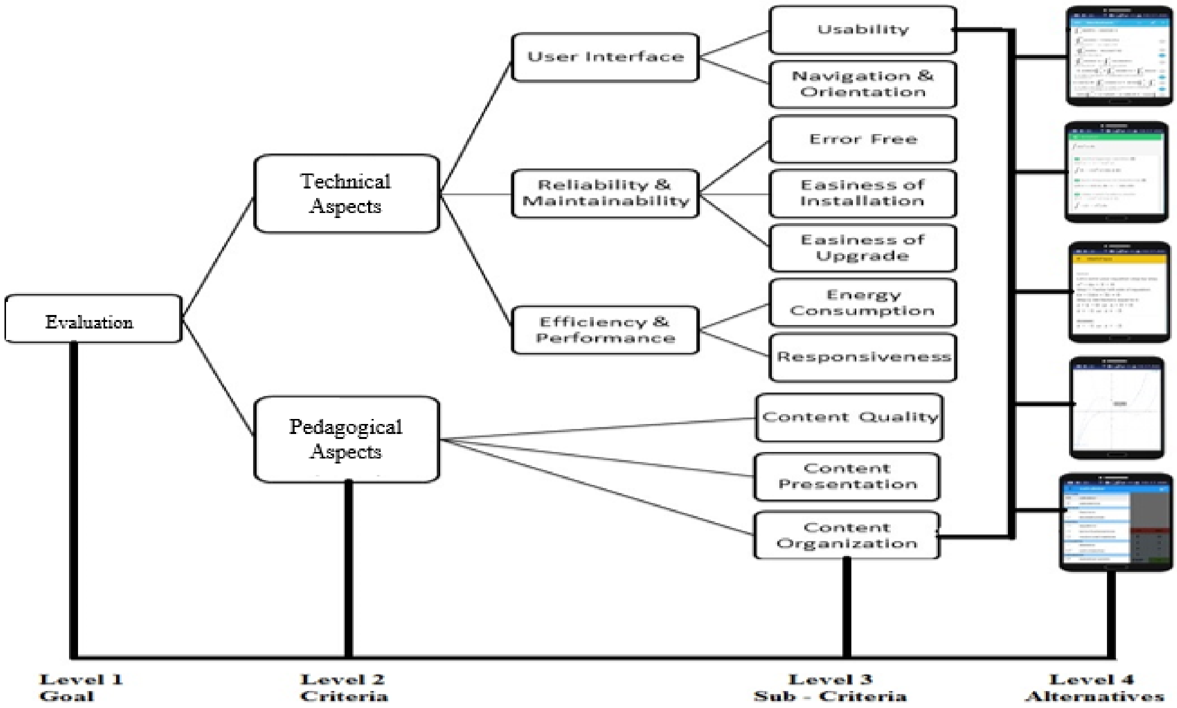


Fig. 3. Framework for ranking.

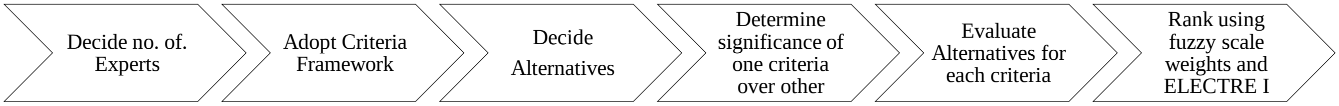

Decide no. of. Experts
Adopt Criteria Framework
Decide Alternatives
Determine significance of one criteria over other
Evaluate Alternatives for each criteria
Rank using fuzzy scale weights and ELECTRE I

Fig. 4. Ranking procedure.

TABLE III. PAIRWISE COMPARISON OF CRITERIA IN TERMS OF LINGUISTIC SCALE

Criteria	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
C1	-	VH	EH	H	H	VH	SH	H	H	VH
C2	VL	-	VH	VH	VH	H	NL NH	H	NL NH	NL NH
C3	EL	VL	-	VH	VH	VH	SH	H	H	H
C4	L	VL	VL	-	NL NH	L	VL	L	L	L
C5	L	VL	VL	NL NH	-	L	VL	L	L	L
C6	VL	L	VL	H	H	-	VL	L	L	L
C7	SL	NL NH	SL	VH	VH	VH	-	SL	SL	SL
C8	L	L	L	H	H	H	SH	-	H	H
C9	L	NL NH	L	H	H	H	SH	L	-	NL NH
C10	VL	NL NH	L	H	H	H	SH	L	NL NH	-

TABLE IV. PAIRWISE COMPARISON OF CRITERIA USING FUZZY SCALE WEIGHTS

Criteria	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	**Sum**	**Weight**	**Normalized weight**
C_1	-	0.825	0.880	0.770	0.770	0.825	0.650	0.770	0.770	0.825	7.085	0.7085	0.158643081
C_2	0.170	-	0.825	0.825	0.825	0.770	0.530	0.770	0.530	0.530	5.605	0.5605	0.125503807
C_3	0.050	0.170	-	0.825	0.825	0.825	0.650	0.770	0.770	0.770	5.605	0.5605	0.125503807
C_4	0.290	0.170	0.170	-	0.530	0.290	0.170	0.290	0.290	0.290	2.2	0.22	0.049261084
C_5	0.290	0.170	0.170	0.530	-	0.290	0.170	0.290	0.290	0.290	2.2	0.22	0.049261084

C_6	0.170	0.290	0.170	0.770	0.770	-	0.170	0.290	0.290	0.290	3.04	0.304	0.068069861
C_7	0.410	0.530	0.410	0.825	0.825	0.825	-	0.410	0.410	0.410	4.645	0.4645	0.104008061
C_8	0.290	0.290	0.290	0.770	0.770	0.770	0.650	-	0.770	0.770	5.08	0.508	0.113748321
C_9	0.290	0.530	0.290	0.770	0.770	0.770	0.650	0.290	-	0.530	4.6	0.46	0.103000448
C_{10}	0.170	0.530	0.290	0.770	0.770	0.770	0.650	0.290	0.530	-	4.6	0.46	0.103000448

TABLE V. EVALUATION OF ALTERNATIVES WITH RESPECT TO CRITERIA

Cri./Alt.	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
Mathematics(A_1)	50	40	40	100	100	50	40	40	40	50
Cymath(A_2)	60	60	50	100	100	50	50	40	30	50
MalMath(A_3)	50	70	50	100	100	50	50	50	60	50
MathPapa(A_4)	80	80	60	100	100	40	70	70	70	50
Math 42(A_5)	70	80	70	100	100	50	70	80	80	80

The decision maker rated alternatives between ranges 0-100 in Table V adopted from SMART strategy used by [27]. Using steps 5 and 6 concordance and discordance matrices were calculated in Tables VI and VII, respectively. The concordance matrix is calculated by pairwise comparison of each alternative's rating with the other alternative for each criteria. If first rating is greater or equal to second rating then the corresponding value would be the sum of normalized weights of the criteria which satisfy this condition divided by the sum of all normalized weights (equals to 1).

TABLE VI. CONCORDANCE MATRIX

Concordance Matrix	A_1	A_2	A_3	A_4	A_5
A_1	-	0.486	0.428	0.270	0.167
A_2	0.897	-	0.658	0.270	0.167
A_3	1	0.841	-	0.373	0.167
A_4	0.932	0.932	0.932	-	0.487
A_5	1	0.937	1	0.841	-

TABLE VII. DISCORDANCE MATRIX

Discordance Matrix	A_1	A_2	A_3	A_4	A_5
A_1	-	0.2	0.3	0.4	0.4
A_2	0.1	-	0.3	0.4	0.5
A_3	0	0.1	-	0.3	0.3
A_4	0.1	0.1	0.1	-	0.3
A_5	0	0	0	0.1	-

TABLE VIII. CONCORDANCE BOOLEAN MATRIX

	A_1	A_2	A_3	A_4	A_5
A_1	0	0	0	0	0
A_2	1	0	1	0	0
A_3	1	1	0	0	0
A_4	1	1	1	0	0
A_5	1	1	1	1	0

TABLE IX. DISCORDANCE BOOLEAN MATRIX

	A_1	A_2	A_3	A_4	A_5
A_1	1	0	0	0	0
A_2	1	1	0	0	0
A_3	1	1	1	0	0
A_4	1	1	1	1	0
A_5	1	1	1	1	1

TABLE X. GLOBAL MATRIX

	A_1	A_2	A_3	A_4	A_5
Mathematics(A_1)	0	0	0	0	0
Cymath(A_2)	1	0	0	0	0
MalMath(A_3)	1	1	0	0	0
MathPapa(A_4)	1	1	1	0	0
Math 42(A_5)	1	1	1	1	0

For discordance matrix, not preferred performance rating values were considered in pairwise comparison. The value is the maximum of differences of rating for each pair wisely compared alternatives with respect to specific criteria. After calculating concordance and discordance matrices 2 threshold values were determined by taking mean of all concordance indices and the mean of all discordance indices. The threshold values obtained from concordance and discordance matrices were calculated as; 0.64 and 0.2, respectively.

Using Steps 6 and 7 concordance and discordance domination matrices were calculated (see Tables VIII and IX). The concordance Boolean matrix is calculated by comparing all indices in the matrix to the threshold. If the value is greater than or equal to threshold then the corresponding value is 1 otherwise it is 0. For the discordance Boolean matrix is calculated by comparing all indices to the threshold value. If the value is less than threshold then it is 1 otherwise it is 0.

Table X represents the multiplication of the corresponding indices from concordance and discordance Boolean matrices were used to calculate the global matrix using Step 8. The value 1 represents first alternative outranks the second alternative in comparison whereas the value 0 represents no preference exists among the two compared alternatives.

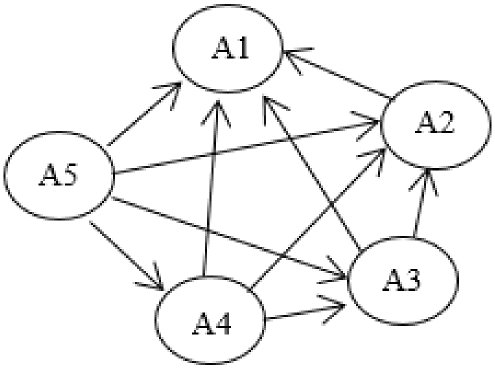


Fig. 5. Decision graph.

From the global matrix, using ELECTRE I method, the most suitable alternative is determined by scanning through the rows and selecting the alternative which has the most number of ones which represent connections. Therefore, according to the evaluation of the alternatives with respect to criteria yields to the ranking; Math42 > MathPapa > MalMath > CyMath > Mathematics. According to the decision graph given in Fig. 5, Math42 (A_5) was ranked as first with respect to the chosen technical and pedagogical criteria by using ELECTRE I method. In addition, ELECTRE I appear to be robust due to understandable and easy to follow steps which are fewer than other multi-criteria method steps with error free and less time consuming.

The decision graph of the ranking was given in Fig. 5.

V. Conclusion

This study adopted technical and pedagogical criteria to rank top 5 highly rated and downloaded mobile applications for Mathematics by using fuzzy scale weights together with ELECTRE I method. Increasing usage rates and the abundance in the number of mobile applications and their pervasive integration to teaching and learning have led users to choose the desired application instantly and with less time consuming efforts. Surprisingly this study is one of the rare studies to apply MCDM methods to an outranking of mobile applications. The only located earlier study was also carried out by the first author where researchers considered technical and non-technical aspects using hybrid MCDM method namely FAHP-TOPSIS to select mobile application for Mathematics. In terms of a higher number of alternatives and criteria ELECTRE I seems more efficient due to a fairly understandable method steps that is not only shorter compared to the other methods but more error free and less time consuming. So far, in this area, no studies have been located to employ ELECTRE I method, this study is a first that implements fuzzy scale weights and ELECTRE I to outrank Android based mobile applications for Mathematics.

In future with the increase usage and integration of mobile applications to teaching and learning, more research should be done using MCDM methods to evaluate the quality and select the desired mobile application. As for the evaluation of mobile applications for Mathematics requires concurrent thought of a few comparative and clashing criteria, MCDM methods are quite practical in handling such problems.

In this study, selecting the most suitable mobile application for Mathematics problem was remedied by using fuzzy scale weights with ELECTRE I method which is an efficient technique to deal with problems involving multiple criteria. The ranking was performed using 5 top rated alternatives with unclear and ambiguous judgment of decision maker's ratings. Finally, this method can be applied to any other disciplines as well.

The main limitations of the study can be listed as the number of decision maker is only one, there are fixed number of alternatives and the number of criteria. Also the judgment of the decision maker is effective in results. Therefore decision maker should have adequate background and level of expertise and should be objective as well.

As for future research thoughts, more number of decision makers, alternatives and criteria could be involved. In addition, the comparison of this single method to other multi criteria methods could be added to identify the effectiveness and efficiency of these techniques. Moreover, either web-based, mobile or stand-alone softwares for MCDM methods could be developed as an aid to decision making situation.

Acknowledgment

The first author would like to thank Öcen Başaran for her unwavering support and financial contribution to make this happen.

References

[1] C. X. N. Cota, A. I. M. Díaz, and M. Á. R. Duque, “Developing a framework to evaluate usability in m-learning systems,” Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality - TEEM ’14, 2014.

[2] J. Traxler, “Defining, Discussing and Evaluating Mobile Learning: The International Review of Research in Open and Distributed Learning, vol. 8, no. 2, Jun. 2007.

[3] M. Wang, R. Shen, R. Tong, F. Yang, and P. Han, “Mobile learning with Cellphones and PocketPCs,” Lecture Notes in Computer Science, pp. 332–339, 2005

[4] M. Prensky, “From Digital Natives to Digital Wisdom: Hopeful Essays for 21st Century Learning,” 2012.

[5] H. Willacy and N. Calder, “Making Mathematics Learning More Engaging for Students in Health Schools through the Use of Apps,” Education Sciences, vol. 7, no. 2, p. 48, Apr. 2017.

[6] R. Martin, T. J. Mcgill, and F. Sudweeks, “Learning Anywhere, Anytime: Student Motivators for M-learning,” Journal of Information Technology Education: Research, vol. 12, pp. 051–067, 2013.

[7] H. Peng, Y. Su, C. Chou, and C. Tsai, “Ubiquitous knowledge construction: mobile learning re-defined and a conceptual framework,” Innovations in Education and Teaching International, vol. 46, no. 2, pp. 171–183, May 2009

[8] A. Drigas and M. Pappas, “A Review of Mobile Learning Applications for Mathematics,” International Journal of Interactive Mobile Technologies (iJIM), vol. 9, no. 3, p. 18, Jul. 2015

[9] M. Bjerede, K. Atkins, and C. Dede. "Ubiquitous mobile technologies and the transformation of schooling." Educational Technology 50.2 (2010): 3-7.

[10] M A. Skillen, “Mobile Learning: Impacts on Mathematics Education”, Proceedings of the 20th Asian Technology Conference in Mathematics, 2015, pp. 205-214.

[11] S. Başaran and Y. Haruna, “Integrating FAHP and TOPSIS to evaluate mobile learning applications for Mathematics,” Procedia Computer Science, vol. 120, pp. 91–98, 2017.

[12] A. Alvaro & E.S. Almeida and S.R.L. Meira, "Towards a Software Component Quality Model," Submitted to the 5th International Conference on Quality Software (QSIC), 2005.

[13] B. H. Boehm, J. R. Brown, H. Kaspar, M. Lipow, E. J. MacLeod and M. J. Merritt, Characteristics of Software Quality, The Netherlands, Amsterdam:North-Holland, 1978.

[14] J. A. McCall. "Factors in software quality." US Rome Air development center reports (1977).

[15] P. Pocatilu and C. Boja. "Quality characteristics and metrics related to m-learning process." Amfiteatru Economic 11.26 (2009): 346-354.

[16] S. Başaran, "Multi-Criteria Decision Analysis Approaches for Selecting and Evaluating Digital Learning Objects," Procedia Computer Science, vol. 102, pp. 251–258, 2016.

[17] M. Sevkli, "An application of the fuzzy ELECTRE method for supplier selection," International Journal of Production Research, vol. 48, no. 12, pp. 3393–3405, Jun. 2010.

[18] D. E. Charilas, O. I. Markaki, J. Psarras, and P. Constantinou, "Application of Fuzzy AHP and ELECTRE to Network Selection," Mobile Lightweight Wireless Systems, pp. 63–73, 2009.

[19] T. Kaya and C. Kahraman, "An integrated fuzzy AHP–ELECTRE methodology for environmental impact assessment," Expert Systems with Applications, vol. 38, no. 7, pp. 8553–8562, Jul. 2011.

[20] Y. Guo. "A decision method for m-commerce partner selection based on AHP/ELECTRE I." Journal of Computational Information Systems 6.9 (2010): 3077-3086.

[21] S. Birgun, E. Cihan, "Supplier Selection Process using ELECTRE Method", Intelligent Systems and Knowledge Engineering (ISKE), pp.634-639, 2010

[22] A. R. Afshari, M. Mojahed, R. M. Yusuff, T. S. Hong, and M. Y. Ismail, "Personnel Selection using ELECTRE," Journal of Applied Sciences, vol. 10, no. 23, pp. 3068–3075, Dec. 2010

[23] J. Buchanan, P. Sheppard, and D. Vanderpoorten. "Ranking projects using the ELECTRE method." Operational Research Society of New Zealand, Proceedings of the 33rd Annual Conference. 1998.

[24] A. A. Economides, "Requirements of Mobile Learning Applications," International Journal of Innovation and Learning, vol. 5, no. 5, p. 457, 2008.

[25] G. W. Soad, N. F. D. Filho, and E. F. Barbosa, "Quality evaluation of mobile learning applications," 2016 IEEE Frontiers in Education Conference (FIE), Oct. 2016.

[26] A. Debnath, M. Majumder, and M. Pal, "Potential of Fuzzy-ELECTRE MCDM in Evaluation of Cyanobacterial Toxins Removal Methods," Arabian Journal for Science and Engineering, vol. 41, no. 10, pp. 3931–3944, Feb. 2016.

[27] S. Chou and Y. Chang, "A decision support system for supplier selection based on a strategy-aligned fuzzy SMART approach," Expert Systems with Applications, vol. 34, no. 4, pp. 2241–2253, May 2008.

Time Series Analysis for Shortened Labor Mean Interval of Dairy Cattle with the Data of BCS, RFS, Weight, Amount of Milk and Outlook

Kohei Arai, Osamu Fukuda,
Hiroshi Okumura
Graduate School of Science and Engineering
Saga University
Saga City, Japan

Kenji Endo
Morinaga Dairy Service Co. Ltd.
1-159 Toyoharaotsu, Nasugun
Nasumachi,
Tochigi 329-3224, Japan

Kenichi Yamashita
The National Institute of Advanced Industrial Science and Technology (AIST), 807-1 Shuku-machi, Tosu
Saga 841-0052, Japan

***Abstract*—Time series analysis for shortened labor mean interval of dairy cattle with the data of BCS (Body Condition Mass), RFS (Rumen Fill Score), Weight, Amount of Milk and Outlook is conducted. Method for shortened the labor mean internal of Japanese dairy cattle based on time-series analysis with the data of visual index of BCS, RFS, Weight, Amount of Milk and Outlook is proposed. In order to shortened the labor mean interval of dairy cattle is the purpose of this research. Through the experiments with 17 Japanese dairy cattle of the 17 Japanese anestrus Holstein dairy cattle, it is found that the combination of weight, BCS and amount of milk is a good indicator for identification of productive cattle. Therefore, the cattle which need hormone treatments can be identified.**

Keywords*—*Body Condition Score (BCS); Rumen Fill Score (RFS); dairy cattle; time-series analysis; cattle productivity

I. Introduction

The labor mean interval is defined as the period between a delivery and the next delivery. The labor mean of dairy cattle is getting longer and longer in the world wide basis. This implies that the total number of dairy cattle is getting down. Therefore, it is very serious problem so that there is strong demand to shortened the labor mean interval.

The labor mean interval of Japanese dairy cattle is approximately 410 days. This means that most of Japanese dairy cattle delivers calf about 410 days after the previous delivery typically. Country’s improved growth target days is 380 days. Therefore, the labor mean interval of Japanese dairy cattle has to be shortened by 30 days (410 to 380 days of labor mean interval would be a goal).

There are so many trials for identify the labor mean interval. Among the many factors of cattle productivity influence, the most influential one is the Body Condition Score: BCS, which is defined as “an effective management tool to estimate the energy reserves of a cow” [1]-[3] and most widely used for herd management. There are many identified system for measuring BCS, which varies according to different countries [1], [2]. Using BCS to evaluate cattle does not require any special equipment and can be conducted anytime during the year. Poor body condition is associated with reduced income per cow, increased postpartum interval, increased dystocia, and lower weaning weight. The most common and widely used (USA and Japan) BCS scale ranges from 1 to 5 with 0.25 increments [3]. Though BCS measured subjectively and its reliability is questioned, it is also evident that BCS have relationship with many other factors of bovine, such as postpartum interval, parity, etc. [3]-[6]. The authors have proposed the method for estrus cycle estimation with three influential factors (BCS, postpartum interval, and parity) for understanding the presence and absence of estrous cycle using a new unique Bayesian Network Model: BNM [7]. It, however, is not possible to consider a relation among the influencing factors.

Although ultrasound diagnostic instruments allow identifying pregnancy (major follicle in ovary) relatively easily, the instruments are not so cheap and not so easy to use for dairy cattle farmers. Therefore, there are strong demands for identification of pregnancy in easy way without any expensive instrument. BCS, RFS, weight, amount of milk and outlook are easy to measure and comprehensive.

In this paper, regressive analysis-based method for estrus cycle estimation is proposed here in this paper in order to consider a relation among the influencing factors. Experiments are conducted with 17 different Japanese Holstein cows observing with their BCS (2.0 to 3.25), hormonal treatments and parity numbers in order to discover the ideal timing for artificial insemination to make them pregnant. These data are acquired from the dairymen in Nasu-machi, Tochigi Prefecture. It is also important to mention that, all these 17 samples found anestrus in their farm.

The aim of the research work is to find out estimation equation of estrous cycle of bovine using regressive analysis for Japanese dairy industries. It is clear from National Livestock Breeding Center: NLBC, Japan that, the overall conception rate of live beef and dairy cattle is decreasing in last 20 years in Japan [8]. Moreover, the findings of relations among influencing factors of the measured BCS, hormone treatments, parity number, and so on are other objectives for improving cattle productivity and herd management. Moreover, using regressive analysis would assist the farm management to find out the presence of estrous cycle more objectively and in an accurate way. Meanwhile, method for productive cattle finding with estrus cycle estimated with BCS and parity number and hormone treatments based on a regressive Analysis is proposed already [9].

In this paper, the following influencing factors for estimation of estrus cycle as well as the labor mean interval, BCS, RFS, weight, amount of milk, and outlook are focused. Detection of major follicles and measurement of blood flow by ultrasonic diagnostic images in the ovaries are effective for estrus detection. It, however, does cost for ultrasonic imager. Meanwhile, aforementioned five factors can be measured relatively easy and do not need expensive instruments at all. Therefore, the five factors are focused in this paper. If productive dairy cattle can be identified, breeding of such cattle. Also, if detection of estrus can be done and if pregnancy cannot be confirmed, then appropriate hormone treatments can be done. This is a basic approach for shortened the labor mean interval.

The following section describes research background followed by preliminary results from the relation between the labor interval and the aforementioned influencing factors. Then, the most influencing factor is determined through time series data analysis followed by conclusion with some discussions.

II. Research Background

A. Body Condition Score

The research reveals to include BCS while considering the estrous cycle identification. BCS is the most significant influential factors in bovine productivity. An organized process for determining BCS was created at the University of Pennsylvania to help achieve consistency and repeatability in BCS. This system finds its accuracy toward the mid-range scores (2.50 to 4.00), which includes most cattle in this investigation. This mid-range is the most critical for making farm management decisions and most influential for the farm nutritionist. BCS outside this range indicate significant problems and varies significantly with respect to each individual inspector/observer. This research considering $BCS_{4.0}$ methods (quarter-point increase) in 17 individual cattle of Morinaga Dairy Service: MDS Co. Ltd. Japan and the following tables describe the meaning of BCS scale. The $BCS_{4.0}$ method (0.25 increase) have good repeatability across and within observers including simplified body scoring as well as have higher value as a diagnostic test.

The BCS process represents the observer’s view into the certain anatomical sites for each cow’s pelvic, loin areas, pin and hook bones, etc. Table I briefly elaborates the observing BCS of 17 individual cows from a farm of Iwate Prefecture, Japan under MDS cooperation.

TABLE I. BCS and it’s General Meaning for 17 Sample Cows

BCS	Meaning (in general)
2.25	No fat pads on pin and hook bones- angular shape
2.5	Palpable fat pads on pin and hook bones- angular shape
2.75	Pin bones- round shape and hook bones- angular shape with less fat pads
3.0	Fat pads on pin and hook bones- round shape
3.25	Visible fat pads on pin and hook bones- round shape

The BCS data is acquired on the time mainly be the fresh time after delivery (30 to 60 days) and the milking period until conception.

B. Rumen Fill Score

RFS is associated with feed intake. Also, it is known that RFS is related to blood parameters (Condition). RFS can be measured visually easily.

It is confirmed that RFS does not change in pregnant cows. For non-pregnant cows, RFS decreases gradually as the calving day approached. After calving, non-pregnant cows showed lower energy status compared with pregnant cows, and some non-pregnant cows showed anovulation and cessation of estrous cycle.

RFS is evaluated with the swelling of the apparent left-hand part (left side flank with the first stomach) to the last. It will be an indicator of satiety. In addition, the cow that is always full of satiety becomes an image like a car with a large volume of rumen and so-called engine displacement (large input and output). Therefore, aim for the highest score of 5 in the dry milk (prenatal) period is attempted. The uterus also gets bigger due to the development of the fetus, and the bait eating falls down, but it is vital to make it eat better. RFS data acquisition time is the same time as the above acquisition of BCS.

C. Hormone Treatments

Hormone treatments can be divided into two categories, CIDR (Vaginal indwelling type luteinizing hormone preparation, and Prepare estrus), and PG (Prostaglandin, and Uterine empyema).

Usually, CIDR is applied to the dairy cattle which has no estrus for a long time for prompt an estrus. If the CIDR does not work, then PG is applied to the cattle. PG has a function to regain corpus luteum. The corpus luteum is a structure on the ovary that produces progesteron. By injection of PG, estrus is induced by follicles growing without closing by losing corpus luteum (progesteron disappears). The purpose of using PG here triggers estrus. (Cow with corpus luteum due to uterine empyema is also used for treatment to release pus in the uterus by administration of PG, but estrous induction is the main in program insemination.

D. Weight

Weight of dairy cattle can also be measured easily. Usually, weight of dairy cattle decreases after delivery and gradually recovered for preparation of the next pregnancy. For body weight measurements, A weight scale is used in this survey. Study farmers do not have a usual scale. Major measures to measure weight by measuring the prescribed chest circumference.

E. Amoount of Milk

Amount of milk is another factor for the dairy cattle. Amount of milk is highly correlated with healthy condition, bait status, productivity of the dairy cattle.

Amount of milk is defined as "Milk amount [kg] of measurement month (-1 without survey)" in this paper.

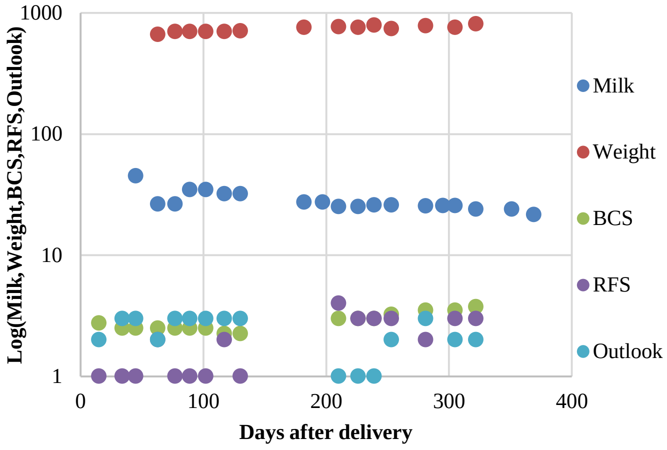
F. Outlook

Outlook is the factor which is reflected with appearance findings (1: Good taste 2: Normal 3: Danger 4: Impossible (difficult recovery) 5: Judgment pending). As a survey destination, these data are acquired as a dairy farmer in Nasu-cho, Nasu-gun, Tochigi Prefecture, Japan.

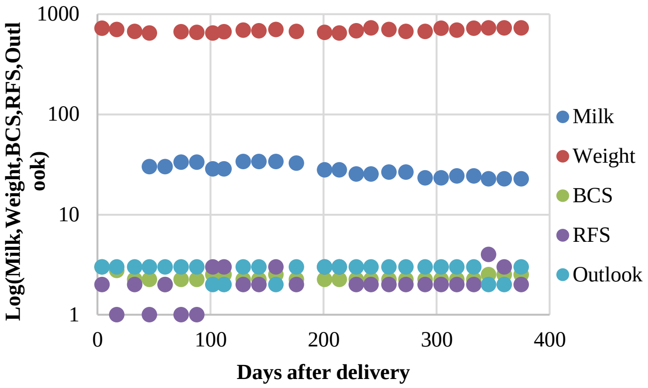
III. Time Series Analysis

A. Features of the Time Series of Raw Data

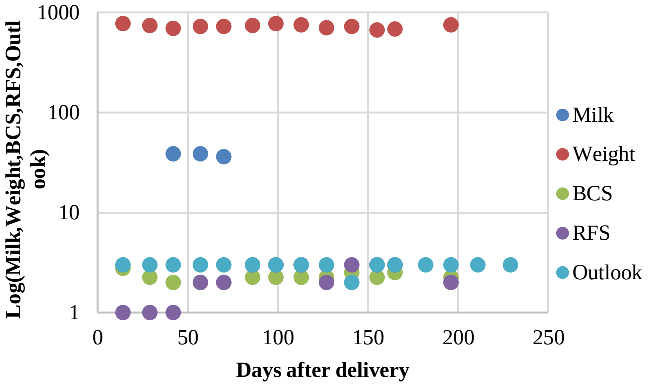
These 17 individual sample data of dairy cattle were collected from a dairy farm of Iwate Prefecture with the cooperation of Morinaga Dairy Service: MDS Co. Ltd., Japan. The BCS were observed in accordance with the UV method of Ferguson [3] by an experienced animal scientist of MDS. The PPI, Parity and other related information is collected from MDS. These 17 individual cattle were Japanese Holstein breed, which were found anestrus in the farm in Iwate Prefecture, Japan. The overall investigation for all these problematic dairy cow is under observation of MDS. Fig. 1 shows time series of data of the aforementioned influencing factors of the dairy cattle which has the data for more than 250 days after the previous delivery. There are 17 of the dairy cattle with the data for much longer than 250 days out of the 17 candidates of the Japanese dairy cattle.



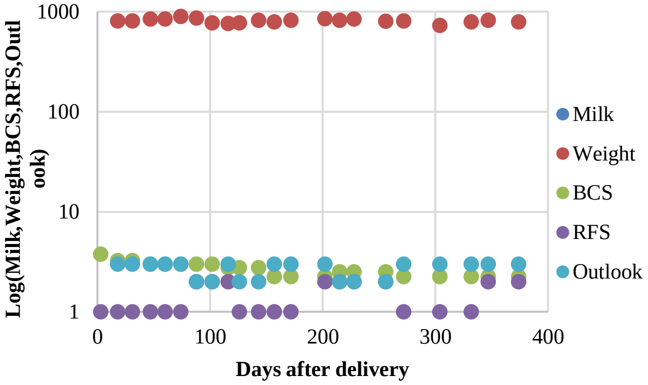
(a)



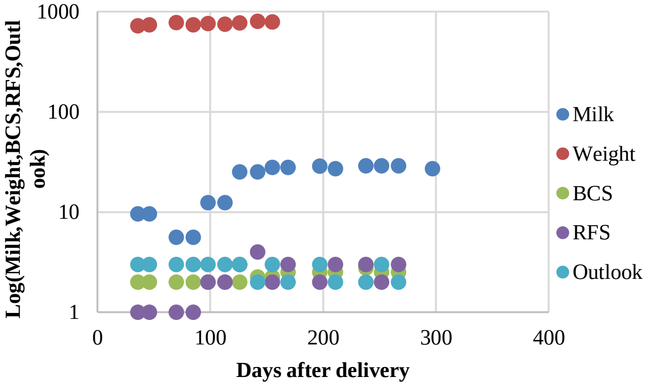
(b)



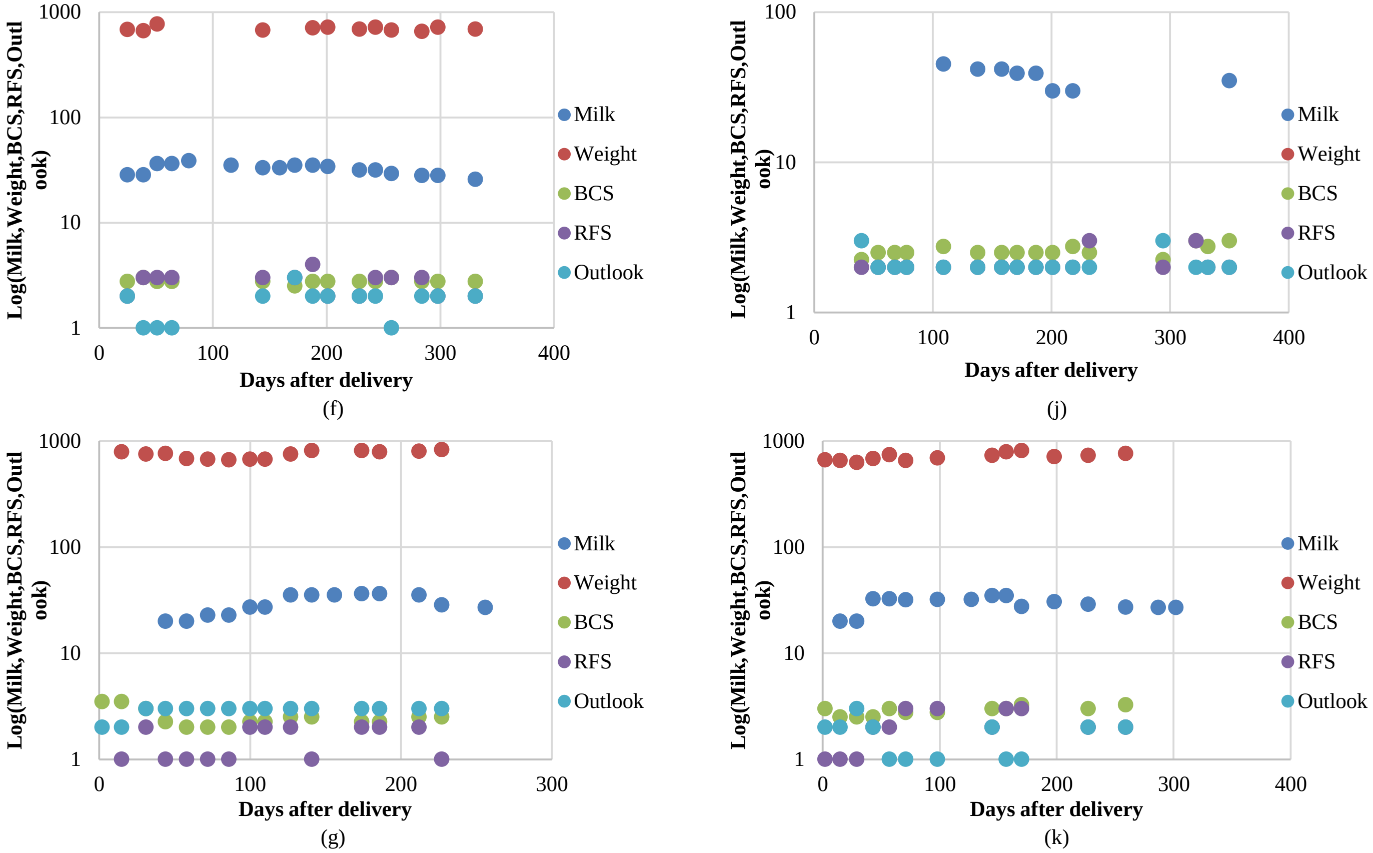
(c)



(d)



(e)

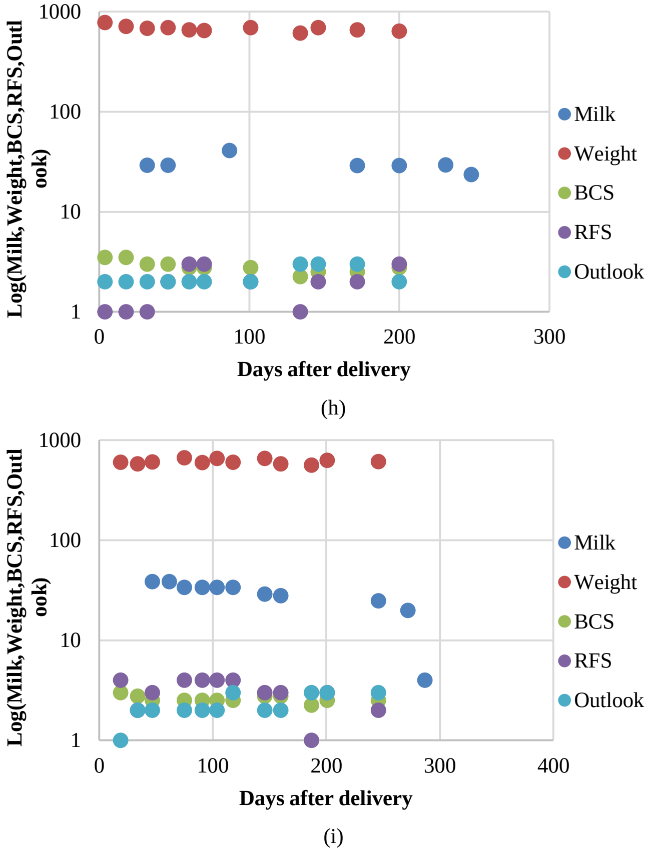


(f)

(j)

(g)

(k)



(h)

(i)

Fig. 1. Time series of data of the influencing factors.

Remarkable features of the time series of data are as follows:

1) Weight of the dairy cattle is getting down just after the delivery. Then, it is recovered gradually.

2) Amount of milk is varied up and down for some dairy cattle. It is getting down for some other dairy cattle.

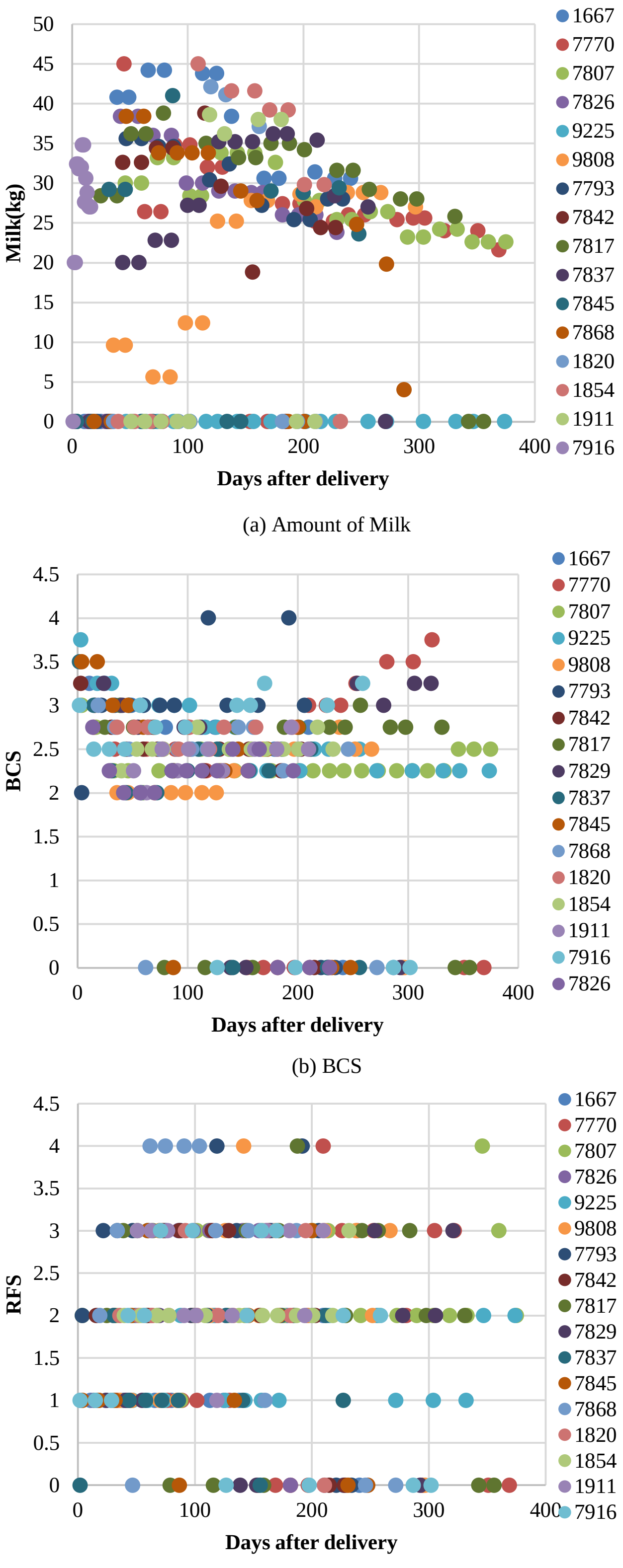
3) BCS is decreased just after the delivery. Then, it is recovered gradually for some dairy cattle. There are some other dairy cattle which show almost no change.

4) RFS is relatively steady and fluctuated randomly a little bit.

5) Outlook is also relatively steady and fluctuated randomly a little bit.

B. Major Results from the Time Series Analysis

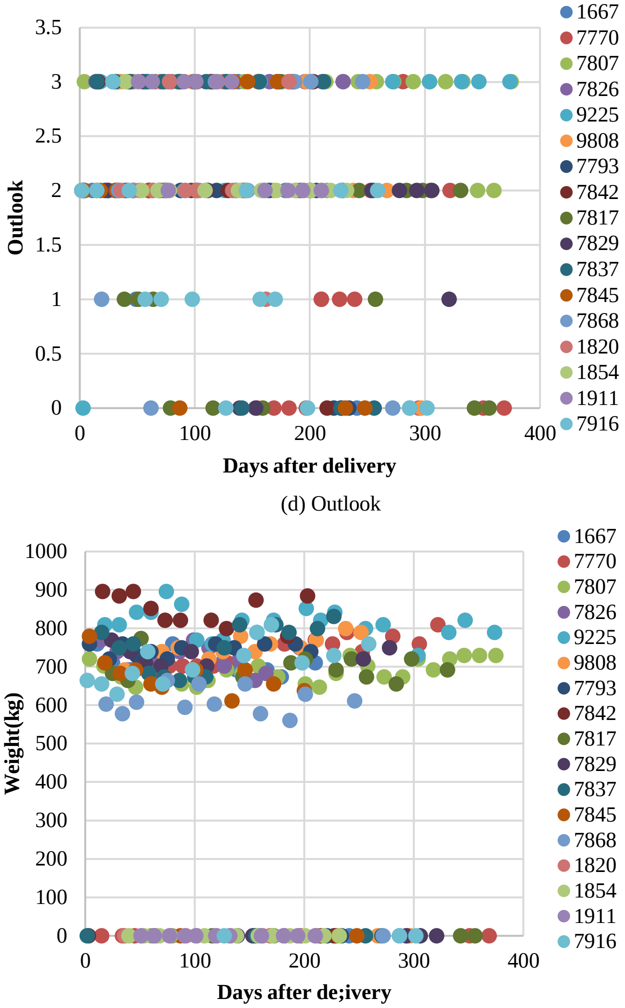
Fig. 2 shows the trends of the aforementioned influencing factors, Amount of milk, Weight, RFS, BCS, and Outlook. Overall trend of the amount of milk shows decreasing except some dairy cattle. Therefore, it cannot be an index of the recovering their readiness of pregnancy. Meanwhile, BCS shows the trend of which their BCS decreases just after their delivery and then the BCS is gradually increased except some dairy cattle. Some of the dairy cattle are not ready for pregnancy one year after the previous delivery for some reasons. Although weight, outlook, RFS of such dairy cattle shows recovering of the readiness of pregnancy, amount of milk of such dairy cattle is not increased Therefore, BCS is a good indicator of the readiness of pregnancy.



(a) Amount of Milk

(b) BCS

(c) RFS



(d) Outlook

(e) Weight

Fig. 2. Trends of Amount of Milk, BCS, RFS, Outlook, and Weight of 17 of the Dairy Cattle.

On the other hand, outlook and RFS are varied up and down randomly. Therefore, both of outlook and RFS are not good indicator for their readiness of pregnancy. Weight of the dairy cattle, meanwhile, is getting down just after the delivery and gradually recovering their weight after that except some dairy cattle of which the dairy cattle are not ready for pregnancy yet nevertheless weight is recovered. Therefore, weight can be a possibility of a good indicator for their readiness of pregnancy.

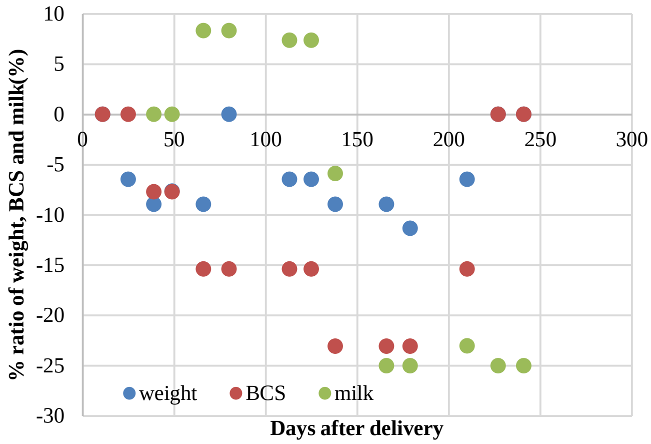
C. Extracting Sensitive Feature

In order to enhance the feature of the time series of data, the percentage ratios r of the influencing factors x is calculated with (1).

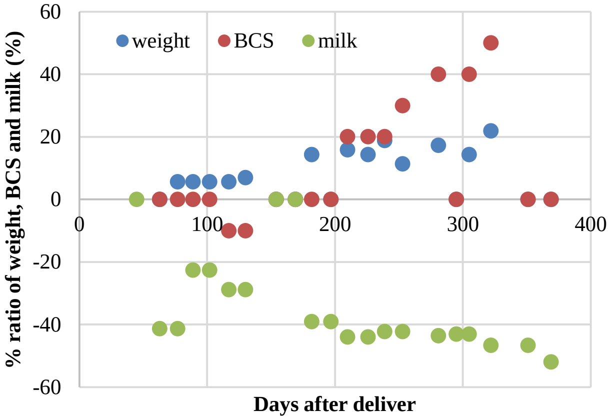
$$r=(x-x')*100/x' \qquad (1)$$

Where x' denotes the first data x at the begging of measurement Therefore, r implies change rate of x.

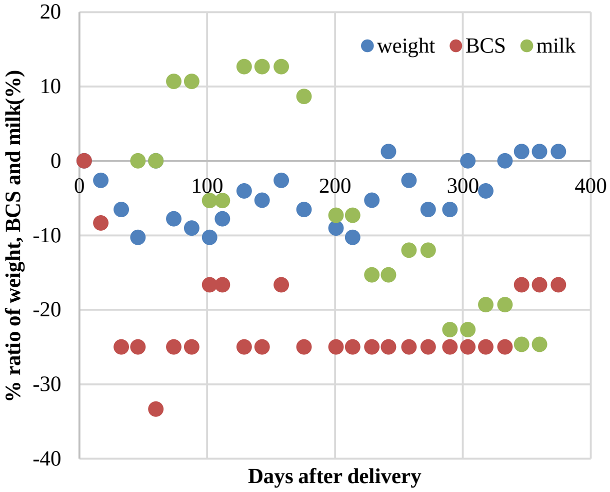
Fig. 3 shows the change rate of the influencing factors, weight, BCS and amount of milk for 17 of the Japanese dairy cattle. As shown in Fig. 3, outlook and RFS are not appropriate factors. Therefore, these two factors are not taken into account. It is much clear that BCS is decreasing just after the delivery and then it is recovered gradually except a few dairy cattle. Also, it is found that amount of milk changes randomly during recovery stage. That is the same thing for weight of the dairy cattle.



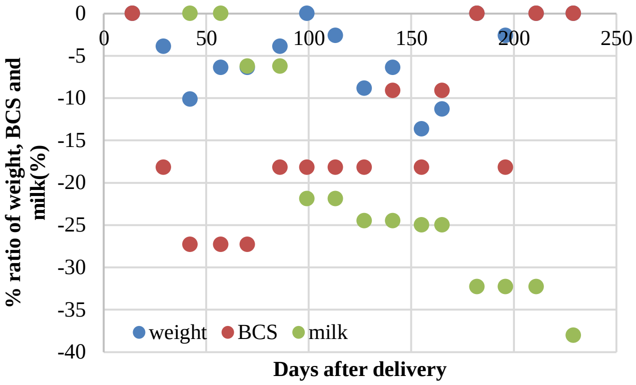
(a)



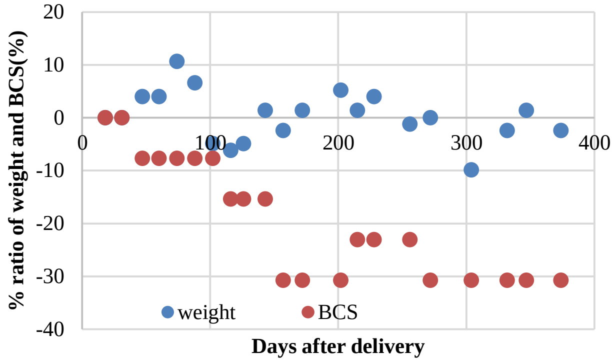
(b)



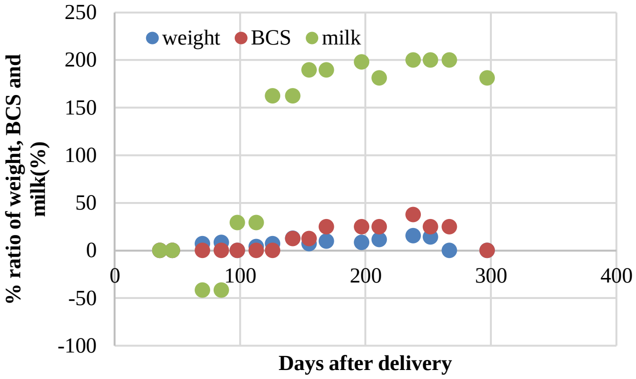
(c)



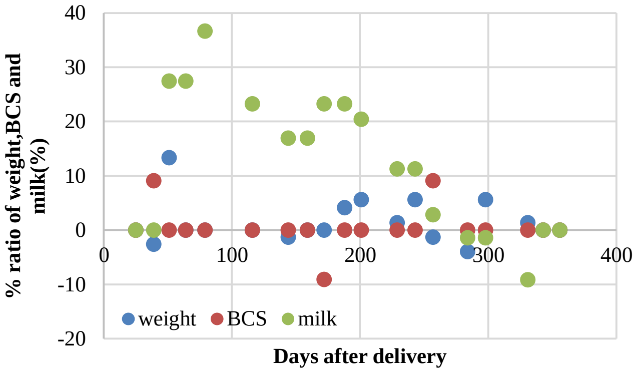
(d)



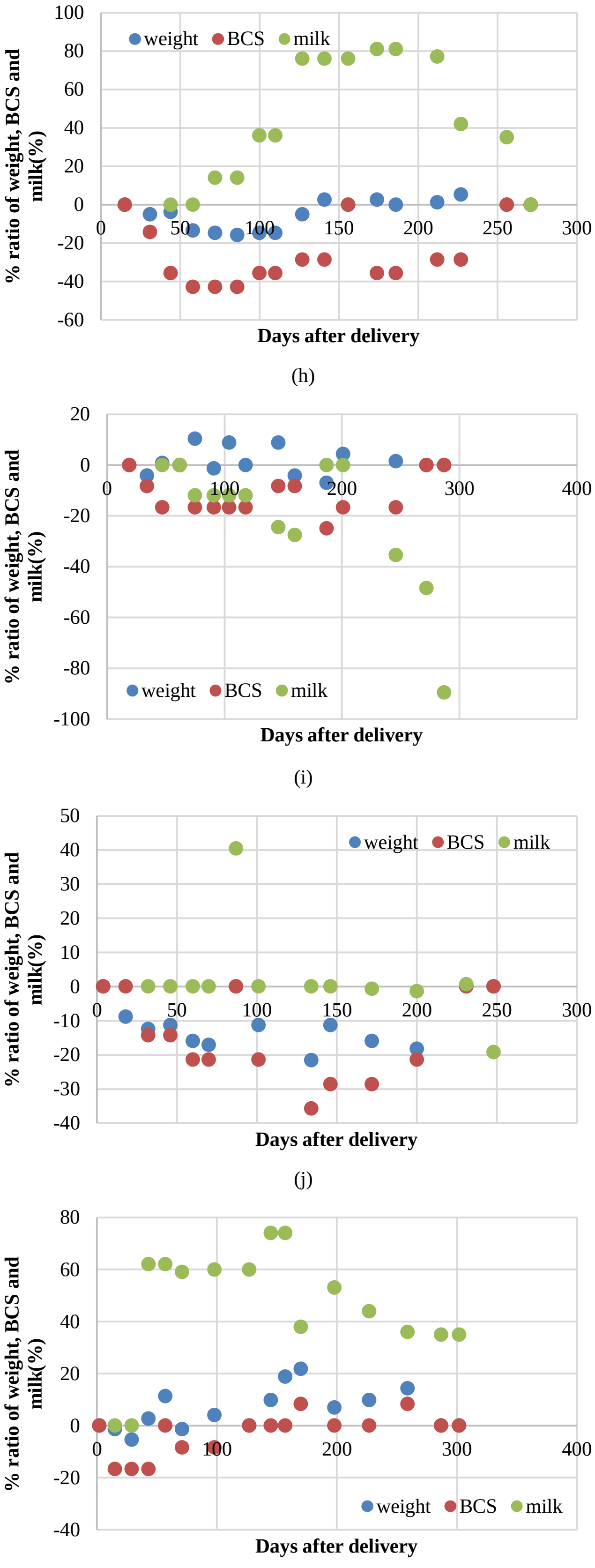
(e)



(f)



(g)



(h)

(i)

(j)

(k)

Fig. 3. Change rate of the influencing factors.

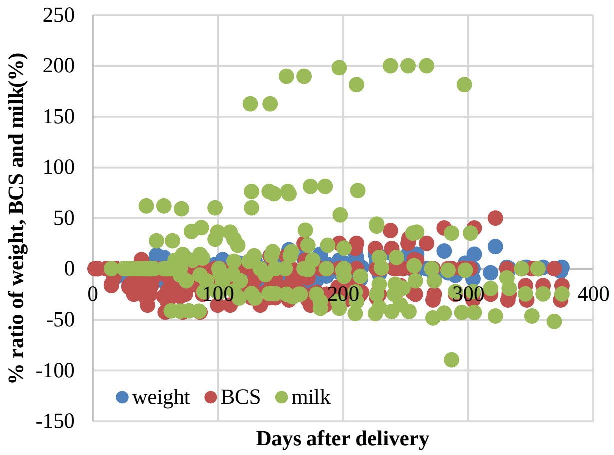


Fig. 4. Summarized results from the time series analysis.

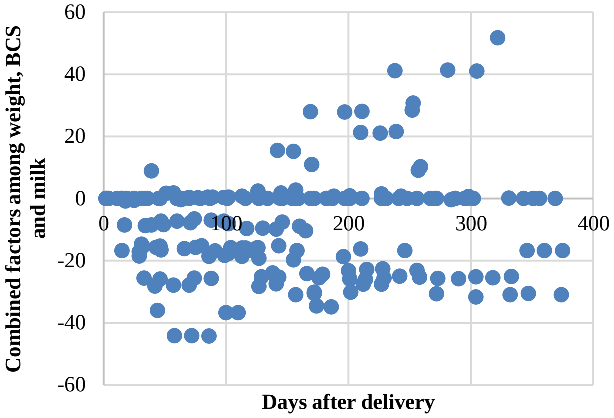


Fig. 5. Trend of the proposed index derived from the linear combination among weight, BCS and amount of milk.

Summarized results are shown in Fig. 4. Depending on influencing factors, the following linear combination of factors is proposed for the index representing recovery status of the dairy cattle after the previous delivery.

$$\text{Index}=C_0\text{*BCS} + C_1\text{*Weight} + C_2\text{*Amount of Milk} \qquad (2)$$

Fig. 5 shows the time series of the proposed index. For the proposed index can be determined with parameterization of coefficients C_i for the linear combination.

It seems a good trend which represents the recovery status of the dairy cattle as shown in Fig. 5.

Consequently, it is found that the proposed index derived from the linear combination among weight, BCS and amount of milk works well. Using this index, it is easily find the dairy cattle which need hormone treatment. Thus, the labor mean interval can be shortened.

IV. Conclusion

Japanese dairy cattle productivity evaluation method based on time-series analysis with the data of visual index of Body Condition Score (BCS), Rumen Fill Score (RFS), Weight, Amount of Milk and Outlook is proposed. Through the experiments with 17 of dairy cattle of the candidates of 17 Japanese anestrus Holstein dairy cows, it is found that the proposed method is useful for identification of productive cattle. Therefore, the cattle which need hormone treatments can be identified. The proposed time series analysis does work for dairy cattles to find the relatively productive dairy cattles for shortened labor mean intervals.

Further study is required for finding much sensitive indicator to the readiness of pregnancy and creates a new method for identifying dairy cattle which need hormone treatment, CIDR and PG.

ACKNOWLEDGMENTS

This research work is supported by Ministry of Agriculture, Forestry and Fisheries: MAFF, Japan under the project name of “Development of technology for enhancement of livestock lifetime productivity by improving fertility through assisted reproduction”. Authors would like to thank MAFF for their support as well as the students in the 4th research group of the Department of Information Science of the Faculty of Science and Engineering of Saga University for their valuable discussions and preparation of cattle data.

REFERENCES

[1] W. Kellogg, “Body Condition Scoring with dairy cattle- FAS4008”, University of Arkansas, USA, Accessed on: January 2016.

[2] G.C. Lamb, C.R. Dahlen, and D.R. Brown, “Reproductive Ultrasonography for monitoring Ovarian Structure Development, Fetal Development, Embryo Survival and Twins in Beef Cows”, The Professional Animal Scientist Symposium, No. 19, 2003, pp. 135-143.

[3] J.D. Ferguson, D.T. Galligan, and N. Thousen, “Principal Descriptor of Body Condition Score in Holstein Cows”, Journal of Dairy Science, No.77, 1994, pp.2695-2703.

[4] P. D. Burns, “The Dairy Cow Heat Cycle”, Colorado State University, Accessed December, 2015.

[5] T.A.Zacarias, S.B. Sena-Natto, A.S. Mendonca, M.M. Franco, and R.A. Figueiredo, “Ovarian Follicular Dynamics in 2 to 3 months old Nelore Calves (Bos Taurus indices)”, Journal of Animal Reproduction, Vol. 12, No.2, June,2015, pp.305-311.

[6] G. A. Perry, O. L. Swanson, E. L. Larimore, B. L. Perry, G. D. Djira, and R. A. Cushman, "Relationship of follicle size and concentrations of estradiol among cows exhibiting or not exhibiting estrus during a fixed-time AI protocol", Journal of Domestic Animal Endocrinology, 48(2014), pp.15-20.

[7] Iqbal Ahmed, Kenji Endo, Osamu Fukuda, Kohei Arai,Hiroshi Okumura, Kenichi Yamashita, Japanese Dairy Cattle Productivity Analysis using Bayesian Network Model (BNM), International Journal of Advanced Computer Science and Applications, 7, 11, 31-37, 2016.

[8] Report of National Livestock Breeding Center, Japan. Website: http://www.nlbc.go.jp/en/, Accessed January, 2016.

[9] Kohei Arai, Narumi Suzaki, Iqbal Ahmed, Osamu Fukuda, Hiroshi Okumura,Kenji Endo, Kenichi Yamashita, Method for Productive Cattle Finding with Estrus Cycle Estimated with BCS and Parity Number and Hormone Treatments based on a Regressive Analysis, IJACSA, 8, 9, 191-196, 2017

AUTHOR’S PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission “A” of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 37 books and published 570 journal papers. He received 30 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.ht

Preference in using Agile Development with Larger Team Size

Ahmed Zia, Waleed Arshad, Waqas Mahmood
Department of Computer Sciences,
Institute of Business Administration - IBA,
Karachi, Pakistan

***Abstract*—Agile software development includes a group of software development methodologies based on iterative development, where requirements and solutions evolve through collaboration between cross-functional self-organizing teams. Different software houses were visited in a developing country to determine the experiences faced by people working on a real world projects using Agile software development methodology following different variants in different team sizes to determine the preference of using Agile software development methodology in larger team sizes. Several people were surveyed out of which few responded with an opinion of not to use agile development in a team sizes exceeding 25 members. According to the experience of people the ideal team size was 5 to maximum 10. Because according to the survey increase in the number of individuals create issues of communication as it is not possible to keep everyone on the same track with larger teams especially in case of scrum meetings which usually held on daily basis, taking responsibilities as everyone becomes reluctant in taking responsibilities believing someone else will take it, sub teams because the more the number of individuals the more will be the sub teams which indirectly increases the dependency among the teams by breaking the tasks into much smaller chunks. The findings also suggest that customer feedback would increase if the team size is less than 25 which in turn says that the Quality of Software is increased. As this study had only focused on the software companies of a developing country it is recommended that further studies should be carried out by surveying the people of other different developed countries**.

Keywords—Agile Development; Ideal Team Size; Larger Team Size Problems

I. Introduction

Many different approaches have been used to develop different software’s by various organizations [1]. They include waterfall, iterative, incremental, spiral, agile and many more. Nowadays the most popular among all of these is agile because of its ability to adapt to change. Change can be of any nature, may it be internal or external i.e. from the customer changing his/her requirements, the primary stakeholders or the secondary ones, including covering any changes in the agile team itself. But when the agile team size starts to grow, different issues start to arise [2]. This research will find out the preference of different organizations to use agile development with larger team size with quality effect on product/project which is being developed.

A large number of authors emphasize that the size of the team, following any variant of agile development methodology, should be kept small. There is a rift between different authors as to how small the team must be to achieve the optimal development results [3] [4]. According previous studies the optimal size of any agile team is up to 15 members while the maximum is around 20 [5]. The decision-making quality suffers in larger teams due to the fact that the decisions will be more unclear which confuses the team in making decisions. Complacency among the team increases as no one starts on the project unless and until they are given direct orders and in a large team it is difficult to keep track of which team member took responsibility of what task. However, all the studies carried out before primarily focus on other factors affecting agile development with a sideline reference to the team size [6].

This research, as mentioned above, finds out the trends of the preference to use agile development with increasing team size and quality effect on product/project. The survey is conducted using quantitative strategy and questionnaires were distributed in different software houses of a developing country. The data collected was analyzed by the SPSS statistics software. However, the main limitation of this research is that only the software houses of a developing country are taken into consideration. And it is also assumed that the software houses surveyed use any form of agile development as their primary process model.

II. Background

Large organizations are now moving towards agile implementation because it supports flexibility and welcome changes at any stage of development. The basic challenge for larger organizations is to retain these features along with maintaining the quality of the product and follow quality assurance principles [1]. As the team size increases, it gets difficult to keep following design and documentation practices [6]. Distributed team increases the risk of multiple styles of documentation which in turn compromises the quality of documentation and design features. With the increase in team sizes, estimation of efforts gets crucial as well. With bigger teams it is challenging to accurately estimate developers’ and QAs’ efforts at the beginning of a certain project [6]. As a result, delivery quality is compromised along with the fact that deadlines are not met accordingly. With bigger teams, it is difficult for the QA to support the entire team, which creates imbalance within the lineup [7]. Agile methods also make automated test case execution troublesome [7]. Whilst in articles [8] and [9] the optimal is considered as 3-7 and 10-15 respectively. However, it is agreed in all of the mentioned papers that the maximum team size should not exceed twenty members. Decrease in lead team size and increasing the morale in the team in agile software development increases the productivity of the whole process [4]. As described in [10] *"A successful globally distributed agile team configuration consists of a smaller number of members to facilitate better certain agile practices, such as the daily stand-up meeting, iteration planning, iteration demos, iteration retrospectives, and user stories".*

On a higher note, methodologies like integrated testing, welcoming and rapidly responding to changes, and people centric approach are not mature enough to be practiced in large development teams [11].

Agile is not only used in co-located teams but also in virtual and distributed teams. The distributed and virtual teams imply that the team members are not physically present in front of each other but are virtually connected to each other via internet. The larger teams create a problem of communication in a virtual environment [3] and a distributed environment [9]. The larger the size of team, higher is the reluctance in accepting a responsibility assuming the job belongs to someone else [3].

The quality of decision making also lacks and suffers in larger teams due to the fact that such decisions will be more complex and unclear resulting in confusion among and between team members. Complacency among the team increases as no one starts on the project unless and until they are given direct orders and in a large team it is difficult to keep track of which team member took responsibility of what task [3]. [3] If the team size is large the sub teams that will be made will have more members than required which increases the dependency on other sub teams as the task assigned to an individual sub team will be much smaller as compared to their size and capability which causes frustration. Participation of individuals also decreases [2].

Moreover, in larger teams especially in virtual and distributed teams where the team members are not co-located different cultural, language and ethnic issues arise [12] [13] [10].

In all the above papers the focus is on other factors influencing the agile software along with sideline references to the team size. This research intends to focus more on the number of individuals working in a team and its overall effect on the agile software development rather than focusing on different agile methodologies and the factors affecting them.

III. Literature Review

Agile development is actually more people-centric rather than being process-centric [11]. Agile basically works on people, collaboration and communications. As a result, a larger team will be problematic to manage with agile development and quality practices which in turn is very much likely to affect the quality of the end product.

Agile also involves a combined team with cross-functional capabilities [7]. With developers and QAs being rapidly communicating with one another, it will be difficult to manage this cross-functionality, along with maintaining the quality, within a larger team. Agility refers to getting closely connected with the customer. Throughout the development of a product, the customer is in collaboration from the beginning till the end [14]. Exploratory and collaborative testing are the essential factors in agile development [6]. With a larger team, it will be unmanageable to keep the team in collaboration in order to perform well. Furthermore, it is very necessary for a company to communicate its common product vision to its QA and developers in order to make them meet the desired results [15]. With larger team size, an unclear vision will cause hurdles in meeting the expected outcomes. Team efficiency is an important aspect to fulfill project on time and most often is dependent on the interaction among the team members and the coordination of the team leader. According to a research, when team size is between 4 to 8 team efficiency goes to the peak and when team size goes beyond 9, efficiency starts to decrease [16]. In a research performed by Elizabeth Whitworth in her research [17] tackled the psychological aspect of Agile development team members to determine how well they perform when working in an Agile team. The results were shocking and very motivational in terms of adopting Agile practices as compared to traditional methods and hence has a positive impact on both personal level and team level productivity.

Agile development requires a proper team headed by a project manager and having skilled team members and managers. Adopting SCRUM methodologies would be beneficial to meet the deadlines and having less team and full coordination among them would result in a user-friendly project fulling all client's requirement. Project failures often occurs due to miscommunications and therefore large team size are actually the bad vehicles of miscommunications [16]. Human Mind works well and have more productive teams in less team size of around 4 to 7 and hence having less communication channels [5].

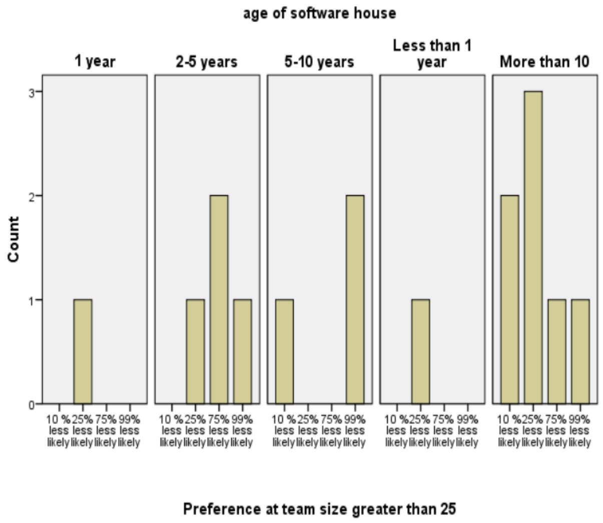


Fig. 1. Software Houses Preference when Team Size Exceeds 25.

The two most important factors for a company pursuing agile methodologies are time and cost [1]. Strict deadlines are to be met in prescribed timelines with minimum expenditure. Larger team size will require more cost and integrating all the personnel will take time, due to which either of the two things might happen; deadlines won't meet, or quality gets compromised. Agile takes iterative development one step further. Minor releases are released instead of one major release [5]. With larger development and QA teams, iterative development will lack collaboration and coordination and as a result quality of those minor releases is compromised. Apart from this, larger teams create dependencies among the developers and QA which in turn leads to resources sitting idle [18]. Development dependencies create delays in timelines. Delayed timelines affect the testing phases. Agile methodologies come with drawbacks such as daily check ins, setting up of daily test environments, regular integrations for the QA, a lot of meetings, and a lot of manual testing [13]. Larger team size will make it unmanageable for the company to track check-ins and fulfill integrations on daily basis. It is difficult to gather a large team for excessive meetings [19] and this results in the reputational loss of both the company/organization and employees and also results in loss which may cause to shut down an organization resulting in the unemployment of many employees and workers [20].

IV. Results

This research focuses on the preference to use Agile Development methodology with an increase in the number of individuals in the respective agile teams. The data was collected from several software houses in a developing country. The questionnaire was distributed among several different individuals. Many different trends for the preference to use agile development were seen when the data was analyzed. The data was analyzed with respect to the age of the software establishment, the type of agile methodology used, the ideal team sizes and the preference to use with larger team size. The following graphs focus on the likeliness of using agile methodology when team size go above 25, 35 and 50 and how different software companies are reacting to this increase in team size with respect to the age and experience of the company.

In figure 1 we can see that when the team size goes above 25, most of the establishments are 50 to 75% less likely to use agile development. We can see that the companies with the more experience are even more reluctant to use agile development. This reluctance increases as we increase the team size further.

When we have team size of above 35 we can see that the reluctance has increased even more. The companies with the more experience are 99 percent less likely to use agile development, while the rest are 25 to 75 percent less likely to use agile development as shown in figure 2.

When the team size is above 50 all of them are 99 or 75 percent less likely to use agile development, disregarding a single exception as shown in figure 3.

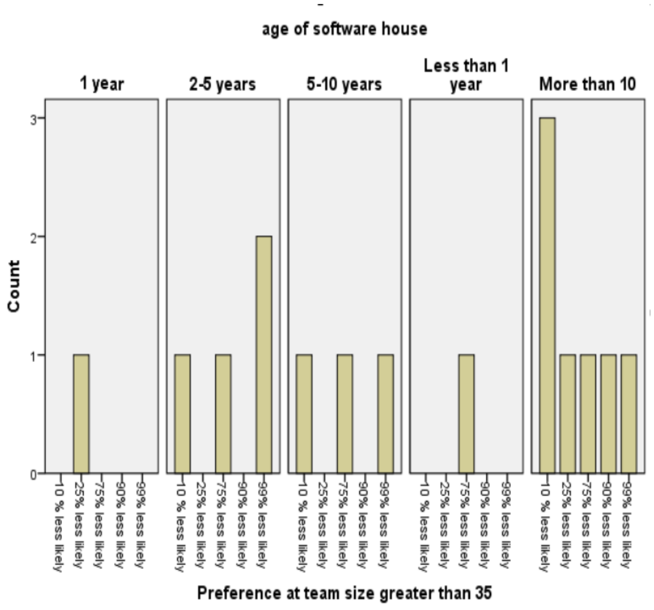


Fig. 2. Software Houses Preference when Team Size Exceeds 35.

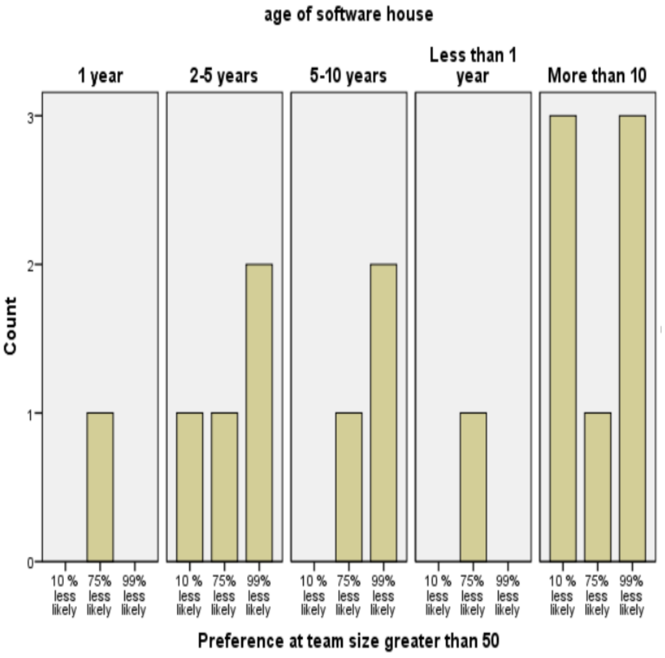


Fig. 3. Software Houses Preference when Team Size Exceeds 50.

The research further shows that there are a lot of variants of agile methodologies and teams including collocated teams, virtual teams distributed teams. The preference to use agile development methodologies is also highly dependent on the optimum team size. In the graphs below the analysis is done from two different angles. In the figures 2.1 and 2.2 Show the ideal team size with respect to the agile methodology used and also with respect to the type of agile team. Figures 2.3 and 2.4 focus on the ideal team sizes with respect to the type of agile teams and irrespective of the type of agile methodology used.

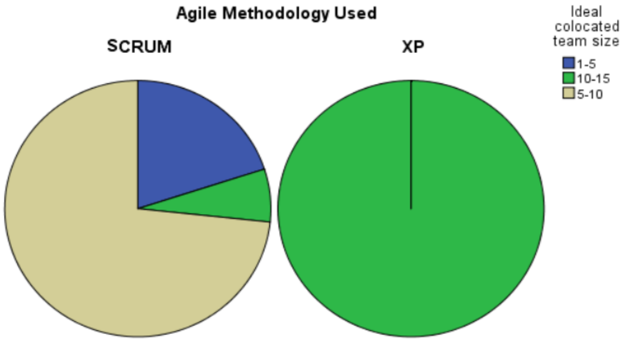


Fig. 4. Ideal Collocated Team Size in Different Agile Methodologies used.

Figure 4 shows that in collocated teams, in case of scrum, the ideal team size is 5 to 10, which is preferred 74% of the times. While a 23% of the companies prefer an even smaller team size of 1-5. In case of XP the ideal team size is 10-15 which is preferred by all of our respondents.

Figure 5 shows that in case of distributed teams or virtual teams the ideal team size is still 5-10 which is preferred by 60% of our respondents, in case of scrum. While in case of XP the ideal team size reduces to 5-10. We can infer from both of these figures that the larger team sizes are less preferred to be used by all the software companies which responded.

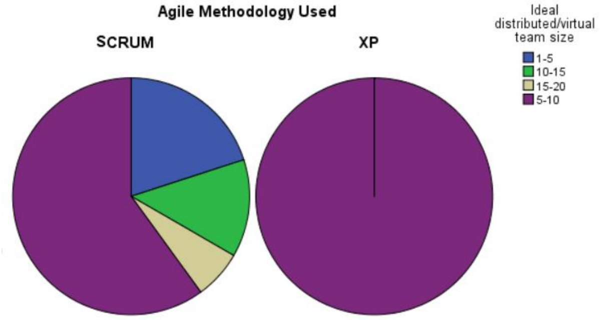


Fig. 5. Ideal Distributed/Virtual Team Size in Different Agile Methodologies used.

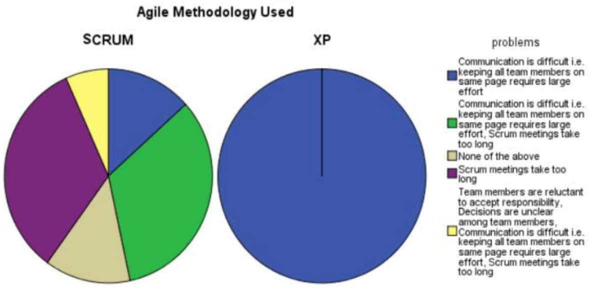


Fig. 6. Problems Faced Due to Larger Team Size in Different Agile Methodologies used.

Figure 6 shows that in SCRUM methodology, SRCUM meetings with large team size is the main problem in Agile development whereas in XP it's the communication as its very difficult to keep all teams members on same page and its requires extra effort if team size is greater than 5.

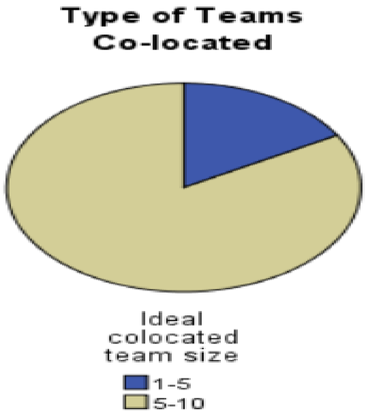


Fig. 7. Ideal Collocated Team Size is 5-10.

Taking into account the ideal team size with respect to the type of team and irrespective of the type of agile methodology used, we can see that from figure 7 that the ideal team size for all the different types of teams is 5-10. Different agile methodologies pose different problems when the team size increases. Furthermore, problems that arise are also vary according to the type of agile methodology.

The above statistics clearly shows that organisations that have adapted SCRUM prefers to work with an ideal team size of 5 – 10 team members whereas organisations that have adapted XP methodology prefers to work with 10 – 15 team members size and this helps them finishing their project or product work with efficiency and within time keeping intact all quality attributes and other necessary attributes required for completion of project or product.

Figure 8 shows that in case of XP, the main problem is communication which is faced by all the respondents. Keeping all the team members, when team size goes large, is very difficult and time consuming to keep all the team members on the same page. In case of scrum, the main problem faced by all of our respondents is that scrum meetings take too long. While other problems include communication and reluctance to accept responsibilities.

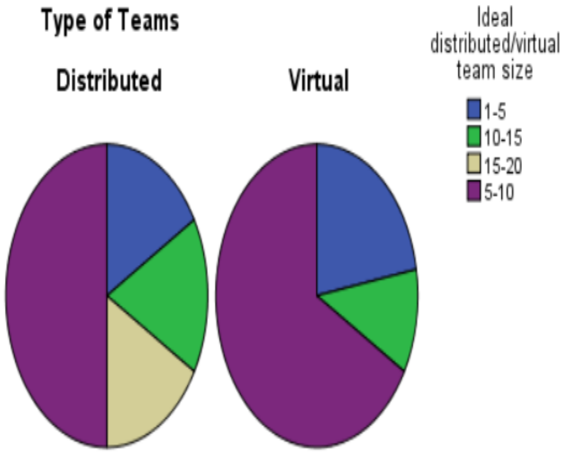


Fig. 8. Different Teams Problems Due to Greater Team Size.

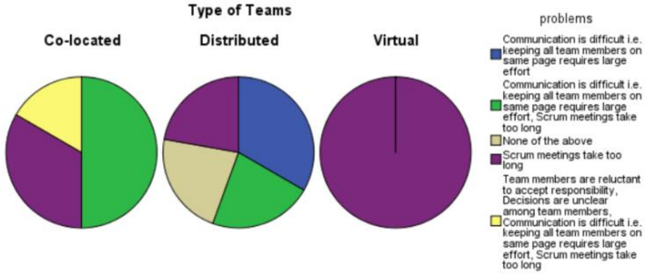


Fig. 9. Problems Faced by Different Types of Teams Due to Greater Size.

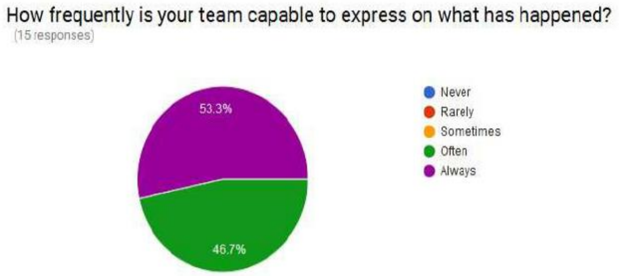


Fig. 10. Capability of Team to Express when Team Size Exceeds.

The above two figures clearly show a major difference in the results. Figure 10 [18] shows that the research conducted previously was on small team sizes, which is why the expressiveness within the team is highly efficient and therefore the quality of the product in agile implementations. Whereas figure 9 shows that the research conducted now has been done on large team sizes, which shows that team efficiency is compromised due to which all the three types of team distributions are lacking quality of communication.

The comparison has major differences due to difference in team sizes. This comparison majorly depicts that with increasing team sizes, it is very much likely that the team expressiveness, collaboration, trust and communication is compromised specially in cases where agile methodologies are implemented and followed.

The current research also shows that means of communication is an issue which can be categorized into the area of following processes. Agile implementations majorly get compromised on following processes. Inter-team's communication is hurdled when agile methodologies are followed in large team structures. Figure 9 also shows that team meetings are a major reason where time is mostly spent, which is a major setback in agile implementation because agile is already very strict in terms of time constraints.

In Figure 9 the main problem faced by collocated teams are Scrum meetings take too long and team members are reluctant to accept responsibilities. In case of Distributed teams, communication is the major problem faced by 55% of the respondents. In virtual teams, Scrum meetings take too long as well as team members are reluctant to accept responsibilities as decisions are unclear among members.

This survey also includes customer feedback related questions to get to know about end user or customer feedback related to greater team size effect of the given requirement of project or product. The comparison is also shown in Table 1. As larger team sizes also affect the software quality, Customer Feedback and the participation of members in a team. Figure 11 shows the participation with larger team size is "worse".

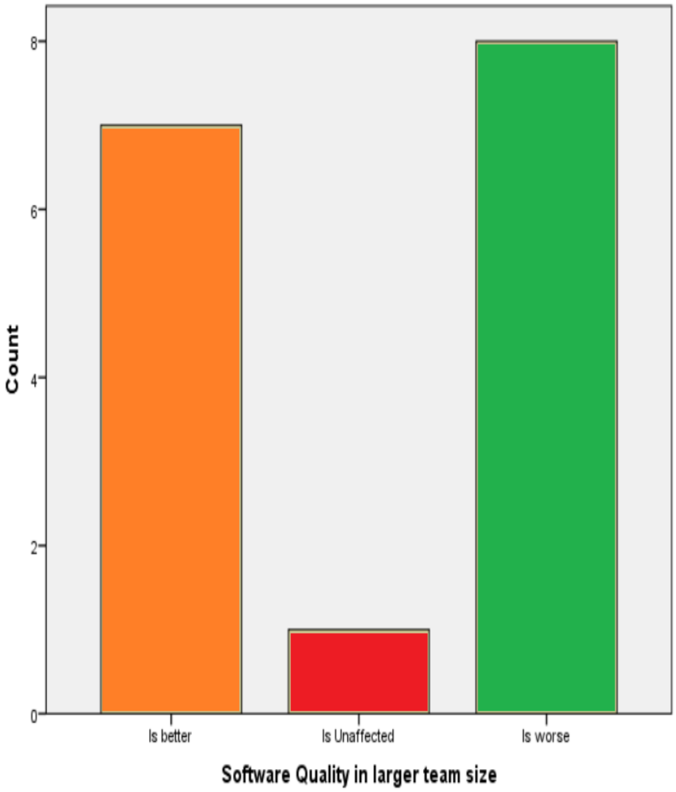


Fig. 11. Software Quality Effects when Team Size is Greater.

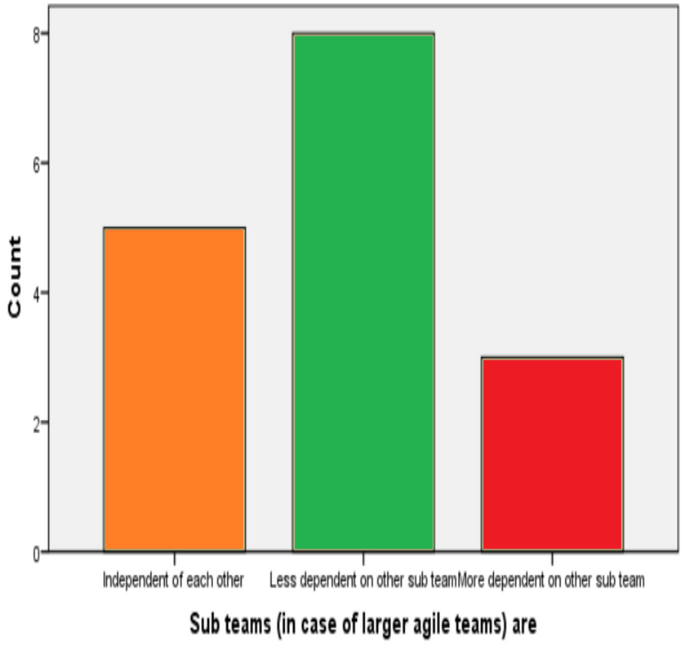


Fig. 12. Team Members Become Dependent Due to Greater Size.

TABLE I. COMPARISON BASED ON DIFFERENCE IN TEAM SIZE

Expressiveness among team members	
Figure 9	**Figure 10**
Communication difficulty (50 %)	Always (53.3 %)
Too long meetings in larger team size (33.3%)	Often (46.7%)
Keeping all team members on the same page (16.6 %)	Sometimes (0 %)

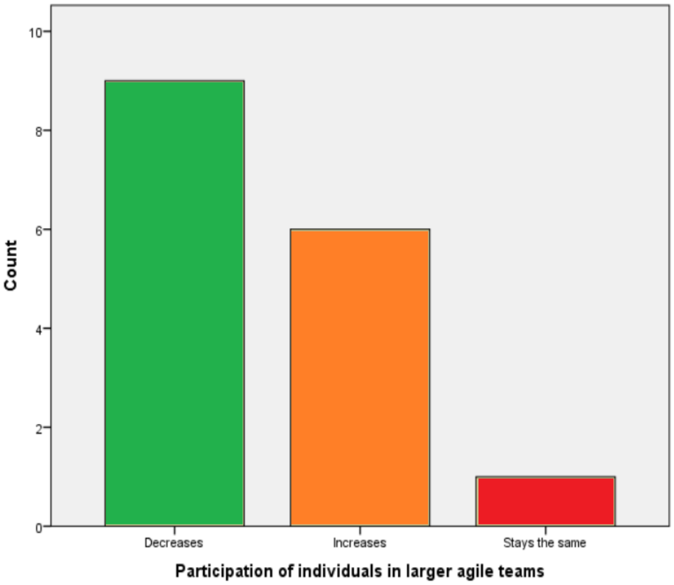


Fig. 13. Team Member's Participation is Less in Larger Team Size.

Figure 12 shows as the team size increases, teams become depend on sub teams.

Figure 13 also shows that team member participation decreases in case of large team size.

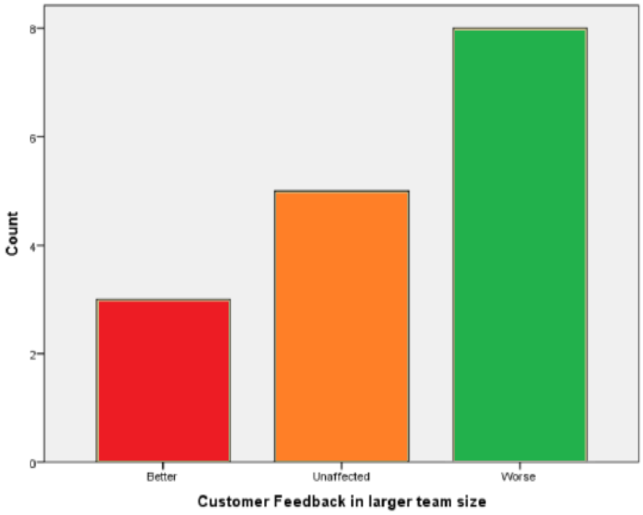


Fig. 14. Customer Feedback in Case of Larger Team Size.

Figure 14 also shows that the customer feedback is also "worse" when team size increases.

V. CONCLUSION

This research is about the preference of using agile development methodology when team size increases. According to the research that has been done before and according to the data collected by surveying people several different software houses in a developing country it shows that larger team size always creates problems of communication and working. As it is proven by the research that in teams having more than 25 members people are more reluctant in taking responsibilities by assuming that someone else will take that, communication is difficult as scrum meetings take too long which would take less time if team size would be of 5-10 and decisions will become clearer. People working in virtual and distributed teams find it difficult to keep all the team members on the same track if the team size goes more than 25. Increase in the team size also increases the number of sub teams and the more the sub teams the more will be the dependency among the teams as work will be distributed in lesser quantity and vice versa in case of team size is less than.

VI. FUTURE WORK AND RECOOMENDATIONS

Due to the problems faced by individuals while working in a team size of more than 25 following different agile methodologies it is highly recommended to keep team size of less than 25. This would increase the performance of the team by effective communication, reducing the dependency among the sub teams, making people more responsible to take responsibilities. People working in a virtual and distributed environment would become more comfortable in keeping every member on the same track. The findings also suggest that customer feedback would increase if the team size is less than 25 which in turn says that the Quality of Software is increased. As this study had only focused the software of a developing country so it is recommended to conduct the same type of survey in other organisations of a developing and developed country.

REFERENCES

[1] G. Rani and J. K. Bajwa, "Mapping and Analysis of Agile Quality Assurance," *International Journal of Advanced Computational Engineering and Networking,* Vol. 5, No. 1, 2017.

[2] H. Hajjdiab and A. S. Taleb, "Adopting agile software development: issues and challenges.," *International Journal of Managing Value and Supply Chains (IJMVSC),* vol. 2, no. 3, pp. 1-10, 2011.

[3] J. Eckstein, Achieving Agile Software Development with Large:Diving into the Deep, Addison-Wesley Professional, 2013.

[4] E. Altameem, "Impact of Agile Methodology on Software Development.," *Computer and Information Science,* vol. 8, no. 2, 2015.

[5] V. Lalsing, S. Kishnah and S. Pudaruth, "People factors in agile software development and project management.," *International Journal of Software Engineering & Applications,* vol. 3, no. 1, p. 117, 2012.

[6] M. Hamid, "Agile Testing.," International Journal of Scientific *and Research Publications,* vol. 5, no. 9, 2015.

[7] P. Rajasekhar and R. M. Shafi, "Agile Software Development and Testing: Approach and Challenges in Advanced Distributed Systems.," *Global Journal of Computer Science and Technology,* vol. 14, no. 1, 2014.

[8] Y. Lindsjørn, D. I. Sjøberg, T. Dingsøyr, G. R. Bergersen and T. Dybå, "Teamwork quality and project success in software development: A

survey of agile development teams.," *Journal of Systems and Software 122 (2016): 274-286,* vol. 12, pp. 274-286, December 2016.

[9] M. Majchrzak, Ł. Stilger and M. Matczak, "Working with agile in a distributed environment.," *Software Engineering from Research and Practice Perspective, L. Madeyski and M. Ochodek, Eds. Polish Information Processing Society Scientific Council,* vol. 2014, pp. 41-54.

[10] J. H. Sharp, S. D. Ryan and V. R. Prybutok, "Global agile team design: An informing science perspective.," *Informing Science: The International Journal of an Emerging Transdiscipline,* vol. 17, pp. 175-187, 2014.

[11] Vijayasarathy and D. Turk, *Agile software development: A survey of early adopters.,* vol. 19, Journal of Information Technology Management, 2008, pp. 1-8.

[12] "The Role of Developers in Agile Teams," 2012. [Online]. Available: http://7bsp1018.wikispaces.com/The+Role+of+Developers+in+ Agile+Teams. [Accessed November 2017].

[13] A Begel and N. Nagappan, "Usage and perceptions of agile software development in an industrial context: An exploratory study.," *In Empirical Software Engineering and Measurement, 2007.,* pp. 255-264, 2007.

[14] B S. Bhasin, "Quality assurance in agile: a study towards achieving excellence.," *In AGILE India,* vol. 2, pp. 64-67, 2012.

[15] A. K. Rai, *Quality of Performance and Performance Risk Evaluation for Agile Software using Fuzzy Inference System,* vol. 9, International Journal of Global Research in Computer Science (UGC Approved Journal), 2018, pp. 1-5.

[16] P. Abilla, "Team Dynamics: Size Matters Redux.," 2006. [Online]. Available: http://www.shmula.com/lost-in-translation-in-large-teams/470/. [Accessed 27 April 2011].

[17] E. Whitworth, Agile Experience: Communication and Collaboration in Agile Software, Carleton University, 2006, p. 466.

[18] W. Mahmood, N. Usmani, S. Farooqui and M. Ali, "Benefits to Organizations after Migrating to Scrum," in *29th International Business Information Management Association Conference*, 2017.

[19] S. Farooqui and W. Mahmood, A Survey of Pakistan's SQA practices: A Comparative Study, Vienna, 2017.

[20] A. Iftikhar and S. Muhammad Ali, *Software quality assurance a study based on Pakistan's software industry.,* vol. 2, 2015: Pakistan Journal of Engineering, Technology & Science.

ANNEXURE

NAME*:

DESIGNATION*:

NAME OF SOFTWARE HOUSE*:

1) The age of your software house establishment

- <1 year
- 1 year
- 2-5 years
- 5-10 years
- >10

2) Your experience at the current organization

- <1 year
- 1-3 years
- 3-5 years
- >5 years

3) What is the variation of agile methodology that is used in your establishment

- SCRUM
- XP
- I-XP
- Agile not used

4) Which tool is being used in Agile Development

- MS Project
- MS Excel
- Scrum Pad
- Version One
- Rally
- Scrum Desks
- Wall and papers
- Storyboard
- XPlanner
- Other _________________

5) What kind of teams, mainly, do you have at your establishment for agile development

- Co-located
- Distributed
- Virtual

6) What is the ideal team size with co-located teams when using agile development

- 1-5
- 5-10
- 10-15
- 15-20
- >20

7) What is the ideal team size distributed or virtual teams when using agile development

- 1-5
- 5-10
- 10-15
- 15-20
- >20

8) What is the main problem you face when team size in agile development is too large

- Team members are reluctant to accept responsibility

- Decisions are unclear among team members
- Communication is difficult i.e. keeping all team members on same page requires large effort
- Scrum meetings take too long
- None of the above

9) Sub teams (in case of larger agile teams) are

- More dependent on other sub teams
- Less dependent on other sub teams
- Independent of each other

10) Participation of individuals in larger agile teams

- Increases
- Decreases
- Stays the same

11) If the team size goes over 25, would you still prefer to use agile development?

- 10% less likely
- 25% less likely
- 75% less likely
- 90% less likely
- 99% less likely

12) If the team size goes over 35, would you still prefer to use agile development?

- 10% less likely
- 25% less likely
- 75% less likely
- 90% less likely
- 99% less likely

13) If the team size goes over 50, would you still prefer to use agile development?

- 10% less likely
- 25% less likely
- 75% less likely
- 90% less likely
- 99% less likely

14) In case of larger team size in agile development software quality of product

- Is better
- Is worse
- Is unaffected

15) In case of larger team size in agile development software quality of project

- Is better
- Is worse
- Is unaffected

16) Customer Feedback in case of larger team size in agile development is

- Better
- Worse
- Unaffected

Green Cloud Computing: Efficient Energy-Aware and Dynamic Resources Management in Data Centers

Sara DIOUANI, Hicham MEDROMI
Engineering research laboratory (LRI)
System Architecture Team
ENSEM, Hassan II University of Casablanca, Morocco

Abstract—**The users of Cloud computing over the last years are constantly increasing since it has become a very important technology in the computing landscape. It provides client decentralized services and a pay-as-you-go model for consuming resources. The growing need for the cloud services oblige the providers to adopt an enlarged sized data center infrastructure which runs thousands of hosts and servers to store and process data. As a result, these large servers engender a lot of heat with visual carbon emission in the air, as well as important energy consumption and higher operating cost. This is why researches in energy economics continue to progress including energy saving techniques in servers, in the network, cooling, and renewable energies, etc. In this paper, we tackled the existing energy efficient methods in the green cloud computing fields and we put forward our green cloud solution for data center dynamic resource management. Our proposed approach aims to reduce the infrastructure energy consumption and maintain the required performances.**

Keywords—Cloud computing; green cloud; data center; energy consumption; resource management

I. Introduction

The recent revolution in information and communication technologies, despite all its benefits for our way of life, has reinforced our dependence on energy. More than 3 billion people use electronic terminals on a daily basis, the electricity consumption of which is strongly linked to the time of use, which is constantly increasing [1]. To support this revolution, huge data centers have sprung up all over the world. Service providers and data hosts are competing for investment in these data centers, real information factories for which profits cannot be counted. However, the power consumption of data centers is a new puzzle that the scientific communities are trying to solve with enormous difficulties.

The issue of energy has emerged in recent years as a central concern that humanity faces with great urgency. Population growth, the gradual exhaustion of previously explored resources, and more recently the rise of information technologies have turned this issue into a challenge that researchers and industry players are tackling [2].

In the field of information technology and in 2010, about 1.5% of the world's electricity has been consumed by computer centers [3]. This is steadily increasing due to the evolution of many domains and especially cloud computing.

In recent years, new problems have emerged in view of the environmental considerations that are increasingly present in our society. In 2014, the power consumption of Data centers exceeded 42 TWh and by 2020 the resulting CO2 production will reach 670 million metric tons annually [4].

Also, so as to face the peak demand of coming requests, servers in the data center, are constantly over-provisioned in working state, which generates wasting a large amount of energy [5]. One of the options to reduce the power consumption of data centers is to reduce the number of idle servers or to switch them into low-power sleep states as part of the green cloud IT vision [6].

Our work is to explore new ways to improve energy efficiency in cloud data centers. Specifically, the goal is to dynamically optimize energy consumption in cloud computing data centers by optimizing the use of its resources by adopting various policies for host and virtual machine (VM) overload detection, migration VM selection and VM placement policies.

The remainder of this paper is arranged as follows. In Section II, a literature review on the energy-aware solution on green cloud computing is presented. In Section III, our solution is proposed. Finally, the conclusions and future works are drawn in Section IV.

II. Related Work on Green Cloud Computing

Generally, the green cloud data centers are related to three principal methods which are: dynamic voltage frequency scaling (DVFS), the scheduling using renewable energy, and dynamic power management (DPM) [7].

So as to have an estimation of the energy consumed by the cloud client application, researchers in [8] have performed a power model under various DVFS policies. In [9], the flow pattern of the cloud tasks is studied and according to the results obtained, the researchers attempt to adjust the incoming VM tasks with in-demand frequency using DVFS.

In [10], the DVFS is adopted to minimize the power consumed in mobile cloud task scheduling, except that this approach does not consider the On and Off control of the servers. In the study [11] the three approaches including DVFS, request dispatching, and dynamic service management are joined so as to reduce the energy consumption. Yet, the limitation is that these researchers have admitted that the servers providing different services are active all the time.

In the SaaS cloud platform, an analytical framework which monitors the states of VMs (idle/busy) is developed. It characterizes and optimizes the power performance tradeoff

[12]. In another search [13] only one type of sleep mode with shutdown are used in a method for reducing the energy consumption, except that this solution is insufficient in the case of quick responses. In [14], they turned off a physical host in cloud computing by the adoption of two parameters which are time and load, so as to save energy consumption. If a VM had its working time exceeds a threshold fixed time, this VM will be displaced. And if a physical host had a load less than a fixed threshold, it will be turned off.

In this research [15], they relied on the processor workload, the disk workload and the ratio of performance degradation as a metrics of a VM so as to reallocate it. In [16], they used an algorithm based on CPU and memory of the VM and the server as parameters to consolidate them. They select VMs which are underloaded regarding the server. Some researchers have come to prove that this server consolidation approach is not adapted for large data centers. Also, multiple meta-heuristic algorithms of optimization are used for consolidating VMs such as those cited in [17]-[19].

According to many previous studies [20], [21], the green cloud computing gathers the efficient management of cloud resources and reducing energy consumption while assuring the quality of service requirements in the service level agreements SLA.

A system view of the green cloud computing where reducing energy is a must can be shown as in Fig. 1. It illustrates that when the cloud user sends an application request, the resource manager controls the resource utilization, and is in charge of the allocation of VMs in physical hosts while ensuring the respect of the SLA. Also, the underloaded servers are turned off while some others may be switched on for consolidating physical hosts as needed (see Fig. 1).

This survey helps to realize that existing limitations in the previous researches and which all of the major energy parameters (e.g. CPU, memory and so forth) necessary to ensure an ideal energy efficiency are not taken into consideration.

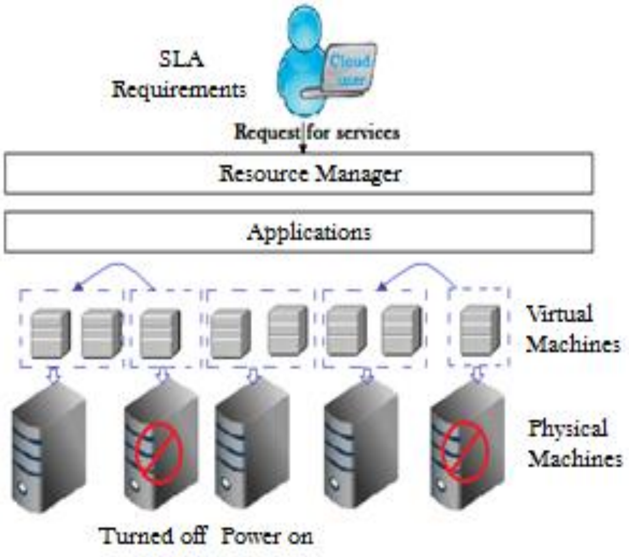


Fig. 1. Overview of green cloud computing.

III. Dynamic and Energy-Aware Solution for Resources Management

Virtual machine consolidation techniques are a means to improve energy efficiency and the utilization of cloud data center resources. However, aggressive VM consolidation approaches lead to physical host over-utilization and generate massive undesired VM migrations, which cause degradation of the performance of both the hosts and the VM [22].

Additionally, it has been a significant challenge to improve energy efficiency and resource utilization in the data center while delivering services with guaranteed quality of service (QoS).

To address this problem, we propose an enhancing energy-efficient and QoS dynamic resource management method, which consists of four principal modules. Our solution considers all major parameters related to the efficiency of the cloud data center energy.

In fact, the relevant energy parameters include the CPUs, the memory amount, the disk storage space, the quantity of transmitted message in the network (bandwidth) and the available amount of input/output operations per second (IOPS) on the physical host.

Besides, the VM placement depends on some precise Service Level Agreement restrictions as follows:

- The affinity constraint: between VMs aimed to get an optimal placement by considering the requirement that two VMs for example, must be placed on the same physical machine (PM). This condition refers to interdependent VMs that use jointly data with each other in prespecified deadlines.
- The security constraint: may be, for example, detaching two VMs on different servers (or even two data centers), so as to ensure their separation.
- The migration constraint: means that it is possible to execute the VM placement only on a set of precisely stated machines, or to maintain a VM on the same host (or same data center).

We defined also other energy parameters which are : the total number of VMs placed on a PM, the total number of PMs used, the number of reallocations of a VM, the period of time of VM interruption in the migration period, the percentage of maximum and minimum use of VMs/PMs and the response time of a task at the level of a VM (SLA).

The aspect of the sustainability of data is also taken into consideration and which refers to ensuring in real time the replication of each data to multiple hosts (such as a primary and backup host).

Our solution uses an infrastructure model that is informed of the state of the system at any moment desired. It is focused on dynamically manage VM allocation and displacement in the data center, in terms of performance, availability system, cost and instantly energy consumption. This approach adopts optimized resource allocation and live migration through a decision and analysis mechanism and with an effective respect of the strict SLA requirements.

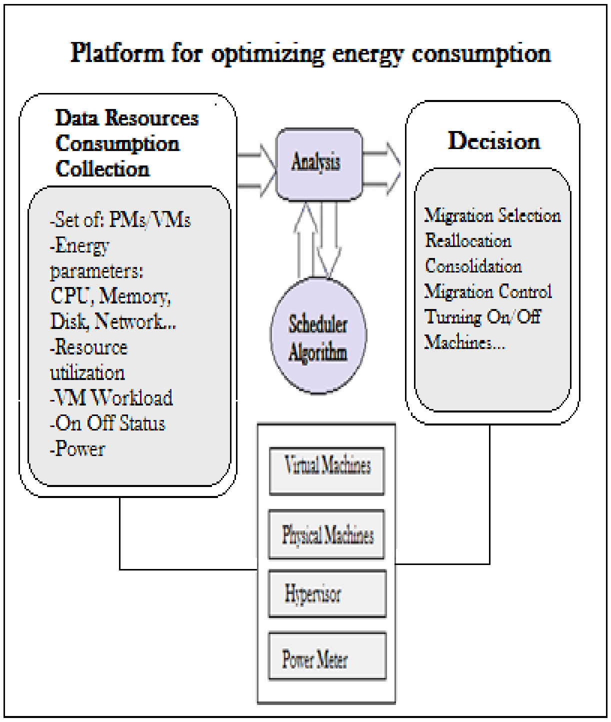


Fig. 2. Overview of the proposed solution.

As Fig. 2 explains, the overview of the proposed optimized cloud model platform and which includes different management stages as following: collecting monitoring data, exploiting these data to calculate a better resources placement, draw a plan to reallocate and dynamically manage resources, and applying the proposed actions (see Fig. 2).

The collection component maintains a periodic monitoring and collection of data related to workload, electrical consumption and the use of resources from the cluster's PMs. These data include CPU, disk storage, memory usage, etc. Also, the electricity collection is done by using power consumption measurement tools (for instance Power Meter or Wattmeter).

The analysis module adopts a built-in scheduler algorithm to perform decisions and which evaluates the results of resource data use. The designed and implemented algorithm does not rely on a particular type of workload and does not need any information about the applications running on the VMs. Its execution involves exact power parameters and high-resolution measurements.

The input of the algorithm is the collection of each PM with its allocated VMs, the different collected resource information, and the specified energy consumption parameters so as to determine which resource to allocate and where to displace it. A new optimized resources placement plan with a set of nodes underloaded and overloaded (to deactivate, turn on later ...) is the output of the algorithm execution.

The decision component aims for the migration, reallocation, and consolidation of the resources. Also, resources can be turned on/off according to the instructions of the analysis model and based on the results previously obtained from the execution of the scheduler algorithm.

IV. Conclusion and Future Works

Today's IT services are using the cloud computing solutions so as to offer to its clients the required services efficiently. Except that the high use of the cloud engenders a large growth in its data center infrastructure. In this case, unfortunately, an enormous amount of electrical energy is consumed and a high amount of carbon dioxide is emitted in the air.

Thus, reducing the energy consumption in cloud data centers while assuring an optimized management of its resources including VMs and servers is becoming a needful aim to achieve. This requirement is related to the green cloud concept by which we can contribute to the environmental protection.

In this paper, various techniques for enhancing the green cloud resources allocation are discussed and which are based essentially on virtualization, migration, and consolidation. Thus, the proposed solution provides an optimized resource management while considering all major energy parameters and major possible constraints of VMs allocation in PMs and which influences on the energy consumed in the cloud computing data center. Also, we focused on taking energy-performance trade-off in concern.

In future, we will detail and implement the scheduler algorithm while respecting the defined Service Level Agreements and the required Quality of services.

References

[1] « Internet Society Global Internet Report 2015 », p. 142.

[2] A. E. I. Ahmed and S.-N. Nada, "Integrated Framework For Green ICT: Energy Efficiency by Using Effective Metric and Efficient Techniques For Green Data Centres".

[3] P. Corcoran, A. Andrae, "Emerging trends in electricity consumption for consumer ICT", 2013.

[4] S. F. Smith, "Is Scheduling a Solved Problem?", in Multidisciplinary Scheduling: Theory and Applications, Springer, Boston, MA, 2005, p. 3‑17.

[5] G. S. Akula and A. Potluri, "Heuristics for migration with consolidation of ensembles of Virtual Machines" in 2014 Sixth International Conference on Communication Systems and Networks (COMSNETS), 2014, p. 1‑4.

[6] S. Agarwal, A. Datta and A. Nath, "Impact of green computing in IT industry to make eco friendly environment," Journal of Global Research in Computer Science, vol. 5, no. 4, pp. 5–10, Apr. 2014.

[7] C. Gu, Z. Li, H. Huang, and X. Jia, "Energy Efficient Scheduling of Servers with Multi-Sleep Modes for Cloud Data Center", IEEE Trans. Cloud Comput., p. 1‑1, 2018.

[8] F. D. Rossi, M. Storch, I. de Oliveira, and C. A. F. D. Rose, "Modeling power consumption for DVFS policies" in 2015 IEEE International Symposium on Circuits and Systems (ISCAS), 2015, p. 1879‑1882.

[9] A. P. Florence, V. Shanthi, and C. B. S. Simon, "Energy Conservation Using Dynamic Voltage Frequency Scaling for Computational Cloud", ScientificWorldJournal, vol. 2016, p. 9328070, 2016.

[10] X. Lin, Y. Wang, Q. Xie, and M. Pedram, "Task Scheduling with Dynamic Voltage and Frequency Scaling for Energy Minimization in the Mobile Cloud Computing Environment", IEEE Trans. Serv. Comput., vol. 8, no 2, p. 175‑186, mars 2015.

[11] Y. Chen, C. Lin, J. Huang, X. Xiang, and X. (Shen, "Energy Efficient Scheduling and Management for Large-Scale Services Computing

Systems", IEEE Trans. Serv. Comput., vol. 10, no 2, p. 217‑230, mars 2017.

[12] Z. Zhou, F. Liu, H. Jin, B. Li, B. Li, and H. Jiang, "On arbitrating the power-performance tradeoff in SaaS clouds" in 2013 Proceedings IEEE INFOCOM, 2013, p. 872‑880.

[13] V. K. M. Raj and R. Shriram, "A study on server Sleep state transition to reduce power consumption in a virtualized server cluster environment" in 2012 Fourth International Conference on Communication Systems and Networks (COMSNETS 2012), 2012, p. 1‑6.

[14] C. C. Lin, P. Liu, and J. J. Wu, "Energy-efficient Virtual Machine Provision Algorithms for Cloud Systems" in 2011 Fourth IEEE International Conference on Utility and Cloud Computing, 2011, p. 81‑88.

[15] M. Sharifi, H. Salimi, and M. Najafzadeh, "Power-efficient distributed scheduling of virtual machines using workload-aware consolidation techniques", J. Supercomput., vol. 61, no 1, p. 46‑66, juill. 2012.

[16] A. Murtazaev and S. Oh, "Sercon: Server Consolidation Algorithm using Live Migration of Virtual Machines for Green Computing", IETE Tech. Rev., vol. 28, no 3, p. 212‑231, mai 2011.

[17] N. Quang-Hung, P. D. Nien, N. H. Nam, N. H. Tuong, and N. Thoai, "A Genetic Algorithm for Power-Aware Virtual Machine Allocation in Private Cloud" in Information and Communication Technology, Springer, Berlin, Heidelberg, 2013, p. 183‑191.

[18] X.-D. Zuo and H.-M. Jia, "An energy saving heuristic algorithm based on consolidation of virtual machines" in 2013 International Conference on Machine Learning and Cybernetics, 2013, vol. 04, p. 1578‑1583.

[19] D. A. Alboaneen, H. Tianfield, and Y. Zhang, "Glowworm Swarm Optimisation Algorithm for Virtual Machine Placement in Cloud Computing" in 2016 Intl IEEE Conferences on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016, p. 808‑814.

[20] J. Huang, K. Wu, and M. Moh, "Dynamic Virtual Machine migration algorithms using enhanced energy consumption model for green cloud data centers" in 2014 International Conference on High Performance Computing Simulation (HPCS), 2014, p. 902‑910.

[21] M. Shaden and A. Heba, "Review of Energy Reduction Techniques for Green Cloud Computing", International Journal of Advanced Computer Science and Applications, 2016:184-195.

[22] J. Sahoo, S. Mohapatra, and R. Lath, "Virtualization: A Survey on Concepts, Taxonomy and Associated Security Issues" in 2010 Second International Conference on Computer and Network Technology, 2010, p. 222‑226.

ASSA: Adaptive E-Learning Smart Students Assessment Model

Dalal Abdullah Aljohany
Computing and Information Technology Department
Al-Hynakiah Community College
Taibah University
Taibah, Saudi Arabia

Reda Mohamed Salama, Mostafa Saleh
Information Technology
Department
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

***Abstract*—Adaptive e-learning can be improved through measured e-assessments that can provide accurate feedback to instructors. E-assessments can not only provide the basis for evaluation of the different pedagogical methods used in teaching and learning but they also can be used to determine the most suitable delivered materials to students according to their skills, abilities, and prior knowledge's. This paper presents the Adaptive Smart Student Assessment (ASSA) model. With ASSA instructors worldwide can define their tests, and their students can take these tests on-line. ASSA determines the students' abilities, skills and preferable Learning Style (LS) with more accuracy and then generates the appropriate questions in an adaptive way then presents them in a preferable learning style of student. It facilitates the evaluation process and measures the students' knowledge level with more accuracy and then store it in the student's profile for later use in the learning process to adapt course material content appropriately according to individual student abilities.**

Keywords*—Adaptive e-learning; e-assessments; adaptive assessment; smart assessment; Learning Style (LS)

I. Introduction

There are different types of e-learning systems; the most recent one is an adaptive e-learning system [1]. In the traditional e-learning system the main criteria in the most cases the implementation of courses follows a "one-size-fits-all" approach, which means all students receive the same content in the same way being unaware of their particular needs. An adaptive e-learning system was proposed to solve this problem. This system tries to adapt to each individual by presenting learning materials dynamically depending on characteristics and learn style of the student [1]. The purpose of adaptive e-Learning is to provide for students the suitable content at the right time, means that the system can determine the knowledge level and organize content automatically for each student [2].

Assessment plays a very important part in any type of the education system, teaching and learning [3]. It aids instructors to evaluate the student's ability level of knowledge. Without an effective assessment, it is impossible to understand the progress of a student, the quality of education that he/she has attained and how effective these courses can be used in his/her future studies [4]. Haken in [5], described assessment as essential in ensuring educational institutions achieve their learning goals. The researcher also found that assessment was crucial in giving the necessary evidence required to seek and maintain accreditation.

Recent advances in computer technology and theories have accelerated the change of test format from traditional tests to Computerized Adaptive Testing (CAT). Traditional tests are typically "fixed-item" tests in which the examinees answer the same questions within a given test. The desire for computerized administration of tests extends beyond using computers to present material and collect responses. It also extends to adding some "intelligence" behind the ways test tasks are presented and the ways they are scored. Within the testing industry, adding intelligence to the selection of test tasks is called "Computerized Adaptive Testing", where the selection of tasks and questions are related to the characteristics and abilities of each individual student [6]. Student's prior knowledge and abilities are important factors to consider in managing the testing and questioning adaptation process. Assessment adaptation takes place on both levels of the test itself and the selected questions of the test [7].

The aim of our paper is to develop an adaptive web-based tool to assist students in the assessment process. In this way, we propose to make the advantages of CATs readily available worldwide, and to allow instructors to define their tests and evaluate exactly their students with a minimum of effort. We have called this tool Adaptive Smart Students Assessment (ASSA) model. The tool that we have developed can be used in two different ways: as a test editor, so instructors can define their tests in an easy way, and as an assessment tool, so students can take the tests online. ASSA determines the students' abilities, skills and preferable learning style with more accuracy and then generates the appropriate questions in an adaptive way, then presents them in a preferable learning style of student.

The rest of this paper is organized as follows. Section II defines background review. Section III describes the proposed assessment model "ASSA". Section IV presents prototype implementation of our model. Finally, this paper concludes in Section V.

II. Background

A. Computer Adaptive Testing (CAT)

CAT is designed in such a way that it makes use of the computers to build adaptive tests in which the selection of each question and the decision to stop the process are dynamically adapted to the student’s performance in the test [8]. The theory used in this system is Item Response Theory (IRT), in which the methodology followed is very straight forward, questions that are posted to the students need not be too difficult or too simple [8]. This is obtained by adjusting the questions to the examinee based on his/her previous answers. The degree of difficulty of the subsequent question is chosen in a way so that the new question is neither too hard, nor too easy for the examinee. When the examinee attends his/her first question the capability of him cannot be predicted but gradually when he/she attempts to the second question the computer can analyze very well the knowledge level of the examinee [9]. The advantages of Computer Based Assessment (CBA) and CAT remain the same, but CAT has certain extra functionality that makes it very efficient and effective compared to the other assessment. There are different platforms available for designing and deploying CAT [10]. Reisinge et al. [10] propose software architecture that enables the development of completely customizable CAT tools with respect to domain-specific item design and visualization as well as deployed CAT algorithms. CAT retains the advantages of CBA, but there are additional benefits that make CAT a very efficient and effective means of assessment [11]:

- Reliable and accurate estimate of the efficiency of the student.
- Reduce the time of testing.
- Avoid easy / difficult items that may cause stress.

There are several types of adaptive tests but all of them are derived with two steps: question selection and score estimation [8]. Question selection is the integral part of the assessment in which the questions are collected and compiled based on the student’s performance level. These questions are forming a pool which will have the collection of questions that was asked to any particular student at a single time. The next step is the score estimation in which the student performance is complied with the responses given by them and it's refined at regular intervals. This allows the questions asked next to be more appropriate still. This cycle continues until either a specified number of questions have been administered or some measure of score precision is reached. Several adaptive testing systems have been developed such as: The SIETTE [12].

Developing of multistage adaptive test (MST) panels is a hot topic today and it has encouraged new developments. The most commonly used approaches for MST: bottom-up and top-down [13]. In the bottom-up approach the whole test divided into several unit, and each unites is constructed first, then all unites are gathered to get the whole test, while the top-down approach trails the reverse direction. Both methods have their pros and cons, and sometimes neither is convenient for practitioners [13]. The advanced mix strategy presented in [13] is to build best multistage adaptive test (MST) panels. In [14], two different approaches, CAT with a shadow test approach and computerized multistage testing have been proposed to ensure the satisfaction of subject matter experts. In the shadow test approach, a full-length test is assembled that meets the constraints and provides maximum information at the current ability estimate, while computerized multistage testing gives subject matter experts an opportunity to review all test forms prior to administration.

B. Adaptive Questions

This technique generates dynamically sequenced questions based on the student’s response. Question selection is based on several predefined rules as well as the student’s current responses [15]. Therefore, highly structured pools of questions are required for rules to have enough information to select questions. As opposed to adaptive testing, this approach presents instructors with more flexibility in including didactical and personal methods. This is done by creating appropriate rules [15]. Several adaptive questions systems/tools have been developed such as, AthenaQTI [15] and CoSyQTI [16].

C. Adaptive Presentation of Questions

This method present questions in the appropriate format and structure that matches student learning style. Students have different way to understand assessment materials this called learning style. More formally, learning styles represent a combination of cognitive, affective and other psychological characteristics that work as relatively stable indicators of the way a student perceives, interacts with and responds to the learning environment [17]. Fortunately, many Adaptive e-Learning researches assumed that personalizing or presentation of the question to match the student's learning preferences, especially the learning style, would aid the student’s understanding. Several systems have been developed that provide adaptive presentation for assessment materials, such as: INSPIRE System [18] and PIAT [19].

III. Adaptive Smart Students Assessment (ASSA) Model

ASSA is one of the components of the Adaptive e-Learning Environment. Instead of building that environment from scratch, the decision was to use an Open Source Learning Management System (LMS). Moodle was chosen because of its popularity as it is used in several universities (https://moodle.net/stats/). Moodle also is known as simple and easy to adapt and customize to the needs of the educational system. We can see in Fig. 1, the basic architecture of the Adaptive e-Learning Environment. Three main engines in the adaptive e-Learning Environment are integrated to the open source Moodle, namely, Authoring, Delivery, and Assessment engines. Each of those main engines works smartly with the aid of the knowledge-base which contains the information about the botanical domain and is being incrementally built using web forms by different users in diverse locations. Each component, including the KB, has an independent web-based interface that allows the whole system to be used both as a learning tool and as an independent assessment tool.

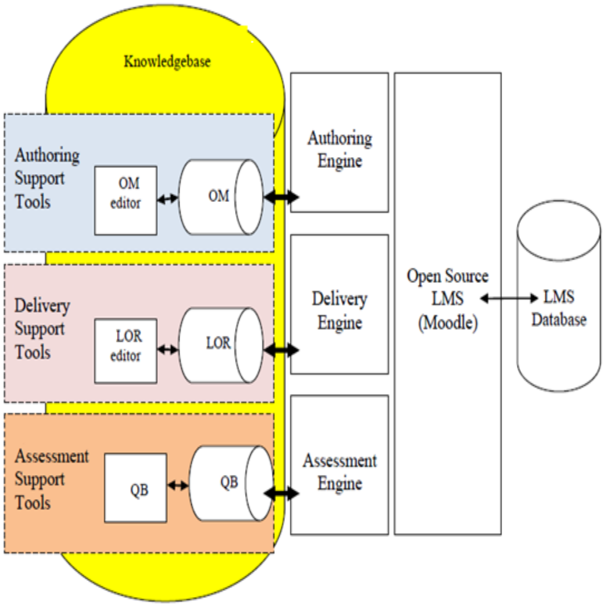


Fig. 1. Architecture for the adaptive e-learning environment.

A. *Knowledgebase*

As shown in Fig. 2, the Knowledgebase is composed mainly of four major components: the LO, QB and domain ontology knowledgebase; student database; and course database. The knowledgebase is composed of the Learning Object Repository (LOR); the Ontology Model (OM); and the Question Bank (QB). The QB is composed of two components: Smart Learning Question Repository (SLQR), and Smart Constituents Learning Objects Repository (SCLOR). The SLQR is data base that hold a collection of Smart Learning Question (SLQ) objects. A question is not an atomic object, but rather question object integrates the question constituents. Those constituents may wholly or partially be drawn from SCLOR. The database is composed of the Student Model (SM) and the Course Model (CM), which themselves are further decomposed. The SM is composed of two components: the student's Learning Style Model (LSM) that is defined in terms of the three dimensions of Felder & Silverman Learning Style Model (FSLSM) and the Student's Background Domain Knowledge (SBDK) representing the knowledge that the student captures with an acceptable cognitive depth for the domain of study. In addition, the Course Model (CM) is composed of three components: The Course Learning Outcomes (CLO), the Course Syllabus, and the Table of Contents (TOC).

The main objective of assessment is to measure knowledge level of student which will be used in adaptive e-learning environment to adaptive learning materials. So, we need to measure score of each concepts/topic to determine knowledge level of student. This is generally done through providing questions to the students to answer them. Once the student submits his/her an answer of a question, this answer will be evaluated. Then constructive feedback and result will be provided to student. ASSA determines the students' abilities, skills and preferable learning style with more accuracy and then produces the suitable questions in an adaptive way, then presents them in a preferable learning style of student. A high number of learning style models was proposed to define learning style of student, we have selected the Felder model as the basis of our research and we focused only on three of the FSLSM's dimensions, namely, Global/Sequential, Sensing/Intuitive and Visual/Verbal. Therefore, there are 8 (2^3). The question can be presented in three different ways to match three dimensions of student learning style. Therefore, if question is not clear, student can ask rephrasing question to match another dimension of his/her learning style.

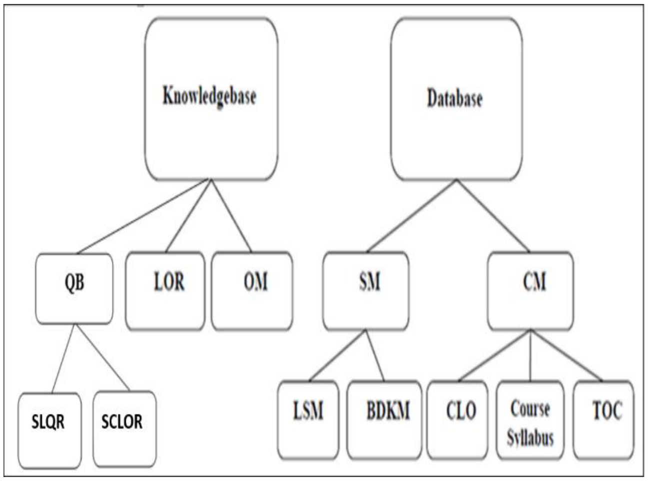


Fig. 2. The Knowledgebase.

ASSA facilitates the evaluation process and measures the knowledge level of students with more accuracy and then store it in the student's profile for later use in the learning process to adapt course material content appropriately according to different student abilities. ASSA used Ontology Model (OM) to direct the assessment process to assess the student's knowledge and skills, and it also uses the Revised Bloom Taxonomy (RBT) to navigate upward and downward the cognition pyramid to determine the student's cognitive level. ASSA can be used in two different ways:

- Instructors and/or domain experts can use ASSA to develop the tests, that is, to define their topics, questions, parameters, and specifications.
- Students can use ASSA to take the tests that are automatically generated according to the specifications provided by the examiner.

B. *The Question Bank (QB) Model*

Question Bank (QB) contains a collection of questions of test that are stored in SLQR, and may be supported by components from SCLOR, and then presented to the student according his/her Learning Style Model (LSM). SLQR is an essential storage (data base) that maintain and hold a collection of Learning Question (LQ) objects. LQ like an LO, is specified through a set of metadata attributes to facilitate question selection and manipulation. LQ metadata which is described with SCORM metadata. The LQ Model has extended the standard metadata model of SCORM by: adding additional attributes essential for supporting the theories it implements, such as Learning Style Model. SCLOR is a repository which holds a collection of objects that support and constituting the question. Some of these objects are multimedia objects in different format. Like SLQR, it is described by SCORM metadata standard and extra attributes.

C. The Question Model

The Question is described in terms of many dimensions: Question type, Question purpose, and Question structure.

1) Question types: Reviewing the different types of questions in all types of tests, the following question types are recognized:

- MCQ,
- T/F,
- Fill-in the spaces (may take one of two forms; free answers or like MCQ, i.e., selecting from a given list of possible answers),
- Matching two groups,
- Problem solving.

Grading is a critical determinant of the types of questions to be used in a test. The number of students, nature of the test, availability of expert graders, etc. are critical determinants. Classical automated grading natively covers the first four types of questions with simple correct/no correct evaluation, while Smart Grader addresses grading of the fifth question type.

2) Question's assessment purpose: Each question serves a purpose in the assessment process. Fig. 3 describes the interrelationship of a question to the different elements of assessment. The question is tied with M-M relationships to the course educational objectives and outcomes, knowledge topics, and skills that it assesses.

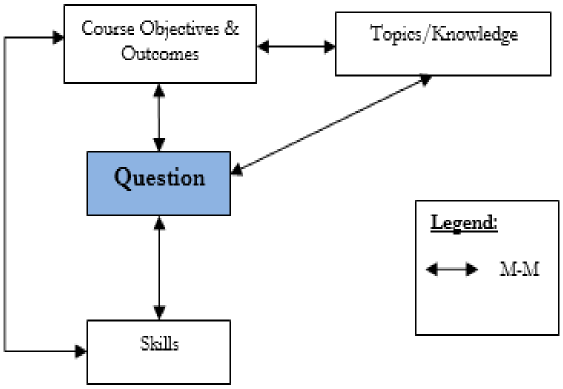


Fig. 3. Question has different assessment elements.

3) Question structure: The structure of a Question is shown in Fig. 4. A question is composed of three main components: Question Header, Question Details, and Question Answers. Each of these components has a different structure for the different question types. This is normal due to the differences in their nature. Noteworthy, the problem-solving type of questions is more complex and requires special intelligent handling. In answering such questions, the student follows an action plan strategy that the grader must recognize for proper evaluation.

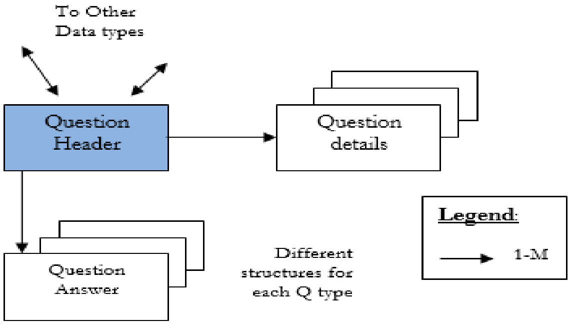


Fig. 4. A question structure.

D. Grading and Assessment Analysis

When student submits his/her answer, assessment engine will pass answer to simple or smart grader depends on question type. A simple grader covers the four types of questions, namely, T/F, MCQ, Matching, and fill in- the- spaces while smart grader developed for assessing problem-solving and proving type of questions. When the grading process is complete, the evaluation engine provides the student with his/her result and appropriate feedback.

Most question types are simply graded except the problem-solving type. Smart Grader (SG) is developed to assess and evaluate problem solving question (such as programming exercise). This is generally done through providing programming exercises for the students to answer them. To assess these exercises effectively, it is necessary for the SG to analyze any solution provided by student and compare it with instructor/expert solutions. Then SG provides constructive feedback and result to student. The programming exercise do not have a unique strategy to solve it, but it can be solved in many ways. Therefore, in evaluation phase matching a program line by line is not effective method of analyzing it. The evaluation process of programming exercise questions requires the instructor to define all possible strategies of solution to compare them with student solution. Therefore, evaluation of problem solving question is complex process and it requires special intelligent handling.

Problem solving is seen as a series of steps that can be used to reach a goal (answer). In each step, multiple actions are possible leading to other step. Therefore, in answering problem-solving questions, the student follows an action plan strategy. To assess these exercises effectively, it is necessary for instructor/expert to define plan that contains all possible strategies of answer. Then, SG detects the solution strategy that the student follows by analyzing his/her action plans in answering the question. As shown in Fig. 5, when student submits his/her solution, SG must first analyze student solution and recognize action plan behind it. Then, SG compares student action plan with instructor/expert action plan of solution. After that, SG provides constructive feedback and result to student.

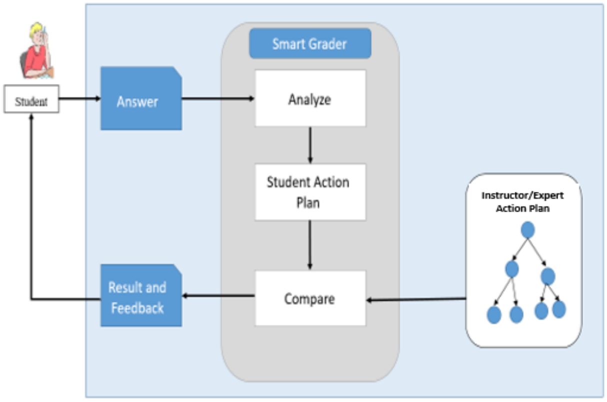


Fig. 5. Evaluation process of problem solving question.

The instructor/expert action plan of possible programming exercise solutions specifies using grammar and production rule and represents using tree where root is goal and internal nodes is non-terminals and leaf represent terminals. The production rules to the goal (answer) is useful to specify instructor/expert action plan of solutions. The first rule starts with goal as following:

Start → Goal (answer)

The goal can be reach using multiple strategy, therefore second rule will be as following:

Goal → Startgy$_1$ | Strategy$_2$ | …. |Strategy$_n$

These strategies can also be divided into other sub-strategies. Each strategy has series of steps that can be used to reach a goal (answer). In each step, multiple actions are possible leading to other step. An action can be decomposed into more actions at a later stage. When student submits his/her answer, SG recognizes action plan that followed by student which is a set of one or more rules that satisfy the goal. If the student action does not match a production rule action, the SG provides immediate feedback to student.

Suppose we have problem P and set of problem solving strategies of the form:

S_1 or S_2 …or S_n.

The solving problem space represent by using and-or tree. The associated and-or tree is a set of labelled nodes such that:

- The root of the tree is a node labelled by P.
- The alternative sets of children corresponding to alternative strategies (S_1 or S_2 … S_n) of solving are grouped together by an "or".
- For every node Si may have a set of children nodes ($S_{i,1}$ … $S_{i,n}$) corresponding to alternative sub-strategies of solving and they are grouped together by an "or".
- For every node (S_i or $S_{i,j}$) there exists a set of children nodes N_1, …, N_n. Each node represents one strategy step of the problem solution. The nodes are conjoined by an arc, to differentiate them from other nodes that might be related with other strategies. Also, these nodes are arranged in order to reflect step order in solution.
- The terminal leaves are the actions.

Suppose we have problem P that can be solved by two strategies (S_1 or S_2). There are three steps (a, b and c) in S_1 and two steps in S_2 (a and b). The solving problem space of P represent by and-or tree as shown in Fig. 6.

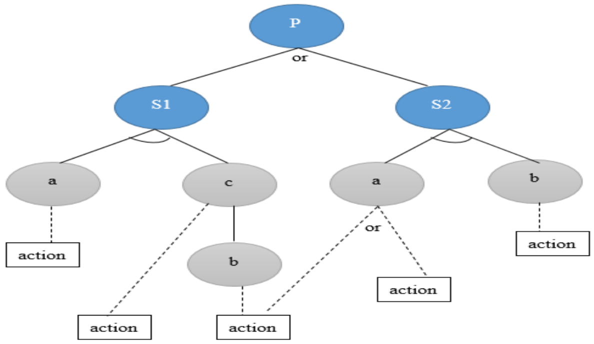


Fig. 6. Tree of solving problem space.

IV. Prototype Implementation

We implement ASSA prototype using PHP and oracle Database. ASSA allows producing and delivering adaptive and smart test through web interfaces. It works inside web-based adaptive e-learning systems as a diagnosis tool. It based on a 2-tier client/server platform for adaptive student assessment. Fig. 7, explains the 4-layer architecture of ASSA platform. Four subsystems have been developed, which accessed by four types of users: admin, experts, instructors and students. Through Web browser (Presentation layer), users communicate with the corresponding Web application on the Application Server where dynamic Web content is created using PHP (Application layer). The system Database is located on the Oracle Database Server (Data layer).

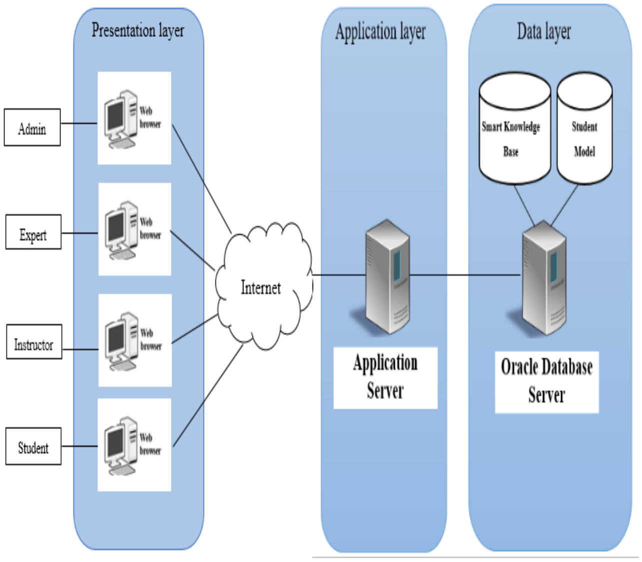
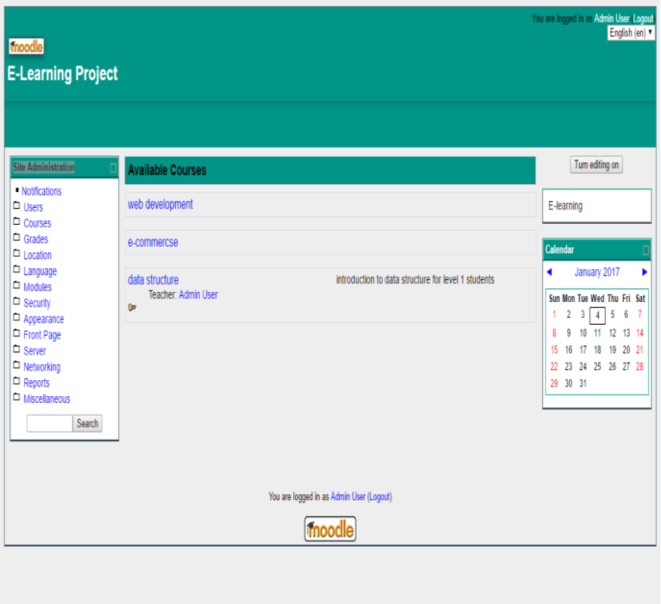


Fig. 7. ASSA prototype architecture.

A. The Admin Sub-System

This sub-system provides the admin with web-based services to add (update or delete) the basic data needed for the other sub-systems (as shown in Fig. 8).



You are logged in as Admin User Logout

web development

e-commercse

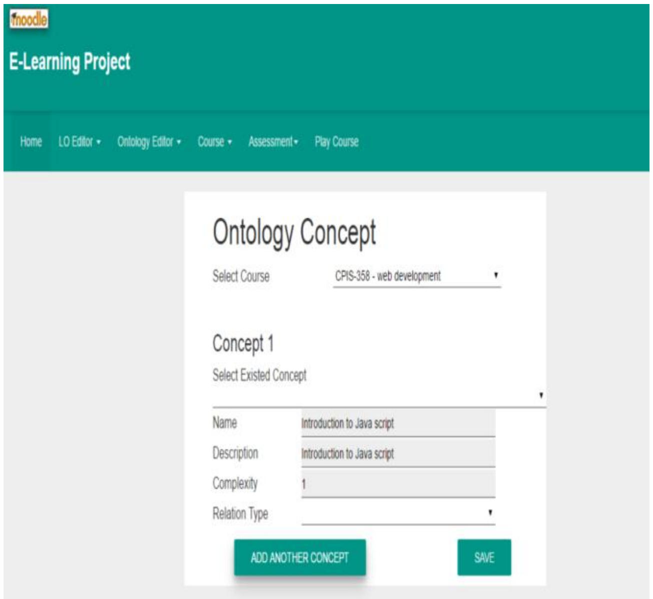
data structure introduction to data structure for level 1 students
Teacher: Admin User

You are logged in as Admin User (Logout)

Fig. 8. Admin sub-system.

B. The Expert Sub-System

This sub-system provides the expert with web-based services to build Ontology Model (OM) by adding concepts/topics and specifying hierarchy and relations connecting them with other. As shown in Fig. 9, the expert can add new concept to the course and store the following data for each concept: name, description, complexity level and he/she can determine and specify relation that connects this concept with other concepts.



Ontology Concept

Select Course CPIS-358 - web development

Concept 1

Select Existed Concept

Name Introduction to Java script

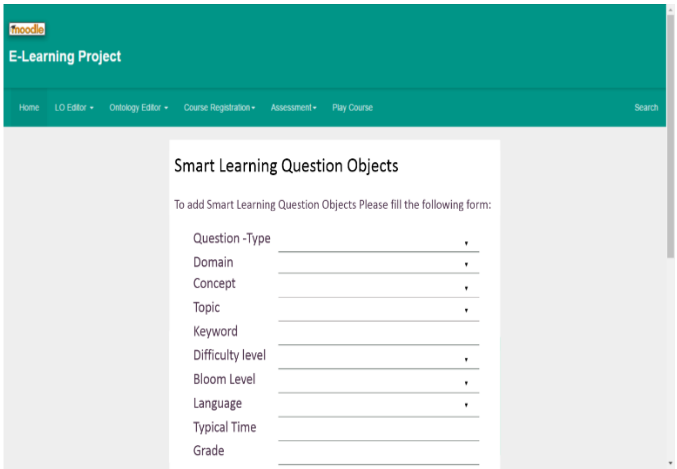
Description Introduction to Java script

Complexity 1

Relation Type

Fig. 9. Add concept interface.

Also, this sub-system provides expert web-based services to builds the Question Bank (QB) based on the Question Model. QB contains collections of SLQ. Adding SLQ requires several steps from expert. SLQ specified through a set of metadata attributes to facilitate questions selection and manipulation**. The first important step** is to create SLQ requires expert to specify this metadata attributes of smart questions (as shown in in Fig.10).



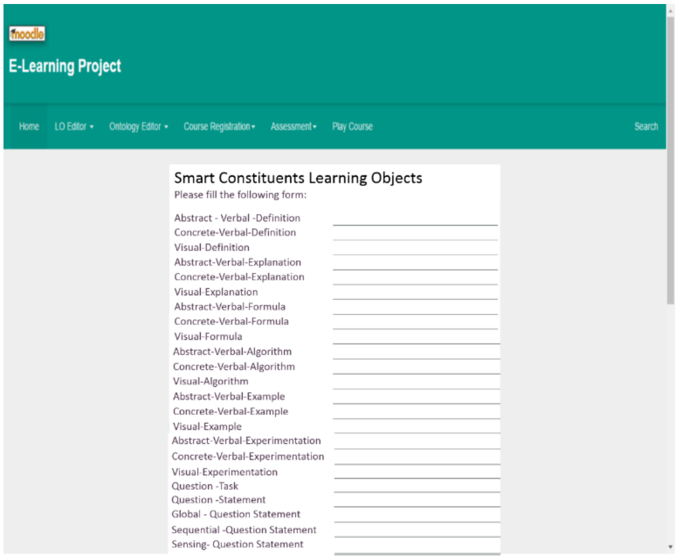
Smart Learning Question Objects

To add Smart Learning Question Objects Please fill the following form:

Question -Type
Domain
Concept
Topic
Keyword
Difficulty level
Bloom Level
Language
Typical Time
Grade

Fig. 10. SLQ metadata attributes interface.

As shown in Fig. 11, creating question constituent's objects is **the second step** to create SLQ. As mentioned before, the question object is an abstract framework that specifies and integrates the question constituent's objects.



Smart Constituents Learning Objects
Please fill the following form:

Abstract - Verbal -Definition
Concrete-Verbal-Definition
Visual-Definition
Abstract-Verbal-Explanation
Concrete-Verbal-Explanation
Visual-Explanation
Abstract-Verbal-Formula
Concrete-Verbal-Formula
Visual-Formula
Abstract-Verbal-Algorithm
Concrete-Verbal-Algorithm
Visual-Algorithm
Abstract-Verbal-Example
Concrete-Verbal-Example
Visual-Example
Abstract-Verbal-Experimentation
Concrete-Verbal-Experimentation
Visual-Experimentation
Question -Task
Question -Statement
Global - Question Statement
Sequential -Question Statement
Sensing- Question Statement

Fig. 11. Question constituent's objects.

The last step to build SLQ requires expert to add details and answers of questions. Our model covers many types of questions as mentioned before, such as: T/F, MCQ, fill-in the spaces, matching two groups and problem solving. Each of these components has a different structure for the different question types. Fig.12 displays structure of MCQ.

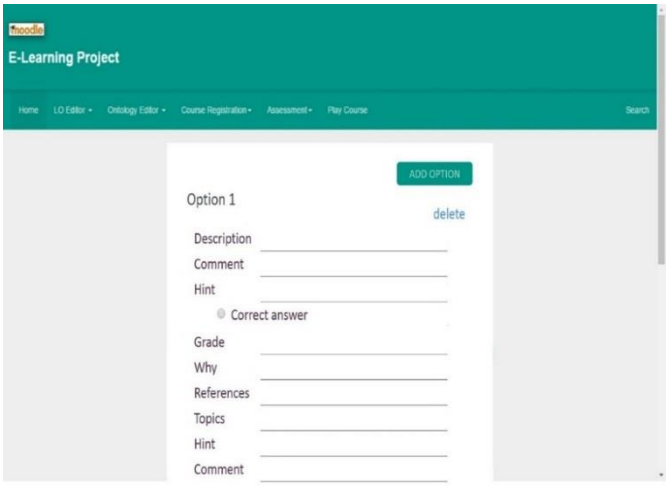


Fig. 12. Structure of MCQ.

Fig. 13 displays structure of fill-in the spaces questions while Fig. 14 presents matching two groups questions structure.

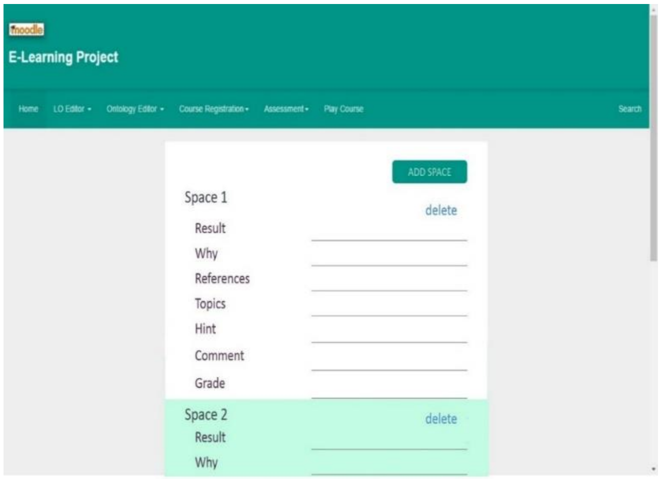


Fig. 13. Structure of fill-in the spaces question.

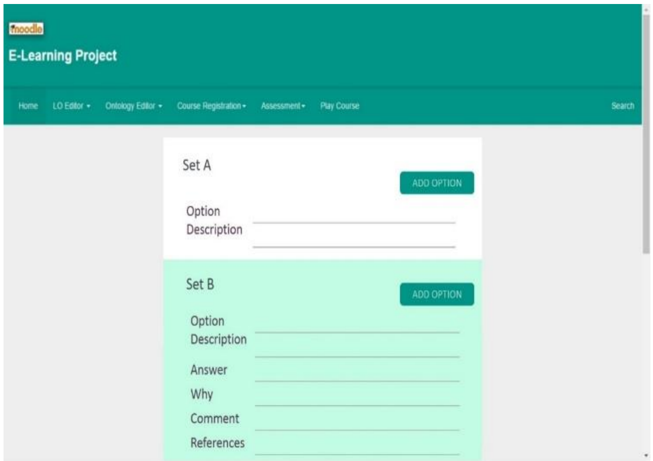


Fig. 14. Structure of matching two groups question.

As shown in Fig. 15, the problem-solving type of questions is more complex and requires special intelligent.

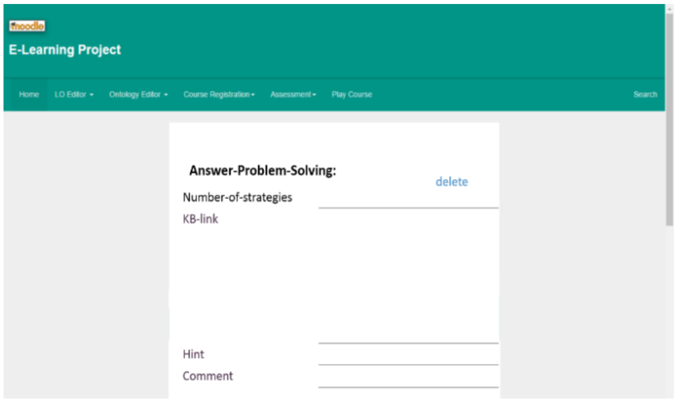


Fig. 15. Structure of problem solving question.

C. *The Instructor Sub-System*

This sub-system provides the instructor with web-based services to create two types of assessments Quiz and Exams with different types of questions from. As shown in Fig. 16, in "Assessment" menu item instructor can find the following options:

- Add Assessment.
- View Course Assessment.
- Add view questions.
- Link questions to Assessment.

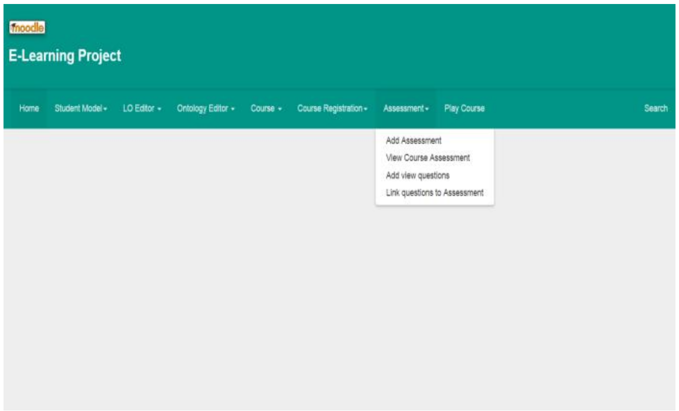
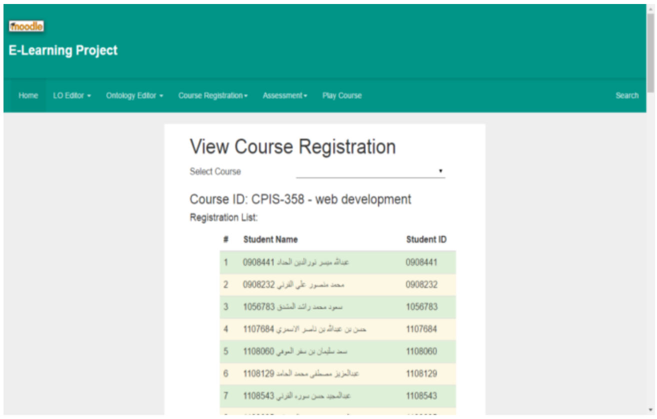

moodle
E-Learning Project
Home Student Model ▾ LO Editor ▾ Ontology Editor ▾ Course ▾ Course Registration ▾ Assessment ▾ Play Course Search
Add Assessment
View Course Assessment
Add view questions
Link questions to Assessment

Fig. 16. Assessment menu.


moodle
E-Learning Project
Home LO Editor ▾ Ontology Editor ▾ Course Registration ▾ Assessment ▾ Play Course Search

View Course Registration

Select Course

Course ID: CPIS-358 - web development

Registration List:

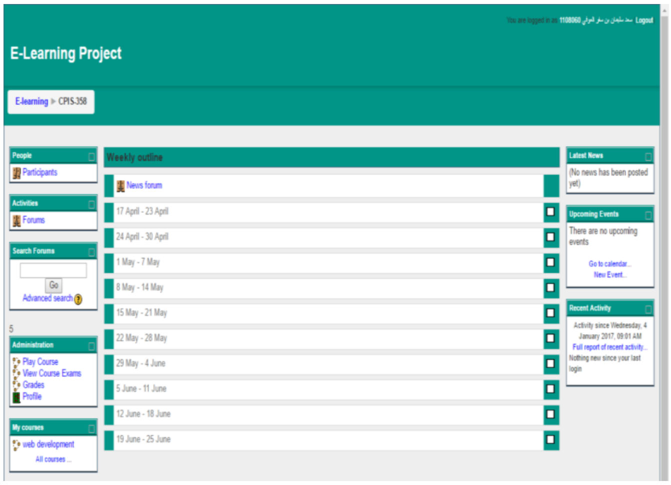
#	Student Name	Student ID
1	0908441 عبدالله ميسر نورالدين الحداد	0908441
2	0908232 محمد منصور علي القرني	0908232
3	1056783 سعود محمد راشد المشدق	1056783
4	1107684 حسن بن عبدالله بن ناصر الاسمري	1107684
5	1108060 سعد سليمان بن سفر العوفي	1108060
6	1108129 عبدالعزيز مصطفى محمد الحامد	1108129
7	1108543 عبدالمجيد حسن سوره القرني	1108543

Fig. 17. View students in course interface.

From Course Registration instructor can view students registered in this course as shown in Fig. 17.

D. The Student Sub-System

Student can open the Moodle application (in browser) and Login with student username and password then select the course as shown in Fig. 18. This sub-system allows student to take quiz that contains different types of questions with different difficulty levels. These questions are presented to the student in a sequential manner commensurate with his abilities and level of knowledge. Also, these questions are presented in a way that suits the student's preferred learning style.



(No news has been posted yet)

There are no upcoming events

Activity since Wednesday, 4 January 2017, 09:01 AM

Nothing new since your last login

Fig. 18. Students interface.

V. Discussion and Conclusion

In adaptive e-learning system the assessment is very important, it is used to measure students' knowledge level to adapt course material's according it. Therefore, this type of e-learning requires smart assessment model that determines student knowledge level with more accuracy. This paper presented Adaptive Smart Students Assessment model, named ASSM that works inside web-based adaptive e-learning systems as assessment tool. ASSA can be used in two different ways: as a test editor, so instructors can define their tests in an easy way, and as an assessment tool, so students can take the tests online. This model simplifies the assessment process and measures the students' knowledge level with more accuracy and then stores it in the student's profile for later use in the learning process to adapt course material content appropriately according to individual student abilities.

The adaptation in this model includes sequences of questions to avoid questions that are much easy or much difficult to student. Also, adaptation include presentation of questions to match student learning style. ASSA determines the students' abilities, skills and preferable learning style with more accuracy and then generates the appropriate questions in an adaptive way, then presents them in a preferable learning style of student.

ASSA differs from other adaptive models in several issues. In this model, the idea for providing adaptive presentation of question statement was developed. Also, in ASSA three dimensions of the FSLSM are considered rather than using only one of them. This allows providing more accurate adaptivity by incorporating different aspects of learning styles as proposed by the learning style model. Also, this model differs from other adaptive models because it contains all types of questions. Simple grader developed to assess simple type of questions, while smart grader developed to assess complex type of questions (such as programing exercises). The smart grader has ability to define many solutions for single programing exercise.

We implement ASSA prototype using PHP and oracle Database. ASSA allows producing and delivering adaptive and smart test through web interfaces. It is based on a 2-tier client/server platform for adaptive student assessment.

The future work includes incorporating other types of questions (such as: comparing, critiquing, discuss, describe etc.) which require language processing researches.

Acknowledgment

This work was supported by King Abdulaziz City of Science and Technology (KACST) funding (Grant No. AT-204-34). We thank KACST for their financial support.

References

[1] H. D. Surjono, "The Design of Adaptive E-Learning System based on Student's Learning Styles," (IJCSIT) International Journal of Computer Science and Information Technologies, vol. 2, no. 5, pp. 2350-2353, 2011.

[2] V. Esichaikul, S. Lamnoi and C. Bechter, "Student Modelling in Adaptive E-Learning Systems," Knowledge Management & E-Learning: An International Journal, vol. 3,no.3, 2011.

[3] M. Phankokkruad, K. Woraratpanya, "Web Service Architecture for Computer-Adaptive Testing on e-Learning," World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering, vol. 3,no. 10, 2008.

[4] "Assessing and Evaluating Student Learning".

[5] M. Haken, "Closing the loop-learning from assessment," In: Presentation made at the University of Maryland Eastern Shore Assessment Workshop. Princess Anne: MD, 2006.

[6] F. Mampadi, et al., "Design of Adaptive Hypermedia Learning Systems: A Cognitive Style Approach," Computers & Education, 2010.

[7] A. Tzanavari, S. Retalis and P. Pastellis, "Giving More Adaptation Flexibility to Authors of Adaptive Assessments," in Adaptive Hypermedia and Adaptive Web-Based Systems, Springer Berlin Heidelberg, 2004.

[8] T. Davey, "A Guide to Computer Adaptive Testing Systems," COUNCIL OF CHIEF STATE SCHOOL OFFICERS, 2011.

[9] S. M. Čisar, D. Radosav, B. Markoski, R. Pinter and P. Čisar, "Computer Adaptive Testing of Student Knowledge," Acta Polytechnica Hungarica, vol. 7, no. 4, pp. 139-152, 2010.

[10] F. Reisinge, A. Eckmaier and C. Helm, "A flexible online platform for computerized adaptive testing," *International Journal of Educational Technology in Higher Education*, vol. 14, no. 2, 20 January 2017.

[11] C.Marinagi, "Web-based adaptive self-assessment in Higher Education," Education in a technological world: communicating current and emerging research and technological efforts, pp. 78-84, 2011.

[12] R. Conejo and et al. "SIETTE: A Web–Based Tool for Adaptive Testing," International Journal of Artificial Intelligence in Education, vol. 14, pp. 29-61, 2004.

[13] X. Xiong, "A Hybrid Strategy to Construct Multistage Adaptive Tests," *Applied Psychological Measurement,* March 26, 2018.

[14] Kimura. T," The impacts of computer adaptive testing from a variety of perspectives," *Journal of educational evaluation for health professions*, vol. 14, no. 12, 29 May 2017.

[15] A. Tzanavari, S. Retalis and P. Pastellis, "Giving More Adaptation Flexibility to Authors of Adaptive Assessments," in Adaptive Hypermedia and Adaptive Web-Based Systems, Springer Berlin Heidelberg, 2004.

[16] P. LALOS1, S. RETALIS and Y. PSAROMILIGKOS, "Creating personalised quizzes both to the learner and to the access device characteristics: the Case of CosyQTI," in 3rd International Workshop on Authoring of Adaptive and Adaptable Educational Hypermedia (A3EH), 2005.

[17] J.W. Keefe, "Learning style: an overview,"NASSP's Student Learning Styles: Diagnosing and Prescribing Programs, pp. 1–17, 1979.

[18] K.A.Papanikolaou and et.al, "Personalizing the interaction in a Web-based educational hypermedia system: the case of INSPIRE," in User-Modeling and User-Adapted Interaction, pp. 213–267, 2003.

[19] L. Al-Rajhi, R. Salama and S. Gamalel-Din, "Personalized Intelligent Assessment Model for Measuring Initial Students Abilities," In Proceedings of the 2014 Workshop on Interaction Design in Educational Environments, p. 41, June 09 - 09, 2014.

Implementation of NOGIE and NOWGIE for Human Skin Detection

M. Omer Aftab, Junaid Javed, M. Bilal, Arfa Hassan
Department of Computer Science
Lahore Garrison University,
Lahore, Pakistan.

M. Adnan Khan
Department of Computer Science
NCBA&E,
Lahore, Pakistan.

***Abstract*—The Digital image processing is one of the most widely implemented fields worldwide. The most applied applications of digital image processing are facial recognition, finger print recognition, medical imaging, law enforcement, cyber-crime investigation, identification of various diseases and criminals, etc. The subject to be discussed in this article is skin detection. Skin detection has solved many serious problems related to digital image process. It is one of the main features in making an intelligent image processing system. The proposed methodology conducts an improved and well enhanced skin detection, the skin and non-skin parts are divided from an input image or video, noise is removed, HSV is applied which also acts as a color model that generates more better results in accordance to RGB or YCbCr for skin and face identification. The algorithms, NOGIE (Noise Object Global Image Enhancement) and NOWGIE (Noise Object with Global Image Enhancement) are applied separately on the input and the results can be compared for better perception and understanding of the applied skin detection techniques, the skin parts are highlighted as "White" while the Non-skin parts are highlighted as "Black". The results are different NOWGIE gives better results than the NOGIE due to the image enhancement technique. This methodology is subjected to be implemented in special security drones for the identification of suspects, terrorists and spy's the algorithms provides the ability to detect humans from a non-skin background making an autonomous and excellent security system.**

Keywords—Skin detection; Digital Image Processing (DIP); Noise Object Global Image Enhancement (NOGIE); Noise Object with Global Image Enhancement (NOWGIE); Hue Saturation and Value (HSV); RGB

I. Introduction

The digital image processing is one of the most widely implemented fields throughout and is a big step towards autonomy. The world is comprising of a tremendous number of applications working on the very foundation of digital image processing that one is surrounded by that cannot be neglected. The skin detection is a branch of digital image processing that have gained quite much importance and is continued to do so. The rapid technological growth causing the implementation of skin detection widely so much making it a complete domain inherited with numerous contraptions, like fingerprint recognition, face detection and recognition, lip reading, pattern recognition, artificial intelligence and much more [1]. Currently it is also subjected to be playing a vital role in various security systems for the reduction in criminal activities, better law enforcement like intelligent traffic video surveillance that uses the skin detection techniques for human facial recognition to identify criminals etc. [2]. The question is how does it work? An intelligent skin detection system should be able to detect human skin when it is provided with some input image or video. If it is accompanied by an image comprising a human, some animals and trees etc. the system must identify the human it detects the human skin by stratification of human skin color and texture as the human skin is not specifically geometric, skin detection leads to facial recognition [3]. The system works by categorizing via color segmentation of the skin and non-skin parts ignoring the non-skin part. On the skin part it is able to identify the human skin as it is provided by the (Hue Saturation and value) **HSV** as an input. The HSV acts as a color model that is more efficient than the RGB, the procedure requires background knowledge about the objects regarding within the image [4]. The color segmentation helps to skim and identify the skin from which can recognize the faces, finger prints and other human body parts.

In this article the proposed methodology is actually an upgraded version of the previous work done that was the application of GIE and without GIE techniques for skin detection [5]. This enhanced version is the application of the NOGIE and NOWGIE algorithms to detect human skin from an image or video. Both of the methodologies basically work on the same principle of human skin detection. In this methodology, These algorithms are very different from the previous and other various skin detection techniques as it promotes autonomy and more improved results. NOGIE stands for (Noise Object Global Image Enhancement) and NOWGIE stands for (Noise Object With Global Image Enhancement). Both of the algorithms are quite related to the GIE and without GIE terms, but these algorithms are applied step wise and gives a more enhanced output for the detection of human skin providing better results. On the input image or video first the NOGIE algorithm is applied after that the NOWGIE algorithm is applied and their results are compared at the end. The application of this methodology is via MATLAB, an image or video is read and then the noise is removed after that object is detected the HSV value is inserted in the algorithm here the HSV acts as a membership function that varies and continue to change it until the value is set for the desired results. After the skin is detected, same procedure is repeated on the input, but equalization is implemented after the object detection. This methodology is subjected to be deployed in special security drones for the identification of criminal suspects, terrorists or spy's. For example, if a spy or terrorist of an enemy have crossed the border or is in a sensitive area these special drones will be released to search the targeted area where the spy, terrorist is expected to hide. The drones will start to look up and will be able to detect the suspect from a non-skin background with ease. Further the drones can be programmed to perform some action after the identification of a spy like to turn on the alarms, call the security teams, etc. making a highly intelligent autonomous fool-proof security system.

II. Literature Review

After studying and reviewing a lot of research papers relative to skin detection there are various techniques being applied in accordance to detect human skin.The HSV is far better than the regular RGB. First is to acquire the optimum results the RGB space is converted into the HSV space, HSV color space is independent of the three-color components of the RGB color space. Then calculate the differences in the Histograms of successive video frames on each color component respectively the calculated Histogram difference is the base for the feature detection, this fuzzy logic based human skin detection technique is proved to be very satisfactory, but the Interpolation methodology have proven to be far better [6], [7]. Applying the skin segmentation along with the fuzzy logic is another technique where the whole frame comprising skin pixels is divided into two parts one is the fuzzy part [6] and the other is skin segmentation, training the system requires images with face and non-face here the image pixels are read in row segments to form the column segments the images consisting of face the system takes then as 1 while the comprising of no face the system takes it as a T-S model based fuzzy training is implemented and the fuzzy function comprising weights are learnt via algorithm training the system requires multiple images of different pixel values, distances and sizes, 69 face and 56 faceless images were provided for system training for accurate results in detecting faces at instant speed [8]. A simple method is applied for the acquisition of skin pixels from RGB images consisting of facial constraints, such RGB image is taken as an input then techniques are applied for the detection of nose, the color pixels of nose's skin tone is extracted the nose is considered as a main region for the identification of the same skin pixels from the facial regions after that skin segmentation is performed and histogram model is constructed by applying fusion strategy via Gaussian Model, the results are obtained and compared both Gaussian model and fusion strategy gives good results but the Fusion strategy provides the best output [9]. The methodology works for the acquisition of facial features from images with faces within automatically by the selection of convolutional neural network (CNN) that is already trained the CNN is one of the most used type of artificial neural network for digital image processing comprising of multi-layer architecture, the trained CNN when provided with an image it divides the pixels and tend to identify the skin toned pixels that leads to the identification of facial structure, this methodology is capable of matching a face when young and when old as the skin tend to get older with time but this technique not only recognize the human face but also identifies a face when young and when it gets older by memorizing the facial constraints of a younger face and calculates the age distance with the same older face of singleton and image data set efficiently [10].

III. Problem Statement

The problem that arises in the lack of accuracy for skin detection and also the mistakes, bugs in the algorithms that affect the output and can also alter the results. If the skin detection is accurate the output displayed is often not very clear or enhanced. In the proposed method the algorithm does an accurate skin detection along with enhancement of the output image that makes it more visible and clearer to understand. Apart from the facial or finger print recognition the skin detection is also applied in medical sciences for the identification of various skin diseases, infections etc. This makes it more complicated for the exact identification of any skin disease. Because if not can cause problems and lead to false treatment and medication as the system is unable to get a proper input so how it will be able to generate an accurate output.

IV. Proposed Method

In this paper the proposed methodology is a complete and improved skin detection technique, the algorithms are self-sufficient to detect human skin from an image and video. Fig. 1 shows a complete diagrammatic depiction of the proposed methodology. When a video is provided as the input the algorithms work by first reading the video and then the framing is done, still frames are selected and the methodology is implemented the procedure is same for an image, images are already static. The detection of a certain object is via cascade object detector, It divides the input into two parts skin and non-skin parts, the skin parts are highlighted as "White" while the non-skin parts as "Black". If a video is provided as input it reads the video frame by frame it selects a static frame apply the skin detection because during a video a change in body posture or face angle it will not work. Fig. 2 shows a noisy image, the algorithm implements a few steps in the noise removal from the input image or the static frame from the video. Fig. 3 shows an image after the noise removal while the Fig. 4 represents the steps done for the removal of noise from the input image. Fig. 5 shows the next step that the face from the input image or video is searched via a function known as the [vision Histogram Based Tracker] the function acts as an object for the skin detection after that it detect the facial constraints like the eyes, nose, skin tone, skin texture this allows it to uniquely identify a human with respect to the non-skin background, but if talk about the nose, it initiates with more precise skin detection as the nose is a skin part with respective skin pixels in background leads to increase in the accuracy of face detection. Fig. 6 briefs about the application of equalization for the NOWGIE algorithm this procedure is also followed by the histogram plotting that is shown in the Fig. 7 that a histogram will be plotted with respect to the input image and the resultant image of skin detection. Fig. 8 shows the original image. Fig. 9 explains the acquisition of skin tone by the HSV that acts as a membership function for the input video and image, the algorithm detects the skin and the HSV values of the input image. Further the next step is the light compensation of the image shown in Fig. 10. Fig. 11 depicts the skin detection via NOGIE, the skin-parts in the input are represented as "White" while the non-skin parts are represented as "Black" while Fig. 12 depicts the skin detection via NOWGIE.

A. Algorithms

The algorithms of proposed methods are as following:

1) *NOGIE*

a) Image/ Video acquisition

b) Noise removal

c) Object detection.

d) HSV

e) Skin detection

2) *NOWGIE*

a) Image/ Video acquisition

b) Noise removal

c) Object detection.

d) Enhancement

e) HSV

f) Skin detection Selection.

B. Flow Chart

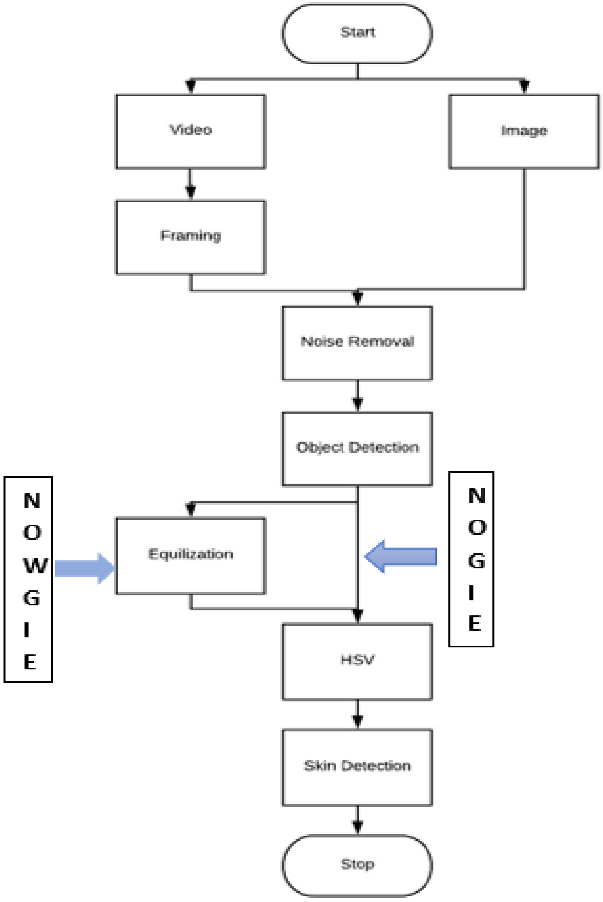


Fig. 1. Flow chart of proposed methodology.

V. Simulation and Results

For simulation and results MATLAB R2017a is used. The simulation results are as following:

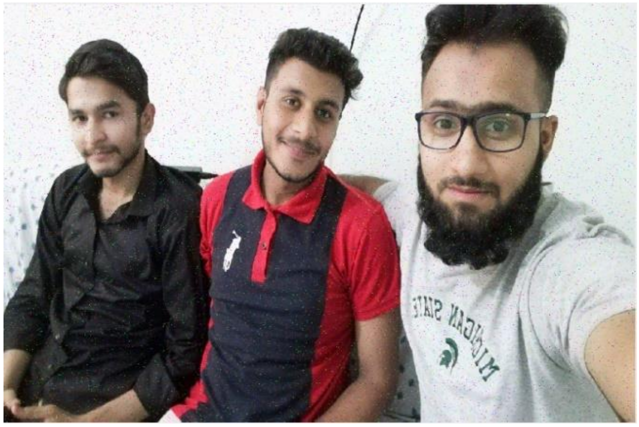


Fig. 2. Noisy image.

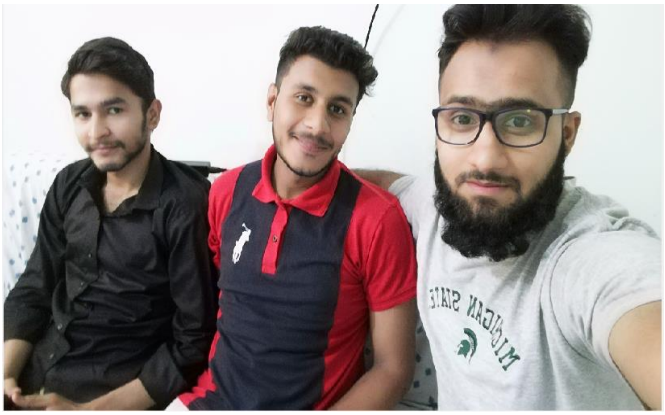


Fig. 3. Image after noise removal.

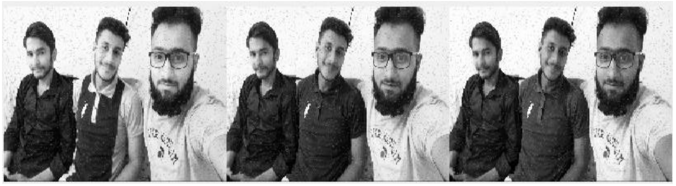


Fig. 4. Steps of noise removal.

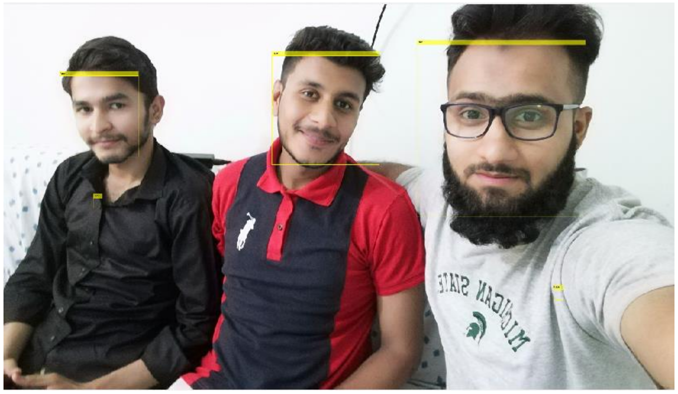


Fig. 5. Face detection.

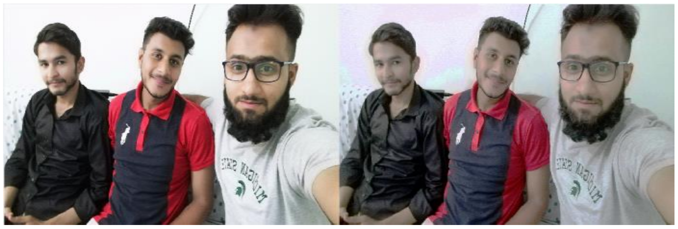


Fig. 6. Original image vs Equalized image.

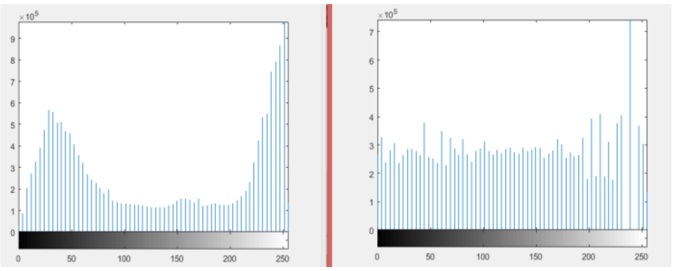


Fig. 7. Histogram plotting on original image vs Equalized image.

Fig. 8. Original image.

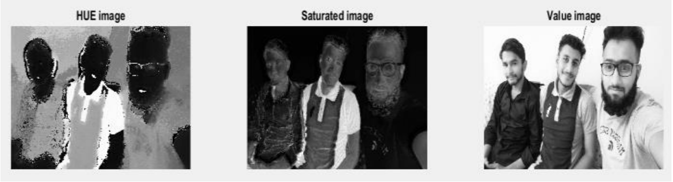
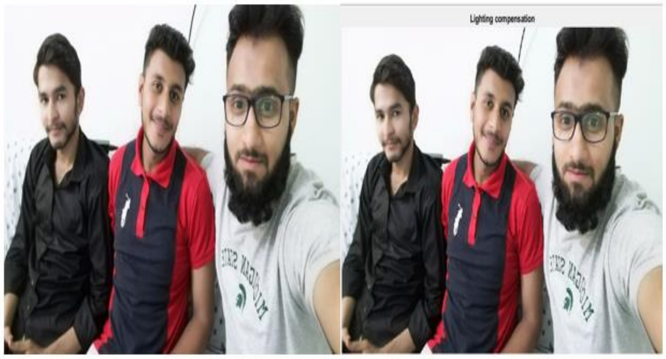


Fig. 9. Hue saturation value (HSV) of original image.



Original image Light compensation image

Fig. 10. Light compensation from original image.

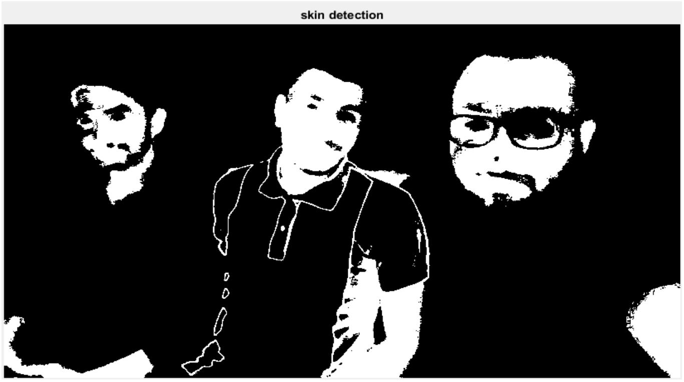


Fig. 11. Skin detection with NOGIE.

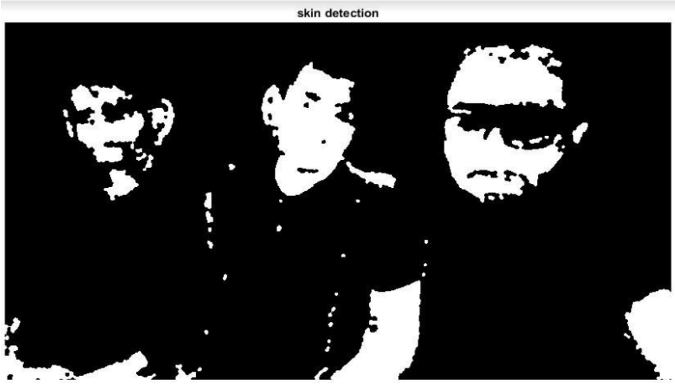


Fig. 12. Skin detection with NOWGIE.

VI. Conclusion

In the digital image processing field, facial recognition or finger print recognition are a part of the skin detection topic. Although there are multiple skin detection techniques applied but the proposed methodology in this article have provided very satisfying results, the methodology implies on the HSV that gives more better output in than RGB or YCbCr. The proposed methodology comprises of the object detection technique which enables it to only detect human skin from the input comprising of the existence of any material. The mentioned algorithms NOGIE and NOWGIE, both of these algorithms are applied on the same input providing different results but they are compared for the better understanding of the methodology. NOWGIE provides more better results in comparison to NOGIE. Due to special image enhancement by the equalization technique is applied on the input in NOWGIE that is not applied in NOGIE but in both algorithms noise is removed that gives improved skin detection this enhanced version of the previous methodology gives far more better and improved results as this methodology comprises of the object detection techniques by which it first detects the human skin that acts as an object apart from any other skin. The skin pixel ratio of this proposed method and previously proposed method are mentioned in Table I. In the with GIE and without GIE techniques the skin pixel values are different for old and new algorithms but when compared the skin pixel percentage in with GIE method is less than the without GIE due to the accuracy of detecting only skin pixels while without GIE is less accurate as it is detecting a bit of the non-skin pixels as well.

TABLE I. ASPECT OF SKIN PIXEL RATIO IN NEW AND OLD METHODOLOGY

Srno	**Algorithm**	**With GIE**			**Without GIE**		
		Total pixels	**Skin pixels**	**Skin pixel percentage**	**Total pixels**	**Skin pixels**	**Skin pixel percentage**
1	old	6005760	1223628	20.37%	6005760	1714242	28.54%
2	new	6005760	858421	14.29%	6005760	994445	16.55%

VII. DISCUSSION AND FUTURE WORK

Skin detection is a difficult task to perform in the area of digital image processing due to the difference in the skin tone of humans from different regions around the world. Although the human skin tone can also resemble to other things too. With the successful and satisfying results of the with GIE technique presented in this methodology. In future this method can be established for the identification of various skin diseases mainly skin cancer.

ACKNOWLEDGMENT

We will like to acknowledge our respected teachers Miss Arfa Hassan and Mr. Adnan Khan who believed in us, gave us the opportunity and played a major role in our supervision, we thank them and pay gratitude for their unconditional support and efforts it would not had had been possible without both of you.

REFERENCES

[1] Bush, Idoko John, Abiyev, Rahib, Sallam Ma\'aitah, Mohammad Khaleel, and Altiparmak, Hamit, “Integrated artificial intelligence algorithm for skin detection,” ITM Web Conf., vol. 16, p. 2004, 2018.

[2] A. A. Zaidan, H. A. Karim, N. N. Ahmad, G. M. Alam, and B. B. Zaidan, “A new hybrid module for skin detector using fuzzy inference system structure and explicit rules,” Int. J. Phys. Sci., vol. 5, no. 13, pp. 2084–2097, 2010.

[3] S. I. Shaikh, “Fusion Technique for Human Skin Detection,” vol. 4, no. 06, pp. 1083–1089, 2015.

[4] A. Bhatia, S. Srivastava, and A. Agarwal, “Face Detection Using Fuzzy Logic and Skin Color Segmentation in Images,” 2010 3rd Int. Conf. Emerg. Trends Eng. Technol., pp. 225–228, 2010.

[5] A. Hassan, U. Tariq, A. Iqbal, and M. A. Khan, “Muhammad Adnan Khan 4,” vol. 2, no. 1, pp. 85–95, 2016.

[6] E. Elbaşi, “Fuzzy logic-based scenario recognition from video sequences,” J. Appl. Res. Technol., vol. 11, no. 5, pp. 702–707, 2013.

[7] P. F. Processor, “ITEE Journal,” vol. 1, no. 1, pp. 30–38, 2012.

[8] M. Rai, R. K. Yadav, and G. Sinha, “Algorithm for Human Skin Detection Using Fuzzy Logic,” pp. 1–6.

[9] B. Poon, M. A. Amin, and H. Yan, “PCA Based Human Face Recognition with Improved Methods for Distorted Images due to Illumination and Color Background,” IAENG Int. J. Comput. Sci., vol. 43, no. 3, pp. 277–283, 2016.

[10] H. El Khiyari and H. Wechsler, “Age Invariant Face Recognition Using Convolutional Neural Networks and Set Distances,” J. Inf. Secur., vol. 08, no. 03, pp. 174–185, 2017.

Relationship Strength Based Privacy for the Online Social Networks

Javed Ahmed[1], Adnan Manzoor[2], Nazar H. Phulpoto[2], Imtiaz A. Halepoto[3], Muhammad Sulleman Memon[3]
[1]Department of Computer Science, IBA Sukkur University, Pakistan
[2]Department of Information Technology, QUEST Nawabshah, Pakistan
[3]Department of Computer Systems Engineering, QUEST Nawabshah, Pakistan

***Abstract*—The trend of communication is changing from mobile messages to the online social networks, for example, Facebook. The social networking applications and websites provide many of the characteristics, such as personal photo sharing. On the positive side by that many individuals form the social relationships. However, the online social networks may lead to the misuse of personal information and its disclosure. The social networks are static and assume equal values for the individuals who are directly connected. On the other hand, in real life the social relationships are dynamic and they are based on different attributes such as location, family background, neighborhood and many more. In order to be secure from the undesirable consequences due to personal information leakage, the effective mechanisms are required. In this paper, a model is proposed for the privacy in online social networks. The proposed model restricts the disclosure of personal information to the individuals. The information of one individual may be disclosed based on the relationship strength and the context. The implementation of this model on the social networks reduces the percentage of information disclosure to the less known individuals.**

***Keywords*—Online social networks; privacy; social relationships**

I. Introduction

Day by day increasing availability of the Internet also increase number of devices that are used for communication, such as mobile phones. These devices help in arranging the online streaming and conferencing. One of the main usages of these devices is the communication through the online social networks (OSNs). The users of phone spent unprecedented time while using the OSN websites. Many of the individuals also use OSNs for the business purpose to advertise the products. However, the in- formation sharing such as location sharing on the social networks may lead to information disclosure. Many of the user leave the privacy settings on of the social networks on default. As the meeting online is very different when compared with the meeting in real life. So, it is very important to protect your data and personal information. Due to which, the security and privacy concerns are getting attention of many networking communities [1]. With features available for the privacy settings, it is mentally fine to put the information on private. However, the social friendship with individuals may leak the information the attackers [2], Liu et al. [3]. Unlike social networks the relationships in real life evolve with time. So it also raises many questions regarding the maintenance of social relationships. In OSNs there is a need for a proper mechanism to manage social relationships of individuals in a dynamic environment with diverse audiences. The main motivation for this research is to develop a model to represent user's diverse social relationships on the basis of relational strength and social context. In everyday life relationship strength and social context are crucial factors to decide what to reveal and whom to reveal. Whereas, current OSNs offer friend as the only possible bidirectional relationship, which lack diversity in the type of social relationships which users form in everyday life. The objective of this research to design a model for social relationships in online social networks which mimic real life relationship forming pattern. More specifically, this paper provides the details of modeling dynamism, asymmetry, relational strength, and contextual integrity in user relationship in OSNs. The following questions are addressed:

- How to model user's relationship in online social networks?
- How to model user's relationship strength in online social networks?
- How to model user's contextual role in online social networks?
- How to model user's interactions in online social networks?

The study on audience segregation was conducted by Leenes et al. [4], where the authors develop an experimental online social network prototype known as the Clique. The Clique is inspired from Goffmans theory of self-presentation and offers the mechanism for audience segregation. The Clique required the users to invest energy and time to perform audience segregation. The study [5] based on the partitioning a users friends has also improved the privacy concerns. Authors in [6], [7] proposed the model for grouping social friends with matching characteristics in order to improve the privacy (see also [8]). The evaluation of privacy on social websites is also in consideration of many researchers [9]-[11].

II. Privacy: Theoretical Framework

One of the simplest definitions of privacy is the individual's claim and rights to control personal information from being access by the unauthorized or public. The information privacy on the social network is control by the individual and it is expected to remain safe from the disclosure [12]. Some of the well-know sources of online data such as:

- Location based applications
- Research and collaboration tools

- Online hospitals
- Online photo sharing
- Open access User profiles

One of the famous frameworks proposed for privacy of information is proposed in [13]. Where the focus was on the elements such as data integrity and privacy. The paper extends the idea over the social websites. Many of the researchers suggest that the information privacy and contextual integrity are related with each other [14], [15]. The authors in [16], [17] suggest the quality of relationship over the social networks plays a vital role in the minimization of information disclosure. In this paper, the proposed theoretical framework merges these social theories to address the multidimensional issue of privacy in social web. In following subsections, we illustrate the deficiency of existing privacy controls with help of problem scenarios that can be motivating factor to adapt our theoretical framework.

A. Contextual Integrity

Lets consider a simple example to understand the contextual integrity. Bob has several friends from his social life and he is also connected with his employer. Bob attain a social gathering near his city. Bob wishes to share the photos of party with his friends but not his office colleagues and employer. Currently all profile information of Bob is available to all his friends equally (by default). Among his friends, Bob also wishes to disclose photos to a limited number of friends depending on relationship context to avoid any embarrassing situation caused by revealing personal information to unintended audiences. One can argue that Bob can manage relationship context by creating lists and circles, whereas managing the appropriateness of these lists and circles is sole responsibility of the user. We know that social relationships are dynamic so maintaining the appropriateness of these static lists and circles is quite difficult and nearly impossible.

B. Disclosure Minimization

In reality social relationships are dynamic, and asymmetric in nature. Let us discuss a scenario to understand relationship dynamism. Alice started friendship with Bob almost five year ago. Alice has a new friend Eve on Facebook. With the passage of time Alice and Bob became the best friends. From the story is it observed that Alices relationship strength with Bob is strong, whereas relationship strength between Alice and Eve is weak. In future it is possible that Alice and Eve become the best friends. Moreover, it is also possible that the friendship between and Alice and Bob may break. Due to which, Alices relationship strength with Bob changes from strong to weak, and with Eve weak to strong. Consider another scenario to illustrate asymmetry in relationships. Bob is friend of his Boss on social networking site. Bob likes and comments positively on each post of his Boss. His Boss never commented or liked his status updates. It might be a mistake to consider Bob as close friend to his Boss. As interaction involve time and effort from participants. Bob has invested a lot of time, whereas, his Boss has invested no time. Boss has high influence on Bob, but Bob has no influence on his Boss. Influence is often asymmetric.

C. User Control

There are several occasions where the privacy of one individual be affected by the others, for example liking a post on Facebook. Photo Tagging is very common example of this phenomenon. The user controls are helpful in the situation. For example, on Facebook there is a control, which prevents others to tag you in a post or photo. The more advanced feature seeks permission from the tagged person before the use of tag. More examples of user controls are the setting of who can see the information you post.

III. Privacy Preserving Relationship Modeling

Our model addresses the issues of context collapse, maximum personal information disclosure, and lack of user control from sociological perspective. The proposed model is a modified version tie strength, contextual integrity, interpersonal boundary management and presentation of self. Such theories contains guidelines for individuals to control their personal information disclosure in face-to-face conservations. The domain of online social networks can also benefits from these foundational concepts of sociology. In following sections, we discuss building block of the model along with its formalism. The detailed description of social theories and their relationship with our model is avoided due to space limitations.

A. Preliminaries of the Model

OSNs are expressed by the number of users, relationship network, data collection and the user activity stream. Multiple criteria for classification of these OSNs entities is used. We benefit from research literature in privacy domain to identify these criteria [18]-[20]. Some of the factors used for classification are tie strength, information sensitivity, interaction intensity and user attitude towards privacy in online social networks.

1) Types of OSN Users: The users can be categorized depending on their behaviour and attitude towards privacy in online social networks. The attitude and the behavior are the key elements towards the information privacy. The privacy risks of each user can be determined by his usual behaviour and attitudes on OSNs. Following are the different types of OSN users [19]:

1. Socializers: The users join OSNs in order to make new friends just for the sake of entertainment. These users have large friend network but most of them are casual friends. The privacy policy suggested for these users is soft privacy.

2. Attention-Seeker: The users join OSNs to present themselves to the world. The users have extensive friend network, but they keep in active conversation with a limited number of friends. Generally, the privacy for these users is soft privacy.

3. Followers: The users join OSNs to keep up with what their peers are doing. The users have medium friend network. The privacy policy suggested for these users is hard privacy.

4. Faithful: The users join OSNs to rekindle old friendships. The users have medium friend network, most of their friends are from school or university. The privacy policy suggested for them is soft privacy.

5. Functionals: The users join OSNs for doing political campaigning, or charity work. The users have large friend network, most of their friends are of casual nature. The privacy policy suggested for them is hard privacy.

2) Types of social contexts: The relationship network of OSNs users is diverse in nature and users play several roles across different social contexts. Ozenc et al. [21] identified that three social contexts are very common among all OSNs users and needed management of intimacy levels within these social context for better social experience in online social networks.

1. Family: This context refers to relatives and can be inferred by analyzing profile attributes such as relationship status.

2. Work: This context refers to professional circle and can be inferred by analyzing profile attributes such as present and past work affiliations.

3. Social: This context refer to friends and can be inferred by analyzing profile attributes such as educational background, interests etc.

3) Types of social interactions: Online social networks provide rich set of user interaction for communication and information sharing and interaction pattern plays vital role to determine the quality of relationships among user in various social contexts.

1. Messaging: This refers to one to one communication method. Each message has sender, receiver, and content.

2. Posting: This refers to one to many communication method. Each post is created by certain user on specific user's wall with specific content, and certain set of audience.

3. Commenting: This is kind of post which is contribution in response to existing topic of discussion.

4. Tagging: This refers to sharing content with stakeholders.

5. Liking: This refers to contribution to existing post.

6. Chatting: This is kind of messaging which include session.

7. Wishing: This is kind of post that may include: creator, wall, data, and audience.

4) Grouping of user data: Since OSNs user share vast variety of multimedia content in their profile pages. Different data items may have different level of information sensitivity. Ho et al. [18] group user data into following categories depending on the sensitivity of information. This categorization can be useful in deciding privacy policy for OSN users.

1. Healthy: These users share data that is not harmful to anyone in terms of privacy.

2. Harmless: It is also like healthy data, which is used by marketing companies for business purpose.

3. Harmful: The disclosure of harmful data to inappropriate audience can create security and privacy risk.

4. Poisonous: The disclosure of poisonous data to audience other than strong ties can create security and privacy risk. This data contains information that can be help to track user or extract his financial information.

Ho et al. [18] also categorize shared data of OSNs users into five groups. All the data shared on OSNs falls into one of these groups. This grouping deals with nature of information contained in the data.

1. Identity: The data such as name or phone number, which is enough to identify a person.

2. Demographic: The data that contains the details such as gender, age, height, etc.

3. Relationships: The data refers to the relationship information of OSN users such as added friends, etc.

4. Activity: The data that shows the activities of a user.

5. Multimedia-content: The data refers multimedia, for example videos shared by the user.

Hu et al. [8] identified four different types of user privileges over data that can be important while assigning privacy policy:

i) Owner: The user is called owner of the data if it is contained in space of the user.

ii) Contributor: The user is called contributor of the data if it is commented or liked by the user.

iii) Stakeholder: The user is called stakeholder of the data if it tags the user.

iv) Disseminator: The user is called disseminator of the data if it is shared by the user.

5) Social relationship based privacy levels: The four privacy levels are suggested on the basis of relationships strength, social context and type of the users:

1. No-Privacy: This privacy policy is very liberal in nature. It allows everyone to access all type of user data.

2. Soft-Privacy: This privacy policy restricts access to poisonous data only to audience with strong ties, whereas healthy and harmless data is accessible to everyone. This policy is suitable for socializers, attention seekers, and faithfuls.

3. Hard-Privacy: This privacy policy allows everyone to access healthy data, whereas access to other types of data is restricted. This policy is suitable for followers and faithful users.

4. Full-Privacy: This privacy policy is very conservative in nature.

Table I represents various entities described in this section and highlights the their influence on each other. We describe privacy preserving social relationship model in next section using these building blocks.

B. Formalization of the Model

An OSN denoted by S is a 5-tuple and it is defined as: Users, Data, Relationships, Interactions, Policy. The description of each is given below.

TABLE I. PRIVACY POLICY FOR OSN USERS

Attributes	Socializer	Attention Seeker	Faithful	Follower	Functional
Friend Network	Large	Large	Small	Medium	Medium
Interaction Type	Photo Posting Photo Tagging Commenting Liking	Photo Posting Commenting Liking Wishing	Messege Chatting Wall Posting –	Commenting Liking Wall Posting –	Wall Posting Liking Commenting –
Relationship Strength	Weak Tie	Weak Tie	Strong Tie	Strong Tie	Weak Tie
Contextual Role	Social & Work	Social	Family & Social	Family & Social	Work & Social
Context Type	Harmful	Harmful Poisonous	Harmless	Harmless	Harmless
Privacy Policy	Soft Privacy	Soft Privacy	Hard Privacy	Hard Privacy	Soft Privacy

1) Users is the tuple: (U, Type, userType, Profile, userProfile, Policy, userPolicy, such as: U{u_1,· · · ,u_n} a finite set of OSN users identifiers. Type={Socializers, Attention-Seekers, Followers, Faithfuls, Functionals} userType= $U = 2^{Type}$ this is the case of assigns for each user at least one social category.

$Profile = \{p_1, , p_m\}$ is a finite set of profiles such that: $m \leq n$.

userProfile: $U - -2^{Profile}$ is a function that assigns for each user at least one profile.

Policy={No-Privacy, Soft-Privacy, Hard-Privacy, Full-privacy}

$userPolicy : Profile- > Policy$ is a function that assigns a privacy policy to each profile.

2) Data is the tuple: (D, Type, dataType, Sensitivity, dataSensitivit) $D = \{d_1, , d_m\}$ a finite set of data items represented by data identifier.

Type={Identity, Demographic, Relationship, Ativity, Multimedia-Content}

$dataType = D- > Type$ is a function that assigns for each data item a type.

Sensitivity={Healthy, Harmless, Harmful, Poisonous}

$dataSensitivity = D- > Sensitivity$ is a function that assigns sensitivity level to each data item.

3) Relationship is the tuple: (U,D,C,S,P,relU2U,relU2D,relD2D), where:

C={Social, Family, Work} is a set representing the relationship context.

S = attr_1: val_1,, attr_n:val_n this set represents relationship strength.

P = Owner, Stakeholder, Contributor, Disseminator represents users privilege over data items.

$relU2U = U \times U- > C \times S$ is a function to determine relationships among users.

$ErelU2D = UD- > 2^P$ is a function to determine relationship among user and data.

$relD2D = D- > 2^D$ is a function to determine relationship among different data resources.

4) Interactions is tuple (U, D, R, , Weight, History): = {Messaging,Posting,Commenting, Tagging,Liking,Chatting,Wishing} is set of actions

$Weight : - > [0, 1]$

$History : U- > 2$

5) Policy: It is an propositional logic formula over the set of parameterized actions.

An OSN is formalized using above mathematical representation that facilitates the system component description and manipulation. We describe formally all what is earlier mentioned in the previous section. The users are described as entities with type, profiles and their associated policies. In our formalism we represent all kind of relationship between the OSN entities and we annotate them with a weight value that characterize the strength of the relationship. The data items are considered as objects with the sensitivity dimension. We also take into consideration all kind of actions that are needed in the interactions between users themselves as well within the existing objects. Finally we describe a policy as a constraint taking the form of a propositional logic formula where the atomic propositions are the OSN entities values.

IV. CONCLUSION AND FUTURE WORK

With the growing number o smart phones as well as the Internet access. Moreover, the trend of using the social networking websites is also increasing. Due to the many social networking websites the data of users is available to the audience. This leads to the privacy concerns and disclosure of personal information. This paper presented a model based on the social relationships on OSNs. The model adopts the well know theories and decides the privacy concerns by defining weak and strong ties. The proposed model proved to minimize the disclosure of personal information. In future, the same work could be performed by using the ontological models for high performance.

REFERENCES

[1] Maritza Johnson, Serge Egelman, and Steven M Bellovin. Facebook and privacy: its complicated. In Proceedings of the eighth symposium on usable privacy and security, page 9. ACM, 2012.

[2] Cuneyt Gurcan Akcora and Elena Ferrari. Graphical user interfaces for privacy settings. In Encyclopedia of Social Network Analysis and Mining, pages 648660. Springer, 2014.

[3] Yabing Liu, Krishna P Gummadi, Balachander Krishnamurthy, and Alan Mislove. Analyzing facebook privacy settings: user expectations vs. reality. In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, pages 6170. ACM, 2011.

[4] Ronald Leenes. Context is everything sociality and privacy in online social network sites. In Privacy and identity management for life, pages 4865. Springer, 2010.

[5] Fabeah Adu-Oppong, Casey K Gardiner, Apu Kapadia, and Patrick P Tsang. Social circles: Tackling privacy in social networks. In Symposium on Usable Privacy and Security (SOUPS), 2008.

[6] Anna Squicciarini, S Karumanchi, Dongyang Lin, and Nicole DeSisto. Automatic social group organization and privacy management. In Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2012 8th International Conference on, pages 8996. IEEE, 2012.

[7] Anna Squicciarini, Sushama Karumanchi, Dan Lin, and Nicole DeSisto. Identifying hidden social circles for advanced privacy configuration. Computers & Security, 41:4051, 2014.

[8] Hongxin Hu and Gail-Joon Ahn. Multiparty authorization framework for data sharing in online social networks. In Data and Applications Security and Privacy XXV, pages 2943. Springer, 2011.

[9] Eric Gilbert and Karrie Karahalios. Predicting tie strength with social media. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 211220. ACM, 2009.

[10] Eric Gilbert. Predicting tie strength in a new medium. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, pages 10471056. ACM, 2012.

[11] Rongjing Xiang, Jennifer Neville, and Monica Rogati. Modeling relationship strength in online social networks. In Proceedings of the 19th international conference on World wide web, pages 981990. ACM, 2010.

[12] Jerry Kang. Information privacy in cyberspace transactions. Stanford Law Review, pages 11931294, 1998.

[13] Katrin Borcea-Pfitzmann, Andreas Pfitzmann, and Manuela Berg. Privacy 3.0:= data minimization+ user control+ contextual integrity. it-Information Technology Methoden und innovative Anwendungen der Informatik und Informationstechnik, 53(1):3440, 2011.

[14] Erving Goffman. The presentation of self in everyday life [1959]. Contemporary sociological theory, pages 4661, 2012.

[15] Helen Nissenbaum. Privacy as contextual integrity. Washington law review, 79(1), 2004.

[16] Irwin Altman. The environment and social behavior: Privacy, personal space, territory, and crowding. 1975.

[17] Mark S Granovetter. The strength of weak ties. American journal of sociology, pages 13601380, 1973.

[18] Ai Ho, Abdou Maiga, and Esma A13 053'fmeur. Privacy protection issues in social networking sites. In Computer Systems and Applications, 2009. AICCSA 2009. IEEE/ACS International Conference on, pages 271278. IEEE, 2009.

[19] Ofcom. Social networking: A quantitative and qualitative research report into attitudes, behaviours and use, 2008.

[20] Peter V Marsden and Karen E Campbell. Measuring tie strength. Social forces, 63(2):482501, 1984.

[21] Fatih Kursat Ozenc and Shelly D Farnham. Life modes in social media. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 561570. ACM, 2011.

Performance Evaluation of Polynomial Pool-Based Key Pre-Distribution Protocol for Wireless Sensor Network Applications

Malek Ben Amira, Mayssa Bouraoui, Noureddine Boulajfen
Research Center for Microelectronics and Nanotechnology,
Technopole Sousse, 4050
Sousse, Tunisia

***Abstract*—In nowadays, wireless sensor network (WSN) has been established as a leading emerging technology in the field of remote area distributed sensing due to its diverse application areas. Key pre-distribution is an important task in WSN because after the deployment of sensor nodes, their neighbors become strange to each other. To secure the communication, neighbor nodes have to generate a secret shared key, or a key-path must exist between these nodes. In this paper, we have discussed and presented various key pre-distribution protocols, namely, the polynomial pool-based key pre-distribution which is a scheme for creating pairwise keys between sensors on the foundation of a polynomial-based key pre-distribution protocol, introducing two effective instantiations: a *random subset assignment* key pre-distribution scheme and a *grid-based* key pre-distribution scheme. Other studied key pre-distribution schemes (KPDS) are Peer Intermediaries Key for Establishment (PIKE) and Group-based key pre-distribution scheme. The performances of these schemes have been assessed through the simulation of different grids under the TinyOS environment.**

Keywords—Key management; wireless sensor network (WSN); key pre-distribution schemes; polynomial pool-based KPDS; PIKE; group-based KPDS

I. Introduction

Wireless Sensor Network is a group of several resource-constrained sensor nodes that can be accessed via a wireless medium. These sensor nodes are favored because they are low-priced, self-organized and simple to deploy. WSNs are used in military applications, such as military surveillance and battlefield supervision, and civilian ones such as medical monitoring, smart agriculture, etc. [1]. The security of WSNs is a very important aspect which has been actively studied by researchers. Different applications need WSNs to exchange delicate information that necessitate a high level of security to succeed. Yet, strong security is difficult to achieve with limited resources of sensor nodes.

Key management is the element key for security in WSNs because it is the foundation of various security services, like encryption and authentication. The principal goal of key management scheme is to provide secure communication between sensors in the network [2]. But, the critical assignment of key management is the establishment of a pairwise key between two nodes in the network. Different researchers proposed many protocols, such as Polynomial Pool-Based Key Pre-Distribution scheme which has two efficient instantiations: a Random Subset Assignment KPDS and a Grid-based KPDS, Peer Intermediaries Key for Establishment, Group-based KPDS, etc.

In these schemes, the sensors' deployment, which can be randomly or uniformly, can improve the key pre-distribution [3], [4]. So, this paper presents and compares the performances of these different schemes in terms of packet loss rate and energy consumption.

The rest of this article is arranged into six sections. Section 2 presents an overview of the different polynomial-based key pre-distribution techniques and Section 3 introduces the general framework of the polynomial pool-based key pre-distribution and a description of the two instantiations. In Section 4, other key pre-distribution techniques are reviewed. The simulation results are introduced in Section 5. Finally, Section 6 concludes this paper.

II. Polynomial-Based Key Pre-Distribution Schemes

Polynomial-based key pre-distribution protocol [5] is the basis of new techniques such as Polynomial pool-based key pre-distribution. This protocol was created for group key pre-distribution.

The security tolerance of the scheme is decided by the size of security threshold to a great extent [6]. However, once the number of compromised nodes is bigger than the security threshold, the network security performance would be rapidly declined. Besides, to improve the resilience against node capture, the scheme is implemented at the expense of network connectivity [7], [8].

III. Polynomial Pool-Based Key Pre-Distribution

A general framework for key pre-distribution based on the scheme was developed to secure the key establishment techniques. It is called *polynomial pool-based key pre-distribution* [11] due to the use of a pool of many random bivariate polynomials.

The main concept of the polynomial pool-based key pre-distribution can be considered as the combination of the polynomial-based KPDS and the key pool idea consumed in [9] and [10].

Liu and Ning [11] created a general framework for polynomial pool-based pairwise key pre-distribution in wireless sensor networks and two possible instantiations for key pre-distribution schemes, namely Random Subset Assignment KPDS and Grid-based KPDS [12].

A. *Random Subset Assignment KPDS*

We introduce in this section, the first possible instantiation of the common framework by employing a random plan for the subset assignment in the set-up phase.

This scheme can be taken as a prolongation to the fundamental probabilistic scheme introduced in [10]. The primary distinctness of this scheme from the basic probabilistic scheme is that it randomly picks polynomials and attributes their polynomial shares to each sensor instead of randomly choosing keys from a big key pool and attributing them to sensors. For that reason, a random subset can be designed as an extension to the fundamental probabilistic scheme [13]. This scheme also differs in the sense that it uses a distinct key for each link [14].

B. *Grid-Based KPDS*

Another instantiation of the general framework introduced in this section is called grid-based KPDS [15]. This scheme has many interesting properties. First of all, it ensures that even when there are no compromised nodes, any two sensors can create a pairwise key between them and the sensor nodes can report to each other. Second, grid-based KPDS is resilient to node compromise. Even if some sensors are captured, there is still a great chance for the key establishment between the uncompromised nodes using this approach. Third, with grid-based KPDS, a sensor node can define whether or not it can create a pairwise key with another node, and if so, which polynomial should be utilized. As a result, there isn't a communication overhead over the polynomial share discovery.

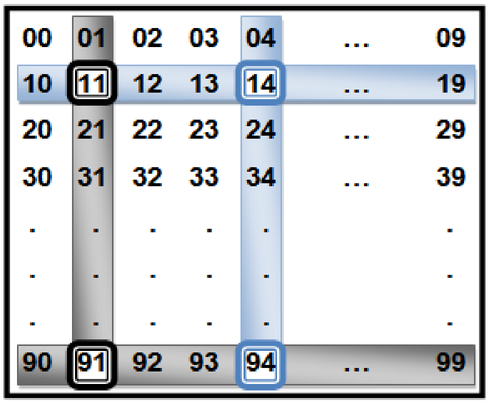
IV. Other Key Pre-Distibution Techniques

Besides the polynomial pool-based key pre-distribution scheme, various key distribution techniques are implemented, such as Peer Intermediaries Key for Establishment and Group-based key pre-distribution scheme. In this section, we present these schemes so that we can compare them with the Polynomial pool-based KPDS.

A. *PIKE*

Chan and Perrig [13] proposed a method called Peer Intermediaries Key for Establishment (PIKE) and dedicated to the key establishment. The basic idea behind this scheme is employing peer sensor nodes like trusted intermediaries. PIKE was created to overcome the absence of scalability of the existing symmetric key distribution schemes. Each node shares another (unique) pairwise key with each of the other nodes (O $\sqrt{n}$) in the network.

Each node in Fig. 1 will be loaded with 18 keys (9 keys for the nodes belonging to its line and 9 keys for the nodes belonging to its column). In general, each node stores 2($\sqrt{n} - 1$) keys and the whole number of unique keys generated is n ($\sqrt{n}$ −1).



00	01	02	03	04	...	09
10	11	12	13	14	...	19
20	21	22	23	24	...	29
30	31	32	33	34	...	39
.	.	.	.	.		.
.	.	.	.	.		.
.	.	.	.	.		.
90	91	92	93	94	...	99

Fig. 1. Virtual space of node identifiers of a network of 100 nodes [16].

B. *Group-Based KPDS*

In Group-based KPDS, sensors are distributed and organized only in groups [17]. The deployment knowledge utilized to increase the performance of key pre-distribution revolves around the observation that the sensor nodes in the same group are distributed close to each other. This assumption is usually true since the sensor nodes in the same group are assumed to be displayed at the same time from the same point. Once the sensor nodes are displayed in the field, they become static.

Based on this deployment model, the sensor nodes in the same deployment group have an important probability of being neighbors. Group-based KPDS uses two methods, *in-group key pre-distribution* and *cross-group key pre-distribution*.

A sensor node can, without difficulty, assess which displayed group or cross-group other sensor nodes appertain to based on their ID as showed in Fig. 2.

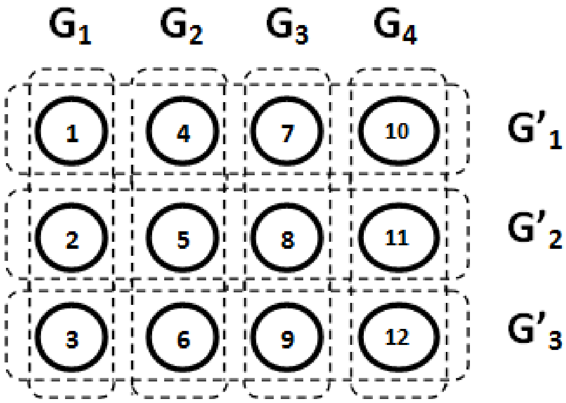


Fig. 2. An example of group construction [17].

V. Simulation and Results

In this section, we assess the performances of Random Subset Assignment KPDS, Grid-based KPDS, PIKE and Group-based KPDS using the TinyOS simulator.

For this purpose, we have simulated random and grid WSNs with 9, 25, 49, 81 and 100 nodes distributed over the field and certain metrics, such as the time of communication between nodes, Packet loss and energy consumption have been measured and compared. The goal behind simulations is to find out the perfect scheme which minimizes the packet loss and energy consumption for each network, and which scheme provides the best probability of establishing a direct and indirect key.

C. Time for Communication between Nodes

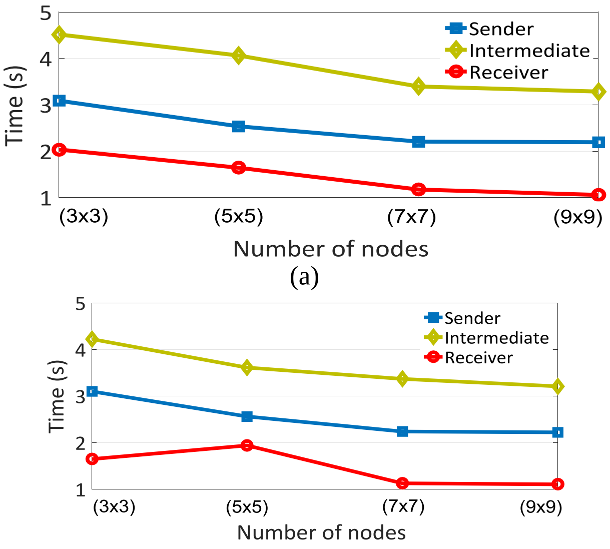
Random Subset Assignment KPDS has a random topology which generates a direct communication between two nodes and consequently a direct pairwise key establishment without using an intermediary node. And for Grid-based KPDS, any sensor node can create a direct pairwise key among two nodes.

So in this section, we have studied the time for communication between any two nodes in the network to find which node (sender, intermediary or receiver) and scheme consumes more time in the communication in only PIKE and Group-based KPDS.

Random Subset Assignment KPDS has a random topology which generates a direct communication between two nodes and consequently a direct pairwise key establishment without using an intermediary node. And for Grid-basedKPDS, any sensor node can create a direct pairwise key between two nodes.

So in this section, we have studied the time for communication between any two nodes in the network to find which node (sender, intermediary or receiver) and scheme consumes more time in the communication in only PIKE and Group-based KPDS.

After several simulations of PIKE and Group-based KPDS, it was noted from Fig. 3 that the time to establish a session for the intermediate node is superior to the Sender and Receiver nodes because it needs more time to communicate with them.



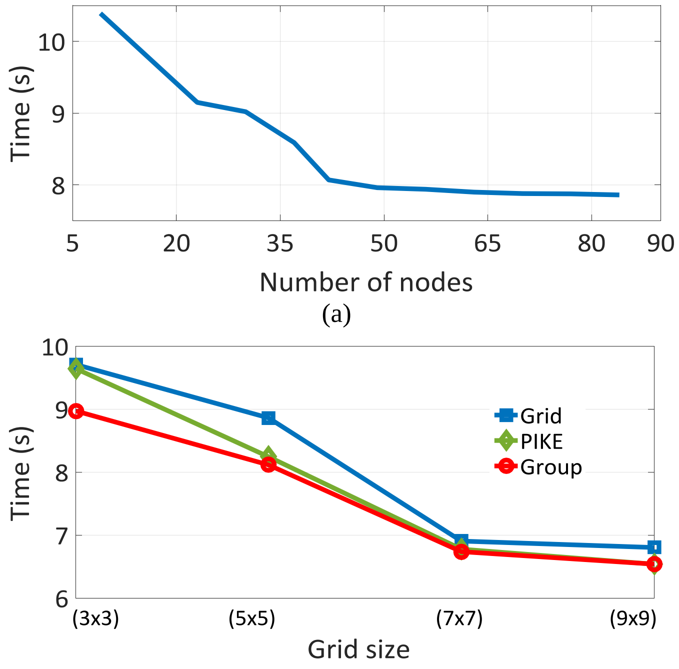
(a)

(b)

Fig. 3. Time for establishing a session between different nodes for (a) PIKE, (b) Group-based KPDS.

Because once intermediary node is chosen, sender node encrypts the new key to be shared with the receiver node using the key it shares with the intermediary and then sends it to intermediary node. Intermediary node decrypts the key and re-encrypts it using the key it shares with the receiver node, and sends it to receiver node.

After several simulations of the different techniques from 9 to 81 nodes during a fixed-time simulation and in the same area, it was noted from Fig. 4 that the time required for establishing a session is the highest for a network with 9 nodes compared to bigger larger networks because the nodes in each scheme are deployed in the same area, so when the network size increases, the time for establishing a session decreases since the distance between the nodes also decreases.



(a)

(b)

Fig. 4. Time for establishing a session for (a) Random subset assignment KPDS, (b) Grid-based KPDS, PIKE and group-based KPDS.

Among the different schemes, the Random Subset Assignment one consumes more time than others. In this scheme, the sender node requires an intermediate node to send its message to the receiver node, hence going through several intermediate nodes to find the suitable one to share a key with. So, this scheme requires more time than Grid, PIKE and Group-based ones.

D. Packet Loss

We have studied the packet loss caused by the different type of nodes (sender, intermediate and receivers nodes) for the four schemes.

Fig. 5 shows the average packet loss rates for the three groups of nodes for Random Subset Assignment KPDS, Grid-based KPDS, PIKE and Group-based KPDS. We have noticed that the average packet loss reaches the highest level in Random Subset Assignment and the lowest level in Group-based.

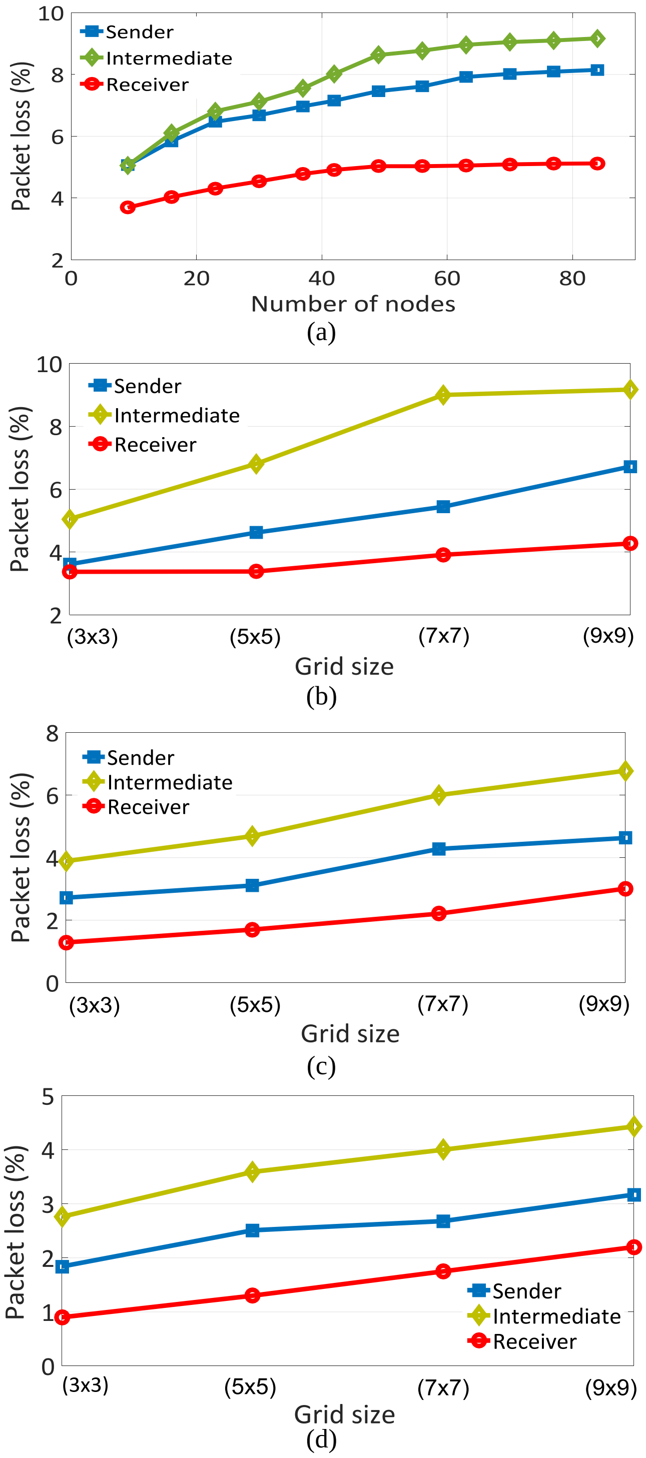


Fig. 5. Average packet loss rate for the “Sender”, “Intermediate” and “RECEIVER” NODes for (a) Random subset assignment KPDS, (b) Grid-based KPDS, (c) PIKE and (d) Group-based KPDS.

We have also noticed that when the size of the network increases, the average packet loss rate increases as well.

In Random Subset Assignment KPDS, the overhead storage is low. In addition, sensors can be added without communicating with nodes already deployed in the network. Given some storage constraints and the necessary probability of sharing the direct keys between sensor nodes, the Random Subset Assignment KPDS can allow a limited number of compromised sensor nodes while polynomials pre-distribution scheme can allow a big fraction of compromised nodes. However, due to the affectation of the nodes in a specific order, Grid-based KPDS allows a perfect distribution of nodes so that the sensor nodes can create direct keys which are adjacent to each other. Thus, it can considerably reduce the overhead communication of the key path establishment, which leads to a better packet loss than that of the Random Subset Assignment KPDS. On the other hand, PIKE is a key-establishment protocol that implicates employing one or many sensor nodes as a trusted intermediary to expedite key establishment. Unlike the other protocols, memory overheads and the communication of this protocol help to achieve a higher security against the compromised node and a restricted probability of packet loss compared to other protocols. As for Group-based KPDS, the deployment model is more realistic than the other models, such as Random Subset Assignment KPDS and Grid-based KPDS, because it requires less effort in the deployment of sensor nodes, while providing an opportunity to improve key pre-distribution and a better probability of packet loss.

From Fig. 5 we can note that the intermediate nodes exhibit the highest packet loss rate compared to the sender and receiver nodes for the 4 studied schemes (Random Subset Assignment, Grid-based KPDS, PIKE and Group-based KPDS).

E. Energy Consumption

Fig. 6 reveals that when the size of the network increases, the average energy consumption increases as well. However, the average energy consumption in Random Subset Assignment KPDS is superior to that of the other schemes. For Grid-based KPDS, PIKE and Group-based KPDS, there is no big difference in the average energy consumption, but Group-based KPDS achieves the lowest values.

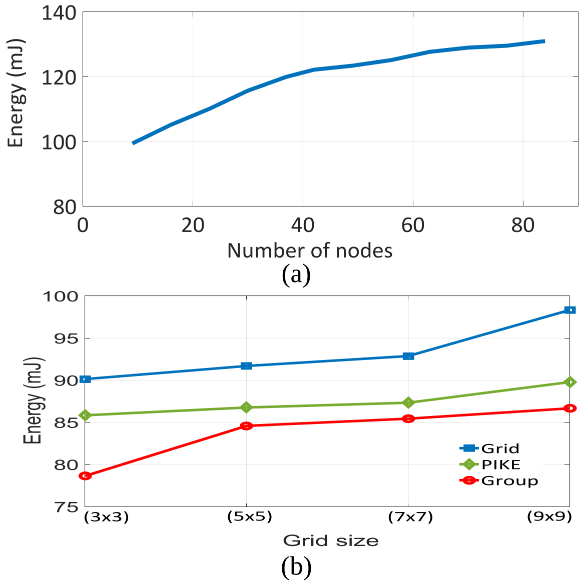


Fig. 6. Average energy consumed by the nodes of the four schemes for (a) Random subset assignment KPDS, (b) Grid-based KPDS, PIKE and group-based KPDS.

Although Grid-based KPDS can guarantee the highest security level, it has certain constraints in terms of the maximum network. Compared to other methods, PIKE minimizes the storage key in the memory nodes before deployment. However, the exchanges of key establishment messages consume time and energy. In PIKE, network nodes are proposed to be used as trusted intermediaries instead of the base station in order to relax nodes close to the base station. However, this solution could be a disadvantage, i.e. if the trusted intermediary nodes *A* and *B* are captured, *A* will no longer share a key with *B*. PIKE has a lower level of memory storage than Random Subset Assignment KPDS, while requiring a communication overhead. This scheme presents many interesting trade-offs in terms of memory and energy overhead compared with the trade-offs available by the other schemes. However, even though the probability of a secure communication between the neighbor "cross-group" is low, Group-based scheme presents a high connectivity and the deployment of sensor nodes is easy and effortless. This scheme presents a strong resilience against attacking nodes, helps to improve key pre-distribution and introduces better energy consumption than the other protocols.

F. Probability of Establishing a Direct and Indirect Key between two Nodes

Fig. 7 shows the results of comparing the four protocols in terms of the direct key establishment. It is noted that when the network size increases, the probability of establishing a direct key between two nodes decreases.

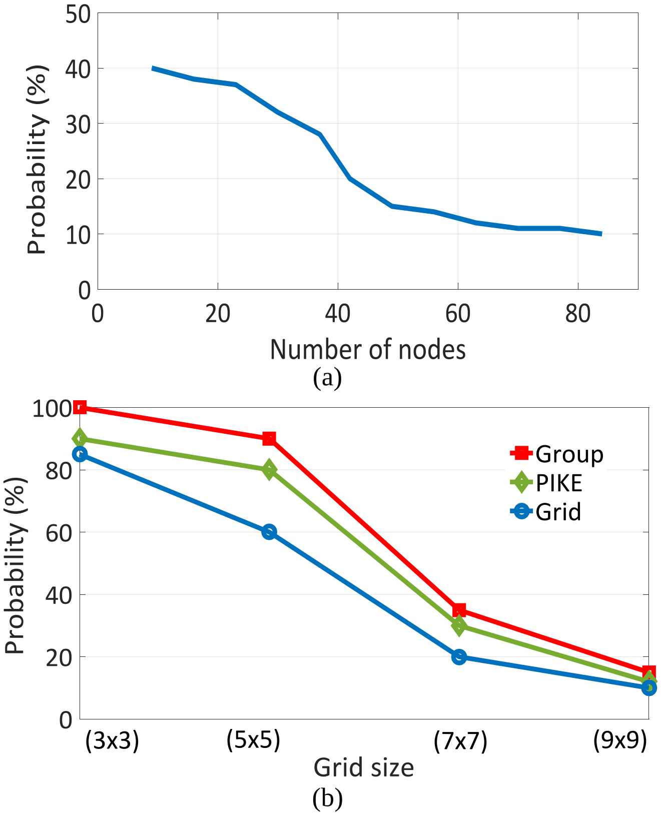


Fig. 7. Probability of establishing a direct key between two nodes for (a) Random subset assignment KPDS, (b) Group-based KPDS, PIKE and grid-based KPDS.

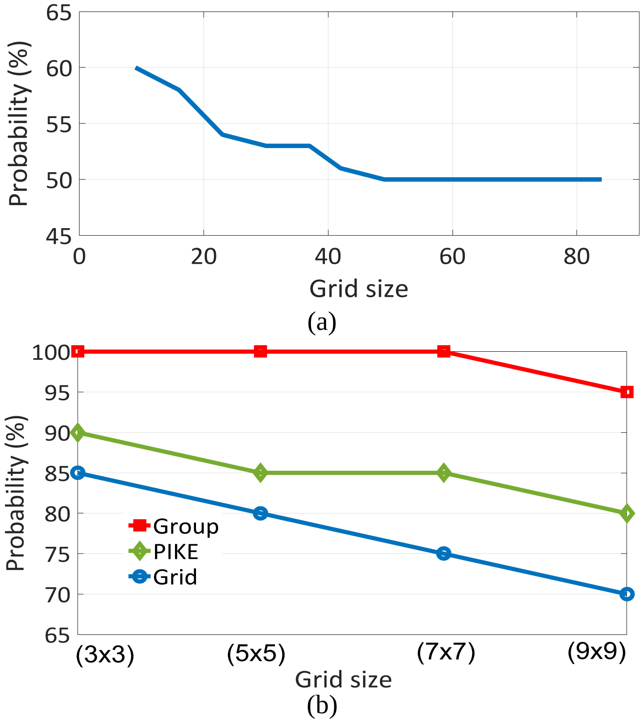


Fig. 8. Probability of establishing an indirect key between two nodes for (a) Random subset assignment KPDS, (b) Grid-based KPDS, PIKE and group-based KPDS.

We can see that Group-based KPDS has certainly a greater probability of establishing a direct key between two sensors than Random Subset Assignment KPDS, PIKE and Grid-based KPDS. This displays that Group-based KPDS can handle an important number of sensor networks with the same network settings.

Although Random Subset Assignment KPDS can be configured to obtain a perfect security, it can support only a restricted number of sensor nodes to guarantee a certain probability of having direct keys between sensor nodes. But, Group-based KPDS can reach a much higher probability to establish direct keys between neighboring nodes than Grid-based KPDS [18]. So, the performance of this scheme is better than that of Grid-based KPDS.

Fig. 8 compares the probability of creating an indirect key between the sensor nodes between the different protocols.

In this part, we consider the probability of an indirect key between two sensor nodes when they cannot create a direct key. We can clearly see from Fig. 8 that the probability of the Group-based KPDS outperforms all other protocols. In other words, as long as the sensor nodes are deployed in groups, Group-based KPDS can be used to obtain high-performance key pre-distribution schemes for sensor networks.

Grid-based KPDS has unique properties which the other schemes do not have. First of all, it is ensured that any two nodes could create a pairwise key either direct or indirect communication and without using an intermediate node when the sensor nodes can be transmitted to each other and in the absence of compromised sensor nodes in the network. In addition to the efficiency in the determination of the key path, the transmission cost is inferior to that of other systems. In the second place, even if there are compromised nodes in the network, there is an important probability that two non-compromised nodes can restore a pairwise key. For PIKE, this protocol has a uniform communication model for the key establishment, which is difficult to be disturbed by an attacker. Contrary to the current popular mechanisms such as random-key pre-distribution, PIKE has the benefit of a non-probabilistic key establishment, thus whatever two nodes are ensured to establish a key. Also, the probability of having a direct key between two adjacent sensor nodes in the Group-based KPDS is much bigger than that in the Random Subset Assignment KPDS and Grid-based KPDS.

Group-based KPDS has a better security performance than Random subset assignment KPDS at the level of both compromised direct key and compromised indirect key between nodes deployed in the same group of the network.

A comparative study of the various key pre-distribution schemes presented in Table I [19], this comparative study considering the type, scalability, computational overhead, communication overhead, storage load, resilience to node capture and security as the parameters.

TABLE I. COMPARASON BETWEEN THE FOUR SCHEMES

	Random Subset Assignment KPDS	**Grid-Based KPDS**	**PIKE**	**Group-based KPDS**
Type	Prob.	Prob.	Det.	Prob.
Scalability	Good	Good	Not Scalable	Good
Computational overhead	Low	Low	Low	Low
Communication overhead	Low	Low	Low	Low
Storage load	Low	Low	Low	Good
Network Resiliency	Maximal	Maximal	Maximal	Maximal
Nodes Compromised	Yes	Yes	Yes	Yes
Security	Normal	Good	Normal	High

VI. CONCLUSION

In this paper, we presented the polynomial pool-based pairwise key pre-distribution in sensor networks, which is based on the basic polynomial-based key pre-distribution and its two instantiations (key pre distribution scheme based on Random Subset Assignment KPDS and the grid-based KPDS). We have also introduced Peer Intermediaries Key for Establishment and Group-based key pre-distribution scheme.

By simulating these schemes using TinyOS simulator for random and grid networks, we showed that Group-based KPDS is more efficient than the other schemes because it has achieved the lowest values in terms of energy consumption, packet loss rate and time for communication between nodes. Also, Group-based KPDS provides a much higher probability in terms of establishing direct and indirect keys.

Future work introduces the new version of Group-based KPDS, which will be evaluated and compared to PIKE and the standard Group-based KPDS.

REFERENCES

[1] S. Muhammad K. Raazi and S. Lee, “A Survey on Key Management Strategies for Different Applications of Wireless Sensor Networks”, Journal of Computing Science and Engineering, Vol. 4, No. 1, Pages 23-51, March 2010.

[2] S. Akhbarfar and A.M. Rahmani, “A Survey on key pre distribution Schemes for security in Wireless Sensor Networks”, International Journal of Computer Networks and Communications Security, Vol. 2, No. 12, 423–442, December 2014.

[3] S. Sibi and A. R Thamizarasi “Key Pre-Distribution Methods of Wireless Sensor Networks” International journal of Scientific & Engineering Research, Vol 4 ,2013.

[4] M. Javanbakht, H. Erfani, H.H. S.Javadi and P.Daneshjoo, “Key Predistribution Scheme for Clustered Hierarchical Wireless Sensor Networks based on Combinatorial Design”, Published online in Wiley Online Library, Vol. 7, No 11, pp 2003–2014, 2014.

[5] C. Blundo, A. Santis, A. Herzberg, S. Kuten, U. Vaccaro and M. Yung, “Perfectly secure key distribution for dynamic conference”. Advances in Cryptology - CRYPTO’92, Lecture notes in Computer Science, Vol. 740, 471-486, Springer-Verlog, New York, 1992.

[6] S. Akhbarifar and A. M. Rahmani, “A Survey on key pre-distribution Schemes for security in Wireless Sensor Networks”. International Journal of Computer Networks and Communications Security, Vol. 2, No. 12, 423–442, Decembre 2014.

[7] A. A. Magar, “A Survey about Key Pre-distribution Scheme in Wireless Sensor Networks”, International Journal of Engineering Research and General Science, Vol. 2, Issue 6, October-November 2014.

[8] S. A. Zade and D. G. Harkut, “Key PreDistribution Model for Wireless Sensor Network”, International Journal of Application or Innovation in Engineering & Management (IJAIEM), Vol 4, Issue 5, May 2015.

[9] H. Chan, A. Perig and D. Song, “Random key pre distribution scheme for sensor networks”, IEEE Symposium Research in Security and Privacy, 2003.

[10] L. Zhu, Z. Zhang, J. Li and R. Zhou, “An Improved Random Key Predistribution Scheme for Wireless Sensor Networks Using Deployment Knowledge”. International Journal of Security and Its Applications Vol. 10, No. 5, pp.225-234, 2016.

[11] L.Mathew, J. K. John, T. Thomas, M. Karthik, “Three Tier Security Scheme in Wireless Sensors Networks with Mobile Sinks Using Grid”, International Journals of Advanced Research in Computers and Communication Engineering. Vol. 2, Issue 10, October 2013.

[12] Y. Xiao, V. Krishna Rayi, B. Sun, X. Du, F. Hu and M. Galloway, “A survey of key management schemes in wireless sensor networks”, Computer Communications 30, Elsevier, 2007.

[13] H. Chan and A. Perrig “PIKE: peer intermediaries for key establishment in sensor networks”, Proceedings of the 24th annual joint conference of the IEEE computer and communications societies (INFOCOM ’05), Miami, FL, USA, March 2005.

[14] D. Liu, P. Ning and W. Du, “Group-Based Key Predistribution for Wireless Sensor Networks”, ACM Transactions on Sensor Networks, Vol. 4, No. 2, Article 11, March 2008.

[15] A. Nisha and N. D. Kale, “A Survey on key Generation and Pre distribution Technique in wireless Sensors Networks”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 2, February 2014.

[16] S. Bala, G. Sharma and A. K. Verma, “Classification of Symmetric Key Management Schemes for Wireless Sensor Networks” International Journal of Security and Its Applications Vol. 7, No. 2, March 2013.

Applications of Data Envelopment Analysis in Development and Assessment of Sustainability Across Economic, Environmental and Social Dimensions

Hamid Hosseini
Department of Mathematics
Kerman Branch, Islamic Azad University
Kerman, Iran

Abbas Ali Noura
Department of Mathematics
University of Sistan & Baluchestan
Iran

Sara Fanati Rashidi*
Department of Mathematics
Shiraz Branch, Islamic Azad University
Shiraz, Iran

***Abstract*—Recently, senior managers are paying much more attention to the environmental aspects of decision-making units. Technically, global economy is inextricably connected to the environment, as it is heavily dependent on extraction and exploitation of natural resources. In this article, we try to propose a number of models for efficiency evaluation that combine the growing concepts in environmental areas along with social and economic subjects. Generally speaking, if economic growth is to be continuous and effective in the long term, it must be based on a combination of economic, environmental and social components. The existing literature on data envelopment analysis (DEA) is often based on economic efficiency. However, due to the environmental pollution at a global level, there have been recent studies in relation to sustainability efficiency with focus on environmental and social aspects; although, these studies were limited and left much room for further research. The present study evaluates the efficiency of decision-making units using social, economic and environmental indicators, and tries to minimize the flaws of DEA in the proposed models by making relative comparisons to previous models.**

Keywords—Data envelopment analysis; desirable and undesirable outputs; strong and weak disposability; sustainability efficiency

I. INTRODUCTION

Data envelopment analysis (DEA) is a linear programming-based method initially proposed by [1]. This method [1] presented a linear programming model (CCR) for efficiency evaluation of decision-making units (DMUs) with multiple inputs and outputs.

Technically, aside from measuring economic efficiency and ranking the units, DEA provides the managers and planners in a given organization with methods for improvement of strategies and all-round growth and development of the organization under study.

Nowadays, a widespread viewpoint is formed suggesting that without consideration to the waste and pollution resulting from industrial activities and production and consumption processes, which would endanger both the environment and the humans, we will not achieve sustainable economic growth.

Therefore, economic, environmental and social policies need to be designed in a way that would effectively improve the efficiency of decision-making units.

So far, a significant amount of research has been dedicated to the applications of DEA; these efforts, however, have been mainly focused on evaluation of DMUs in areas of science and engineering irrelevant to the environment [2] and [3].

In recent years, a number of studies have combined DEA with the life cycle of DMUs in order to evaluate the environmental efficiency of various systems [4]-[8]. Nonetheless, these studies only cover the environmental and economic aspects of sustainability and ignore the social dimension altogether. Some other researchers have used DEA to assess all three of the environmental, social and economic indicators, but they still faced certain challenges combining these three indicators [9]-[12].

Despite its advantages, data envelopment analysis has two major limitations that play a fundamental role during evaluation of sustainability efficiency:

DEA tells us if a unit is efficient or inefficient, but does not discriminate between efficient units (i.e., it does not rank the efficient units). Now, since all efficient units are assigned an efficiency score of one, it would be difficult to choose an alternative in absence of a ranking scheme [13].

When we are faced with large input and output sets based on the number of units, a flaw occurs in the efficiency evaluation and a large number of units are deemed efficient [14].

In this article, we try to eliminate these two flaws through our proposed models. The rest of the article is structured as follows:

In Section 3, we discuss the axioms holding for undesirable outputs (wastes). In Section 4, we engage in efficiency evaluation based on economic, environmental and social

*Corresponding Author

indicators. In Section 5, we present and solve a numerical example using our suggested models, and make a comparison of results in the end. Our research results will be described in the sixth section.

Technically, the pollution and waste harming the environment are called undesirable outputs.

Undesirable outputs are outputs produced along with the main outputs [15]; due to the nature of undesirable outputs, it is often difficult to determine their prices in the market, something that is usually done by experienced experts using shadow prices.

For instance, paper is produced using the four inputs of pulp, capital, workforce and energy; however, alongside the produced paper, outputs such as biochemical oxygen, suspended solids, sulfur oxides and particles are produced as well, which are impossible to put a market price on [15]. Note that by undesirable output prices, we refer to the costs imposed on us for production of such outputs. Some other examples of such costs would be environmental contamination, disease prevalence and the expenses related to waste management.

II. Axioms Holding for Undesirable Outputs

A. Null-Joint

Undesirable outputs are null-joint when

□□$Y^D . Y^U) \in PPS \quad . \quad Y^U = 0 \rightarrow Y^D = 0$

This shows that when the desirable output has a positive value, the undesirable outputs will definitely have positive values as well.

For instance, it would be impossible to produce paper without production of biochemical oxygen, sulfur and so on.

B. Weak Disposability

In most processes, undesirable outputs are produced alongside desirable outputs.

This axiom states that along with the reduction of desirable outputs, undesirable outputs will decrease as well [15]. In other words, a relative reduction in desirable outputs would require the reduction of undesirable outputs to the same proportion.

Therefore, considering the weak disposability axiom and the standard production set, our new production possibility set (PPS), denoted by $P^W\ (X)$, will be as follows:

$\Pi\Pi\Sigma = P^W\ (X) = \{(Y^{DK} . Y^{UK}) : Y^{DK} \le Y^D\ \lambda . Y^{UK} = Y^U \lambda . X\lambda \le X . \lambda \in R^n_+\ \}$

Model (1) presents the linear programming form of the PPS. Note that j=1,2,⋯,n represents the number of decision-making units that use the input vector $X \in R^m$ to produce the desirable output vector $Y \in R^s$ and the undesirable output vector $U \in\ R^+$.

$Max \qquad \varphi_0$

s.t

$\sum_{j=1}^{n} \lambda_j\, x_j \le x_0 \qquad\qquad j=1,2,\ldots,n$

$\sum_{j=1}^{n} \lambda_j y_j \ge \varphi_0 y_o \qquad\qquad j=1,2,\ldots,n \qquad (1)$

$\sum_{j=1}^{n} \lambda_j\, u_j = u_o \qquad\qquad j=1,2,\ldots,n$

$\lambda_j \ge 0 \qquad\qquad j=1,2,\ldots,n$

C. Strong Disposability

Similar to the previous topic, this axiom puts certain constraints on the model and states that it is possible to increase the desirable outputs, while preventing undesirable outputs from increasing to the same degree; in some cases, we could even bring the production of undesirable outputs close to zero.

From a profitability perspective, this axiom only cares about increasing desirable outputs and doesn’t take a serious look at the production rate of undesirable outputs.

According to the strong disposability axiom and the standard conditions of production possibility set,

PPS $=\left\{(x . y)\ \middle|\, x \ge \sum_{j=1}^{n} \lambda_j\, x_j \,.\, y \le \sum_{j=1}^{n} \lambda_j y_j \,.\, \lambda_j \ge 0 \,.\, j = 1. \ldots . n\right\}$

$P^S\ \ (X)$ denotes the new production possibility set as presented in the following:

PPS $= P^S\,(X) = \{(Y^{DK} . Y^{UK}) : Y^{DK} \le Y^D\ \lambda . Y^{UK} \le Y^U \lambda . X\lambda \le x . \lambda \in R^n_+\}$

Model (2) provides the nonlinear programming problem for this PPS. In this model, we have n DMUs that use the input vector $X \in R^m$ to produce the desirable output vector $Y \in R^s$ and the undesirable output vector $U \in\ R^+$.

Max $\quad \varphi_o$

s.t

$\sum_{j=1}^{n} \lambda_j\ \ x_j\ \ \le\ \ x_o \qquad$ j=1,2,…,n

$\sum_{j=1}^{n} \lambda_j\ y_j\ \ \ge\ \varphi_o\ y_o \qquad$ j=1,2,…,n $\qquad$ (2)

$\sum_{j=1}^{n} \lambda_j\ \ U_j\ \ \ \ge\ \frac{1}{\varphi_o}\ \ u_o \qquad$ j=1,2,…,n

$\lambda_j\ \ \ \ge 0 \qquad\qquad$ j=1,2,…,n

III. Efficiency Calculation based on Economic, Environmental and Social Indicators

The existing literature on data envelopment analysis is mostly based on economic efficiency. Now, since global economy and efficiency evaluation are both influenced by ecological, social and economic components, environmental pollution has influenced the sustainability of efficiency measurement [4].

Since our objective in this research is to measure efficiency in several different dimensions and then combine them together, we can use the following linear programming problem to obtain the sustainability efficiency.

Considering the presence of both desirable and undesirable outputs in real-world situations [15], we propose the following model for assessment of efficiency through economic, environmental and social indicators, which are denoted by d = 1,2,3, respectively.

Model (3) is used for calculation of efficiency in the mentioned dimensions. In this regard, consider n decision-making units that use the input vector $X \in R^m$ to produce the desirable output vector $Y \in R^s$ and the undesirable output vector $U \in R^+$:

$$\begin{aligned}
&\text{Min} \quad \theta_o^d \\
&\quad \text{s.t} \\
&\textstyle\sum_{j=1}^{n} \lambda_j x_j \le \theta_o^d x_o && j=1,2,\ldots,n \\
&\textstyle\sum_{j=1}^{n} \lambda_j y_j \ge y_o && j=1,2,\ldots,n \qquad (3) \\
&\textstyle\sum_{j=1}^{n} \lambda_j u_j = u_o && j=1,2,\ldots,n \\
&\lambda_j \ge 0 && j=1,2,\ldots,n
\end{aligned}$$

θ_o^d measures the efficiency of DMU_O in dimension d. Note that units with $\theta_o^d = 1$ are efficient and the ones with $\theta_o^d < 1$ are considered inefficient.

The difference between model (3) and models (1) and (2) is that in model (3), we evaluate the DMUs based on their inputs. This is due to the fact that in many situations, like the example used in this study, certain indicators such as economic and social indicators don't have any undesirable outputs, in which case the constraints related to undesirable outputs are removed. Furthermore, the environmental indicator doesn't have any desirable outputs and thus, we can remove the respective constraints; in this case, we can fixate the undesirable outputs and evaluate our DMUs based on their input levels.

In today's world, where most countries, especially industrial countries, put a strong emphasis on the environment and pollution prevention, environmental efficiency is an emerging subject of interest.

Technically speaking, any decision-making unit can affect the whole society in one way or another. For instance, imagine a factory that causes a great amount of environmental pollution despite being economically efficient; undoubtedly, although economic efficiency is an important factor, environmental efficiency is as important to say the least. Another example would be a factory producing socially undesirable products such as tobacco or alcoholic drinks.

In this study, in addition to economic efficiency, we try to consider the social and environmental aspects as well and consider each of them as an indicator or dimension of efficiency.

Technically, the $\theta_{o_j}^d$ value obtained from model (3) is the efficiency of DMU_O in dimension d. [16] proposed the following formula for measurement of mean efficiency:

j=1,2,…,n (4)

$$\theta_o^{sust} = \frac{\sum_{d \in D} \theta_{o_j}^d}{|D|}$$

We must note that the CCR model introduced by [1] was used for efficiency calculation, as presented in [16]'s study; due to the previously mentioned reasons, model (3) was used in the prior sections.

IV. Numerical Example

In the current example, 21 Iranian industries are evaluated, where the total capital employed in the industry is considered as the input, value added is the desirable economic output, job creation level is the desirable social output and airborne contaminant levels represent the undesirable environmental output [17].

TABLE I. INPUTS AND DESIRABLE AND UNDESIRABLE OUTPUTS FOR THE 21 IRANIAN INDUSTRIES UNDER STUDY

Code	**Industry**	**Capital Employed (input)**	**Value Added (desirable economic output)**	**Number of Workers (desirable social output)**	**so_2 (tons) (undesirable environmental output)**	**co_2 (tons) (undesirable environmental output)**	**SPM (tons) (undesirable environmental output)**
1	Food	*10.05*	*8.34*	*15.41*	*37.75*	*11.47*	*16*
2	Textile	*4.53*	*3.24*	*9.32*	*7.94*	*2.44*	*3.8*
3	Garment	*0.07*	*0.61*	*0.61*	*0.13*	*0.07*	*0.08*
4	Leather	*0.16*	*0.32*	*0.85*	*0.53*	*0.13*	*0.26*
5	Wood	*0.36*	*0.36*	*0.67*	*1.27*	*0.39*	*0.62*
6	Paper	*0.88*	*0.83*	*1.63*	*0.22*	*1.1*	*1.3*
7	Publishing	*0.37*	*0.47*	*1.19*	*0.1*	*0.01*	*0.09*
8	Coke	*4.01*	*11.19*	*1.46*	*13.79*	*13.2*	*10.25*
9	Material	*33.56*	*14.85*	*7.13*	*7.54*	*11.9*	*8.89*
10	Rubber Products	*3.15*	*2.7*	*4.71*	*1.93*	*0.93*	*1.15*
11	Non-metallic Mineral Products	*10.51*	*8.63*	*14.37*	*13.22*	*39.6*	*49.94*
12	Basic Metals	*16.54*	*17*	*6.44*	*7.79*	*15.1*	*1.07*
13	Fabricated Metal Products	*1.92*	*3.27*	*6.5*	*0.27*	*1.15*	*1.63*
14	Machinery	*3.4*	*4.34*	*7.52*	*2.53*	*0.14*	*1.71*
15	Office Machinery	*0.22*	*0.12*	*0.23*	*0.02*	*0.01*	*0.01*
16	Power Generating Machinery	*1.59*	*2.78*	*4.82*	*1.06*	*0.57*	*0.7*
17	Broadcasting	*0.35*	*0.51*	*0.81*	*0.01*	*0.05*	*0.05*
18	Medical Instruments	*0.26*	*0.45*	*1.04*	*0.3*	*0.11*	*0.17*
19	Vehicles	*5.82*	*17.54*	*11.09*	*2.49*	*1.18*	*1.62*
20	Transport Vehicles	*1.24*	*1.84*	*2.56*	*0.47*	*0.2*	*0.3*
21	Furniture	*1.01*	*0.61*	*1.64*	*0.64*	*0.25*	*0.36*

TABLE II. COMPARISON OF MEAN Θ_0^{SUST} EFFICIENCY AND Φ_0 EFFICIENCY

DMU	*Economic efficiency via model (3)*	*Social efficiency via model (3)*	*Environmental efficiency via model (3)*	*Mean θ_o^{sust} efficiency via formula (4)*	*φ_o efficiency via model (1)*	*φ_o efficiency via the nonlinear model (2)*
1	*0.10*	*0.18*	*1*	*0.42*	*1.00*	*1*
2	*0.08*	*0.24*	*0.47*	*0.26*	*1.00*	*1.342*
3	*1*	*1*	*0.52*	*0.84*	*1.00*	*1*
4	*0.23*	*0.61*	*0.90*	*0.58*	*1.00*	*2.414*
5	*0.11*	*0.21*	*0.96*	*0.42*	*1.0525*	*3.749*
6	*0.11*	*0.21*	*0.33*	*0.21*	*1.0545*	*3.291*
7	*0.15*	*0.37*	*0.09*	*0.20*	*1.00*	*2.821*
8	*0.32*	*0.04*	*1*	*0.45*	*1.00*	*1*
9	*0.05*	*0.02*	*0.10*	*0.05*	*1.00*	*1.074*
10	*0.10*	*0.17*	*0.17*	*0.14*	*1.3631*	*2.389*
11	*0.09*	*0.16*	*1*	*0.41*	*1.00*	*1*
12	*0.12*	*0.04*	*0.25*	*0.13*	*1.00*	*1*
13	*0.20*	*0.39*	*0.18*	*0.25*	*1.2960*	*1.635*
14	*0.15*	*0.25*	*0.21*	*0.20*	*1.5400*	*1.515*
15	*0.06*	*0.12*	*0.02*	*0.06*	*1.00*	*8.137*
16	*0.20*	*0.35*	*0.19*	*0.24*	*1.2049*	*1.698*
17	*0.17*	*0.27*	*0.04*	*0.16*	*1.00*	*2.658*
18	*0.20*	*0.46*	*0.32*	*0.32*	*1.00*	*1*
19	*0.35*	*0.22*	*0.12*	*0.23*	*1.00*	*1*
20	*0.17*	*0.24*	*0.11*	*0.17*	*1.4510*	*2.095*
21	*0.07*	*0.19*	*0.18*	*0.14*	*1.00*	*4.837*

Note that by airborne contaminants, we mean the CO_2 (carbon dioxide), SPM (suspended particulate matter) and SO_2 (sulfur dioxide) resulted by the use of fossil fuels, as first investigated from an environmental aspect in [17] in Iranian production industries. In this article, their combined efficiency is investigated via the mentioned models based on economic, environmental and social indicators. Data are presented in Table I.

After solving model (1) for the 21 DMUs using MATLAB software and solving model (2) via GAMS, we proceed to solve model (3); Table II provides the results, i.e. mean efficiency scores per economic, social and environmental indicators

V. CONCLUSION

In this paper, we made an evaluation of two efficiency calculation methods for decision-making units with undesirable outputs. The first approach involved the models presented by [15], and the second approach employed a modified version of [16] method. Results produced by the two methods were compared within the framework of a numerical example.

According to the results, our proposed approach was able to provide a better ranking among efficient decision-making units.

Our suggested method has improved the discriminatory power of standard DEA by categorizing the inputs through economic, environmental and social indicators. The capabilities of the approach and its applications were demonstrated in a real-world case study, and relative comparisons were made to previous methods.

As can be observed in Table II, 14 DMUs were introduced as efficient based on the weak disposability axiom [15] and 7 units were found efficient under assumption of strong disposability. Meanwhile, using the combined method, only one unit was considered efficient in the economic and social dimensions and three units were deemed efficient based on the environmental indicator; the units were also uniquely ranked using (4). As previously mentioned, this demonstrates the high discriminatory power of our proposed models in respect to efficient units.

REFERENCES

[1] Charnes A, Cooper WW, and Rhodes E (1978). Measuring the efficiency of decision making units. European journal of operational research, 2(6): 429-444.

[2] Liu JS, Lu LY, Lu WM, and Lin BJ (2013). A survey of DEA applications. Omega, 41(5): 893-902.

[3] Liu JS, Lu LY, Lu WM (2015).Research fronts in data envelopment analysis. Omega, 58: 33-45.

[4] Begum H, Siwar C, Er AC, and Alam ASAF(2016). Environmentally friendly practices of oil palm cultivators. International Journal of Advanced and Applied Sciences, 3(2): 15-19.

[5] Salem SMA, Rahman NAA (2016). The effect of bank-specific factors and unstable macroeconomic environment on bank efficiency: evidence from FCC. International Journal of Advanced and Applied Sciences, 3(9): 97-102.

[6] Dutsenwai HS, Ahmad BB, Tanko AI, and Mijinyawa A (2017). Spatio-temporal analysis of vegetation and oil spill intensity in Ogoniland. International Journal of Advanced and Applied Sciences, 4(4): 81-90.

[7] Mohammadi A, Rafiee S, Jafari A, Keyhani A, Dalgaard T, Knudsen MT, Nguyen TL, Borek R, and Hermansen JE (2015). Joint life cycle

assessment and data envelopment analysis for the benchmarking of environmental impacts in rice paddy production. Journal of Cleaner Production, 106: 521-532.

[8] Vázquez-Rowe Iand Iribarren D (2014). Review of life – cycle approach coupled with data envelopment analysis: Launching the CFP+DEA method for energy policy making. The Scientific World Journal, 2015: Article ID 813921, 10 pages. https://doi.org/10.1155/2015/813921

[9] Chang DS, Kuo LR, and Chen y (2013). Industrial changes in corporate sustainability performance – an empirical overview using data envelopment analysis. Journal of Cleaner production, 56: 147-155.

[10] Khodakarami M, Shabani A, and Farzipoor Saen R (2014). A new look at measuring sustainability of industrial parks: a two-stage data envelopment analysis approach. Clean Technologies and Environmental Policy, 16(8): 1577-1596.

[11] Reig-Martínez E, Gómez-Limón JA, and Picazo-Tadeo AJ (2011). Ranking farms with a composite indicator of sustainability. Agricultural economics, 42(5): 561-575.

[12] Tajbakhsh A, Hassini E (2014). A data envelopment analysis approach to evaluate sustainability in supply chain networks. Journal of Cleaner Production, 105: 74-85.

[13] Cook WD and Seifore LM (2009). Data envelopment analysis (DEA)-thirty years on. European journal of operational research, 192(1): 1-17.

[14] Avkiran NK (2002). Productivity analysis in the service sector with data envelopment analysis. In: Avkiran NK (Ed.), 2nd edition, Camira, Qld.

[15] Fare R, Grosskopf S, Lovell CAK, and Pasurka C (1989). Multilateral productivity comparisons when some outputs are undesirable: a nonparametric approach .The Review of Economics and Statistics,71: 90-98.

[16] Galan-Martin A, Guillen-Gosalbez G, and Stamford L (2016). Enhanced data envelopment analysis for sustainability assessment. Computers and Chemical Engineering, 90: 188-200.

[17] Nasrollahi Z, Sadeghi Aarani Z, and Ghafari Goolak M (2012). Modeling undesirable factors (environmental pollutants) in efficiency evaluation using DEA: A case study of IRAN'S manufacturing industries. Nameh-ye Mofid Journal, 90: 87-110.

An Improved Bat Algorithm based on Novel Initialization Technique for Global Optimization Problem

Waqas Haider Bangyal
Member IEEE SMC
Department of Computer Science,
Iqra University, Islamabad, Pakistan

Jamil Ahmad
Senior Member IEEE, Professor Computer Science
Department of Computer Science,
Kohat University of Science and Technology (KUST),
Kohat, Pakistan

Hafiz Tayyab Rauf
Department of Computer Science,
University of Gujrat, Gujrat, Pakistan

Sobia Pervaiz
Department of Software Engineering,
University of Gujrat, Gujrat, Pakistan

Abstract—**Bat algorithm (BA) is a nature-inspired metaheuristic algorithm which is widely used to solve the real world global optimization problem. BA is a population-based intelligent stochastic search technique that emerged from the echolocation features of bats and created from the mimics of bats foraging behavior. One of the major issue faced by the BA is frequently captured in local optima while handling the complex real-world problems. In this study, a new variant of BA named as improved bat algorithm (I-BAT) is proposed. Improved bat algorithm modifies the standard BA by enhancing its exploitation capabilities, and secondly for initialization of swarm, a quasi-random sequence Torus has been applied to overcome the issue of convergence and diversity. Population initialization is a vital factor in BA, which considerably influences the diversity and convergence of swarm. In order to improve the diversity and convergence, quasi-random sequences are more useful to initialize the population rather than the random distribution. The proposed strategy is applied to standard benchmark functions that are extensively used in the literature. The experimental results illustrate the superiority of the proposed technique. The simulation results verify the efficiency of proposed technique for swarm over the benchmark algorithm that is implemented for the function optimization.**

Keywords—Bat algorithm; local optima; exploration and exploitation; quasi-random sequence

I. Introduction

Optimization of the process that involves searching a vector from a function creates an optimum solution. All possible values are considered as available solutions, while the exceptional value referred as the optimum solution. Generally, optimization algorithms are used to resolve the local and global search optimization issues. Optimization algorithms have two categories: stochastic algorithms and deterministic algorithm [1]. Deterministic algorithms use gradient and generate same solutions for all iterations, which are initiated with the same starting point. Thus, stochastic algorithms generate distinct solutions even if the starting points are same and never uses gradient. Although, the final values which are slightly different are supposed to give the same optimum solutions within a given precision [2]. Stochastic and population-based algorithms have two further parts: Heuristics and Meta-Heuristics [3]. Swarm Intelligence (SI) is one of the nature-inspired meta-heuristic algorithm that is frequently used solve the complex optimization problems. Some traditional neural networks and evolutionary algorithms [4] also use for data classification and optimization.

To handle complex real-world optimization problems, SI nature stimulated technique has been used for many years. Beni [5] was the first who introduced SI, which is inspired with the behavior of birds, fishes, and insects, and their exclusive capability to handle a complex nature of problems in the fashion of swarms. Thus, the same condition would seem complicated if they work individually instead of swarms. Individual bees, ants [6], fishes and birds have limited intelligence, however, when they cooperate with each other for social interaction and interact with the environment, they are capable to accomplish tough tasks, e.g. to get a food source from shortest path and organization of their nests [7].

BA is one of the most famous SI based algorithm was introduced by Xin-She Yang [8], which is inspired by echolocation of micro-bats. Bats produce some echo in the environment, during hunting or flying. By producing an echo, they get an accurate image of the environment and exact location of their prey, due to this reason bats can find their prey in complete darkness [9]. There is a wide range enhanced and improved versions of BA that have been introduced recently. In addition to this, BAT algorithm diversely implemented in various applications of different fields like image processing, engineering design, feature selection and many more [10].

Choosing an initial configuration to initialize the population is one of the primary tasks in evolutionary computing. The performance of evolutionary algorithm may vary due to the different fashion of firing the individuals into the search space

[11]. The swarm covers the more search space; the more there is the fair chance of reaching an optimal solution, Random Initialization of population is usually employed when there is no candidate solution available. The final solution can be improved by selecting the most suitable distribution for population initialization.

Swarm convergence has been considered as the dominant issue for the researchers. So far many researchers have focused to determine whether bat concentrates to the same curve or not. They are assumed to point out the feature that performs the significant role in swarm convergence. Performance of BAT algorithm is intensely concerned by the premature convergence of bats [12]. BA may stuck in local minima due to premature convergence before a global optimum found. To overcome local optima problem, many researchers have proposed various improve methods [13] in which disparate mutations are performed on relevant parameters such as velocity, pulse rate, frequency, loudness and swarm size [14], that help bats to move into a new area of search space. Mutation in BA helps the bats to avoid premature convergence around local optimal [15].

Premature convergence is the main problem with BAT algorithm and other optimization techniques including Evolutionary Algorithms (EAs) such as Gene expression programming (GEP), Genetic Algorithm (GA), Differential Evolution (DE) and Genetic Programming (GP) [16]. Diversification (exploration) and intensification (exploitation) plays an important role in heuristics. Exploitation referred to the local search ability while exploration represents the global search ability of any population-based algorithm [17]. The performance of swarm-based algorithms is highly dependent on the balance between exploitation and exploration. Premature convergence is caused by excessive exploitation and less exploration, while greater exploration and less exploitation may provoke difficulties to reach the optimal solution [18].

The local search ability (exploration) of standard BA is better than the global search ability (exploration). Therefore, to improve exploration capability, we have carried out an improved version of BA called improved bat algorithm (I-BAT). For ensuring the integrity of proposed technique I-BAT is compared with the original BAT on nine well-known benchmark test functions. Experimental result shows that proposed variant has performed well as compared to original BA on specific test functions.

The rest of the paper is structured as: Section 2 presents related work, while working of original Bat algorithm is described in Section 3. Methodology is presented in Section 4 and Section 5 contains discussion and final results for the proposed method. Section 6 presents the conclusion and future work.

II. Related Work

In the field of medical science, P. Kora et al. [19] implemented a new modified BA to extract main features of each cardiac beat. After the extraction of these best features, they are embedded as an input in neural architecture classifier. According to the exhaustive analysis results, it is illustrated that by using optimization on main features, the execution of classifier is significantly improved. Moreover, a novel method of BA described by authors in [20], for parameter estimation in the nonlinear dynamic biological system. The authors included the impact of both Levy Fight and Chaotic dynamics. The optimization is performed on the parameters of secondary system with the use of introduced Chaotic Levy Fight BA to follow the dynamics related to the primary system. Statistical results illustrate the efficiency and stability of proposed algorithm in biological systems. On the other hand, for image recognition J. Zhang in [21] tried to solve image recognition problem that is the reason he proposed a new method Bat Algorithm with Mutation (BAM). In BAM, a modification was embedded during the process of updating the BAs in which BA mutate for the optimal solution.

Paiva *et al.* [22] proposed a new version of BA algorithm with name Modified Bat having the Cauchy mutation and elite opposition Based learning. The objective of this proposed version is to expand convergence velocity and produce the diversity of algorithm. A comparison is conducted with all recent research of BA, and four standard benchmark functions are implemented in the proposed version for the sake of comparison. After the exhaustive analysis, the excellence of proposed version is proved. An Accelerated Bat Algorithm (ABATA) proposed by the authors in [23], where the author used Nelder-Mead approach for local search, to refine optimal best solution in all iterations. Nelder-Mead approach performs a well-defined local search, and able to improve the exploitation abilities in ABATA. The working of ABATA is verified through seven integer programming problems, as well as, compared with four standard algorithms. The results illustrated that ABATA could obtain the optimal global solution in less computational time.

Enhanced method of BAT introduced by authors in [20], for parameter estimation in the nonlinear dynamic biological system. The authors included the impact of both Levy Fight and Chaotic dynamics. The optimization is performed on the parameters of secondary system with the use of introduced Chaotic Levy Fight BAT, to follow the dynamics related to the primary system. Statistical results illustrate the efficiency and stability of introduced algorithm in biological systems.

To sort out the global numerical optimization problem, a robust hybrid metaheuristic optimization method (HS/BA) provided by the authors in [24] that is considered as an improved version of the traditional BA. In the proposed method, a mutation operator was adjusted for maximizing convergence speed. The defined method was verified through fourteen test functions, which described that proposed method outperforms. In [25], authors proposed a new approach of BAT Algorithm with named Hybrid BA Algorithm (HBA). In this approach, the authors merged traditional BAT with Differential Evolution (DE). Standard functions are used for implementation. According to the experimental results, HBA provides improved results than a traditional BAT.

An advanced version of Bat Algorithm called Modified Bat Algorithm (MBA) was proposed in [26]. They introduced modified Bat with enhancement of exploration methodology, in which loudness and pulse emission rate of BATs were changed. Experimental analysis was performed on 15 standard functions, which showed that MBA gives a quality solution related to optimization problems. The authors in [27] introduced a new approach Hybrid Self-Adaptive Bat Algorithm (HSABA). The new approach was carried out by combining Self-Adaptive Bat Algorithm (SABA) and various Differential Evolution (DE) strategies. This new approach was implemented as a heuristic for local search (a modified operator). A comparison was performed with well-defined comparative studies as well as with other standard algorithms. The comparison results certified that HSABA works adequately for enhancing the impact of Population-Based Algorithms.

In [28], authors proposed a new method in the computational intelligence field, which was an improved version of BAT named as IBACH, and used to solve the problems of integer programming. The introduced algorithm implements chaotic behavior to produce BAT solution behavior like the acoustic monopoly. A numerical analysis was conducted with the comparison of other algorithms: PSO, traditional BA, and various harmony search algorithm. Although, the ability of this introduced algorithm was obtained when it computes optimal solution in fewer computations. To avoid from pre mature convergence, directional echolocation was proposed with respect to traditional BAT in [29], which improves the exploitation and exploration abilities of the traditional BA. For improving the performance of BAT, three other enhancements have been included in the BAT. The proposed approach Directional BA Algorithm (dBA), has been verified through various functions belongs to a CEC'2005 standard suite. Similarly, the algorithm was compared with various BAT versions and additional ten algorithms. The results concluded that dBA is better than others.

III. BAT Algorithm

BAT algorithm is a nature-inspired algorithm belongs to SI family, proposed by Xin-She Yang [8]. Bat algorithm works on the echolocation of micro bats and used echo of bats for seeking of food. Yang focused on three rules for the implementation of the bat: Firstly, to measure the distance to the specific point, all bats use echolocation. Secondly, bats fly randomly with fixed frequency towards specified location with specific velocity, however, the loudness and wavelength can vary. Thus, bats automatically adjust their wavelengths according to their target. Thirdly, the author considered that loudness is varied from maximum to minimum rather than any other way.

In bat algorithm each bat of population reveals a candidate solution. Each candidate solution is illustrated with the help of vector $x_i = (x_1, \dots x_i)^t$ with real value elements x_{ij} ,for $i = 1 \dots N_p$and the interval for each element is taken from $x_{ij} \in [x_{lb} \dots x_{ub}]$. While, x_{lb} and x_{ub} determines the upper and lower bounds, however, N_p represents the size of population [27]. The major components of algorithms are initialization, variation operation, local search, evaluation of a solution, and replacement.

Step 1: In initialization, the parameters of an algorithm are initialized, after that it generates an initial population using random distribution, and at last, the best solution is illustrated from that initial population.

Step 2: By using natural rules of bat echolocation, the variation operator is used to move and represent virtual bats in search space and generate new solutions by following equation:

$$f_i^{(t)} = f_{min} + (f_{max} - f_{min})U(0,1) \tag{1}$$

The value of f_{max} and f_{min} depends upon the problem nature, where U $(0,1)$ is uniform random number generator.

Where the updated velocity of particles can be represented by the following equation:

$$v_i^{(t+1)} = v_i^t + (x_i^t - best)f_i^t \tag{2}$$

$best$ is the current globally best location and x_i^t represents current bat position at iteration t .

Bats are moved towards the bats new position in Dimension D with the following equation:

$$x_i^{(t+1)} = x_i^t + v_i^{t+1} \tag{3}$$

Step 3: In local search, current best solution is modified by using random walk with direct exploitation, where following equation is used:

$$x_{new} = best + \epsilon A_i^t \tag{4}$$

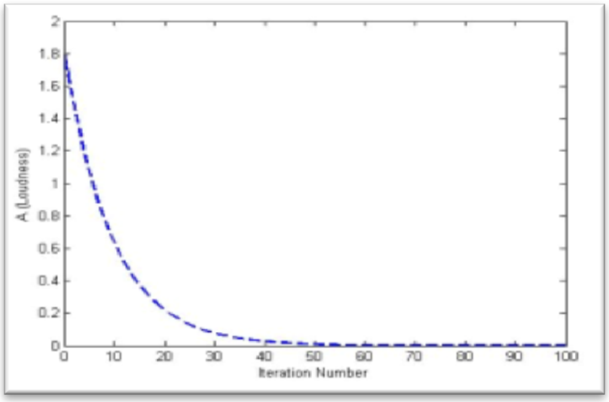


Fig. 1. Changing loudness with respect to iterations.

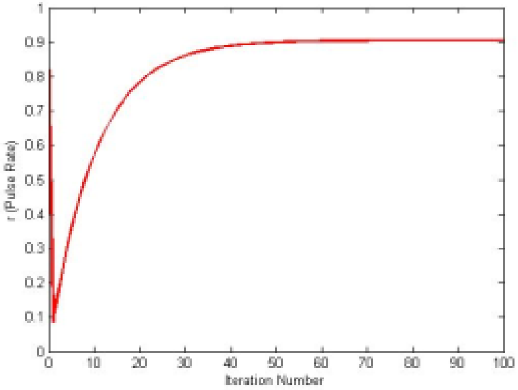


Fig. 2. Changing loudness with respect to iterations.

Here, ${}_i^{(t)}$ is used for loudness and ϵ is a random number between $[-1,1]$ used for scaling factor.

Step 4: Probability of pulse rate r_i is carried out to launch pulse rate. To accept a new best solution, the probability depends upon the loudness A_i^t. Basically, the loudness A_i^t and the pulse rate r_i two parameters are used to control the standard BAT algorithm. Normally, when the population reaches closer the local optimum, then the loudness A_i^treduces (decrease) and pulse rate $r_i$enlarge (increase) (Fig. 1 and 2). When the bat finds its prey, and the loudness increases and pulse rate decreases these both features simulate the natural bats. The equations for decreasing loudness and increasing pulse rate are as follows:

$$ {}_i^{(t+1)} = \alpha A_i^{(t)} \tag{5}$$

$$ r_i^{(t)} = r_i^0 \, [1 - \exp(-\gamma^t)] \tag{6}$$

In above equations, α and γ are constants where α is that parameter which handles the convergence, it is same as simulated annealing algorithm's cooling factor. Algorithm 1 presents the pseudo code for the original Bat algorithm (Fig. 3).

_Algorithm_1: Standard_Bat_Algorithm

```
    Input: $x_i = (x_1, \dots x_i)^t$ →Bat population
    Output: $x^{best}$ & $\min(f(x))$ →optimal solution and minimal objective function value
$x_i$ =Rand_Bat_Init($x_i$); population initialization
Evaluate($x_i$) ;evaluate newly generated population
$Compute(x^{best})$;find current global best
While ($t \le t_{max}$ ) do
        for $i = 1 \dots N_p$ do
          Update  frequency using eq.1
          Update velocity  using eq.2
          Update  position using q.3
            if rand(0,1) > $r_i^{(t)}$ then
            $\boldsymbol{x_j^t}$ = improve solution using eq.4
            end
    if $f_{new}$ = Compute $\boldsymbol{x_j^t}$; generate new solution
    evaluate = evaluate +1;
    if $f_{min}^{new} < f_{old}$ and Rand(0,1) < $A_i^{(t)}$ then
    $x_i = \boldsymbol{x_j^t}$; $f_{old} = f_{min}^{new}$;
    end if
    $f_{min} = \min(f(x^{best})))$;
   end for
end while
```

IV. Methodology

As it has been mentioned above, we have made two major contributions to this study. First, we have introduced one novel methods of initialization of population using low discrepancies sequence that uses the torus quasi-random sequence to create the initialization of the swarm rather random distribution. Second, we have proposed a new strategy of searching for standard bat algorithm by improving the exploitation and convergence capability with controlled parameters.

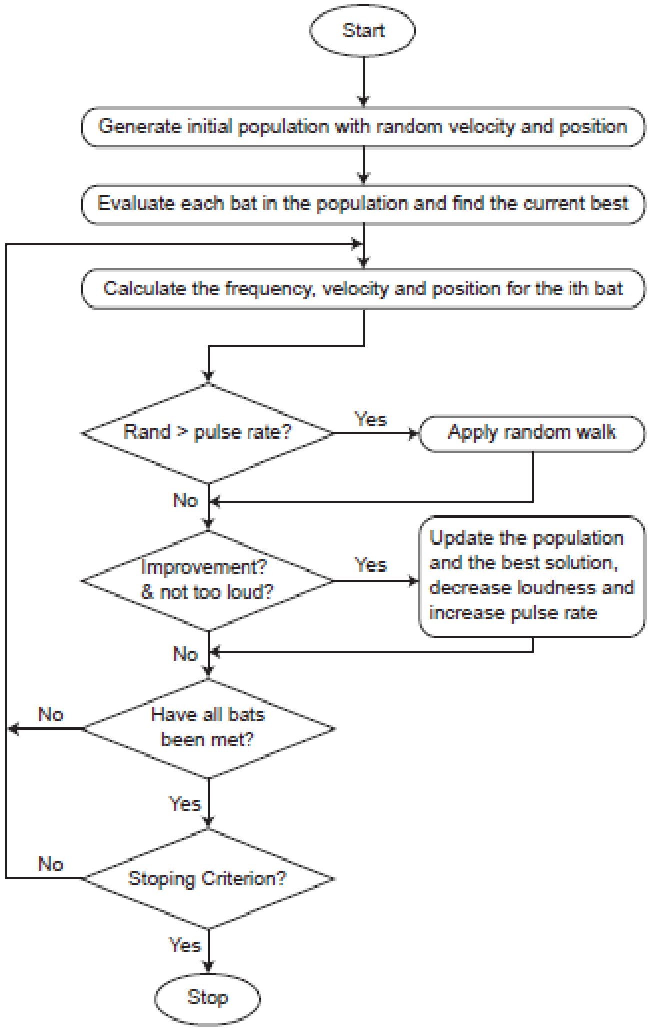

Start
Generate initial population with random velocity and position
Evaluate each bat in the population and find the current best
Calculate the frequency, velocity and position for the ith bat
Rand > pulse rate?
Yes
Apply random walk
No
Improvement? & not too loud?
Yes
Update the population and the best solution, decrease loudness and increase pulse rate
No
No
Have all bats been met?
Yes
No
Stoping Criterion?
Yes
Stop

Fig. 3. Flow chart for standard bat algorithm.

A. Random Number Generator

The function, $Random(a_{min,}a_{max})$ of the built in library is used to generate uniform numbers at random points [30]. Influence of regularity on any sequence is recognized by the probability density function of a constant uniform distribution. Given below is the equation of probability density function:

$$ f(p) = \begin{cases} \frac{1}{q-r} \, f o r \; q < p < r \\ 0 \quad for \; t < q \; or \; p > r \end{cases} \tag{7}$$

Where q and r denotes the features of maximum likelihood. The importance of $f(p)$ is worthless at the edge of q and r, because of 0 impact at the integrals of $f(p)dp$ across any interval. Fig. 4 contains the graphical representation of random number generation following uniform distribution. The probability function of evaluation estimates the evaluation of the parameter of maximum likelihood using the equation below:

$$ P(q, r|p) = nlog(r - q) \tag{8}$$

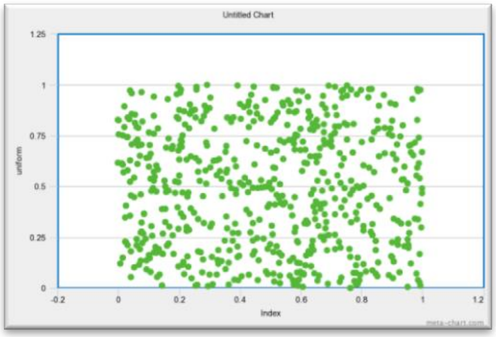


Fig. 4. Random Data Generation using Uniform Distribution [30].

B. The Torus Sequence

The authors in [31] first time, introduced the geometric term Torus for generating a torus mesh that is needed for the geometric correlative system. Torus mesh is generally utilized in the game development community and can be produced using the left-hand or right-hand correlative system. The torus can be represented at 1d, 2d, and 3d by the circle, donut, and 2d quadrilateral sequentially. The equations for torus 3d representation are given below:

$$a(\alpha,\rho) = (d + r_c cos\theta)cos\rho \quad (9)$$

$$b(\alpha,\rho) = (d + r_c cos\theta)sin\rho \quad (10)$$

$$c(\alpha,\rho) = r_c sin\rho \quad (11)$$

α,ρ are their angles of circles and d representing the distance to torus center from tube center, r_c indicates to the circle radius. The author used, R studio has been carried out with the latest version of 3.4.3 using the package "Rand toolbox" to produce torus distribution based random data series. The mathematical representation of torus distribution is as follows:

$$x_z = \left(f\left(z\sqrt{p_1}\right), \dots, f\left(z\sqrt{p_d}\right)\right) \quad (12)$$

f is a fraction computed by $f = x - floor(x)$ where p_1 indicates the sequences of ith prime number. Prime parameters allow only 100,000 dimensions, for more than 100,000 dimensions a manual configuration will be needed. Fig. 5 contains the graphical representation of random number generation following uniform distribution.

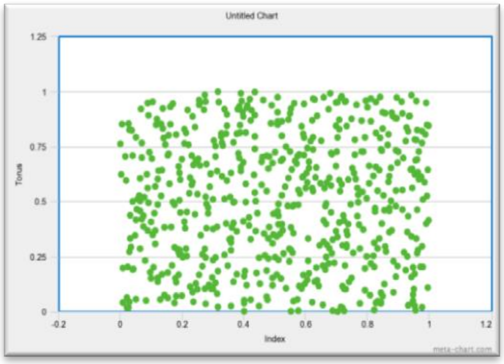


Fig. 5. Random data generation using torus distribution [31].

C. Improved Bat Algorithm

To enhance the local search capability and to sustain the divergence of population so that the algorithm will avoid to trap in local optima, there is no bound to variate the BA through mutation adoption, parameter selection or hybridized the bat algorithm with other algorithms. Although, the process of modification in BA remained with the same issue as it never ensures a modified BA each time achieves a global optimum solution. To reduce this issue, in this paper, proposed improved bat algorithm (I-BA), which enhance the exploitation ability of bats and sustain the divergence of population to acquire consistent results. In this study is presented two modifications with objective to improve the exploration and exploitation abilities of bat algorithm for the improvement of its performance.

- *Novel Initialization Approach Torus*

In this study, we have proposed following one novel method of population initialization approach using low discrepancies sequence Torus named as (TO-BA) that uses the torus quasi-random sequence to create the initialization of the swarm. It can be seen that selection of robust distribution for population initialization may enhance the convergence rate. For population creation, random uniform distribution usually used to initialize the swarm. Generating Torus random population at random location as $\mathrm{Torus}(\mathrm{P_n}, \mathrm{P_d})$.

Where P_n *and* P_d population size and population are dimension respectively. The addition in the standard bat version is as:

x_i =Torus_Bat_Init (x_i); (13)

- *Enhanced Local Search Method*

We introduced next modification by focusing on local search method. In conventional BA, swarm's bats are granted to travel from their present position to new random position by applying local random walk. Local search pattern plays an important role to determine the local optimum. In (4), the local search will only focus on the neighbors of best position obtained by the entire swarm. However, if this position is far from the global optimal solution, then some local search capabilities will be useless due to the inefficient exploitation. To overcome this, we modify random walk by the following equation:

$$x_{new} = best + \varepsilon A_i^t * (0.1 * N(0,1) * (x_i^{(t)} - Iter * best)) \quad (14)$$

Where A_i^t representing average loudness and ε is random number over the interval of [-1, 1], 0.1 is controlled factor of exploitation. $N(0,1)$ represents random number generated by Gaussian distribution over the interval of $[0,1]$. $Iter$ is described as current iteration in Jth dimension and $best$ is global best position of bat.

Our enhanced local search approach is proposed for overcoming this exploitation problem, i.e., improves exploitation ability around the best solution of bats and preserves population diversity and obtains a steady result. The algorithm for I-BAT can be found in Algorithm 2.

Algorithm_2: Improved_Bat_Algorithm

```
Input: $x_i = (x_1, ... x_i)^t$ →Bat population
Output: $x^{best}$ & $\min(f(x))$ →optimal solution and minimal objective function value
$x_i$ =Torus_Bat_Init($x_i$); population initialization
Evaluate($x_i$) ;evaluate newly generated population
$Compute(x^{best})$;find current global best
While ($t \leq t_{max}$ ) do
    for $i = 1 ... N_p$ do
        Update frequency using eq.1
        Update velocity using eq.2
        Update position using q.3
        if rand(0,1) > $r_i^{(t)}$ then
            $\boldsymbol{x_j^t}$ = im
prove solution using eq.14
        end
    if $f_{new}$ = Compute $\boldsymbol{x_j^t}$; generate new solution
    evaluate = evaluate +1;
    if $f_{min}^{new} < f_{old}$ and Rand(0,1) < $A_i^{(t)}$ then
    $x_i = \boldsymbol{x_j^t}$; $f_{old} = f_{min}^{new}$;
    end if
    $f_{min} = \min(f(x^{best}))$;
  end for
end while
```

V. Results and Discussion

The proposed improved Bat algorithm (I-BAT) is implemented on the machine with the specification of 2.3 GHz Core (M) 2 Duo CPU processor. To assure the robustness and integrity of proposed algorithms, a collection of nine benchmark test functions has been employed to perform the comparison of proposed I-BAT with Standard Bat algorithm. These are the standard nonlinear benchmark functions usually carried out to investigate the performance of any population-based algorithm in terms of convergence speed, exploitation, and exploration capability. Table I represents the definition of benchmark test function and the properties of these functions. In Table I, $\boldsymbol{f}^*$ shows global minimum of the objective functions$\boldsymbol{f}$, where $\boldsymbol{x}^*$ represents the possible minimum values. The experimental results for proposed techniques are presented in Table II.

A. Parameter Setting

Simulation parameters are fixed as: Population size is 40 where the dimensions for all functions are set to 10, 20 and 30. A number of iterations for 10, 20 and 30 dim are 1000, 2000 and 3000 respectively. For relatively fair and effective results, all techniques have been implemented to similar parameters. All techniques were examined for 30 runs to compare the performances.

B. Analysis

The objective of this research work is to improve the collective performance of bat algorithm by applying two basic modifications in standard bat algorithm. We have modified the method of population initialization and also improved the exploration capability of bat algorithm. The goal of the research is to find how the nature of simulation results relies on dimensions of the benchmark functions for the optimization. For this, dimensions taken for different functions presented in Table I are D=10, D=20, and D=30.

TABLE I. Definitions of Bench mark Test Functions and Properties of Bench-mark Functions

$\boldsymbol{f}$	Function name	Definition	$\boldsymbol{f}^*$	$\boldsymbol{x}^*$	Domain
f_1	Sphere	$Min\, f(x) = \sum_{i=1}^{n} x_i^2$	0.00	(0,0,0…,0)	$-5.12 \leq \mathrm{x_i} \leq 5.12$
f_2	Axis parallel hyper-Ellipsoid	$Min\, f(x) = \sum_{i=1}^{n} i.x_i^2$	0.00	(0,0,0…,0)	$-5.12 \leq \mathrm{x_i} \leq 5.12$
f_3	Schumer Steiglitz	$Min\, f(x) = \sum_{i=1}^{n} x_i^4$	0.00	(0,0,0…,0)	$-5.12 \leq \mathrm{x_i} \leq 5.12$
f_4	Schwefel 1.2	$Min\, f(x) = \sum_{i}^{D} \left(\sum_{j=1}^{i} x_j \right)^2$	0.00	(0,0,0…,0)	$-100 \leq \mathrm{x_i} \leq 100$
f_5	Rotated hyper-Ellipsoid	$Min\, f(x) = \sum_{i}^{n} \left(\sum_{j=1}^{i} x_j \right)^2$	0.00	(0,0,0…,0)	$-65.536 \leq \mathrm{x_i} \leq 65.536$
f_6	Moved axis parallel hyper-Ellipsoid	$Min\, f(x) = \sum_{i=1}^{n} 5i.x_i^2$	0.00	(5*i)	$-5.12 \leq \mathrm{x_i} \leq 5.12$
f_7	Sum of different power	$Min\, f(x) = \sum_{i=1}^{n} \lvert x_i \rvert^{i+1}$	0.00	(0,0,0…,0)	$-1 \leq \mathrm{x_i} \leq 1$
f_8	Sum Squares Function	$Min\, f(x) = \sum_{i=1}^{n} i.x_i^2$	0.00	(0,0,0…,0)	$-10 \leq \mathrm{x_i} \leq 10$
f_9	Zakharov Function	$Min\, f(x) = \sum_{i=1}^{n} x_i^2 + (1/2 \sum_{i=1}^{n} i.x_i)^2 + (1/2 \sum_{i=1}^{n} i.x_i)^4$	0.00	(0,0,0…,0)	$-5 \leq \mathrm{x_i} \leq 10$

TABLE II. COMPERISON OF STANDARD BAT AND IMPROVED BAT

			Standard-BAT					**I-BAT**				
Functions	***Iterations***	***DIM***	***Best***	***Worst***	***Mean***	***Median***	***Std. Dev.***	***Best***	***Worst***	***Mean***	***Median***	***Std. Dev.***
F1	1000	10	7.70E-08	2.16E-07	1.52E-07	1.49E-07	4.08E-08	3.26E-11	8.46E-11	**5.76E-11**	5.47E-11	**1.56E-11**
	2000	20	8.22E-08	2.40E-07	1.38E-07	1.29E-07	4.22E-08	8.30E-10	3.14E-09	**2.19E-09**	2.28E-09	**5.85E-10**
	3000	30	1.29E-07	2.37E-07	1.60E-07	3.25E-08	1.49E-07	3.33E-09	6.77E-09	**5.34E-09**	5.38E-09	**8.78E-10**
F2	1000	10	8.22E-08	1.18E-06	5.93E-07	4.65E-07	2.52E-07	1.64E-09	4.79E-09	**3.03E-09**	2.78E-09	**9.44E-10**
	2000	20	9.14E-07	2.78E-06	1.57E-06	1.41E-06	2.56E-07	8.74E-09	2.74E-08	**1.37E-08**	9.96E-09	**5.61E-09**
	3000	30	2.25E-06	3.23E-06	**3.53E-06**	5.26E-06	**1.18E-06**	6.56E-08	4.26E-05	8.75E-06	9.49E-08	1.62E-05
F3	1000	10	2.84E-15	1.25E-14	7.84E-15	6.97E-15	3.43E-15	1.99E-20	1.11E-18	**3.68E-19**	3.71E-19	**3.03E-19**
	2000	20	1.47E-15	4.51E-15	2.68E-15	2.48E-15	7.72E-16	2.49E-19	1.22E-18	**6.53E-19**	5.12E-19	**2.88E-19**
	3000	30	1.35E-15	3.54E-15	2.31E-15	1.99E-15	7.38E-16	1.85E-18	9.07E-18	**4.41E-18**	3.10E-18	**2.38E-18**
F4	1000	10	1.34E+03	7.31E+03	3.57E+03	2.49E+03	2.06E+03	1.31E-02	9.37E-02	**6.00E-02**	6.62E-02	**2.46E-02**
	2000	20	1.13E+04	2.37E+05	8.69E+04	4.53E+04	7.99E+04	7.28E+02	1.75E+03	**1.38E+03**	1.39E+03	**2.89E+02**
	3000	30	4.50E+04	5.43E+05	2.48E+05	1.45E+05	1.80E+05	4.42E+03	9.64E+03	**7.32E+03**	7.58E+03	**1.53E+03**
F5	1000	10	3.49E+03	7.30E+06	1.49E+06	1.90E+05	2.32E+06	8.59E-19	8.89E-08	**2.37E-08**	5.70E-18	**3.66E-08**
	2000	20	3.94E+07	2.43E+08	1.44E+08	1.43E+08	5.67E+07	5.81E-17	2.51E-07	**8.17E-08**	5.94E-08	**7.79E-08**
	3000	30	3.00E+08	1.06E+09	7.18E+08	7.04E+08	2.06E+08	7.91E-07	1.56E-06	**1.27E-06**	1.26E-06	**2.21E-07**
F6	1000	10	4.11E-07	5.90E-06	3.21E-06	2.32E-06	1.88E-06	8.21E-09	2.39E-08	**1.52E-08**	1.39E-08	**4.72E-09**
	2000	20	4.57E-06	1.39E-05	7.94E-06	7.06E-06	2.67E-06	4.37E-08	1.37E-07	**6.83E-08**	4.98E-08	**2.80E-08**
	3000	30	1.13E-05	2.63E-05	1.72E-05	1.62E-05	4.69E-06	7.05E-07	1.92E-06	**1.22E-06**	9.46E-07	**4.27E-07**
F7	1000	10	1.93E-04	7.91E-04	5.00E-04	4.77E-04	1.70E-04	4.77E-07	3.54E-06	**1.47E-06**	1.06E-06	**1.07E-06**
	2000	20	5.68E-04	1.32E-03	9.24E-04	8.23E-04	2.05E-04	1.57E-07	2.65E-06	**1.36E-06**	1.36E-06	**7.13E-07**
	3000	30	7.14E-04	1.81E-03	1.24E-03	1.11E-03	3.52E-04	4.33E-07	2.86E-06	**1.59E-06**	1.31E-06	**7.53E-07**
F8	1000	10	8.99E-08	1.02E-06	6.77E-07	6.10E-07	2.73E-07	1.64E-09	6.19E-09	**4.01E-09**	4.09E-09	**1.58E-09**
	2000	20	7.98E-07	1.75E-06	1.44E-06	1.43E-06	2.81E-07	2.57E-08	1.03E-07	**5.28E-08**	4.63E-08	**2.10E-08**
	3000	30	2.17E-06	5.87E-06	3.85E-06	3.39E-06	1.05E-06	6.84E-07	1.29E-06	**9.39E-07**	9.27E-07	**1.58E-07**
F9	1000	10	1.11E-07	4.36E-07	3.27E-07	3.35E-07	1.06E-07	3.57E-09	9.61E-09	**5.31E-09**	4.98E-09	**1.56E-09**
	2000	20	1.91E-07	2.62E-02	2.62E-03	3.06E-07	7.87E-03	9.34E-06	1.34E-05	**1.22E-05**	1.21E-05	**1.14E-06**
	3000	30	1.51E+00	1.86E+02	3.46E+01	1.23E+01	5.14E+01	2.08E-05	1.05E-04	**5.89E-05**	5.31E-05	**2.61E-05**

To compare an algorithm with other algorithm for the same nature of problem, its true value will be seen. Hence, I-BAT is compared with standard BA. Both algorithms were used to solve same standard test functions presented for Table I and utilized the same parameter settings explained in Section V. The best simulation results are highlighted as bold. For the fair comparison, the proposed technique improved BA (I-BAT), compared with the standard bat algorithm on nine well-known benchmark test functions. The comparative results are given in Table II. From Table II, we can see that the performance of I-BAT is better as compared to standard Bat algorithm in terms of exploration capability of bats. After a brief analysis on the basis of results, we can conclude that the proposed methods are enough robust in nature to be used for the purpose of numerical optimization problems. The experimental results show that I-BAT outperforms over standard BA in all nine bench mark test functions. Below is the graphical representation of proposed methods on all benchmark test functions. From Fig. 6 to 14 the graphical representation of detailed comparison for $\boldsymbol{f_1}$to $\boldsymbol{f_9}$ has been shown.

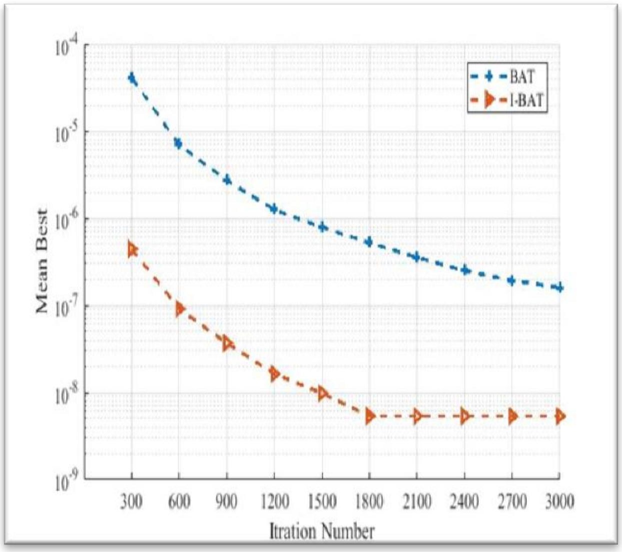


Fig. 6. I-BAT Curve on $\boldsymbol{f_1}$.

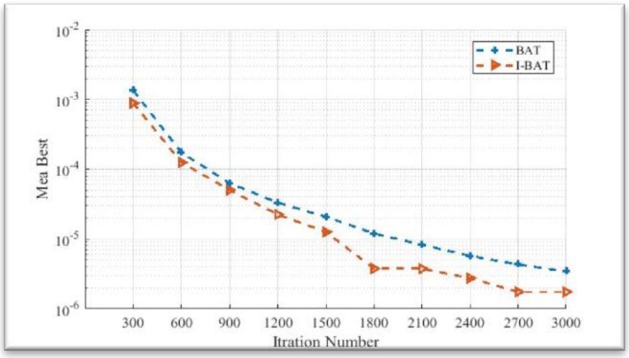


Fig. 7. I- BAT Curve on f_2.

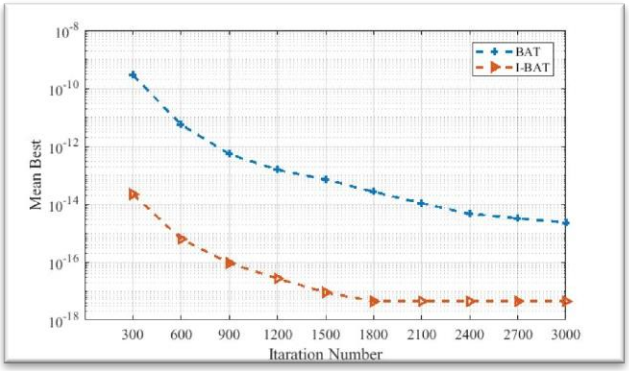


Fig. 8. I- BAT Curve on f_3.

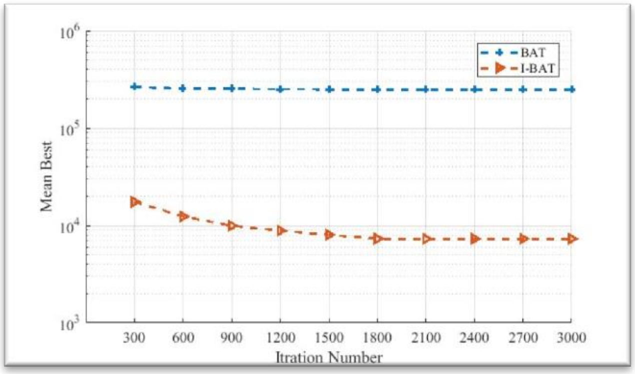


Fig. 9. I- BAT Curve on f_4.

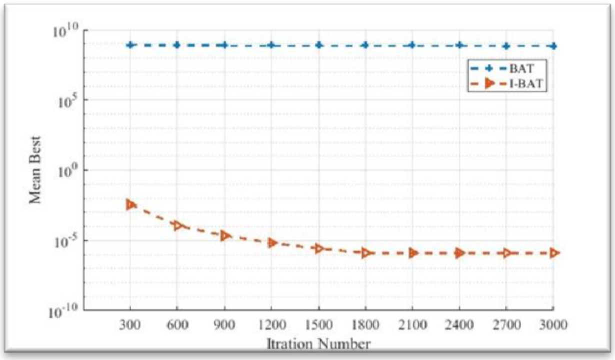


Fig. 10. I- BAT Curve on f_5.

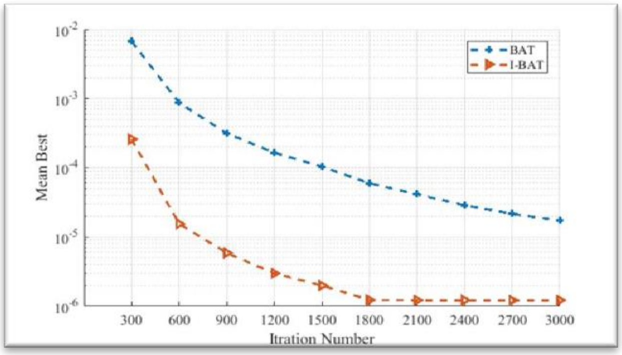


Fig. 11. I- BAT Curve on f_6.

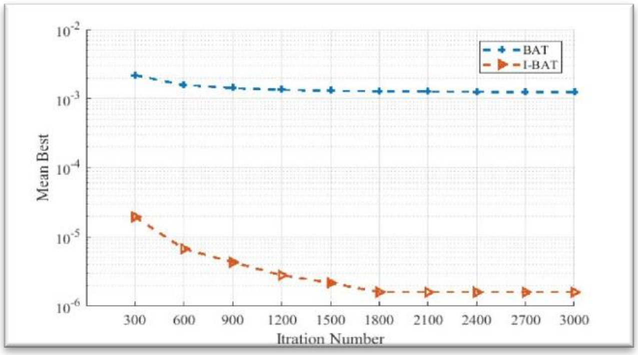


Fig. 12. I- BAT Curve on f_7.

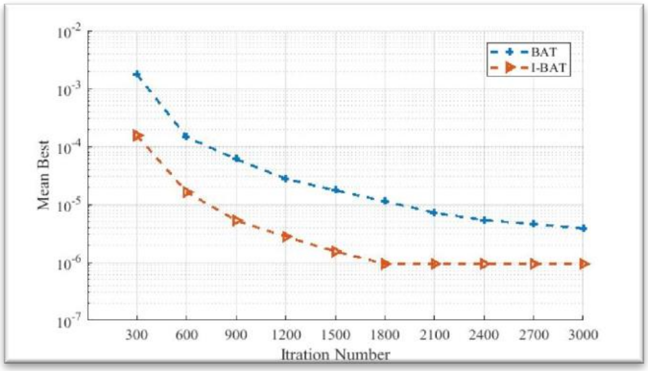


Fig. 13. I- BAT Curve on f_8.

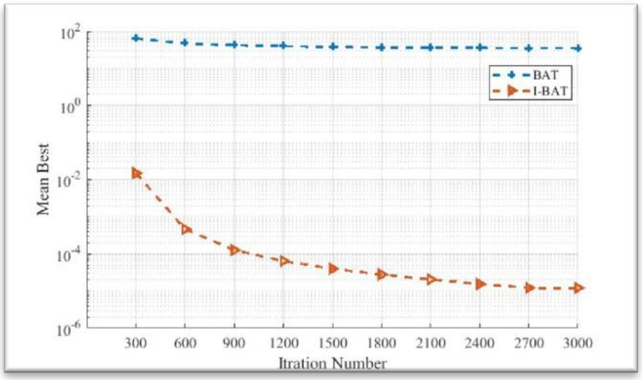


Fig. 14. I- BAT Curve on f_9.

VI. Conclusion

To overcome the issues of exploitation and exploration capabilities of conventional BA, a novel variant of Bat algorithm I-BAT is proposed that would also enhance the searching ability to avoid the local optimum. The novel variant comprises standard BA with strong searching capability joined with novel quasi-random sequence Torus for initialization of swarm and employed on function optimization problems. The proposed strategy maintains the diversity of the swarm and improves the local searching capability. The simulation result shows that the proposed technique has better convergence accuracy and can escape from premature convergence successfully. It also depicts that our designed technique is much better, when it is compared with the standard BA. For future consideration it is appealing to check the performance of new initialization proposed approaches for higher dimensional problems.

References

[1] Fogel, "An introduction to simulated evolutionary optimization," IEEE transactions on neural networks, vol. 5, no. 1, pp. 3-14, 1994.

[2] T. P. Runarsson and X. Yao, "Stochastic ranking for constrained evolutionary optimization," IEEE Transactions on evolutionary computation, vol. 4, no. 3, pp. 284-294, 2000.

[3] X. S. Yang, "Review of meta-heuristics and generalised evolutionary walk algorithm," International Journal of Bio-Inspired Computation, vol. 3, no. 2, pp. 77-84, 2011.

[4] W. H. Bangyal, J. Ahmad, I. Shafi, and Q. Abbas, "A forward only counter propagation network-based approach for contraceptive method choice classification task," Journal of Experimental & Theoretical Artificial Intelligence, vol. 24, no. 2, pp. 211-218, 2012.

[5] G. Beni and J. Wang, "Swarm intelligence in cellular robotic systems," in Robots and Biological Systems: Towards a New Bionics?, Springer, Berlin, Heidelberg, 1993, pp. 703-712.

[6] Christian Blum and Xiaodong Li, "Swarm Intelligence in Optimization," in Swarm Intelligence., 2008, pp. 43-85.

[7] Kolias, G. Kambourakis, and M. Maragoudakis, "Swarm intelligence in intrusion detection: A survey," SciVersa ScienceDirect, vol. 30, pp. 625-642, 2011.

[8] X. S. Yang, "A new metaheuristic bat-inspired algorithm," in Nature inspired cooperative strategies for optimization (NICSO 2010) , Springer, Berlin, Heidelberg, 2010, pp. 65-74.

[9] Xin-She Yang and Amir Hossein Gandomi, "Bat algorithm: a novel approach for global engineering optimization," Engineering Computations: International Journal for Computer-Aided Engineering and Software, vol. 29, pp. 464-483, 2012.

[10] Xin-She Yang, "Bat Algorithm for Multi-objective Optimisation," Int. J. Bio-Inspired Computation, Vol. 3, No. 5, pp.267-27, vol. 3, pp. 267-274, 2012.

[11] S.M. Rahnamayan, H. R. Tizhoosh, and M. Salama, "A novel population initialization method for accelerating evolutionary algorithms," Computers & Mathematics with Applications, vol. 53, no. 10, pp. 1605-1614, 2007.

[12] Nazir Mohd Nawi, M. Z. Rehma, Abdullah Khan, Haruna Chiroma, and Tutut Herawan, "A Modified Bat Algorithm Based on Gaussian Distribution for Solving Optimization Problem," Journal of Computational and Theoretical Nanoscience, vol. 13, pp. 706-714, 2016.

[13] G. G. Wang, B. Chang, and Z. Zhang, "A multi-swarm bat algorithm for global optimization," in Evolutionary Computation (CEC), 2015 IEEE Congress on IEEE, 2015, pp. 480-485.

[14] X. S. Yang and A. Hossein Gandomi, "Bat algorithm: a novel approach for global engineering optimization," Engineering Computations, vol. 29, no. 5, pp. 464-483, 2012.

[15] M. W. U. Kabir and M. S. Alam, "Bat algorithm with self-adaptive mutation: a comparative study on numerical optimization problems," International Journal of Computer Applications, vol. 100, no. 10, 2014.

[16] Russell, Eberhart, and Y. H. Shi, "Comparison between genetic algorithms and particle swarm optimization," in International conference on evolutionary programming. Berlin, 1998, pp. 611-616.

[17] S. Yilmaz and E. U. Kucuksille, "Improved bat algorithm (IBA) on continuous optimization problems," Lecture Notes on Software Engineering, vol. 1, no. 3, p. 279, 2013.

[18] G. Rudolph, "Self-adaptive mutations may lead to premature convergence," IEEE Transactions on Evolutionary Computation, vol. 5, no. 4, pp. 410-414, 2001.

[19] Padmavathi Kora and Sri Ramakrishna Kalva, "Improved Bat algorithm for the detection of myocardial infarction," SpringerPlus, vol. 4, p. 666, 2015.

[20] Jiann-Horng Lin, Chao-Wei Chou, Chorng-Horng Yang, and Hsien-Leing Tsai, "A Chaotic Levy Flight Bat Algorithm for Parameter Estimation in Nonlinear Dynamic Biological Systems," Journal of Computer and Information Technology, vol. 2, pp. 56-63, 2011.

[21] Jiawei Zhang and Gaige Wang, "Image Matching Using a Bat Algorithm with Mutation," Applied Mechanics and Materials, vol. 203, pp. 88-93, 2012.

[22] Fábio A. P. Paiva, Marcos H. F. Marcone, Cláudio R. M. Silva, and Izabele V. O. Leite, "Modified Bat Algorithm With Cauchy Mutation and Elite Opposition-Based Learning," in Computational Intelligence (LA-CCI), 2017 IEEE Latin American Conference on., 2017, pp. 1-6.

[23] Ahmed Fouad Ali, "Accelerated Bat Algorithm for Solving Integer Programming Problems," Egyptian Computer Science Journal, vol. 39, pp. 25-40, 2015.

[24] Gaige Wang and Lihong Guo, "A Novel Hybrid Bat Algorithm with Harmony Search for Global Numerical Optimization," Journal of Applied Mathematics, vol. 2013, 2013.

[25] Iztok Fister Jr., Dusan Fister, and Xin-She Yang, "A Hybrid Bat Algorithm," arXiv preprint arXiv:1303.6310, 2013.

[26] S. Yılmaz, E. Ugur Kucuksille, and Y. Cengiz, "Modified Bat Algorithm," vol. 20, pp. 71-78, 2014.

[27] Iztok Fister Jr., Simon Fong, Janez Brest, and Iztok Fister, "A Novel Hybrid Self-Adaptive Bat Algorithm," The Scientific World Journal, 2014.

[28] Osama Abdel-Raouf, Mohamed Abdel-Baset, and Ibrahim El-Henawy, "An Improved Chaotic Bat Algorithm for Solving Integer Programming Problems," IJCEM International Journal of Computational Engineering & Management, vol. 17, pp. 2230-7893, 2014.

[29] Asma CHAKRI, Rabia KHELIF, Mohamed BENOUARET, and Xin-She YANG, "New directional bat algorithm for continuous optimization problems," Expert Systems with Applications, vol. 69, pp. 159-175, 2017.

[30] S. Kotz, N. Balakrishnan, and N. L. Johnson, Continuous multivariate distributions, Models and applications ed.: John wiley & sons, 2004, vol. 1.

[31] V. V. Nikulin and I. R. Shafarevich, Geometries and groups.: Springer Science & Business Media, 2012.

Effects of Modulation Index on Harmonics of SP-PWM Inverter Supplying Universal Motor

Asif A. Solangi
Mehran UET SZAB Campus,
Khairpur Mir's Sindh Pakistan

Rameez Shaikh
Sukkur IBA University, Sindh
Pakistan

Noman Khan Pathan
Mehran UET SZAB Campus,
Khairpur Mir's Sindh Pakistan

Mehr Gul
Balochistan University of
Information Technology,
Engineering and Management
Sciences, Balochistan Pakistan

Farhana Umer
The Islamia University of
Bahawalpur, Punjab Pakistan

Zeeshan Anjum Memon
Mehran UET SZAB Campus,
Khairpur Mir's Sindh Pakistan

***Abstract*—This manuscript presents the effects of changing modulation indices on current and voltage harmonics of universal motor when it is supplied by single phase PWM (SP-PWM) inverter, the effect has been analyzed with simulation and experimental setup. For variable speed applications universal motor can be controlled either by phase angle control drive or by SP-PWM inverter drive. SP-PWM inverter-fed drive is common technique that is used to adjust the voltage applied to motor, so that variable speed can be obtained. With the application of SP-PWM inverter-fed drive, harmonics are generated because of power electronic devices. According to the IEEE standard 519, the total harmonic distortion (THD) must be within 5%. In this paper, the effect of modulation index (MI) is used to analyze THD content, and its variation alters the harmonic content. However, the effects are also analyzed through experimental setup in order to validate the system performance. In future work, keeping modulation index constant, different PWM strategies can be employed in order to decrease harmonics.**

***Keywords*—Harmonics; modulation index; SP-PWM inverter; universal motor**

I. Introduction

In recent years speed control of universal motor by using SP-PWM inverter fed drive is widely used because of its efficient control [1], [22]. In operation, universal motor is very similar to series dc motor, but unlike dc series it can be operated on ac voltage also. Universal motor is designed to operate on either dc or ac. Universal motor is different from dc series motor construction wise. Direction of torque is same for any current polarity as well as ac current. Universal motor have some good features due to its more power related to its size and weight. As compared to induction motor universal motor is popular because of its high speed ranging from (1500 to 2000 RPM). Universal motor is suitable for washing machines, drills and dust extractors. Universal motor have also some drawbacks because of sparking produced by commutator segments. And also lower life time and loud noise as compared to induction motor. Due to advancement in power electronics technology, power inverters are widely used in power systems, household appliances, transportation, specifically in variable frequency drives [2]. Speed of universal motor can be controlled by using two types of controllers i.e. phase angle control (using TRIAC or Thyristor) or by with PWM converters (Using IGBT) [3]. Using such type of convertors speed of universal motor can be controlled by adjusting output voltage of converter. As universal motor is widely used in home appliances so the controllers are designed to operate on ac voltage [4]. SP-PWM inverter -fed drive is common technique that is used to adjust the voltage applied to motor [5]. With PWM inverter-fed drive harmonics are generated because of application of power electronic devices used in the circuitry. The harmonics produced must be within allowable limit of 5% as recommended by IEEE standard 519, so that other loads may not be affected by harmonics produced. Different methods are proposed to decrease THD content generated by SP-PWM inverter-fed drive supplying universal motor. Voltage source inverters are implied in Sinusoidal pulse width modulation techniques, VSI are remained area of interest for researchers. In SP-PWM inverter -fed drive, some parameters effect the production of harmonic content, out of these parameters modulation index should be changed to control amplitude and speed. So variation of modulation index is analyzed [6]. With the application of single phase PWM inverter-fed drive the iron loss in electric sheets of stator and rotor can be reduced significantly [5]. The increased awareness of harmonics in recent years is the result of concerns that harmonic distortion levels are increasing on many electrical power systems [7], however, this is the hot research direction now-a-days, the emphasis is on the cost effective exploitation regarding power quality and reliability [8]. Today’s one of the most serious issue due to which industrial consumers are suffering financial losses enforced by utility regulations is power quality. Out of several PQ disturbances THD is most common type of disturbance, thus it’s necessary to analyze THD content due to usage of power electronic components used in household appliances to guarantee the more pure form of power to customers [9]. Due to these harmonics, problems are created which include interference with communication lines, degradation of insulation levels of equipments, heating of winding, excessive currents, increased power losses which ultimately reduce life of equipment. Harmonics affect power quality and increase system losses up to 20% out of which 27% may be attributed to harmonics [10-11]. Recent studies reveal that the additional power losses in distribution networks may be in range of 4–8.5% for various harmonic levels [12]. Power electronic converters are used in VFDs and ultimately causes elevated harmonics into the system. Design engineers and semiconductors suppliers are interested in energy efficient and low cost variable frequency drives. PWM based variable switching frequency technique is proposed to reduce total harmonic distortions [13], in addition, the use of very high range of switching frequency is not suitable for semiconductor switches of real grid-connected inverters [14]. As per the IEEE standards 519, current and voltage total harmonic distortion must be restricted lower than 5%. PWM controllers operating at elevated index of modulation causes decline in THD. In universal motor, steady speed control system is supplied by pulse width modulation controlled through ac chopper, additionally providing enhanced range of speed and ultimately reacting even healthier under abrupt alterations in load [15]. In this research work development of simulation models of universal motor and PWM controlled single phase inverter-fed drive is presented by using Matlab/Simulink/Simscape software. A simulation model is developed for harmonic analysis of universal motor when it's fed by SP-PWM controlled single phase inverter-drive. Same work is done experimentally, by using a reference variable generator, PWM controller, single phase IGBT driver set. Harmonic analysis is done with different modulation indices. Experimental and simulation results are observed nearly equal. Harmonic contents are lower at higher modulation indices. Section III contains equations related to voltage and current for universal motor. Different speed control techniques are also discussed in this section. Simulation model for Universal motor supplied by single phase inverter is described in Section IV. Detail of experimental setup for harmonic analysis of the system is also given in this section. Results are compared and discussed. Finally paper is concluded in Section V.

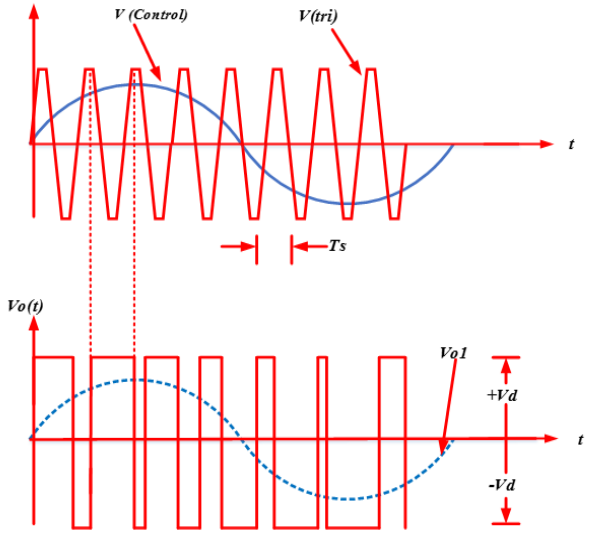


Fig. 1. Sinusoidal pulse width modulation.

II. Sinusoidal Pulse Width Modulation

Many industrial applications use variable speed drives, and PWM is mostly common technique used in these variable speed drives. Among different PWM techniques sinusoidal pulse width modulation is widely used. In sinusoidal pulse width modulation width of gating pulse is changed in accordance with the width of reference signal. The control signals are generated by comparing sinusoidal wave with triangular or squire wave as shown in Fig. 1, the frequency of reference signal determines the inverter output frequency, in result it controls the modulation index whereas number of pulses per half cycle depends upon the frequency of carrier signal. DC input voltage is fed to SP-PWM for producing sinusoidal wave at prescribed frequency. There are two SP-PWM techniques for H-bridge inverter which are unipolar and bipolar switching [16] SPWM technique has reduced power loss, because of its switching devices remain almost off (Low current means low power) and remain hardly on (Low voltage, low power) PWM signals can easily be generated by using modern microcontrollers [17].

III. Speed Control of Universal Motor

Universal motor is mostly supplied by single phase ac voltages [18], [21]. Fig. 2 shows schematic diagram for universal motor. System of equations for field voltages and voltage of armature are shown in (1).

$$\begin{pmatrix} v_f \\ v_a \end{pmatrix} = \begin{pmatrix} R_f + L_f P & 0 \\ -M_\omega & R_a + L_a P \end{pmatrix} \begin{pmatrix} i_f \\ i_a \end{pmatrix} \quad (1)$$

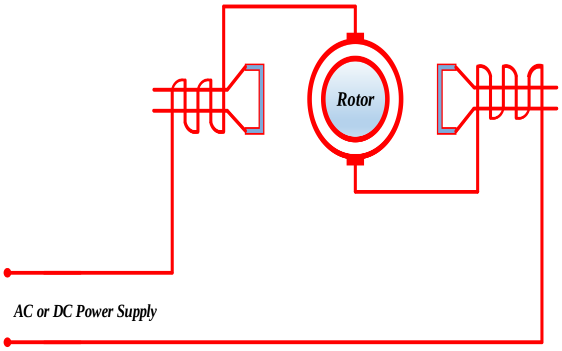


Fig. 2. Schematic diagram of universal motor.

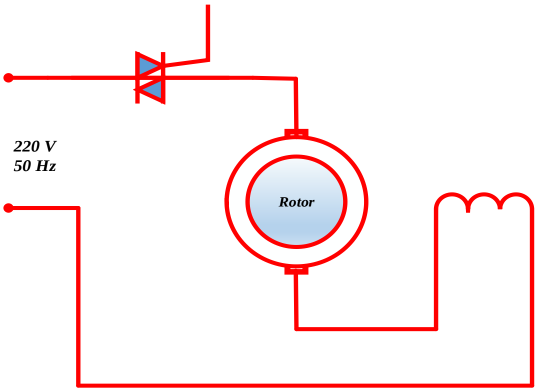


Fig. 3. Speed control of motor using TRIAC.

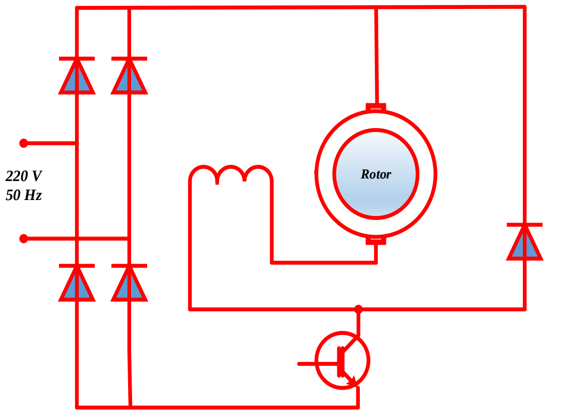


Fig. 4. Speed control of motor using chopper.

Where, the field and armature currents are related by (2).

$$\begin{pmatrix} i_f \\ i_a \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} [i(t)] \tag{2}$$

The voltage across the winding are related by

$$v(t) = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} v_f \\ v_a \end{pmatrix} \tag{3}$$

By using (2) and (3), the voltage equation can be rewritten as

$$v(t) = \begin{pmatrix} R & LP & M_{\omega} \end{pmatrix} i(t) \tag{4}$$

Where

$$R = R_a + R_f \tag{5}$$

$$L = L_a + L_f \tag{6}$$

Speed of universal motor is mostly controlled by using ac chopper and phase angle control method. TRIAC is used in phase angle control technique to control the amount of voltage supplied to the universal motor as shown in Fig. 3. In the chopper control application, voltage applied to the motor is adjusted by using PWM technique. Amount of voltage supplied to motor is changed by variations in duty cycle of PWM signal [19]. Carrier frequency also affects the THD [20]. AC power chopper needed one more stage and it consists of power switch rectifier and fast switching power diode. Simplified chopper control technique using IGBT is shown in Fig. 4. Industrial applications requiring superior performance are using PWM based VFDs. Improvement in power electronics control circuits is caused by recent developments in power semiconductor industry. Hence, different circuit configurations, namely, PWM inverters have become popular. A number of PWM schemes are used to obtain variable voltage and frequency supply. The most widely used PWM scheme for voltage source inverters is sinusoidal PWM [16]. In accordance with the sine function, the inverter output voltage and frequency can be controlled by using SPWM technique. In this scheme the width of each pulse is varied in proportion to the amplitude of a sine wave evaluated at the center of same pulse. The gating signals are generated by comparing a sinusoidal reference signal with a triangular carrier wave of frequency f_c. The frequency of reference signal f_r determines the inverter output frequency f_o; and its peak amplitude A_r controls the modulation index M, and then in turn the RMS. The output voltage V_o if δm is the width of m the pulse the output RMS voltage can be determined by

$$V_o = \left(\sum_{m=1}^{2p} \frac{\delta m}{\pi} \right) \tag{7}$$

IV. Simulation and Experimental Details

Fig. 5 shows the simulation model of single phase pulse width modulator inverter fed supplying universal motor. A physical system in Simulink environment is introduced by the model consisting of Simscape. Simscape is used to make control block diagrams of universal motor, while single phase pulse width modulator is built in Simulink. Pulse width modulator generator helps in altering index of modulation and switching frequency. The fast fourier transform is presented as a tool for analysis for various indices of modulation and frequency of switching. Ideal block of rotational motion sensor represents ideal mechanical rotational sensor. Ideal block of rotational motion sensor helps in conversion of variable (across) between 2 nodes of mechanical rotation into angular velocity in proportion to signal of control. The block of ideal torque source denotes ideal mechanical source. The simulation parameters to model universal motor are J=2e-4; B=1e-6, R_a+Rs=90Ω; L_a+L_s=181 mH

Experimental setup of harmonic analysis of universal motor when it's fed by SPWM controlled single phase inverter-fed drive is shown in Fig. 6. Starting from the left dc Power supply module that is required to provide dc power supply is to reference variable generator for single phase PWM controller. The output voltage of reference variable generator is used as command input voltage of the SP-PWM controller. SP-PWM controller module is used to generate gating signals of IGBT driver set module. IGBT driver set produces dc at its output terminals from dc voltage at its input. Frequency of output voltage depends upon the frequency of command voltage and the peak value of reference signal controls the average output voltage. Power quality analyzer is used to measure THD for current and voltages, it also display harmonic orders with respect to fundamental frequency. Harmonic spectrum response is taken at different M.I. Fig. 7 shows FFT analysis for current of system under discussion for carrier frequencies of 1 kHz with modulation index of 0.6. It shows a THD$_i$ of 50.10%. Both even and odd harmonics are present in harmonic spectrum. As shown in Fig. 8 waveform is not pure sinusoidal because it's obtained at lower modulation index.

Fig. 9 shows experimental results for harmonic spectrum of system with 1 kHz carrier frequency with modulation index of 0.6. It shows THD$_i$ of 56.1% which is nearly equal to simulation results, quite acceptable considering system parameter variations with temperature, parasitic effects in motor as well as single phase PWM inverter circuit, wiring and supply voltage.

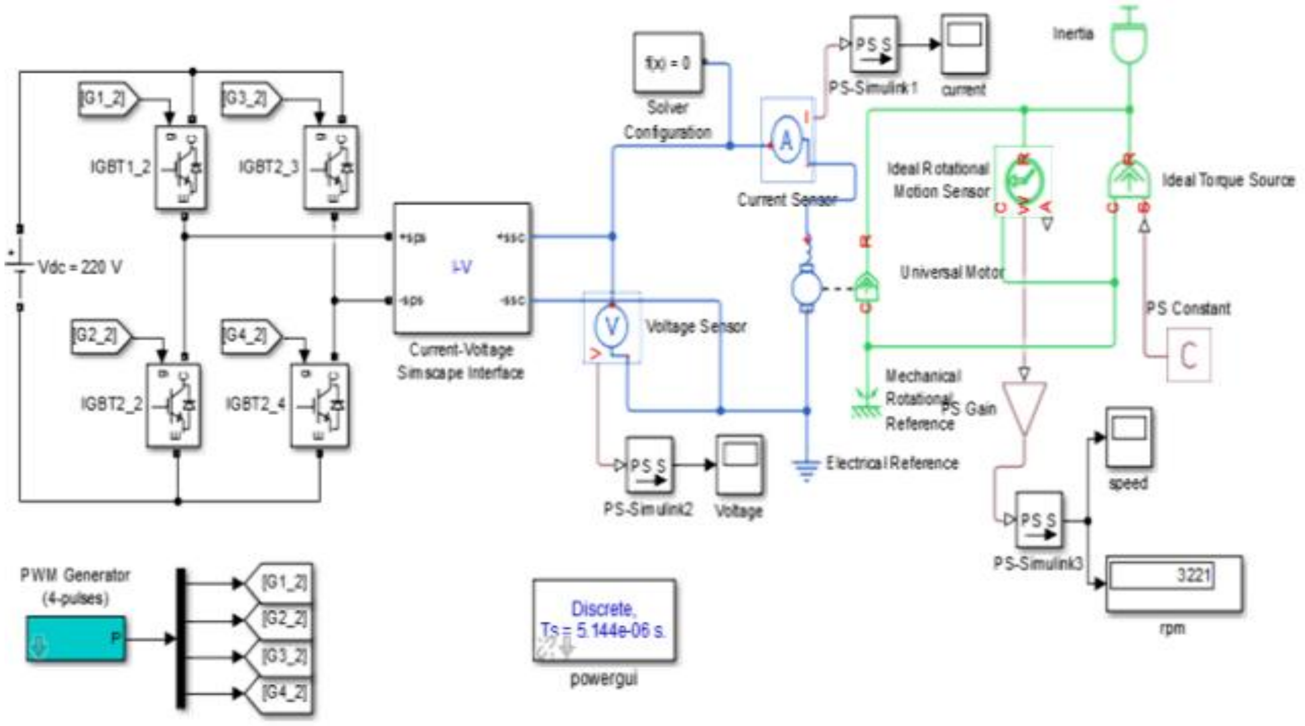


Fig. 5. Matlab based simulation model of universal motor fed by single phase PWM inverter.

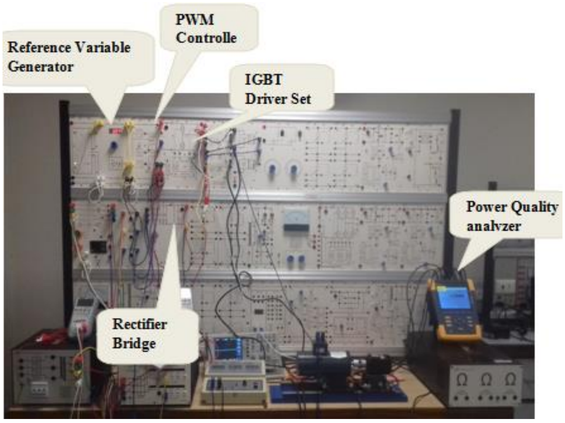


Fig. 6. Experimental setup for universal motor fed by SP-PWM inverter.

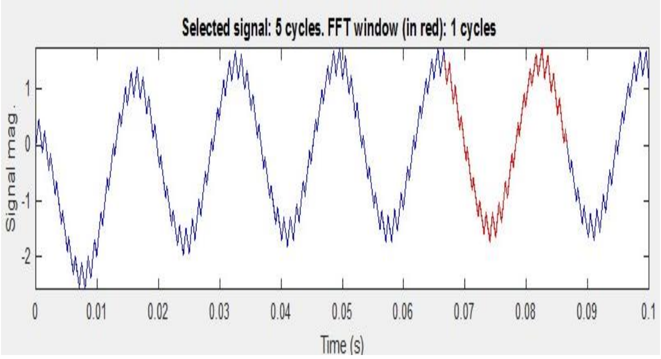


Fig. 8. Output current waveform using simulation at modulation index of 0.6 with 1 kHz carrier frequency.

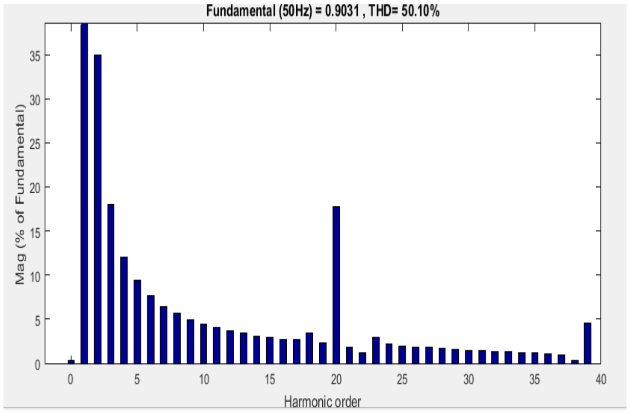
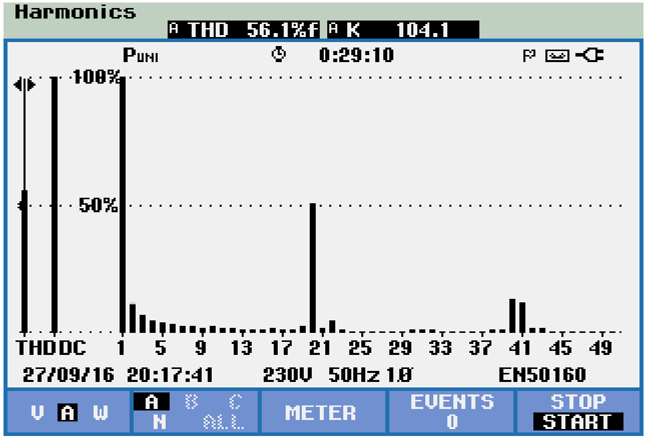


Fig. 7. Simulation based current harmonic spectrum response at 1 KHz frequency with M.I of 0.6.


Harmonics
A THD 56.1%f A K 104.1
PUNI 0:29:10
100%
50%

27/09/16 20:17:41 230V 50Hz 1Ø EN50160

Fig. 9. Experimental based current harmonic spectrum response at 1 Khz frequency with M.I of 0.6.

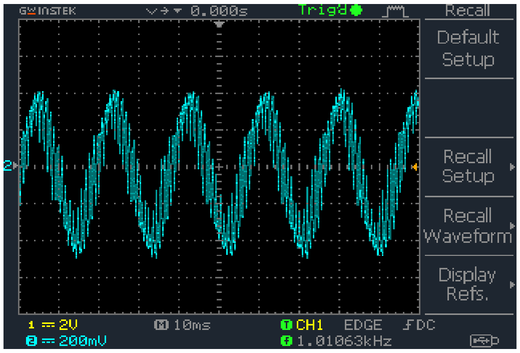


Fig. 10. Output current waveform using experimental setup at modulation index of 0.6 with 1 KHz carrier frequency.

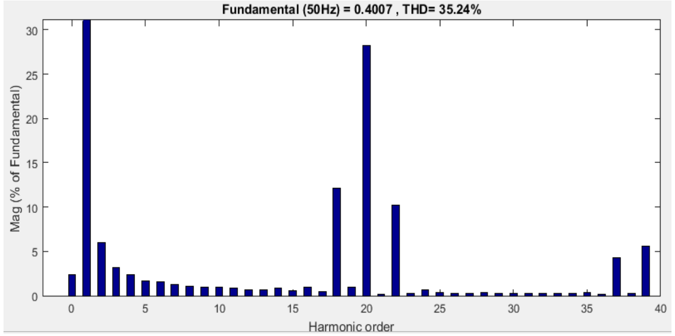


Fig. 11. Simulation based current harmonic spectrum response at 1 kHz frequency with M.I of 0.9.

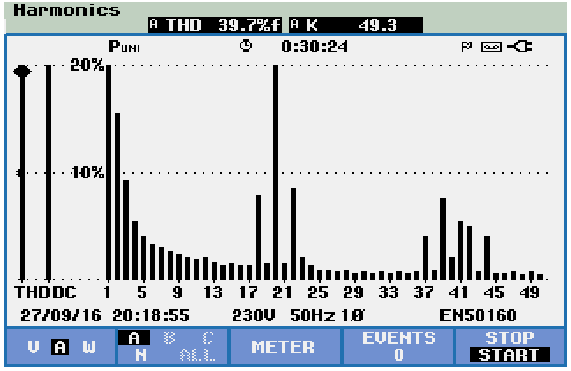


Fig. 12. Experimental based current harmonic spectrum response at 1 kHz frequency with M.I of 0.9.

Fig. 10 shows output current wave form with experimental setup at modulation index of 0.6 with distorted waveform by further increasing modulation index at value of 0.9, and simulation results for current harmonics are obtained, Fig. 11 and 12, respectively show simulation and experimental harmonic analysis for current of system under discussion with modulation index of 0.9 and it is observed that harmonics are reduced considerably with increasing modulation index when its compared with FFT analysis of current at modulation index of 0.6. Increased modulation index i.e. 0.9 affects the output waveform current; it is shown in Fig. 13, that simulation based output current waveform is more sinusoidal then current waveform observed at modulation index of 0.6. Experimental output waveform form for current is shown in Fig. 14.

Fig. 16 and 17, respectively show simulation and experimental harmonic analysis for current of system under discussion with modulation index of 1.

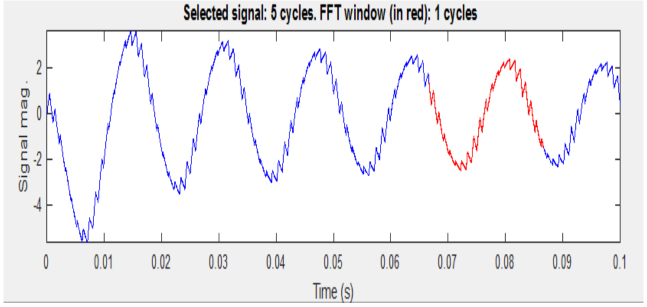


Fig. 13. Output current waveform using simulation at modulation index of 0.9 with 1 kHz carrier frequency.

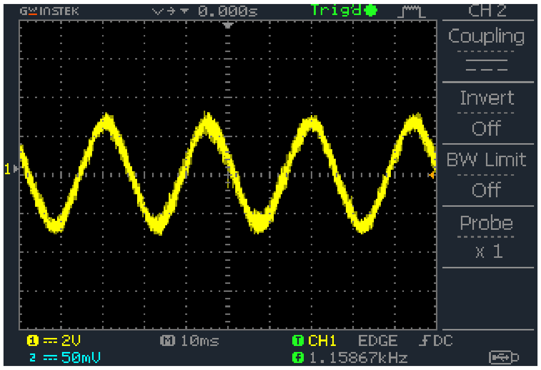


Fig. 14. Output current waveform using experimental setup at modulation index of 0.9 with 1 kHz carrier frequency.

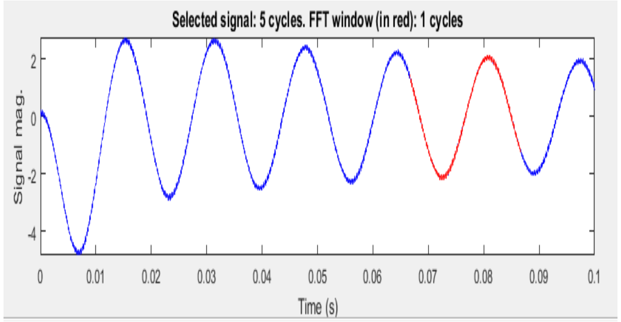


Fig. 15. Simulation based current harmonic spectrum response at 1 kHz frequency with M.I of 1.

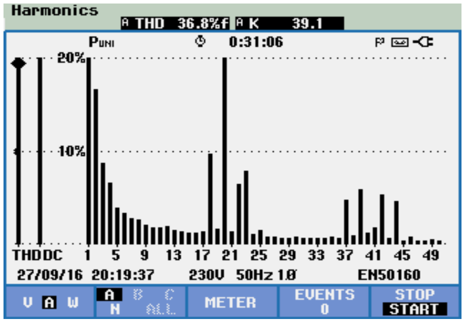


Fig. 16. Experimental based current harmonic spectrum response at 1 kHz frequency with M.I of 1.

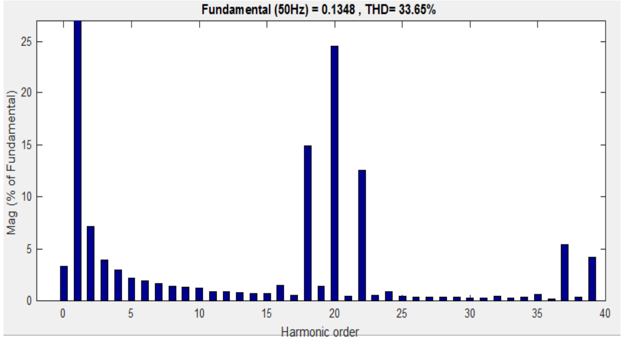


Fig. 17. Output current waveform using simulation at modulation index of 1 with 1 kHz carrier frequency.

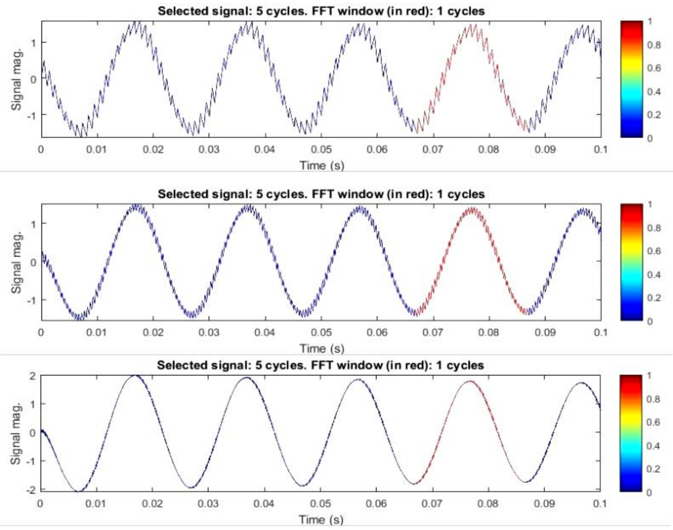


Fig. 18. Simulation based output current wave form at three different modulation indices.

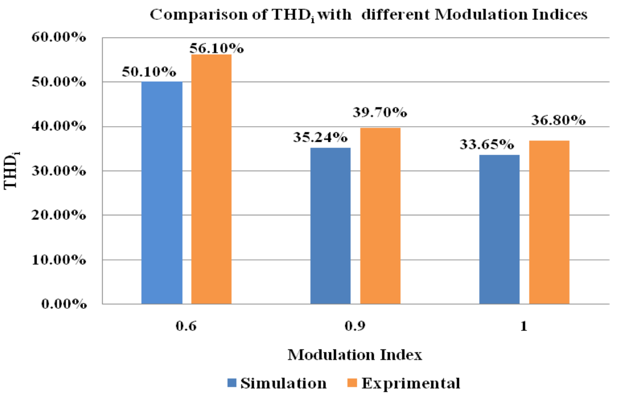


Fig. 19. THD$_i$ at different modulation indices at constant carrier frequency of 1 kHz.

At modulation index of 1, output current waveform is pure sinusoidal as shown in Fig. 15. Comparison of output current waveforms with different modulation indices is shown in Fig. 18 and it's observed that when M.I is fixed at 1, then output current waveform is more sinusoidal. Results of THD$_i$ for different modulation indices are shown in. Carrier frequency is fixed at 1 KHz. It is clear that increasing modulation index current harmonics are decreasing; it's also observed that best output signal is produced by increasing the modulation index graphically it's shown in Fig. 19. Current harmonics will cause harmonics in voltage. These voltage harmonics will affect operation of other appliances connected to the supply network. THDv for three modulation indices were also observed using simulation and experimental arrangements. Fig. 20 shows FFT spectrum of voltage harmonics at modulation index of 0.6 with keeping the carrier frequency constant at 1 kHz and modulation index of 0.6, experimental results of voltage harmonics are shown in Fig. 21.

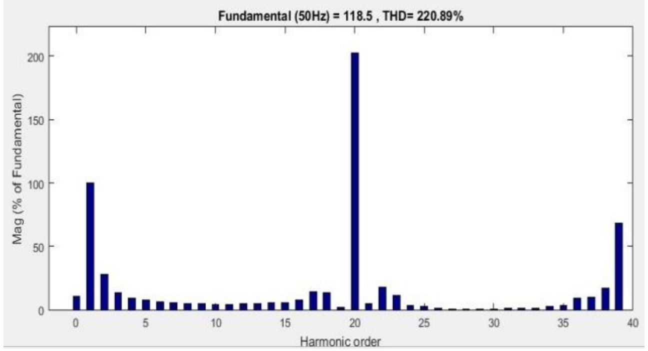


Fig. 20. Simulation based voltage harmonic spectrum response at 1 kHz frequency with M.I of 0.6.

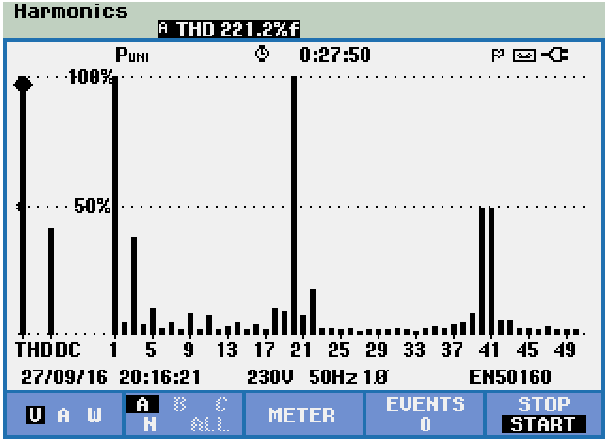

Harmonics
THD 221.2%f
PUNI
0:27:50
100%
50%
THDDC 1 5 9 13 17 21 25 29 33 37 41 45 49
27/09/16 20:16:21 230V 50Hz 1Ø EN50160
V A W
A N
B C ALL
METER
EVENTS 0
STOP START

Fig. 21. Experimental based voltage harmonic spectrum response at 1 kHz carrier frequency with M.I of 0.6.

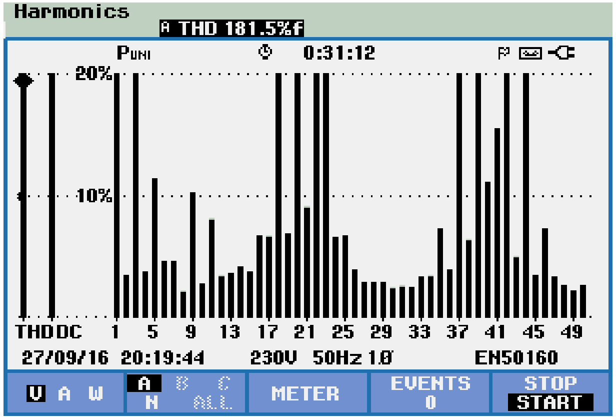

Harmonics
THD 181.5%f
PUNI
0:31:12
20%
10%
THDDC 1 5 9 13 17 21 25 29 33 37 41 45 49
27/09/16 20:19:44 230V 50Hz 1Ø EN50160
V A W
A N
B C ALL
METER
EVENTS 0
STOP START

Fig. 22. Experimental based voltage harmonic spectrum response at 1 kHz frequency with M.I of 0.9.

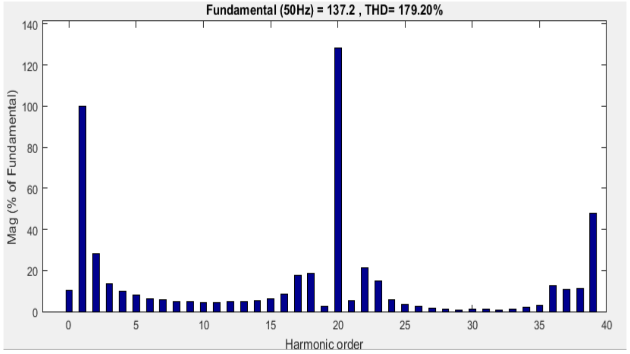


Fig. 23. Simulation based voltage harmonic spectrum response at 1 kHz frequency with M.I of 0.9.

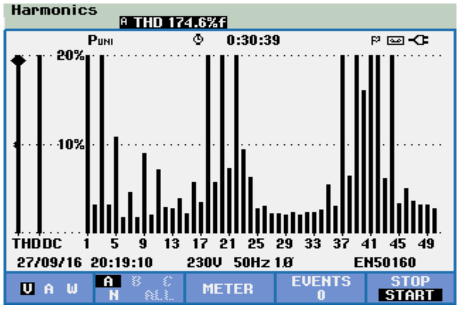


Fig. 24. Experimental based voltage harmonic spectrum of PWM based inverter fed universal motor with 1 kHz carrier frequency with modulation index of 1.

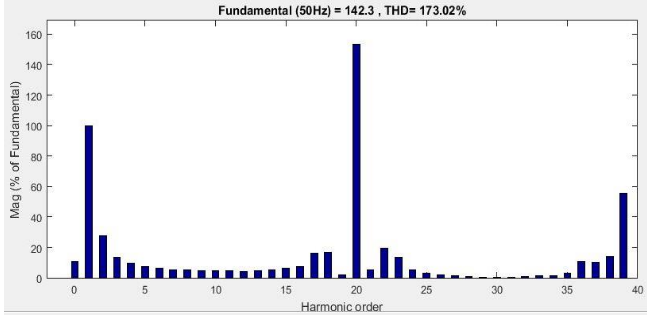


Fig. 25. Simulation based voltage harmonic spectrum response at 1 kHz frequency with M.I of 1.

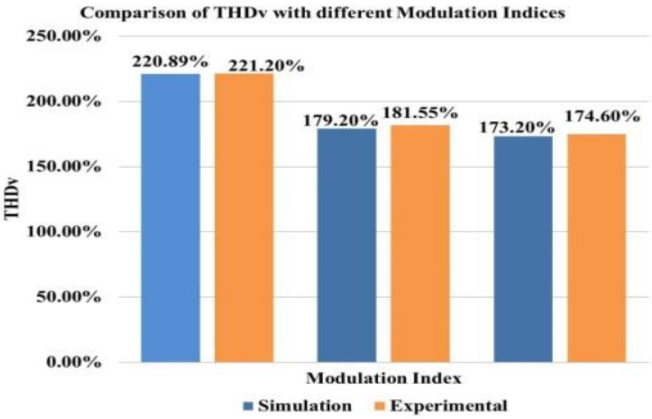


Fig. 26. THD$_v$ at different modulation indices at constant carrier frequency of 1 kHz.

With further increasing modulation index it can be observed that harmonic contents are reducing significantly Fig. 22 and 23 shows the harmonic content using simulation and experimental setup at modulation index of 0.9 but with same carrier frequency with further increase of modulation index, the voltage harmonic content are reducing which can be observed from Fig. 24 and 25 for both simulation and experimental setup at modulation index of 1. Comparison of THDv for different modulation indices are graphically shown in Fig. 26.

V. Conclusion

Most of household appliances use universal motor in their operation and some of applications in industries. Its output torque is controlled through variations in supply voltage. When universal motor is operated for variable speed or high torque applications by incorporating power electronic converters in result harmonics are generated. Nature of harmonics generated imparts adverse effects on performance of universal motor and also other connected equipment. In this research work, harmonic analysis of single phase pulse width modulator inverter with various indices of modulation supplying universal motor is the focus of study. However at very higher switching frequencies losses will be more, this limits the work. A single phase pulse width modulation inverter supplying universal motor is modeled in MATLAB. In order to validate simulated results experimental arrangement is also established. Results show that content of voltage and current harmonics are more than standard limit of 5% as set by IEEE Standard 519. It is also concluded that THD decreases with increasing modulation index. In future work, keeping modulation index constant, different PWM strategies can be employed in order to decrease harmonics.

Acknowledgments

This work is highly acknowledged by Mehran University of Engineering & Technology SZAB Campus, Khairpur Mirs for their support regarding reserch conducted at Labaorory, and we would like to say thanks to Engr. Dr. Mazhar Hussain Baloch Assistant Professor (Electrical Engineering), Mehran University of Engineering & Technology SZAB Campus, Khairpur Mirs for their support regarding review our paper and some suitable guidence about paper publication.

References

[1] Tahir, S., Wang, J., Baloch, M. H., & Kaloi, G. S. (2018). "Digital control techniques based on voltage source inverters in renewable energy applications" *A Review on Electronics*, *7*(2), 18.

[2] J. Song, X. Zhang, L. Zheng, Y. Gao, and Y. Song (2015) “Simulation and experiment of three-phase voltage SPWM inverter” *Proceedings of IEEE 10th Conference on Industrial Electronics and Applications*" pp. 1324-1329.

[3] Y. S. Lai (2003) “Machine modeling and universal controller for vector-controlled induction motor drives” *IEEE Trans. on Energy Conversion*, vol. 18, no. 1, pp. 23-32

[4] Kaňuch, J., & Višnyi, P. (2009) "DC drive for universal motor" *Maszyny elektryczne: Zeszyty problemowe* pp. 7-11.

[5] Bodur, H., A.F. Bakan, and M.H. Sarul, (2000) “Universal motor speed control with current controlled PWM AC chopper by using a microcontroller” *Proceedings of IEEE International Conference on Industrial Technology, Goa, India*. pp. 394-398.

[6] Hassan Feshki, F., & Sarabadani, H. (2011). "Modulation index effect on the 5-level SHE-PWM voltage source inverter" *Engineering*.

[7] Kumar, M., Memon, Z. A., Uqaili, M. A., & Baloch, M. H. (2018). "An Overview of Uninterruptible Power Supply System with Total Harmonic Analysis & Mitigation: An Experimental Investigation for Renewable Energy Applications" *IJCSNS*, *18*(6), 25.

[8] Baloch, M. H., Wattoo, W. A., Kumar, D., Kaloi, G. S., Memon, A. A., & Tahir, S. (2017). "Active and Reactive Power Control of a Variable Speed Wind Energy Conversion System based on Cage Generator" *International Journal of Advanced Computer Science and Applications*, Vol *8*(9), pp.197-202.

[9] Halepoto, I.A, F.R. Abro, A.R. Chachar, Adeena, (2016) "Power Quality Assessment of Compact Fluorescent Lamps," *Sindh University Research Journal-SURJ (Science Series),* pp. 407-412

[10] Kharlov, N.N., Borovikov, V.S., Ushakov, V.Y., Tarasov, E.V. and Bulyga, L.L (2016) "Calculation of steady non-sinusoidal modes and electric power losses in complex electrical networks" *International Conference on Power Electronics and Motion Control (PEMC)* pp. 336-341

[11] Neagu, B. C., Georgescu, G., & Ivanov, O. (2016). "The impact of harmonic current flow on additional power losses in low voltage distribution networks" (2016) *International Conference and Exposition on* Electrical *and Power Engineering (EPE)* pp. 719-722

[12] Kalair, A., Abas, N., Kalair, A.R., Saleem, Z. and Khan, N (2017) "Review of harmonic analysis, modeling and mitigation techniques" *Renewable and Sustainable Energy Reviews*, pp.1152-1187

[13] Mao Xiaolin, Ayyanar Rajapandian, Krishnamurthy Harish K. (2009) "Optimal variable switching frequency scheme for reducing switching loss in single-phase inverters based on time-domain ripple analysis" *IEEE Trans Power Electrical* pp. 991–1001

[14] Tran, Quang-Tho, Anh Viet Truong, and Phuong Minh Le (2016) "Reduction of harmonics in grid-connected inverters using variable switching frequency." *International Journal of Electrical Power & Energy Systems* pp.242-251.

[15] Wang, R.H., and R, T.Walter, (2000) "Modeling of universal motor performance and brush commutation using finite element computed inductance and resistance matrices," *IEEE Transactions on Energy Conversion*, pp. 257-263.

[16] Aboadla E.H., S. Khan, M.H. Habaebi, T. Gunawan, B.A. Hamidah, and M.B. Yaacob (2016) "Effect of modulation index of pulse width modulation inverter on Total Harmonic Distortion for Sinusoidal," *In IEEE International Conference on Intelligent Systems Engineering (ICISE), Islamabad, Pakistan,* pp. 192-196.

[17] Namboodiri, A., & Wani, H. S. (2014). "Unipolar and bipolar PWM inverter" *International Journal for Innovative Research in Science & Technology*, pp. 237-243.

[18] Mirafzal, B., G.L. Skibinski, R.M. Tallam, D.W. Schlegel, and R.A. Lukaszewski, (2007) "Universal induction motor model with low-to-high frequency-response characteristics," *IEEE Transactions on Industry Applications*, pp. 1233-1246

[19] Barge, S.A., and D.R. Jagtap, (2013)"Harmonic Analysis of Sinusoidal Pulse Width Modulation," *International Journal of Advanced Electrical and Electronics Engineering,* pp. 13-16.

[20] Solangi, A., Sahito, A., Soomro, S., Khatri, S., & Memon, M. (2016). "Harmonic Analysis of Universal Motor Supplied by Single Phase PWM Inverter-Fed Drive". *Sindh University Research Journal-SURJ (Science Series)*, Vol. 48 (4) pp.769-774.

[21] Memon, A. A., Shah, S. A. A., Shah, W., Baloch, M. H., Kaloi, G. S., & Mirjat, N. H. (2018). "A Flexible Mathematical Model for Dissimilar Operating Modes of a Switched Reluctance Machine". *IEEE Access*, *6*, 9643-9649.

[22] Tahir, S., Wang, J., Kaloi, G. S., & Baloch, M. H. (2017). "Robust digital deadbeat control design technique for 3 phase VSI with disturbance observer". *IEICE Electronics Express*, *14*(13), 20170351-20170351.

Data-driven based Fault Diagnosis using Principal Component Analysis

Shakir M. Shaikh[1], Imtiaz A. Halepoto[2], Nazar H. Phulpoto[3], Muhammad S. Memon[2], Ayaz Hussain[4], Asif A. Laghari[5]

[1]Department of Control Science and Engineering, HIT Harbin, China
[2]Department of Computer Systems Engineering, QUEST Nawabshah, Pakistan
[3]Department of Information Technology, QUEST Nawabshah, Pakistan
[4]Department of Electrical Engineering, BUETK, Khuzdar, Pakistan
[5]School of Computer Science & Technology, HIT Harbin, China

***Abstract*—Modern industrial systems are growing day by day and unlikely their complexity is also increasing. On the other hand, the design and operations have become a key focus of the researchers in order to improve the production system. To cope up with these chellenges, the data-driven technique like principal component analysis (PCA) is famous to assist the working systems. A data in bulk quanitity from the sensor measurements are often available in such industrial systems. Considering the modern industrial systems and their economic benifits, the fault diagnostic techniqes have been deeply studied. For example, the techniques that consider the process data as the key element. In this paper, the faults have been detected with the data-driven approach using PCA. In particular, the faults have been detected by using T^2 and Q statistics. In this process, PCA projects large data into smaller dimensions. Additionally it also preserves all the important information of process. In order to understand the impact of the technique, Tennessee Eastman chemical plant is considerd for the performance evaluation.**

Keywords—Fault Diagnosis; Principal Component Analysis; Multivariate Statistical Approach; Tennessee Eastman Chemical Plant Introduction

I. Introduction

Industrial process managemen is one the key and emering issue in the small as well large industrial systems. Modern industrial services are in large-scale and they are extremely complex. In addition, the control of process management is carried out with a great number of parameters under the system. In the industries like manufacturing, there is a pressure to produce excellence in end-products, which is in bulk quantit. In parallel, it is also important to minimize the losses of rejection rates and to satisfy the ecological rules. To fulfill the demands, up-to-date industrial systems cover a huge number of parameters working under closed-loop controllers. For that, a data-driven design of system is one of the great interest both in research and academia. Engineering systems such as aircraft controllers, industrial processes, manufacturing systems, transportation systems, electric and electronic system are becoming more complicated to lead the failure. It will directly related to the system reliability, availability, safety and maintainability. One the other hand, such factor are very important for a good industrial system. Many of such systems rely on human efforts and the availablity. In order to improve the performance and industrial systems it is necessary to work on the automation. It reduces the human efforts and the cost so it effects the economic conditions. Nowadays, the demand for the automated systems is also increasing in the market. In this research area, there is need to study different operating constraint and applications of industrial automation that explore and elaborate the process of automation. In automated systems in it necesssary to implement techniques and policies for the fault diagnosis and repair. The fault diagnosis system tries to assure that the plant is safe by identifying unwanted events. It highlights the key issue that may degrade the overall performacne of the system. This information is necessary for plant engineer so that a quick action may be performed. So that an immidieate rescue could be performed for the safety of the industrial system. There are many techniques are available for the performance monitoring and control. PCA is one of the basic technique in the pool of famous techniques.

In Section II related work is discussed. Methology has been discussed in Section III in which implementation of techniques is done stepwise. In Section IV, PCA technique is applied in the industrial benchmark process. In Section V results have been discussed. Section VI concludes the work and provides the guidelines for the future.

II. Related Work

Multivariate statistical methods largely depend upon the huge quantity of past data to define the fluctuations in the process. Multivariate statistical process monitoring has the advantage of easy to design and make the analysis of process industries entire simple, due to this property it is most popular in industrial fault diagnosis systems while in detecting the abnormal operation from the process. The technique which has the capability to retain the major information and significant knowledge in a unique dataset that generates from the industrial process is PCA. There are many approaches are available for the fault diagnosis. These apporoaches use different parameters in order to detect the faults in the control systems. For example, the study in [1] highlited the fault diagnosis based on the neural netowrks. They combined such neural networks by an observer technique. Also, multivariate statistical approaches have been researched to deal with process monitoring [2]. PCA is first appeared in 1889 until now research is ongoing and applications are still in study.

These methods have been successfully implemented in many industrial processes. MacGregor implemented PCA based process monitoring both in continuous and batch process and conclude PCA methods are capable of treating processes with a large correlated process data and can handle easily missing data [1]. Raich and Cinar [3] proposed a diagnosis method based on angle discriminant using PCA. Due to some difficulties, its applicability is not so large instead of that it fits in so many fields and successfully implemented. Though PCA could affect if it is applied in nonlinear problems because real systems are mostly nonlinear in nature, and this technique takes account linear combination due to its linear method. Kramer [4] has generalized PCA to the nonlinear case by using a neural network. These chemical engineering applications, are mostly nonlinear but the method is linear. Application of PCA in the real system have been applied at Dupont and other companies, published in many conferences and journals several types of researches have performed similar case work on data collected simulator of process [5][6]. For sake of easiness many dimensions of dataset proposed to get more from data in different views and plot in single dimension [7], on taking this step that will helps the operator to get information from more than multidimensional data [8]. In some cases multidimensional data acquire is quite difficult due to nonlinearities so an automation process proposed for process monitoring in [9]. The application of PCA in these type of problems motivated by three features. Number 1, PCA can develop a method which takes all the data in low dimension which helps out to get meaning from entire data from the training set by use of all dimensional data. Number 2 the data in structured format with help of PCA help to identifying the affected variables. Number 3, PCA can isolated the space which the variables contain useful information that variables have process information and rest in another subspace which contain noise. In this fault could occur in any subspace primarily [10], this step can increases the sensitivity of process monitoring to detect faults. As an effective data-driven process monitoring technique PCA can adapt complicated conditions according to rules of statistics. It is classical projection methods of multivariate statistical process monitoring which is then applied to train model beneath nominal conditions. Thus it detects online faults [11]. These techniques are highly demanded based on measurements [12]. In practical industrial outliers that is difficult to handle in spite of so many advantages and easy to design model [13]. Sensor failure, network transmission error, machine malfunction, database software, and data recording errors are mainly cause for irregularities produce data with noise [14]. Such cases outliers smoothed by mean and averaging [10].

III. Methodology

PCA is one of the famous and widely used technique. It has been effectively used in various areas including image processing, signal analysis, pattern recognition, data compression and process monitoring. This techniques is simple and efficient and have capability to process industrial data. It is familiar as influential tool for process monitoring. For this purpose, it is used in the process industry for process monitoring. PCA algorithm is a founding technique of automated process monitoring. These advanced PCA methods for example recursive, adaptive and kernel. These techniques extract the useful information from data in keeping view this it is a widely used area in fault diagnosis. It extracts orthogonal vectors in sets, known as loading vectors. It tells the amount of variance known to orthogonal vectors. Consider a process measurement matrix $X \in \mathbb{R}^{n \times m}$, where m is variables and n is observations in measurement matrix.

A. PCA based Fault Detection

Step 1: Pretreatment of data is done in this step. Normalize columns of X

Step 2: Obtain the Covariance of measurement matrix by

$$C = \frac{1}{n-1} X^T X \tag{1}$$

Step 3: In this step, the loading vector is extracted by obtaining the Singular Value Decomposition (SVD) from the above equation:

$$C = \frac{1}{n-1} X^T X = U \Sigma V^T \tag{2}$$

Where $\Lambda = diag(\lambda_1 \geq \cdots \geq \lambda_m \geq 0)$

Step 4: To calculate PCs (principal component) a divide V into the score and residual matrices.

$$\Lambda = \begin{bmatrix} \Lambda_{pc} & 0 \\ 0 & \Lambda_{res} \end{bmatrix}$$

$$\Lambda_{pc} = diag(\lambda_1, \dots \lambda_a) \quad \Lambda_{res} = diag(\lambda_{a+1}, \lambda_{a+2}, \dots \lambda_m$$

$$V = \begin{bmatrix} V_{pc} V_{res} \end{bmatrix} \quad V_{pc} \in \Re^{m \times a} \quad V_{res} \in \Re^{m \times (m-a)}$$

The value of L can be taken from V_{pc}:

$$L = V_{pc} \tag{3}$$

In the above equation V_{res} is the residual space.

Step 5: To obtain the matrix T, the following equation is used:

$$T = XL \tag{1-4}$$

B. PCA and the Faults

In the process, T^2 can be obtained bu the following equation:

$$T^2 = x^T L \Lambda_{pc}^{-1} L^T x \tag{4}$$

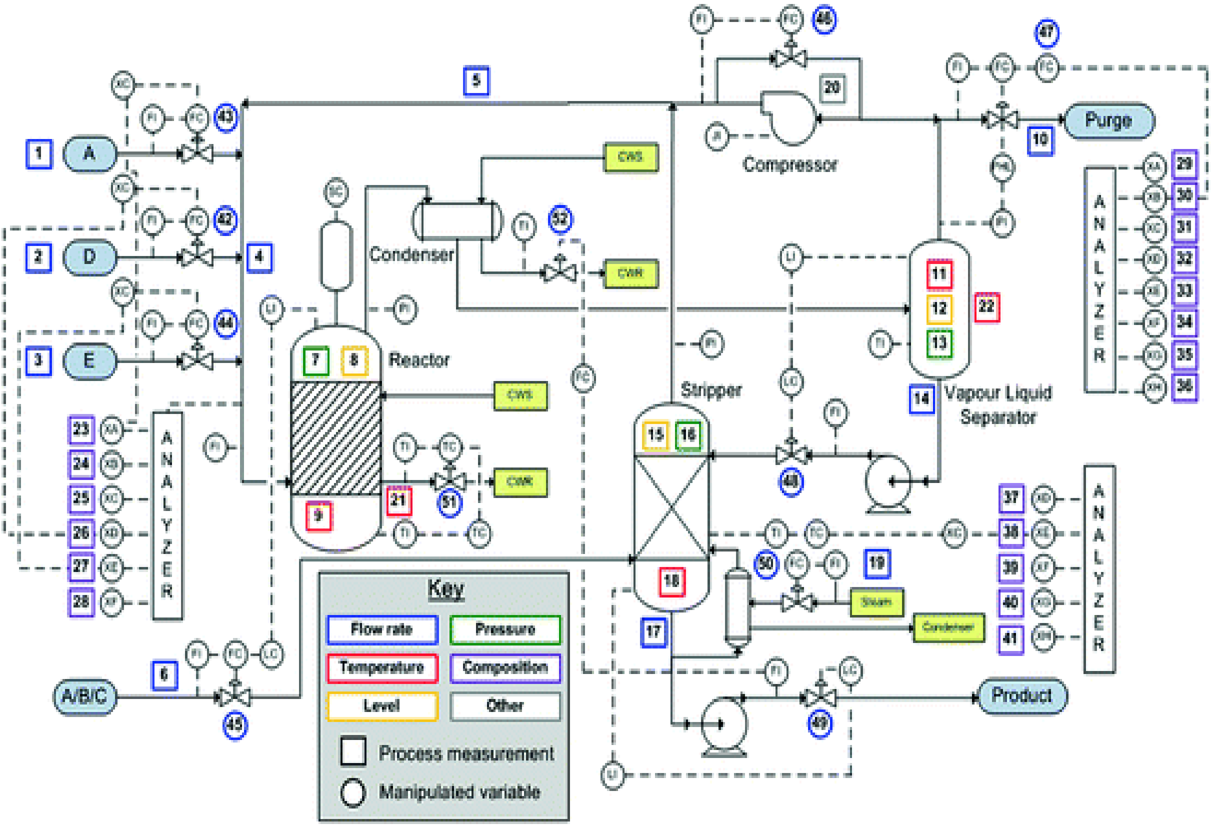


Fig. 1. Tennessee Eastman Process.

L represent the set of loading vectors for large singular variance. The simplified equation will be:

$$j_{th,}T^2 = \frac{a(n-1)(n+1)}{n(n-1)} F_\alpha\left(a, n-a\right) \qquad (5)$$

Where $F_\alpha(a, n-a)$ is the f-distribution. If the threshold value from equation no 1-5 exceeds the test statistics in equation no: 1-6 fault occurs. As T^2statistics is demonstrated on the basis of loading vectors singular values, so it does have a problem to some inaccuracies [15] in the residual part values. The square prediction error is then used which utilizes the residual space [16].

$$Q = x^T V_{res} V_{res}^T x \qquad (6)$$

Q −Statistics threshold is achieved by:

$$j_{th}, SPE = \left(\frac{\theta_1\left(h_0 c_\alpha \sqrt{2\theta_2}\right)}{\theta_1} + 1 + \frac{\theta_2 h_0\left(h_0 - 1\right)}{\theta_1^2} \right)^{1/h_0}$$

Where

$$\theta_i = \sum\nolimits_j^n = a + 1\lambda_j^{2i} \quad \text{and} \quad h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2} \qquad (7)$$

Where c_α represents the standard deviation of the distribution parallel to the $(1-\alpha)$ percentile. Hence, the confidence level for the Q may be determined with the help of equation (7) in order to cope with abnormalities.

IV. The Benchmark Process

In this part to verify the algorithm like PCA, simulation is carried out in order to diagnosis the faults. It is computer-oriented simulator many types of research used it for comparison of different data-driven algorithms as well as model-based. It is like realistic simulator which mimic the original behavior of the typical chemical plant. TEP served as the desired benchmark to assessment algorithms for many techniques such as process observing control and fault diagnosis. TEP utilized to examine multivariate statistical process monitoring (MSPM) methods. It facilitates with different operating regimes. Figure 1 presents the flow of the process. A number of connected modules are present. These include a condenser, reactor, separator, stripper, and compressor. It consists with four in number(1-8)of reactants or input and two products or output, along with by-product, and an inert by symbolically represented as A to H. There are 52 measurements by each process. Among them, 41 are the output variables and the rest are input variables. The output and input variables are depicted in Table 1[8]. Further, the process variables (from XMV(1) to XMV(11)) are used with the standard values (as in accordance with [8]). The researchers in [17] performed the simulation of TE process. There has been successfully worked done in these techniques reported in. Training data set produced for working on this which includes both faulty and normal data set. The work in [18] proposed a control scheme which is thus applied in TEP. Flowchart of an industrial plant is demonstrated in figure 1. The gaseous components are presented as A, C, E, D and the inert B. It first feeds to the reactor. Formation of G and H component as a product from these inputs feed into a reactor. A simulator has been developed details can be found in [19]. Following equations are input-output relations of the process.

$$A(k)+C(k)+D(k)\rightarrow G(liq)$$

$$A(k)+C(k)+D(k)\rightarrow H(liq) \qquad (8)$$

$$A(k)+C(k)+D(k)\rightarrow H(liq)$$

$$D(k)\rightarrow F(liq)$$

TABLE 1. LIST OF PROCESS VARIABLES AND MEASURED VARIABLES [8]

Tag	Description
XMV(1)	D feed flow
XMV(2)	E feed flow
XMV(3)	A feed flow
XMV(4)	A and C feed flow
XMV(6)	Purge Valve
XMV(7)	Separator pot liquid flow
XMV(8)	Stripper liquid product flow
XMV(10)	Reactor Cooling water flow
XMV(11)	Condenser Cooling water flow
XMEASV(1)	A feed (Stream 1)
XMEASV(2)	D feed (Stream 2)
XMEASV(3)	E feed (Stream 3)
XMEASV(4)	A and C feed
XMEASV(5)	Recycle flow
XMEASV(6)	Reactor feed rate
XMEASV(7)	Reactor Pressure
XMEASV(8)	Reactor level
XMEASV(9)	Reactor temperature
XMEASV(10)	Purge rate
XMEASV(11)	Separator temperature
XMEASV(12)	Separator level
XMEASV(13)	Separator pressure
XMEASV(14)	Separator underflow
XMEASV(15)	Stripper level
XMEASV(16)	Stripper pressure
XMEASV(17)	Stripper underflow
XMEASV(18)	Stripper temperature
XMEASV(19)	Stripper steam flow
XMEASV(20)	Compressor work
XMEASV(21)	Reactor water temperature
XMEASV(22)	Separator water temperature

In this equation the Component F is a by-product, that process is exothermic and cannot reversible. They are in first-order with concentrations higher temperature happened due to fastest reaction of component G over reaction of H component. Separator has vapors from reaction which is recycled again and again that is input to the compressor. The stream generated from the process keep for the use of by product and inert. Stripper “stream 10” is driven by a separator which is condensed.

C. *Main Process Variables*

The process has 41 calculated variables and 11 input variables. They are listed in Table 1. Out of which 22 variables which sample every three minutes. There are 21 process faults as described in the work [17].

D. *Simulated Faults in TEP*

There are 21 faults in TEP process listed in Table 1. These faults affect mostly in chemical process parameters like process variables, kinetics, feed concentration and different types of actuators in the chemical process like pump valves. Data-driven approaches require online and offline training data, in this simulator there are 22 online test data sets, their duration for 48 hours plant operation and 960 samples of data are generated during simulations in [5], where faults are added after 160 data sample.

V. RESULTS AND DISCUSSION

There are a number of faults as depicted in the Table 2. All the values of Table 1 and 2 are directly taken from [8]. One of the first steps is to implement the PCA in order to detect such configured faults. The default dataset from the TE simulator is used for the experimentation and evaluation. The input and output variables are configured. Such as, XMEAS(1-22) and XMV(1-11). In real life the reason for the faults is unknown, similarly the TE also introduce faults at different subspaces.

According to Table 2, the fault in the condenser cooling water inlet temperature is represented by IDV2. It occurs after 160 data samples. In figure 2 PCA of IDV 2 is shown. IDV (6) involves in step type of fault, it is simulated by a sudden change in the reactor cooling water inlet temperature. The results of PCA diagnosis the affected variable from “A” feed loss in TE process are shown in figure 3. The algorithm for process monitoring is developed with the help of 960 samples taken from the ordinary process operations. Similarly, IDV (18) is unknown type of fault listed in Table 2, affect the process variable, it influence on the unknown variable. PCA-based statistics identifying the unknown variable shown in figure 4.

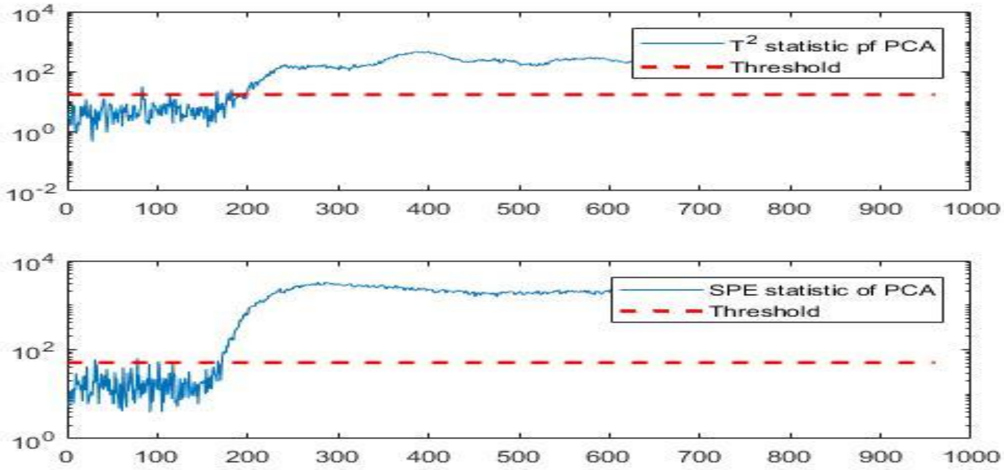


Fig. 2. Detection Result for Fault Scenario 2 PCA.

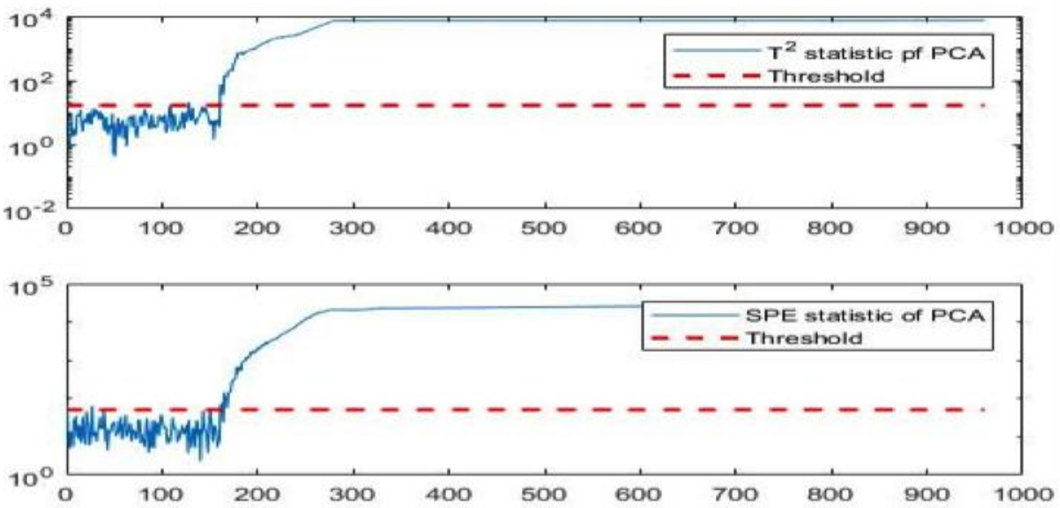


Fig. 3. Detection Result for Fault Scenario 6 PCA.

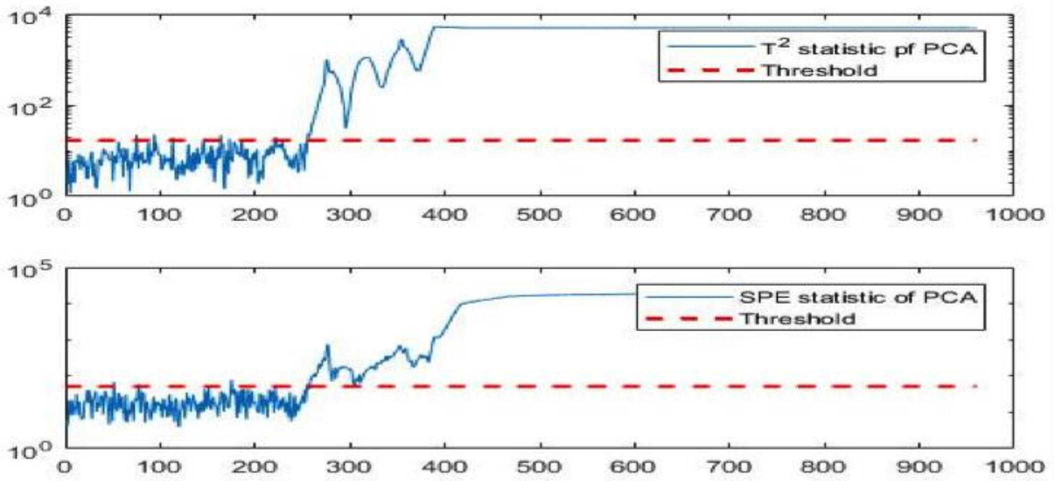


Fig. 4. Detection Result for Fault Scenario 18 PCA.

TABLE 2. DEFINITION OF FAULTS [8]

Fault number	Process Variable	Type
IDV(1)	A/C feed ratio, B composition constant	Step
IDV(2)	B composition, A/C ration constant	Step
IDV(3)	D feed temperature	Step
IDV(4)	Reactor cooling water inlet temperature	Step
IDV(5)	Condenser cooling water inlet temperature	Step
IDV(6)	A feed loss	Step
IDV(7)	C header pressure loss-reduced availability	Step
IDV(8)	A, B C feed composition	Random variation
IDV(9)	D feed temperature	Random variation
IDV(10)	C feed temperature	Random variation
IDV(11)	Reactor cooling water inlet temperature	Random variation
IDV(12)	Condenser cooling water valve	Random variation
IDV(13)	Reaction Kinetics	Slow Drift
IDV(14)	Reactor cooling water valve	Sticking
IDV(15)	Condenser cooling water valve	Sticking
IDV(16)	--	--
IDV(17)	--	--
IDV(18)	--	--
IDV(19)	--	--
IDV(20)	--	--
IDV(21)	Steady state position	Fixed

VI. CONCLUSION

The fault diagnosis is very important for the optimized systems. Specifically, the data-driven techniques are getting famous. This work presents a detailed study on the data-driven technique. The design of a data-driven technique using the PCA is proposed. PCA is simple efficient and easy to design due to these properties it is frequently used in fault diagnosis techniques. The industrial benchmark, Tennessee Eastman process is used for the simulation and analysis of the data-driven technique, which successfully detects the faults.

The major objective of further investigation is to analysis of non-Gaussian process data, since in industries mostly system are non-linear in nature. In this work it is assumed that data which are under consideration is Gaussian. A framework should be established that will directly constructed from process data for construction of fault tolerant architecture.

REFERENCES

[1] ZHOU D, LI G, QIN S J. Total projection to latent structures for process monitoring[J]. AIChE Journal, Wiley Online Library, 2010, 56(1): 168–178.

[2] DE JONG S. SIMPLS: an alternative approach to partial least squares regression[J]. Chemometrics and intelligent laboratory systems, Elsevier, 1993, 18(3): 251–263.

[3] LI G, LIU B, QIN S J . Quality relevant data-driven modeling and monitoring of multivariate dynamic processes: The dynamic T-PLS approach[J]. IEEE transactions on neural networks, IEEE, 2011, 22(12): 2262–2271.

[4] YIN S, ZHU X, KAYNAK O. Improved PLS focused on key-performance-indicator-related fault diagnosis[J]. IEEE Transactions on Industrial Electronics, IEEE, 2015, 62(3): 1651–1658.

[5] CHIANG L H, PELL R J, SEASHOLTZ M B. Exploring process data with the use of robust outlier detection algorithms[J]. Journal of Process Control, Elsevier, 2003, 13(5): 437–449.

[6] FRANK P M. Analytical and qualitative model-based fault diagnosis–a survey and some new results[J]. European Journal of control, Elsevier, 1996, 2(1): 6–28.

[7] YIN S, DING S X, HAGHANI A . A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process[J]. Journal of Process Control, Elsevier, 2012, 22(9): 1567–1581.

[8] STEVEN X. Data-driven Design of Fault Diagnosis and Fault-tolerant Control Systems [M]. SPRINGER, 2014.

[9] AGRAWAL V, PANIGRAHI B K, SUBBARAO P M V. Review of control and fault diagnosis methods applied to coal mills[J]. Journal of Process Control, Elsevier, 2015, 32: 138–153.

[10] TSCHUMITSCHEW K, KLAWONN F. Incremental quantile estimation[J]. Evolving Systems, Springer, 2010, 1(4): 253–264.

[11] TRACY N D, YOUNG J C, MASON R L. Multivariate control charts for individual observations[J]. Journal of quality technology, Taylor & Francis, 1992, 24(2): 88–95.

[12] BRITTO R da S. Detecção de falhas com PCA e PLS aplicados a uma planta didática[J]. Pós-Graduação em Engenharia Elétrica, 2014.

[13] BISHOP C M. Pattern recognition and machine learning (information science and statistics)[J]. Springer-Verlag New York, Inc., Secaucus, NJ, 2006.

[14] CHEN T, MORRIS J, MARTIN E. Probability density estimation via an infinite Gaussian mixture model: application to statistical process monitoring[J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), Wiley Online Library, 2006, 55(5): 699–715.

[15] CHEN Z, ZHANG K, DING S X . Improved canonical correlation analysis-based fault detection methods for industrial processes[J]. Journal of Process Control, Elsevier, 2016, 41: 26–34.

[16] CHEN Z, DING S X, ZHANG K . Canonical correlation analysis-based fault detection methods with application to alumina evaporation process[J]. Control Engineering Practice, Elsevier, 2016, 46: 51–58.

[17] WISE B M, GALLAGHER N B. The process chemometrics approach to process monitoring and fault detection[J]. Journal of Process Control, Elsevier, 1996, 6(6): 329–348.

[18] PIOVOSO M J, KOSANOVICH K A, PEARSON R K. Monitoring process performance in real-time[C]//American Control Conference, 1992. IEEE, 1992: 2359–2363.

[19] Chen, Zhiwen, et al. "Improved canonical correlation analysis-based fault detection methods for industrial processes." Journal of Process Control 41 (2016): 26-34.

A Practical Approach for Evaluating and Prioritizing Situational Factors in Global Software Project Development

Kanza Gulzar[1,2,3], Jun Sang[1,2], Adeel Akbar Memon[4], Muhammad Ramzan[5],
Xiaofeng Xia[1,2] and Hong Xiang[1,2]
[1]Key Laboratory of Dependable Service Computing in Cyber Physical Society of Ministry of Education,
Chongqing University, Chongqing, 400044, China
[2]School of Software Engineering, Chongqing University, Chongqing,401331, China
[3]University Institute of Information Technology, PMAS Arid Agriculture University,
Rawalpindi, 46000, Pakistan
[4]Department of Software Development, Henan 863 Software Co. Ltd,
Henan, Zhengzhou 450001, China
[5]College of Computing and Informatics, Saudi Electronic University,
Riyadh 11673, Saudi Arabia

Abstract—**There has been an enormous increase in globalization that has led to more cooperation and competition across boundaries. Software engineering, particularly distributed software development (DSD) and global software development (GSD), is evolving rapidly and presents several challenges, such as geographical separations, temporal differences, cultural variations, and management strategies. As a result, a variety of situational factors (SFs) arise that causes challenging problems in software development. Both literature and real world software industry study revealed that the extent of the effect of SFs may vary subject to a certain software project. Project executives should need to concentrate on the right SFs for the successful development of a specific project. This work first examines the optimal and most well-balanced GSD-related SFs and then presents a mechanism for prioritizing the SFs to better understand the extent to which an SF generally affects the GSD. A set of 56 SFs in 11 categories is identified and analyzed in this research. A fuzzy set theory based, multi criteria decision making (MCDM) technique, fuzzy analytical hierarchy process (FAHP) was proposed to extract the SFs that have the strongest effects on GSD. The proposed technique is intelligent and automated and can be customized to suit specific conditions and environments. Thus, it can provide support for a much-needed variation that is the hallmark of such software development environments. A case study of a global company working in collaboration on a project JKL was selected to identify and prioritize the most challenging SFs. A sensitivity analysis is carried out to evaluate the extent of the impact for highly ranked SFs related to JKL project.**

Keywords*—*Global software development (GSD); Situational Factors (SFs); Fuzzy Analytic Hierarchy Process (FAHP); Multi criteria decision-making (MCDM); fuzzy set theory; sensitivity analysis

I. Introduction

Software development has become one of the main businesses due to the growing demand for high-quality software and the huge investments in software projects. This trend has led to distributed and globally disseminated software development companies, making the development process even more complex. Distributed software development (DSD) and global software development (GSD) appear to be more plentiful due to the shorter time span and lower costs, and benefit from the most competent programmers and developers from around the world [1], [2]. Both DSD and GSD consist of companies and staffs from different geographic sites working together to achieve the stated mission [3]. Although GSD strategies are very advantageous in providing quality software products, there are many difficulties [4], [5] in managing software development activities.

Requirement engineering (RE) is the first stage in software development, and it plays a critical role in the development of a successful software project [6]. The vital objective of RE is to recognize and realize the stakeholder's concerns and develop high-quality software projects. In a rapidly changing business environment, requirements change continuously to fulfil a user's demand, depending on their location. An organization must have RE that addresses the organization's business policy and manages the requirements according to that policy [7]. Basically, the most commonly followed policy in the software development process is that it should fit the needs of the project [8]. In practice, the generation of concrete requirements for a software project is not separate from the development process. However, the developers (team) with certain technical skills (knowledge) are capable of authenticating the definition of the specified requirements. In turn, the requirements are dependent on the project's nature, the project's content and the prevailing organizational structure. A variety of external circumstances, such as statutory requirements and technical limitations, are enforced by superficial authorities (stakeholders) [9]-[11]. These factors are observed to have a strong influence on the software development and may be linked to the organization, management, human capital, techniques/tools, and social and economic aspects. In the literature, these factors are termed "situational factors" (SFs).

Every project has different working environment and surroundings, so there are different SFs that influence the development of the project. A rich body of literature has been observed in different areas of requirement engineering, and there are many SFs that influence the software development process as a whole [12]. SFs must be evaluated before making any decision about the best process to use for software development [13]. Commonly observed results from the previous literature indicate that situational factors have a strong impact on GSD. It is possible that different projects have the same influencing SFs, but the extent to which a factor can affect a project may vary. To date, the research works have merely listed SFs based on the literature without depicting the extent of their importance, and they lack industrial evidence (refer to Section 2). These factors captivate many of the developmental efforts and create many disputes, such as those in project planning, management, and implementation issues. Therefore, identifying these factors' importance (i.e., the extent to which a particular factor is important) at a low level and making them visible are necessary to enhance their manageability and measurability. It results in a detailed, up-to-date and contextually sound software development activity management that supports for the fruitful development of software projects at a global level.

The aims of this paper are to provide a framework by employing Fuzzy Analytic Hierarchy Process (FAHP) that determines the relative importance of each situational factor for better decision making and to strengthen the findings of the existing research. Analytic Hierarchy Process (AHP) is used to compute the relative importance of the SFs and an in-depth evidence is produced in an organized way. Although AHP has been extensively employed to solve multi criteria decision-making (MCDM) problems, it has some limitations. Therefore, fuzzy logic is introduced into the pairwise comparison to compensate for the discrepancies in the conventional AHP. The significant contribution of the fuzzy modeling is that it is able to build and formulate to impersonate the real world comportment even when data provided is ambiguous and hardly precise. Conversely, for the SFs accompanying uncertainty with them, rational values can be calculated. The use of MS Excel eradicates the need for extra software, so the process is cheaper to implement. Our work reports that the practical endorsements that arise from it will aid in accessing the base data required for decision makers to develop policies to enhance GSD.

The rest of the paper is organized as follows. Section II gives an overview of the literature on SFs. Section III gives a comprehensive explanation of the AHP and FAHP approaches. Section IV elaborates on the research procedure employed. Section V describes the results inferred by applying FAHP to a case study to determine the extent of SFs and validates the results through a sensitivity analysis. Section VI precisely concludes the paper with future perspectives.

II. Literature Review

Over the last decade, a great change in software development has been observed, as has a strong emphasis on the significance of inconsistent situations for effective SD [14]. Ghosh et al. [15] stated that it is complicated to manage requirements at a global level. According to Hanisch et al. [16], considering a virtual domain makes requirement engineering more complex. Therefore, effective requirement handling and quality software development must focus on the software development context. Michael et al. [17] emphasized three contexts: the organization, the environment and the project. In their study, they illustrate the need to consider these contexts in order to have a quality outcome. Their work makes other researchers think in these terms, Cameron argued that the software projects need to organize their developmental processes by considering various factors that subsidize the variation in projects and he pinpointed five tailoring factors [18].

In [19], the authors classify the software development environment factors in four categories: Project, Team, External Stakeholders, and Organization. They also suggested a model for tailoring software process that laid the foundation and set the dimensions for the classification of environmental factors. Ghosh et al. [20] followed the work done in [19] and examined the factors that influence the tailoring of the software decision process. Another study [21] investigated and identified the SFs within an organization. In another research work [22], the authors observed the influence of SFs from the aspects of technology-based self-service and the attitudes toward the service. Daniela [23] generated a list of problems, challenges and affected RE phases across multi-sites. Clarke et al. [12] created a list of SFs that influence software development through a systematic literature review (SLR) that use the data coding techniques employed in Grounded Theory [24]. There is no doubt that their framework is an initial move towards advanced development in this direction, but it discussed the facts generally. Since the aspects of the GSD are missing in this framework, there are many SFs that may be particularly geographically site dependent and need to be identified, and assessing their influence on the accomplishment of a software project is imperative [25], [26].

In a more recent research work on SFs in GSD, Huma et al. followed the work done by Daniela et al. [23] and Clarke et al. [12] and performed an SLR to identify situational factors [9]. They also used data coding techniques based on Grounded Theory to create a list of 37 SFs that was considered sufficiently comprehensive to define the RE process in the GSD environment. However, the limitations of the study were that the SFs were identified on the basis of the reviewed literature and that it lacked recommendation from the software industry. Therefore, many SFs remain to be identified, both within the organization and globally.

So far, it is obvious from the above literature that many works (given in the Appendix) have considered the importance of SFs in software project development. Researchers have suggested that it is necessary for scholars and experts alike to focus on improving the understanding of the situational context while developing software projects [27]. However, not a single study has comprehensively covered the identification and the importance of all the SFs in a GSD environment. The scarcity of such framework is the cause of the limited recognition of the primary constraints and features of GSD. Thus, in this work, efforts are made to determine the most important SFs in a GSD environment that were not found in [28] and [9]. Hence, in view of the importance of SFs in software project development, we consider both the literature and the industrial perspective to prioritize them and devise a more comprehensive framework with the latest technique, which is indeed a need.

The authors of this work have long worked to incorporate artificial intelligence into various areas of requirement engineering [29], [30]. Now, they are focusing on implementing intelligent frameworks for SFs by integrating fuzzy logic with the AHP. The AHP was first introduced by Saatay [31] and is a widely accepted MCDM tool for successful weight estimation. However, traditional AHP restricts to consider rationalism in human decisions [32]. Therefore, of the important information related to vital factors cannot be precisely determined. However, these values could be measured in a more accurate way by replacing the crisp numbers with fuzzy numbers. Zadeh [33] proved logically that fuzzy logic permits us to transform linguistic measures into crisp measures with the help of membership functions, and many researchers have shown that fuzzy logic gives more adequate results. Fuzzy AHP is the extension of Saaty's theory and was proposed by Van Laarhoven and Pedrycz [34]. They employed triangular fuzzy numbers and the logarithmic least squares method (LLSM) to calculate the fuzzy weights to determine priority. Buckley [35] determined the fuzzy comparison ratios by using trapezoidal fuzzy numbers. Xu [36] proposed a fuzzy least squares priority method (LSM). Mikhailov [37] introduced a fuzzy preference programming method (PPM) to calculate the crisp weights from fuzzy comparison matrices. Wang [38] used FAHP to select the best maintenance strategies. Whenever there is uncertainty in prioritizing one factor over another, fuzzy logic be incorporated with AHP to deal with the complexity. Sun Chia-Chi integrated the fuzzy AHP with Fuzzy TOPSIS [39]. Many other researchers have shown that the fuzzy AHP gives more adequate explanations in the decision-making practice compared to the traditional AHP technique [40]. In the next section, we concisely discuss the AHP and fuzzy AHP approaches.

III. AHP Approach and Fuzzy AHP Approach

A. AHP Approach

AHP is a vigorous decision-making technique for computing the preferences among given criteria by matching the alternatives for each criterion to calculate an inclusive rank among the given alternatives. The basic AHP method consists of the following steps:

- Construct a hierarchical structure that shows the fundamental components of the problem and the associations between them.
- Make a pairwise comparison to determine the relative weights of the factors of the decision criteria by eliciting experts' opinions. The Saaty scale (Table I) is used for this purpose.
- After finalizing the pairwise comparison, calculate the local priorities from the judgment matrices; these judgments are denoted with meaningful numbers.
- Find the Consistency Ratio (CR) that is the measure of the consistency of judgments made by the experts.
- Finally, obtain the ultimate or global priorities by incorporating the numbers gained in the preceding step to calculate the final priorities of the components of the hierarchy.

TABLE I. Saaty's Scale for Pairwise Comparison [31]

Saaty's scale	The Relative Significance of Elements
1	Equally important (both are equal)
3	Moderately important (one over another)
5	Strongly important
7	Very strongly important
9	Extremely important
2;4;6;8	Intermediate values

Because AHP is limited in that it does not appropriately handle the vagueness associated with the experts' judgments, fuzzy logic comes into play.

B. Fuzzy AHP Approach

To address the ambiguity issues, the Fuzzy AHP is a credible blend of the pairwise comparison matrix of experts' opinion and fuzzy set theory. Hence, it has become famous for solving multi-attribute decision making (MADM) problems and providing a more precise ranking methodology [41]. All the steps involved in fuzzy AHP are the same as in AHP except for the fuzzy representation of the pairwise comparison by triangular fuzzy numbers (TFNs) [42]. To represent TFNs in the pairwise comparison, three real numbers are stated as a triple (l,m,u), whereas in the fuzzy AHP method, despite having distinct numbers, the numbers 1-9 symbolize triangular fuzzy numbers that handle the vagueness and imprecision in the pairwise priority values of the criteria involved. The fuzzy set definition of the five triangular fuzzy numbers described by the compatible membership function is shown in Fig. 1 and Table II. In fuzzy set theory, if an entity has a membership, then it is symbolized by 1, and if it has no membership, then it is symbolized by 0. Suppose that the universe of discourse is represented by u and that *l(x)* is the membership function that lies in [0, 1]. Mathematically, the membership function for the triangular type fuzzy numbers is defined as (1).

$$o\mu(x) = \begin{cases} 0 & x<1 \\ \frac{x-1}{m-1} & l \leq x \leq m \\ \frac{u-x}{u-m} & m \leq x \leq u \\ 0 & x>u \end{cases} \tag{1}$$

In (1), u and l stand for the upper and lower limits of the fuzzy number M, respectively, and m is the medium value of M. If a TFN can be represented by $M = l_{ij}$, m_{ij}, u_{ij}, we suppose that $l_{ij}<m_{ij}<u_{ij}$ when $i \neq j$, where i and $j = 1,2,3,,n$. When $i = j$, then $M_{ij} = M_{ii} = (1\ 1\ 1)$. Therefore, an accurate priority vector $w = (w_1, w_2, \ldots, w_n)^T$ that is a consequence of the judgment matrix can essentially satisfy the inequities. The addition of two fuzzy triangular numbers $M_i = (l_i, m_i, u_i)$ and $M_j = (l_j, m_j, u_j)$ is shown below, and other operations are done in the same way.

$$\begin{aligned} M_i \oplus M_j &= (l_i, m_i, u_i) \oplus (l_j, m_j, u_j) \\ &= (l_i + l_j, m_i + m_j, u_i + u_j) \end{aligned} \tag{2}$$

Chang proposed extent analysis in [44]. By adapting his formula, the extent analysis values for each element can be calculated as follows:

$$c_{ij}^t = [c_{ij}^t, c_{ij}^t, c_{ij}^t], i, j = 1, 2, \ldots, n_k, t = 1, 2 \tag{3}$$

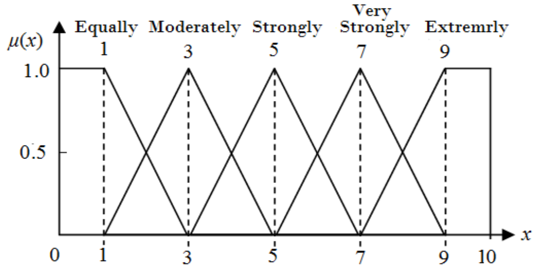


Fig. 1. Triangular fuzzy numbers.

TABLE II. Degree of Importance and Fuzzy Numbers [43]

Intensity of Importance	Triangular Fuzzy Scale	Intensity of Importance	Triangular Fuzzy scale
1	(1,1,1)	1/1	(1/1,1/1, 1/1)
3	(1,3,5)	1/3	(1/5,1/3, 1/1)
5	(3,5,7)	1/5	(1/7,1/5 1/3)
7	(5,7,9)	1/7	(1/9,1/7, 1/5)
9	(7,9,11)	1/9	(1/11,1/9, 1/7)

'T' is a TFN given by the t^{th} expert and k^{th} formula.

$$c_{ij}^{k} = \frac{1}{T} \bigotimes (c_{ij}^{1} + c_{ij}^{2} + \cdots + c_{ij}^{T}) \tag{4}$$

By introducing TFNs, we can describe the steps for fuzzy AHP as follows:

Step I: The value of the fuzzy synthetic extent with respect to the i^{th} object S is calculated, and the triangular fuzzy comparison matrix is articulated as

$$\tilde{A} = \begin{bmatrix} (1,1,1) & \tilde{C}_{12} & \cdots & \tilde{C}_{1n} \\ \tilde{C}_{21} & (1,1,1) & \cdots & \tilde{C}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{C}_{n1} & \tilde{C}_{n2} & \cdots & (1,1,1) \end{bmatrix} \tag{5}$$

Then, the value of the fuzzy synthetic extent with respect to the i^{th} object is given by

$$S_j^k = \sum {}_{j=1}^{n} C_{ij}^k \bigotimes (\sum {}_{i=1}^{nk} \sum {}_{j=1}^{nk} C_{ij}^k)^{-1}, \quad i = 1,2,\ldots,n_k \tag{6}$$

Step II: After getting the synthetic extent value, calculate the degree of possibility of one fuzzy number that is determined to be greater than the other. It is computed as follows:

$$\begin{aligned} V(M_1 \geq M_2) &= sup_{x \geq y}(min(\mu M_1(x), \mu M_2(y))) \\ V(M_1 \geq M_2) &= 1 \ if \ m_1 \geq m_2 \end{aligned} \tag{7}$$

$$\begin{aligned} V(M_1 \geq M_2) = hgt(M_1 \cap M_2) &= \mu M_1(d) \\ V(M_1 \geq M_2) &= hgt(M_1 \cap M_2) \\ &= \frac{l_1 - u_2}{(m_2 - u_2) - (m_1 - l_1)} \end{aligned} \tag{8}$$

Step III: Determine the minimum degree of possibility [38].

$$\begin{aligned} V(M \geq M_1, M_2, \cdots, M_k) = V(M \geq M_1) \ and \ (M \geq M_2) \\ and \cdots and (M \geq M_k) = min \ V(M \geq M_i), \ i = 1,2,\ldots,k \end{aligned} \tag{9}$$

if

$$d'(S_i) = minV(S_i \geq S_k) \tag{10}$$

$then$

$$W' = (d'(S_1), d'(S_2), \ldots, d'(S_n))^T \tag{11}$$

Where $S = (1,2,3,\ldots,n)$ means a matrix having n elements.

Step IV: Normalized weight vectors are obtained by dividing the elements in each column by the sum of that column, adding the elements in each consequent row and dividing this sum by the number of elements in the row using the formula given below:

$$W = (d(S_1), d(S_2), \ldots, d(S_n))^T, \tag{12}$$

The ultimate weight of each element is determined by the multiplication of the criteria, and the matrix is achieved by computing each alternative with respect to each element.

IV. Research Procedure

A. Research Strategy

Design research involves analyzing situations and the effect that they have on the design of system artifacts. This process can help software engineers and developers discover the fundamental restraints and aspects of software development. In addition to the literature survey, the overall process used here consisted of two studies that were conducted in two sessions. The first occurred in Pakistan, and the second was conducted in China. These included interviews, and the data were gathered using online surveys on different survey platforms. The final aim is to incorporate all the research studies and bring an overall perspective to the following research questions:

RQ1:: What are the SFs involved in GSD?

RQ2:: Determine the extent (relative importance) to which each SF (depending on their sub-factors) influences a particular software project development to find the most important SFs.

The research design series begins with a problem statement. A research proposal is made to conduct research. In the first stage, the literature review and online survey are conducted, on the basis of which the salient SFs are listed and then confirmed by the opinions of experts from the software industry and the scientific research community. In the online questionnaire and interviews, the participants specify the values against each SF based on their experience in their working environment (co-located, distributed, and global). Fig. 2 shows the research procedure.

B. Data collection and Analysis

In this paper, through an in-depth study of the available literature on SFs and surveys and interviews with managers and experts from the companies involved in software development, we prepared an inventory by collecting experts' and researchers' perspectives on the SFs related to GSD. The research's statistical sample consisted of 84 total experts, including middle managers, experts and practitioners who responded to the questionnaire survey. Among these, 40 experts from four companies, COEUS and Oveucs in Pakistan and Premier BPO and Henan 863 Software Incubator in China, with at least 5-10 years of experience were interviewed. The

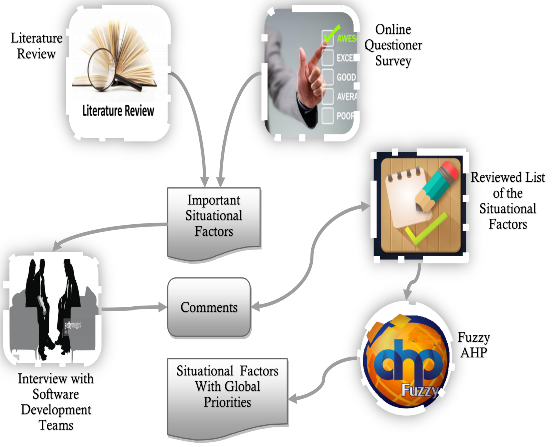

Literature Review
Literature Review
Online Questioner Survey
AWES
EXCE
GOOD
AVERA
POOR
Important Situational Factors
Reviewed List of the Situational Factors
Interview with Software Development Teams
Comments
Fuzzy AHP
ahp
Fuzzy
Situational Factors With Global Priorities

Fig. 2. Research procedure.

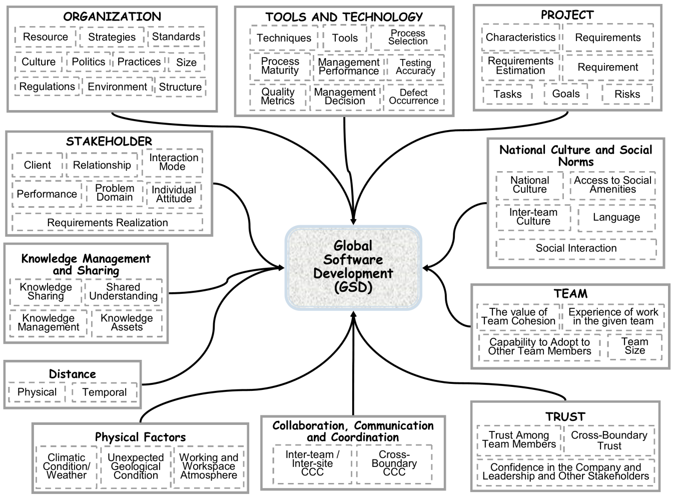


Fig. 3. Situational factors.

demographics of the experts are given in Table III. After defining the 11 key SFs, we mined 56 sub-factors, which are presented in Fig. 3 (answer to RQ1) criteria defined in the AHP. They selected which SF was more important in a pairwise comparison and rated the degree of significance on a scale of 1 to 9 (answer to RQ2). The SFs discussed are not particularly associated with a specific development methodology, but they can generally be applied to any project, regardless of the projects structure or the developmental practice. In the next section, a case study is demonstrated, which was conducted in a multi-site software company in Pakistan for which the most influencing SFs are determined by applying Fuzzy AHP methodology.

TABLE III. RESPONDENT'S DEMOGRAPHICS

Linguistic input		Number
Gender	Male	72
	Female	12
Total		**84**
Designation	Project Manager	10
	Project Team Leader	09
	Software Developer	19
	Software Analyst	10
	Software Architect	10
	System Designer	26
Total		**84**
Working Experience	Project Manager	10
	1-5	20
	5-10	26
	11-15	16
	Less than a year	17
	More than 15 years	5
Total		**84**

V. Case Study

The case study is carried out in a global company working in collaboration with two other global companies to determine the extent to which an SF can affect the SD (RQ2). The selected company is amongst the foremost development companies in Pakistan, China and Germany and develops a wide range of software products for assistance in the real world. We consider a project say JKL which is a collaboration effort between three working sites residing in Lahore, Pakistan and Zhengzhou China, and Berlin Germany. The time and resource constraints had bound us to visit two countries, Pakistan and China. The project consists of four modules, divided among three sites for completion, separately and integrated after finishing.

The motivation of GSD that hiring skilled experts at low cost, mainstream of development activity, including designs and codes for three modules is taking place at Lahore site as wages in Pakistan are less. Some of Pakistani are also working in Berlin and Zhengzhou due to extraordinary earnings. The Chinese company involves in the requirements phase and one module is developed in Berlin that is sent to Lahore after completion. The development process starts after getting a finalized requirements specification document from Zhengzhou. Requirement specification (RS) is analyzed and comprehensive design is planned in Lahore. For coordination among different site the way of communication is the only means of email and Skype and WeChat. After sketching the detailed design the communication among sites is off and on regarding major concerns. For integration the server at Zhengzhou and testing for integration took place in Lahore and Berlin for modules assigned to them respectively.

A. Results of the Implication of Fuzzy AHP Method to JKL Project

Initially, we employed the above SF model (Fig. 3) for the selected case study. The adaptability of the model is confirmed practically through in-depth conversations with the project manager and development team. Among the gathered data are the experts' perspective about all the activities affected by the SFs at each site. The SFs are prioritized on the base of the importance values given by the experts. The Stepwise Fuzzy AHP procedure to calculate the global weights of SFs for JKL project in the GSD is described below.

The Stepwise Fuzzy AHP procedure to calculate the global weights of SFs for a project GSD is described below.

Step 1: For the expert opinions are concerned, the relative influences are gathered. The first step is to break the problem down into a hierarchical structure for the prioritization of SFs (shown in Fig. 4). The expert opinions are accommodated in a hierarchical structure by employing the following terms for

given SFs:
ORG = Organization
DIS = Distance
KMS = Knowledge Management and Sharing
TRU = Trust
SH = Stakeholder
CCC = Collaboration, Communication and Coordination
PRO = Project
TNT = Tools and Technology
TM = Team
NCASN = National Culture and Social Norms
PF = Physical Factors

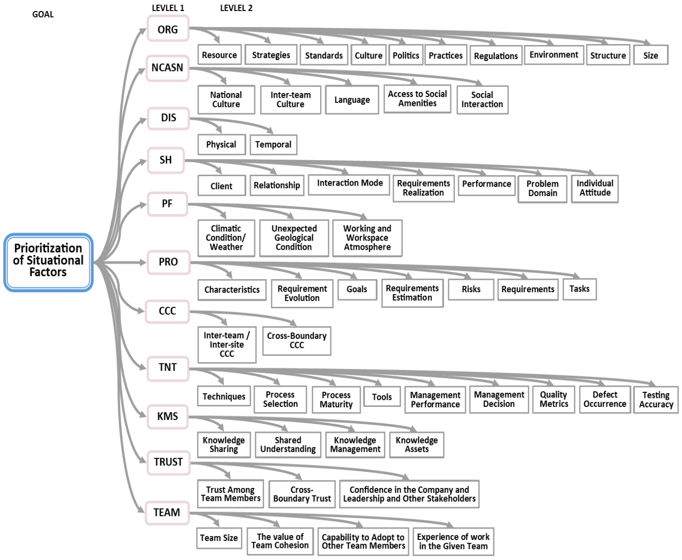


Fig. 4. Analytic hierarchy structure.

The three levels of an AHP model for the prioritization of SFs affecting the JKL project development are shown in Fig. 5, which also presents the structural relationship between the SFs. The first level states the main factors (Table IV) of this research, which are attained by estimating the impact of the sub-factors of the second (Table V) or successive levels. Comparison matrices are created in the same manner for all other criteria and sub-criteria (factors and sub-factors). (Matrices are shown only for main criteria and a sub-criteria under organization due to space limitations.)

TABLE IV. Initial Comparison Matrix Representing the Key Factors

	ORG	DIS	KMS	TRU	SH	CCC	PRO	TNT	TM	NCASN	PF
ORG	1.00	3.00	1.00	1.00	2.00	1.00	1.00	2.00	1.00	1.00	3.00
DIS	0.33	1.00	0.33	0.33	0.50	0.33	0.33	0.50	0.33	0.33	1.00
KMS	1.00	3.00	1.00	1.00	2.00	1.00	1.00	0.50	1.00	1.00	3.00
TRU	1.00	3.00	1.00	1.00	2.00	1.00	1.00	2.00	1.00	1.00	3.00
SH	0.50	2.00	0.50	0.50	1.00	0.50	0.50	1.00	0.50	0.50	2.00
CCC	1.00	3.00	1.00	1.00	2.00	1.00	1.00	2.00	1.00	1.00	3.00
PRO	1.00	3.00	1.00	1.00	2.00	1.00	1.00	2.00	1.00	1.00	3.00
TNT	0.50	2.00	2.00	0.50	1.00	0.50	0.50	1.00	0.50	0.50	2.00
TM	1.00	3.00	1.00	1.00	2.00	1.00	1.00	2.00	1.00	1.00	3.00
NCASN	1.00	3.00	1.00	1.00	2.00	1.00	1.00	2.00	1.00	1.00	3.00
PF	0.33	1.00	0.33	0.33	0.50	0.33	0.33	0.50	0.33	0.33	1.00

Step 2: After making a hierarchical structure, an initial comparison matrix is generated. The consistency check proves the validity of the data because the overall consistency is less than 0.1.

TABLE V. Initial Comparison Matrix Representing the Key Factors

	ORS	OST	OSTD	OC	OP	OPR	OREG	OE	OSTR	OS
ORS	1.00	3.00	3.00	1.00	3.00	1.00	3.00	1.00	3.00	3.00
OST	0.33	1.00	1.00	0.33	1.00	0.33	1.00	0.33	1.00	1.00
OSTD	0.33	1.00	1.00	0.33	1.00	0.33	1.00	0.33	1.00	1.00
OC	1.00	3.00	3.00	1.00	3.00	1.00	3.00	1.00	3.00	3.00
OP	0.33	1.00	1.00	0.33	1.00	0.33	1.00	0.33	1.00	1.00
OPR	1.00	3.00	3.00	1.00	3.00	1.00	3.00	1.00	3.00	3.00
OREG	0.33	1.00	1.00	0.33	1.00	0.33	1.00	0.33	1.00	1.00
OE	1.00	3.00	3.00	1.00	3.00	1.00	3.00	1.00	3.00	3.00
OSTR	0.33	1.00	1.00	0.33	1.00	0.33	1.00	0.33	1.00	1.00
OS	0.33	1.00	1.00	0.33	1.00	0.33	1.00	0.33	1.00	1.00

Step 3: A fuzzy judgment matrix for every factor at each level is established. On the basis of the expert opinion, the linguistic terms (Table I) are assigned to the pairwise comparisons using Table II [43]. According to which a scale of fuzzy numbers representing a membership function by employing three arguments is described showing the range for which the function is defined (Tables VI and VII).

Step 4: Calculate the sum of the rows of factors and sub-factors based on different criteria to give a composed fuzzy column matrix.

Step 5: Calculate the integrated fuzzy expansion to determine the synthetic extent by employing (6). The same method applies to each SF and sub-factor at each level to compute the fuzzy synthetic extent. The combined output of Steps 4 and 5 are shown in Tables VIII and IX.

Step 6: When the synthetic extent is assessed, the synthetic value or the degree of possibility that one fuzzy number is greater than the other can be determined by using (7) to (10). The resulting Composed Crisp Matrix that screens the degree of possibility is shown in Tables X and XI.

Step 7: Determine the minimum degree of possibility, and normalize the values. A comparison of all the synthesized values under the main criteria is made. The minimum value of each factor is calculated. The sum of each factor divided by the sum of the column will determine the preference of that factor on that level. By applying (12), the normalized values are calculated, as shown in Tables XII and XIII.

Step 8: According to [12], using the normalization process, we can get the final weights, as shown in Fig. 5 and 6 for level 1 and level 2, respectively.

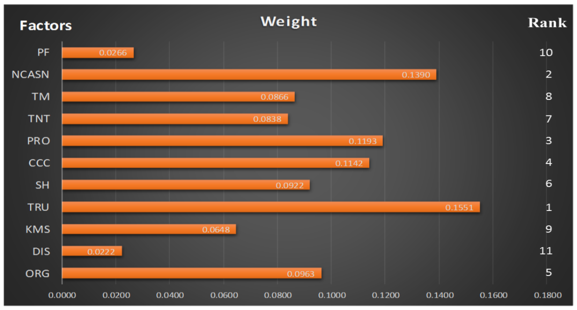


Fig. 5. Overall weight score of the key SFs in GSD.

These results are highlighted by identifying the most significant SFs. Through a precise assessment of the extent of the factors, sorted from highest to lowest, on software project development, we can accelerate the process of improving

TABLE VI. INTEGRATED FUZZY COMPARISON MATRIX AT LEVEL 1 FOR MAIN CRITERIA

	ORG			DIS			KMS			TRU			SH			CCC			PRO			TNT			TM			NCASN			PF		
ORG	1	1	1	1.1722	1.8503	2.4416	1.0085	1.2279	1.4941	0.4491	0.6261	0.9578	0.7534	0.9941	1.3041	0.6812	0.8531	1.0937	0.6482	0.8240	1.1161	0.7340	1.0000	1.3625	0.7340	0.9659	1.2897	0.5468	0.7030	0.9857	1.1420	1.7052	2.3446
DIS	0.4096	0.5405	0.8531	1	1	1	0.5234	0.6888	1.0059	0.3280	0.4131	0.5530	0.4643	0.5842	0.7887	0.4114	0.5237	0.7131	0.3653	0.4842	0.7383	0.4468	0.5777	0.8007	0.4268	0.5643	0.8123	0.3480	0.4313	0.5876	0.7450	1.0053	1.3501
KMS	0.6693	0.8144	0.9915	0.9941	1.4517	1.9105	1	1	1	0.4307	0.5501	0.7426	0.6335	0.8218	1.0746	0.5394	0.6988	0.9330	0.5728	0.6750	0.8531	0.5683	0.7688	1.0565	0.5758	0.7754	1.0845	0.4794	0.5911	0.7913	1.0032	1.4180	1.8821
TRU	1.0441	1.5971	2.2264	1.8084	2.4208	3.0483	1.3465	1.8180	2.3219	1	1	1	1.1487	1.7017	2.1633	1.0353	1.3741	1.6535	0.9603	1.3741	1.7826	1.0873	1.5878	2.0477	1.1096	1.4517	1.7561	0.8481	1.1487	1.4689	1.7653	2.3365	2.8297
SH	0.7668	1.0059	1.3273	1.2679	1.7118	2.1537	0.9306	1.2168	1.5784	0.4623	0.5876	0.8706	1	1	1	0.6224	0.8123	1.1555	0.6711	0.8240	1.0781	0.7754	0.9659	1.2207	0.6808	0.9521	1.3822	0.5631	0.7277	0.9857	1.1154	1.6141	2.2045
CCC	0.9143	1.1722	1.4680	1.4023	1.9094	2.4307	1.0718	1.4310	1.8541	0.6048	0.7277	0.9659	0.8654	1.2311	1.6066	1	1	1	0.7340	1.0000	1.3625	0.8960	1.1722	1.4603	0.9603	1.1161	1.3153	0.6675	0.8531	1.1161	1.2738	1.8325	2.5405
PRO	0.8960	1.2136	1.5427	1.3545	2.0651	2.7378	1.1722	1.4814	1.7457	0.5610	0.7277	1.0414	0.9276	1.2136	1.4902	0.7340	1.0000	1.3625	1	1	1	0.7754	1.1892	1.6973	0.8311	1.1722	1.5743	0.7277	0.9089	1.1227	1.2772	1.8821	2.5519
TNT	0.7340	1.0000	1.3625	1.2490	1.7310	2.2382	0.9466	1.3007	1.7596	0.4884	0.6298	0.9197	0.8192	1.0353	1.2897	0.6848	0.8531	1.1161	0.5892	0.8409	1.2897	1	1	1	0.7866	0.9521	1.2272	0.5743	0.7071	0.9799	1.2200	1.6180	2.0221
TM	0.7754	1.0353	1.3625	1.2311	1.7721	2.3431	0.9221	1.2897	1.7366	0.5695	0.6888	0.9013	0.7235	1.0503	1.4689	0.7603	0.8960	1.0414	0.6352	0.8531	1.2033	0.8149	1.0503	1.2712	1	1	1	0.5548	0.7174	1.0205	1.2813	1.7063	2.1545
NCASN	1.0145	1.4226	1.8287	1.7017	2.3186	2.8736	1.2638	1.6917	2.0858	0.6808	0.8706	1.1791	1.0145	1.3741	1.7758	0.8960	1.1722	1.4981	0.8907	1.1002	1.3741	1.0205	1.4142	1.7412	0.9799	1.3940	1.8026	1	1	1	1.5225	2.2062	2.7950
PF	0.4265	0.5864	0.8757	0.7407	0.9947	1.3422	0.5313	0.7052	0.9968	0.3534	0.4280	0.5665	0.4536	0.6196	0.8965	0.3936	0.5457	0.7851	0.3919	0.5313	0.7830	0.4945	0.6180	0.8197	0.4641	0.5861	0.7805	0.3578	0.4533	0.6568	1	1	1

TABLE VII. INTEGRATED FUZZY COMPARISON MATRIX AT LEVEL 2 FOR THE SUB CRITERIA ORGANIZATION

	ORS			OST			OSTD			OC			OP			OPR			OREG			OE			OSTR			OS		
ORS	1	1	1	0.8531	1.0880	1.3660	0.7030	0.9227	1.2136	0.8409	0.9833	1.1487	1.3007	1.6465	2.0224	0.8123	1.0259	1.3007	1.0000	1.1406	1.3083	0.6693	0.8027	0.9857	1.0503	1.3161	1.6013	1.1256	1.4333	1.7978
OST	0.7320	0.9191	1.1722	1	1	1	0.6929	0.8734	1.1096	0.7426	0.8960	1.0937	1.2821	1.5465	1.8503	0.8409	1.0000	1.1892	0.9466	1.2143	1.5157	0.6693	0.7825	0.9330	1.0000	1.1746	1.3741	1.1031	1.3333	1.5878
OSTD	0.8240	1.0838	1.4226	0.9013	1.1450	1.4432	1	1	1	0.8409	1.0565	1.3195	1.5919	1.9509	2.2914	0.9330	1.1450	1.3940	1.1096	1.3110	1.5247	0.7174	0.8960	1.1323	1.1892	1.3730	1.5468	1.1892	1.5069	1.8601
OC	0.8706	1.0170	1.1892	0.9143	1.1161	1.3465	0.7579	0.9466	1.1892	1	1	1	1.4641	1.8609	2.2440	0.9659	1.0838	1.2136	1.0205	1.1945	1.3940	0.7426	0.8481	0.9857	1.1096	1.3501	1.6013	1.1722	1.4450	1.7514
OP	0.4945	0.6074	0.7688	0.5405	0.6466	0.7800	0.4364	0.5126	0.6282	0.4456	0.5374	0.6830	1	1	1	0.4900	0.5920	0.7426	0.5405	0.6525	0.8170	0.3956	0.4913	0.6335	0.5827	0.6694	0.8007	0.6120	0.7467	0.9330
OPR	0.7688	0.9748	1.2311	0.8409	1.0000	1.1892	0.7174	0.8734	1.0718	0.8240	0.9227	1.0353	1.3465	1.6891	2.0410	1	1	1	0.9659	1.1791	1.4142	0.6929	0.7825	0.9013	1.0000	1.2097	1.4432	1.0873	1.4140	1.7768
OREG	0.7643	0.8767	1.0000	0.6598	0.8235	1.0565	0.6559	0.7628	0.9013	0.7174	0.8371	0.9799	1.2239	1.5325	1.8503	0.7071	0.8481	1.0353	1	1	1	0.5359	0.6257	0.7734	0.9276	1.0880	1.2564	0.9659	1.2202	1.5247
OE	1.0145	1.2457	1.4941	1.0718	1.2780	1.4941	0.8832	1.1161	1.3940	1.0145	1.1791	1.3465	1.5784	2.0353	2.5276	1.1096	1.2780	1.4432	1.2930	1.5981	1.8661	1	1	1	1.2311	1.5920	1.9714	1.3660	1.6891	2.0118
OSTR	0.6245	0.7598	0.9521	0.7277	0.8513	1.0000	0.6465	0.7283	0.8409	0.6245	0.7407	0.9013	1.2490	1.4938	1.7162	0.6929	0.8267	1.0000	0.7960	0.9191	1.0781	0.5072	0.6281	0.8123	1	1	1	0.9466	1.1791	1.4432
OS	0.5562	0.6977	0.8884	0.6298	0.7500	0.9066	0.5376	0.6636	0.8409	0.5710	0.6920	0.8531	1.0718	1.3392	1.6341	0.5628	0.7072	0.9197	0.6559	0.8195	1.0353	0.4971	0.5920	0.7320	0.6929	0.8481	1.0565	1	1	1

TABLE VIII. COMPOSED FUZZY COLUMN MATRIX AND INTEGRATED FUZZY EXPANSION AT LEVEL 1

	Sum of Rows			**Sum of Columns**		
ORG	8.8694	11.7497	15.3899	0.0524	0.0894	0.1533
DIS	5.4686	6.8131	9.2028	0.0323	0.0518	0.0916
KMS	7.4666	9.5651	12.3198	0.0441	0.0728	0.1227
TRU	13.1536	17.8105	22.2984	0.0777	0.1355	0.2221
SH	8.8557	11.4183	14.9567	0.0523	0.0869	0.1490
CCC	10.3900	13.4453	17.1199	0.0614	0.1023	0.1705
PRO	10.2565	13.8539	17.8666	0.0606	0.1054	0.1779
TNT	9.0919	11.6679	15.2047	0.0537	0.0888	0.1514
TM	9.2680	12.0591	15.5032	0.0547	0.0918	0.1544
NCASN	11.9848	15.9644	19.9541	0.0708	0.1215	0.1987
PF	5.6075	7.0684	9.5027	0.0331	0.0538	0.0946

TABLE IX. COMPOSED FUZZY COLUMN MATRIX AND INTEGRATED FUZZY EXPANSION AT LEVEL 2

	Sum of Rows			**Sum of Columns**		
ORS	9.3551	11.3591	13.7446	0.0749	0.1084	0.1566
OST	9.0095	10.7397	12.8256	0.0722	0.1025	0.1461
OSTD	10.2964	12.4681	14.9345	0.0825	0.1190	0.1702
OC	10.0177	11.8621	13.9150	0.0802	0.1132	0.1585
OP	5.5376	6.4560	7.7869	0.0444	0.0616	0.0887
OPR	9.2438	11.0452	13.1038	0.0740	0.1054	0.1493
OREG	8.1577	9.6146	11.3776	0.0653	0.0917	0.1296
OE	11.5621	14.0114	16.5487	0.0926	0.1337	0.1885
OSTR	7.8149	9.1271	10.7441	0.0626	0.0871	0.1224
OS	6.7751	8.1094	9.8665	0.0543	0.0774	0.1124

TABLE X. COMPOSED CRISP MATRIX AT LEVEL 1

	ORG	DIS	KMS	TRU	SH	CCC	PRO	TNT	TM	NCASN	PF
ORG	1	1	0.621	1	0.877	0.853	1	0.977	0.720	1	1
DIS	0.511	0.694	0.143	0.529	0.375	0.367	0.295	0.480	0.231	0.968	1
KMS	0.809	1	0.418	0.833	0.675	0.656	0.497	0.782	0.516	1	1
TRU	1	1	1	1	1	1	1	1	1	1	1
SH	0.975	1	1	0.594	0.850	0.827	0.631	0.951	0.693	1	1
CCC	1	1	1	0.736	1	0.972	1	1	0.839	1	1
PRO	1	1	1	0.769	1	1	1	1	0.870	1	1
TNT	0.994	1	1	0.612	1	0.869	0.541	0.970	0.712	1	1
TM	1	1	1	0.637	1	0.898	0.558	1	0.738	1	1
NCASN	1	1	1	0.896	1	1	1	1	1	1	1
PF	0.543	1	0.727	0.172	0.561	0.407	0.233	0.539	0.512	0.261	1

TABLE XI. COMPOSED CRISP MATRIX AT LEVEL 2 FOR SUB FACTOR ORG

	ORS	OST	OSTD	OC	OP	OPR	OREG	OE	OSTR	OS
OST	0.923	0.794	0.860	1	0.961	1	0.302	1	1	1
OSTD	1	1	1	1	1	1	0.419	1	1	1
OC	1	1	0.929	1	1	1	0.368	1	1	1
OP	0.228	0.288	0.098	0.141	0.251	0.437	0.000	0.506	0.686	1
OPR	0.961	1	0.831	0.899	1	1	0.319	1	1	1
OREG	0.767	0.843	0.634	0.697	1	0.803	0.216	1	1	1
OE	1	1	1	1	1	1	1	1	1	1
OSTR	0.690	0.766	0.556	0.618	1	0.725	0.449	0.390	1	1
OS	0.547	0.616	0.419	0.473	1	0.578	0.371	0.260	0.837	1

TABLE XII. NORMALIZATION AT LEVEL 1

	Minimum Degree of Possibility	Normalization
ORG	0.621	0.0963
DIS	0.143	0.0222
KMS	0.418	0.0648
TRU	1.000	0.1551
SH	0.594	0.0922
CCC	0.736	0.1142
PRO	0.769	0.1193
TNT	0.541	0.0838
TM	0.558	0.0866
NCASN	0.896	0.1390
PF	0.172	0.0266

TABLE XIII. NORMALIZATION AT LEVEL 2 SUB FACTOR ORG

	Minimum Degree of Possibility	Normalization
ORS	0.352	0.0970
OST	0.302	0.0832
OSTD	0.419	0.1157
OC	0.368	0.1015
OP	0.000	0.0000
OPR	0.319	0.0880
OREG	0.216	0.0595
OE	1.000	0.2758
OSTR	0.390	0.1076
OS	0.260	0.0717

development, which leads to competitiveness. The proposed research considers 56 sub-factors that are classified into 11 categories and prioritized by the fuzzy MCDM model. Fig. 6 shows the priority of the main factors and Fig. 5 shows the results for the sub-factors of the Organization factor . Trust is graded as the highest priority because it is the most important factor for improving the yield of the software industry. National Culture and Social Norms is the second highest ranked factor. Project-related factors are 3^{rd}, communication and collaboration are 4^{th}, and Organization and Stakeholders are 5^{th} and 6^{th}, respectively. Previously, geographical distances were considered very significant, but this factor is suggested to not be crucial for the growth of the software industry because the means of communication and collaboration (which is the 4^{th} most important category) have altered the ways of GSD. Other SFs that are also important are listed in Fig. 6. The final assignment of the priority order for the SFs rationally agrees with the experts' opinion. To prove the consistency of our research work, we performed a sensitivity analysis in the next subsection.

B. Comparative and Sensitivity Analysis

In this section, we present a sensitivity analysis of the strengths of the proposed work. Along with the sensitivity

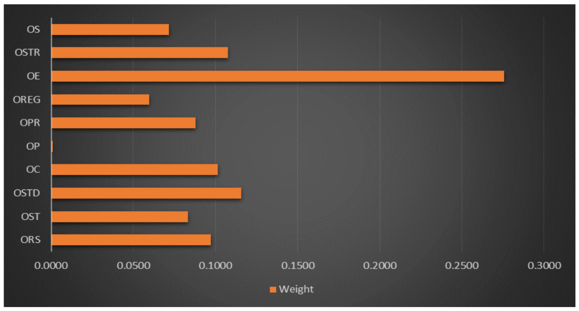


Fig. 6. Overall weights of the subfactors with respect to ORG.

analysis a comparison of the two latest research works on SFs is given in Table XIV. For sensitivity analysis the values of the criteria are systematically changed to understand that how the outcome is affected as a result of the change. Decision makers often have to face uncertain situations, which create hurdles in making the right decisions and implementing the appropriate strategies. Prioritizing SFs by FAHP gives an experimental assessment of the varying situations and their importance for a particular project in GSD. Numerous aspects of different SFs' variations have concerned researchers in general, and to date, the outcomes in the GSD environment have been either contradictory or insufficient. For example, some researchers have considered that cultural differences are important because they have a strong impact on GSD and others did not consider culture to be as important (shown in Appendix A). However, in our case, cultural differences have the second highest rank. As the effect of the SFs varies from country to country and organization to organization, so to prove the consistency of ultimate decision, sensitivity analysis is accompanied with MCDM [45]. This is to examine the effect of variation on criteria causing to affect the final outcome.

TABLE XIV. IDENTIFIED SFS BASED ON A REVIEW OF PROMINENT LITERATURE ON SFS

Research works Identified SFs	Paul Clarcke	Huma Hayat	Our Work Contribution
Organization	✓	✓	Identify all the important SFs with respect to literature and industry point of view and devised a precise mechanism to prioritize them.
Project	✓	✓	
Team	✓	✓	
Knowledge sharing	✓	✓	
Culture	✓	X	
Stakeholder	✓	✓	
Tools & Technology	X	✓	
Communication, collaboration and coordination	X	X	
Trust	X	X	
Distance	X	X	
Workspace	X	✓	

Considering the conflicting issues related to two countries' working atmosphere, the most significant issues with Pakistan's software industry are the lack of Trust and issues related to language, lunch breaks, and vacation schedules (cultural aspects). Pakistani staff complained about the bad English-language skills of Chinese workers, but these workers were very good in other aspects. The major problems with the Chinese workers were related to the lack of Trust and Cultural differences. As discussed above, the language is also a sub-factor of the SF Culture. Therefore, it received the second highest rank, following Trust that is graded at 1st position. Employees at all companies considered project requirement estimation and evaluation to be critical, thus leading the Project category to the 3rd highest rank. From above, the Trust, Culture and Project are the first three to influence the project development. The results of sensitivity analysis are showing the influence of the variation in frequency of these three SFs (Fig. 7) and the results obtained by defining the matrix between criteria variable (Trust) and effort (Fig. 8), the validity of proposed work is proved.

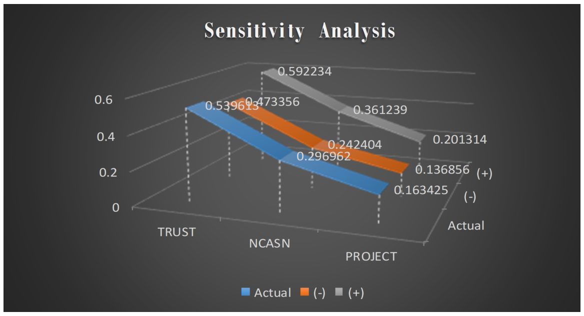


Fig. 7. Sensitivity analysis results.

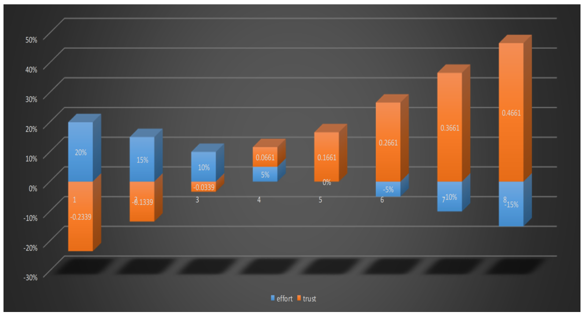


Fig. 8. Sensitivity analysis of trust on total effort.

The main objective of our research is to automate SF evaluation, and this is the first step towards it. The manual assessment of SFs with respect to a certain location is a critical task and should be handled vigorously. An automated and intelligent technique can help to make correct and precise estimates. It should be efficient to reduce the time and human resources compared to human-based techniques. The recommendations from global industry experts and sensitivity analysis proved that the proposed method is the best fit in many cases, as it is infeasible to devote significant efforts to each SF. Our work is limited to only four multinational software companies located in Pakistan and China. The strength of the approach can be tested by conducting numerous case studies in the software industry and comparing the results with those obtained in other countries to reach an ultimate strategy finding.

VI. CONCLUSION

In the global software development environment, due to the complexity of varying situations, the software industry must recognize the situational perspective to get a competitive advantage. From the literature and a software industry survey, we conclude that SFs are the key concerns in GSD and that their proper identification and prioritization can help to reduce the unnecessary complexity of software projects. However, manual decision-making is a challenging and time-consuming process that involves many ambiguous and imprecise factors. The vital importance of this work is the intelligent prioritization of SFs, which will enable managers to understand the relative preference among the factors and design an improved methodology for the software industry to proceed at the global level. As this is a part of our continued research work, we are working to explore the hidden relationships among the SFs and their associated risks and to present the essential competency provisions. We also aim to develop an intelligent, automated and real-time recommender system to keep an up-to-date precept-ability of the SFs and their associated risks to address them in a dynamic environment and support practitioners with the necessary groundwork to realize the booming future of GSD.

Appendix

See extended literature reviewed.

Acknowledgment

This work was supported by National Key R&D Program of China under Grant 2017YFB0802400.

References

[1] J. D. Herbsleb, “Global Software Engineering: The Future of Socio-technical Coordination,” in Future of Software Engineering (FOSE ’07), 2007, pp. 188-198.

[2] F. Lanubile, “Collaboration in Distributed Software Development,” in Software Engineering: International Summer Schools, ISSSE 2006-2008, Salerno, Italy, Revised Tutorial Lectures, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 174-193.

[3] H. Holmstrom, E. O. Conchuir, P. J. Agerfalk, and B. Fitzgerald, “Global Software Development Challenges: A Case Study on Temporal, Geographical and Socio-Cultural Distance,” in 2006 IEEE International Conference on Global Software Engineering (ICGSE’06), 2006, pp. 3-11.

[4] J. D. Herbsleb and D. Moitra, “Global software development,” IEEE Softw., vol. 18, no. 2, pp. 16-20, 2001.

[5] M. Niazi, S. Mahmood, M. Alshayeb, A. A. B. Baqais, and A. Q. Gill, “Motivators for adopting social computing in global software development: An empirical study,” J. Softw. Evol. Process, vol. 29, no. 8, p. e1872, Aug. 2017.

[6] H. F. Hofmann and F. Lehner, “Requirements engineering as a success factor in software projects,” IEEE Softw., vol. 18, no. 4, pp. 58-66, Jul. 2001.

[7] L. Passos, K. Czarnecki, S. Apel, A. Wasowski, C. Kastner, and J. Guo, “Feature-oriented software evolution,” in Proceedings of the Seventh International Workshop on Variability Modelling of Software-intensive Systems - VaMoS ’13, 2013, p. 1.

[8] G. H. Subramanian, G. Klein, J. J. Jiang, and C. L. Chan, “Balancing four factors in system development projects,” Commun. ACM, vol. 52, no. 10, p. 118, Oct. 2009.

[9] H. H. Khan, “Factors generating risks during requirement engineering process in global software development environment,” Int. J. Digit. Inf. Wirel. Commun., vol. 4, no. 1, pp. 63-78, 2014.

[10] P. D. Chatzoglou, “Factors affecting completion of the requirements capture stage of projects with different characteristics,” Inf. Softw. Technol., vol. 39, no. 9, pp. 627-640, Jan. 1997.

[11] R. Prikladnicki and J. L. N. Audy, “Managing Global Software Engineering: A Comparative Analysis of Offshore Outsourcing and the Internal Offshoring of Software Development,” Inf. Syst. Manag., vol. 29, no. 3, pp. 216-232, Jun. 2012.

[12] P. Clarke and R. V. O’Connor, “The situational factors that affect the software development process: Towards a comprehensive reference framework,” Inf. Softw. Technol., vol. 54, no. 5, pp. 433-447, May 2012.

[13] S. Dallman, L. Nguyen, J. W. Lamp, and J. L. Cybulski, “Contextual Factors Which Influence Creativity in Requirements Engineering,” in 13th European Conference on Information Systems (ECIS), 2005, pp. 1734-1745.

[14] P. J. Agerfalk and J. Ralyte, “Situational Requirements Engineering Processes: reflecting on method engineering and requirements practice,” Softw. Process Improv. Pract., vol. 11, no. 5, pp. 447-450, Sep. 2006.

[15] S. Ghosh, A. Dubey, and S. Ramaswamy, “C-FaRM: A collaborative and context aware framework for requirements management,” in 2011 4th International Workshop on Managing Requirements Knowledge, 2011, pp. 29-30.

[16] J. Hanisch, T. Thanasankit, and B. Corbitt, “Understanding the cultural and social impacts on requirements engineering processes: Identifying some problems challenging virtual team interaction with clients,” in ECIS 2001 Proceedings. 43, 2001.

[17] C. Michael and Kang. Kyo, “Issues in Requirements Elicitation,” Software Engineering Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, Technical Report CMU/SEI-92-TR-012, 1992.

[18] J. Cameron, ”Configurable development processes,” Commun. ACM, vol. 45, no. 3, pp. 72-77, Mar. 2002.

[19] P. Xu and B. Ramesh, “Software Process Tailoring: An Empirical Investigation,” J. Manag. Inf. Syst., vol. 24, no. 2, pp. 293-328, Oct. 2007.

[20] T. W. Ferratt and B. Mai, “Tailoring software development,” in Proceedings of the 2010 Special Interest Group on Management Information System’s 48th annual conference on Computer personnel research on Computer personnel research - SIGMIS-CPR ’10, 2010, p. 165.

[21] B. Dede and I. Lioufko, “Situational Factors Affecting Software Development Process Selection,” University of Gothenburg, 2010.

[22] P. A. Dabholkar and R. P. Bagozzi, “An attitudinal model of technology-based self-service: Moderating effects of consumer traits and situational factors,” J. Acad. Mark. Sci., vol. 30, no. 3, p. 184, Jun. 2002.

[23] D. E. Damian and D. Zowghi, “RE challenges in multi-site software development organisations,” Requir. Eng., vol. 8, no. 3, pp. 149160, Aug. 2003.

[24] L. B. Lempert, “Asking Questions of the Data: Memo Writing in the Grounded Theory Tradition,” in The SAGE Handbook of Grounded Theory, 1 Oliver’s Yard, 55 City Road, London England EC1Y 1SP United Kingdom: SAGE Publications Ltd, 2007, pp. 245-264.

[25] W. Bekkers, I. van de Weerd, S. Brinkkemper, and A. Mahieu, “The Influence of Situational Factors in Software Product Management: An Empirical Study,” in 2008 Second International Workshop on Software Product Management, 2008, pp. 41-48.

[26] A. M. Soderberg, S. Krishna, and P. Bjorn, “Global Software Development: Commitment, Trust and Cultural Sensitivity in Strategic Partnerships,” J. Int. Manag., vol. 19, no. 4, pp. 347-361, Dec. 2013.

[27] B. L. Marcolin, “Spiraling Effect of IS Outsourcing Contract Interpretations,” in Information Systems Outsourcing, Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 277-310.

[28] P. Clarke and R. V. O’Connor, “Changing Situational Contexts Present a Constant Challenge to Software Developers,” in Systems, Software and Services Process Improvement, 2015, pp. 100-111.

[29] K. Gulzar, J. Sang, and M. Ramzan, “Conflict Identification among Usability Requirements Using Fuzzy Logic,” Int. J. Comput. Theory Eng., vol. 9, no. 4, pp. 268-272, 2017.

[30] K. Gulzar, J. Sang, M. Ramzan, and M. Kashif, “Fuzzy Approach to Prioritize Usability Requirements Conflicts: An Experimental Evaluation,” IEEE Access, vol. 5, pp. 13570-13577, 2017.

[31] T. L. Saaty, “The analytical hierarchical process.” New York, USA: McGraw-Hill, 1980.

[32] C. C. Yang and B. S. Chen, “Key Quality Performance Evaluation Using Fuzzy AHP,” J. Chinese Inst. Ind. Eng., vol. 21, no. 6, pp. 543-550, Jan. 2004.

[33] L. A. Zadeh, “Fuzzy sets,” Inf. Control, vol. 8, no. 3, pp. 338-353, Jun. 1965.

[34] P. J. M. van Laarhoven and W. Pedrycz, “A fuzzy extension of Saaty’s priority theory,” Fuzzy Sets Syst., vol. 11, no. 1-3, pp. 229-241, 1983.

[35] J. J. Buckley, “Fuzzy hierarchical analysis,” Fuzzy Sets Syst., vol. 17, no. 3, pp. 233-247, Dec. 1985.

[36] R. Xu, “Fuzzy least-squares priority method in the analytic hierarchy process,” Fuzzy Sets Syst., vol. 112, no. 3, pp. 395-404, Jun. 2000.

[37] L. Mikhailov, “A fuzzy programming method for deriving priorities in the analytic hierarchy process,” J. Oper. Res. Soc., vol. 51, no. 3, pp. 341-349, Mar. 2000.

[38] L. Wang, J. Chu, and J. Wu, “Selection of optimum maintenance strategies based on a fuzzy analytic hierarchy process,” Int. J. Prod. Econ., vol. 107, no. 1, pp. 151-163, May 2007.

[39] C. C. Sun, “A performance evaluation model by integrating fuzzy AHP and fuzzy TOPSIS methods,” Expert Syst. Appl., vol. 37, no. 12, pp. 7745-7754, Dec. 2010.

[40] E. Bulut, O. Duru, T. Kececi, and S. Yoshida, “Use of consistency index, expert prioritization and direct numerical inputs for generic fuzzy-AHP modeling: A process model for shipping asset management,” Expert Syst. Appl., vol. 39, no. 2, pp. 1911-1923, Feb. 2012.

[41] A. Nieto-Morote and F. Ruz-Vila, “A Fuzzy AHP Multi-Criteria Decision-Making Approach Applied to Combined Cooling, Heating, and Power Production Systens,” Int. J. Inf. Technol. Decis. Mak., vol. 10, no. 3, pp. 497-517, May 2011.

[42] Y. C. Erensal, T. ncan, and M. L. Demircan, “Determining key capabilities in technology management using fuzzy analytic hierarchy process: A case study of Turkey,” Inf. Sci. (Ny)., vol. 176, no. 18, pp. 2755-2770, Sep. 2006.

[43] A. Sarfaraz, P. Mukerjee, and K. Jenab, “Using fuzzy analytical hierarchy process (AHP) to evaluate web development platform,” Manag. Sci. Lett., vol. 2, no. 1, pp. 253-262, Jan. 2012.

[44] D. Y. Chang, “Applications of the extent analysis method on fuzzy AHP,” Eur. J. Oper. Res., vol. 95, no. 3, pp. 649-655, Dec. 1996.

[45] I. Syamsuddin, “Multicriteria Evaluation and Sensitivity Analysis on Information Security,” Oct. 2013.

Evaluation and Analysis of Bio-Inspired Optimization Techniques for Bill Estimation in Fog Computing

Hafsa Arshad[1], Hasan Ali Khattak[1] *, Munam Ali Shah[1], Assad Abbas[1] and Zoobia Ameer[2]
[1]Department of Computer Science, COMSATS University Islamabad, Islamabad 44500, Pakistan
[2]Department of Physics, Shaheed Benazir Bhutto Women University, Peshawar 25500, Pakistan

***Abstract*—In light of constant developments in the realm of Information Communication and Technologies, large-scale businesses and Internet service providers have realized the limitation of data storage capacity available to them. This led organizations to cloud computing, a concept of sharing of resources among different service providers by renting these resources through service level agreements. Fog computing is an extension to cloud computing architecture in which resources are brought closer to the consumers. Fog computing, being a distinct from cloud computing as it provides storage services along with computing resources. To use these services, the organizations have to pay according to their usage. In this paper, two nature-inspired algorithms, i.e. Pigeon Inspired Optimization (PIO) and Binary Bat Algorithm (BBA) are compared to regulate the effective management of resources so that the cost of resources can be curtailed and billing can be achieved by calculating utilized resources under the service level agreement. PIO and BBA are used to evaluate energy utilization by cloudlets or edge nodes that can be used subsequently for approximating the utilization and bill through a Time of Use pricing scheme. We appraise above-mentioned techniques to evaluate their performance concerning the bill estimation based on the usage of fog servers. With respect to the utilization of resources and reduction in the bill, simulation results have revealed that the BBA gives pointedly better results than PIO.**

***Keywords*—Cloud computing; fog computing; bio-inspired algorithms; pricing; cloudlets**

I. Introduction

In order to enhance the efficiency and performance in distributed computing, components of a software system are distributed or shared among multiple systems. Cloud computing is regarded as a type of distributed computing that comprises the services available to users from distant locations. It is an evolving computing architecture that counts on shared computing resources to handle applications in spite of having local servers. The users can utilize through cloud computing, several services and resources such as processing and storage through internet. The on-demand delivery of the Information Technology (IT) resources is guaranteed by charging the services through the pay as you go model. The cloud customers pay to the service providers for providing the services directly to the end users. One can acquire as many resources as required by remitting the resources used. In fact, the cloud computing has either fully removed or has substantially reduced the costs of developing and maintaining the IT infrastructure for small and medium sized enterprises (SME's). Due to some intrinsic issues, many applications cannot work effectively in the cloud environment. For example, because of the low bandwidth the data transmission to the cloud can not be at the same rate at which it is generated. Therefore, noteworthy delays can be faced which are unacceptable in certain cases. Fog computing has established its efficacy to overcome several issues of distributed computing, including inefficient resource management, Quality-of-Service (QoS), security, and privacy issues. The data in the fog computing environment is processed locally in a virtual platform at a much faster pace as compared to a centralized cloud server.

The term Fog computing was introduced by Cisco Systems for the first time, which is also known as edge computing. It facilitates in wireless data transfer in the Internet of Things (IoT) model by taking the computing power near to the Edge of the network so that the devices have an easy and vigorous access. The key idea of fog computing is to boost the efficiency and minimize the data transmissions in predominance to cloud for analysis, storage, and processing purposes. One of the key benefits is the lesser dormancy for devices and lesser network load on the internet mainstay. There are numerous application domains of fog computing, including software-defined networks, vehicular networks, smart grid, smart cities and smart buildings [1]. The long-term analytic processing is carried out in the cloud environment whereas short term data analysis takes place at the fog servers or edge nodes. It is also worth mentioning that the fog computing cannot be completely replaced by the cloud. It only strengthens the cloud computing by decreasing the data sent to the cloud for processing through the mainstay Internet.

As stated above, the fog computing lessens the bulk of data sent to the cloud, subsequently conserving the network bandwidth. Furthermore, it also advances the system response time by keeping the data close to the edge of the network and making the data available promptly. Moreover, the fog computing supports agility and minimizes the dormancy or latency.

The organizations rent the services from the fog service providers. However, it is imperative to efficiently manage and utilize the available resources to reduce the bills. One way to reduce the costs is to optimally utilize the resources through the Time of Use (ToU) pricing model. To illustrate the scenario, an example of smart homes is considered where smart meters are placed for measuring the energy consumed that is subsequently used for bill calculation [2], [3]. Fig. 1 shows general three layered architecture of edge nodes or cloudlets. In the top-

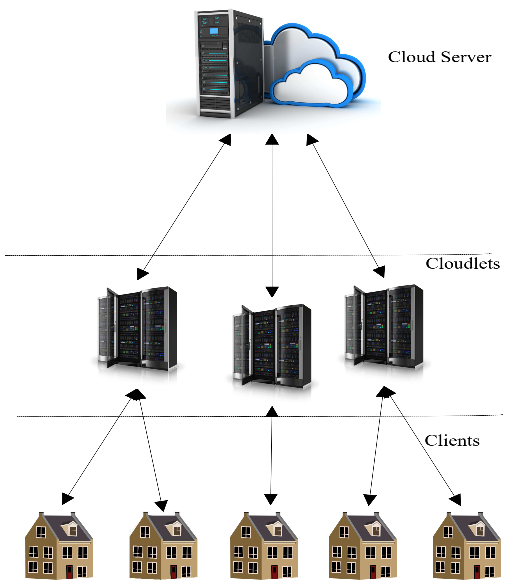


Fig. 1. General 3-layered architecture.

most layer, cloud server is placed. The second (middle) layer is the application layer that comprises of cloudlets while, the bottom layer which in fact is the third layer contains the clients or consumers of the resources. The consumers or clients can manage the cost and energy consumed, by optimal scheduling. With the increasing demand of cloud services, the necessity to estimate the cost of the consumed resources has increased to large extend. Here, our objective is to estimate the bills on the basis of the energy consumed by the cloudlets by using the bio-inspired techniques, i.e., Pigeon Inspired Optimization (PIO) and the Binary Bat Algorithm (BBA). As a result, the resource utilization can be adjusted in response to the pricing scheme.

In this paper, we employ two bio-inspired optimization techniques, namely, PIO and BBA, to determine the optimal solution for calculating the bills based on the usage of the cloudlets. Along with bio-inspired techniques, ToU pricing signal is used that actually favours the efficient utilization of the system along with the reduction in the overall costs that will eventually benefit both the customers as well as service providers. Through ToU rating scheme, one can get to know when and how much energy is being consumed and likewise consumers also get benefit as they can save their operational costs. The echolocative behaviour of bats is used in the BBA [4]–[6] while homing behaviour of pigeons is considered in the PIO [7]–[9] to obtain the optimal solution. The reason of selection of these two techniques is that they perform better than many other techniques. BBA performed better than PSO and GA [4], [5] and PIO outpaced EDE in [7], [8]. These optimization algorithms perform better than other algorithms in terms of approximating the cost and resource which are being utilized.

The paper below has been presented as follows. Section II presents the related work and the proposed model has been explained in Section III. In Section IV, results and simulations of the proposed system have been explained and the Section V settles the paper and highpoints the future work.

II. Related Work

Being a nascent computing technology, fog computing and cloudlet resource management are under consideration by many scholars. The literature review done in this paper is purely based on cost estimation techniques which are utilized for determining the bills based on the resource usage of cloudlet. State-of-the-art work is given in Table I, stating the contributions of the researchers in this regard. Cost, resource management, scheduling and latency are the key features which are considered along with several parameters including servers, energy consumption, time, memory and throughput.

The authors in [10], proposed a cost makespan aware scheduling heuristic which is a cost scheduling algorithm. The objective of this work is to acquire the balance between the cost for the utilization of cloud resources and the performance of application execution. However, this proposed scheduling can be extended for large scale applications by keeping energy efficiency in mind. In [11], the authors proposed greedy heuristics for Basic Service Placement Problem (BSPP) and Cost-aware Service Placement Problem (CSPP). The main aim of this work is to help service providers to minimize the access latency and reduce the cost of service providers, including resource usages and service placements updating cost on cloudlets. However, privacy and security of the users mobility is ignored. The authors in [12], have proposed a Cost Deadline Based (CDB) task scheduling algorithm in order to schedule tasks with cloudsim. The main goal of this work is to reduce the rate of missed deadline and help to define cost in the perspective of both user and provider. Efficient utilization of resources is not considered in this work.

The authors in [13], have proposed a heuristic based workflow scheduling scheme using cloud-pricing model. Scheduling approaches include two phases; VM packing and Multi Request to Single Resource (MRSR) scheme. The proposed work reduces the number of entailed VM instances and obtain a cost saving while guaranteeing the users SLA. The cost is reduced by 30% by using this scheme as compared to other conventional workflow scheduling schemes. However, processing and communication cost of tasks are not considered. In [14], priority-based load balancing approach is proposed that schedules and migrates the VM corresponding to the weights assigned to each virtual machine. The goal of this work is to provide enhanced services to the users who are giving more revenue to the service provider. The drawback of this work is that fixed threshold values are used. Cost saving super professional executor (Suprex) with auto-scaling mechanism is proposed by Aslanpour et al. in [15]. The aim of this work is to provide an executor with the capability to isolate the surplus VM till the billed hour is terminated in order to overcome the challenge of postponed VM startup (Suprex). Suprex executor, reduces the renting cost of VM by 7%. However, in some situations, this executor results in lower utilization. In [16], a framework for enhanced resource allocation and prediction on the basis of traits and characteristics of a customer is proposed by a cloud broker in the cloud federation environment. On the basis of past record or data of each customer, the resources and prices are determined. However, QoS is ignored in this work. The authors in [17], have proposed a Science Gateway Cloud (SGC)platform and a cost adaptive resource management scheme along with workflow scheduling scheme with a division policy. Although, deadline assurance and cost reduction can be gratified simultaneously, however, cost performance degradation occurs when resource pool management is inefficiently done.

The aim of [18], is to lessen the cost for each queue and the sum cost for both mobile user and the CSP. Algorithms for energy cost minimization, while verifying finite processing delay are proposed in this work. Since, single CSP is used in this work hence, quality of offloading service is sacrificed which can further be enhanced by using multiple CSPs. The authors in [19], have proposed dual side dynamic control algorithms for cost-delay trade-offs of mobile users and CSP in MCC for which cooperation and non-cooperation scenarios are considered.In non-cooperation scenario,the users and the CSP are minimizing their own cost for prescribed delay constraints while in cooperation scenario they are minimizing the social cost for stated processing delay constraints. However, QoS and deadlines ignored in this work.

Cheng et al. in [20], have proposed a deep reinforcement learning (DRL) based system with resource provisioning and task scheduling to minimize the energy cost for CSPs with huge amount of user requests and large-scale data centers. ToU and RTP are the pricing signals applied in this work along with Pay-As-You-Go billing contract or agreement. Energy cost efficiency is improved 320% and 144% runtime reduction is achieved. However, dependencies are involved while dealing with large amount of user requests. In [21], the authors have proposed a mathematical framework for dynamic on demand pricing model using IaaS cloud service instances by considering users and providers utility. Genetic algorithm is used for optimized estimation and minimized execution cost. Comparative scrutiny of the intended framework with a utility based pricing model in terms of minimum bill calculation is performed in this work. However, dynamic behaviour of network is not analyzed. The authors in [22], have proposed a load balancing with optimal cost scheduling algorithm to minimize cost and processing power. Already accommodated requests are rescheduled to make space for new upcoming requests for cost optimization at CSP. This algorithm does not operate when all the VMs are busy in the data centers and the upcoming new requests are in waiting condition.

In [23], a negotiation-based iterative approach for task scheduling (NBTS) is proposed to minimize the bill under dynamic energy pricing. The proposed algorithm ends when total energy cost is not diminished. Up to 51.8% improvement is achieved in electric bill reduction. The paper [24] is the extension of [23] as they both are addressing the same problem. The authors in [24], have proposed a negotiation-based cost minimization (NBCM) algorithm in order to minimize the energy cost of users. Along with that task scheduling (NBTS) and energy storage control (NBSC) systems are also proposed. The main aim of this work is to schedule electricity consumption in a way that the electric bill of the users can be minimized. The total energy cost reduction of 64.22% is achieved as compared to the baseline methods. Both [23] and [24] have used dynamic energy pricing including ToU and total power consumption-dependent.

The authors in [25], have proposed a model to optimize the resource utilization and cost reduction using dynamic VM provisioning in the cloud. A cost reduction of 16% is achieved using this model. However, pricing condition applied to spot instances can be further improved. In [26], an admission cost model is proposed for modelling different resource consumptions. The goal of this work is to enhance the system throughput in a cloudlet environment. A storage extension for the existing cloudsim framework to enable simulations of storage as-a-Service (STaaS) components are proposed by Sturm et al. in [27]. For validation of the extension, resource utilization and the cost that arises due to the usage of STaaS clouds are evaluated. However, there is no mechanism for dealing with complex SLAs and realistic pricing models are not used. The authors in [28], have proposed a VM placement scheme to resolve the cost optimization problem along with that VM reallocation grounded on resource utilization-aware activities is also proposed. The objective of this work is to lower the operating cost so that the performance degradation is lessened than the threshold. However, it does not reflect overall trends i.e., temporary resource utilization is done. In [29], Cost oriented model (CoM) is proposed to optimally allocate cloud computing resources for demand side management. The aim of this model is to reduce the rental cost of cloud resources. Amazon cloud computing service-based pricing scheme along with ToU and RTP for demand side management is used in this paper. QoS metric-based resource provisioning technique is proposed in [30] for the cloud computing environment, in order to minimize the cost of cloud workloads and execution time along with other QoS parameters. However, resource utilization can be further improved. The authors in [31], have proposed an efficient online algorithm for dynamic joint VM pricing, scheduling of job and server provisioning containing geo-distributed data centers in order to maximize profit of a cloud provider. Though the backup resources are expensive or job drop penalty is high.

The algorithm proposed in [32], balances the load effectively among VMs by mapping tasks on the basis of foraging behaviour of honey bees. It also minimizes the cost of consuming virtual machine instances. Although, the load of independent tasks is considered for balancing, yet the load of dependent tasks is not considered in this work. In [33], a provably-efficient online dynamic scheduling and pricing (Dyn-SP) algorithm for delay tolerant batch services is proposed to amplify the profit of service provider. It produces close-to-optimal profit. This algorithm is limited to 2 service classes although multiple service classes can be used. The authors in [34], have proposed bi-level cost-wise optimization approach to schedule the consumption level of customers to attain the optimal performance regarding energy minimization cost. However, security is ignored. In [35], the authors have proposed cost optimization algorithm, determining that which resources should be taken on lease from public cloud to accomplish the work ow execution within deadline. Level based scheduling algorithm also known as sub-deadline of work ow is also proposed. Cost based resource allocation strategy is proposed in [36], that will optimize the providers profit along with load balancing technique. The objective of this work is to reduce the total completion time and execution time and increase the gain of providers along with the satisfaction of users by assigning priorities to the users who have paid the cost for the used resources to the service provider. One of the drawback of this work is that static threshold values are used. In [37], the authors have proposed a dynamic resource allocation framework for NFV-enabled MEC consisting of fast online heuristic-based incremental allocation mechanism and a re-optimization solution. The aim of this paper is to efficiently allocate all the resources among NFV-enabled MECs so that low latency and cost efficiency are achieved. Although, 33% of the cost can be saved by the proposed work as compared to existing solutions with xed-location MEC however, this framework does not assist multiple services of different performance requirements. In [38], a mathematical model of cloud computing is proposed by Ibrahim et al., in the economic fractional dynamic system. In the proposed model, agent will separately optimize the entire cost involving two mechanisms; the implementation cost of cloud computing pattern and the effort cost of shifting to the cloud computing pattern. However, hope bifurcation of economic systems is dependent on several other rules. In [39], the authors have proposed a cloud computing simulation model of the smart grid for hosting smart grid applications, determining VM/ cloudlet parameters, along with cost prediction and processing time for the smart grid tasks. However, security policies for smart grid cloud are ignored.

It can be seen clearly from this state-of-the-art work that there exist a lot of gaps which needs to be filled. In this regard, this work is based on minimizing cost while utilizing maximum amount of resources.

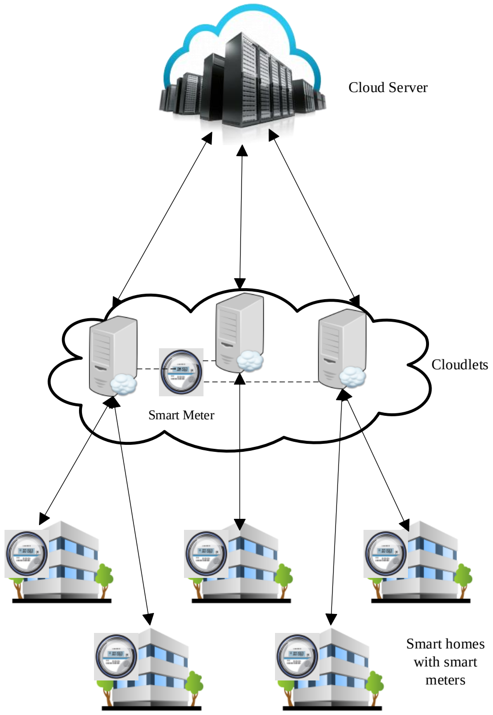


Fig. 2. Proposed cloudlet architecture.

III. Proposed Methodology

In the proposed system, we apply PIO and BBA along with the ToU pricing scheme to estimate the bill of the cloudlet utilization. The task is accomplished by determining the energy consumption. Fig. 2 shows the proposed architecture of the cloudlets. The cloudlets and smart homes are connected with the smart meter to determine the energy consumed which is further used for bill estimation. The information about the amount of energy consumed by the smart homes is gathered through the smart meter and transmitted to the cloudlets. Eventually, the optimization technique is applied and optimal schedule of power consumption of all of the smart homes and cloudlets is examined. The cost is determined by estimating the amount of resources utilized by the clients by using pay as you go pricing model. The consumers can control the cost and energy consumed by optimal scheduling. For this, two bio-inspired optimization techniques, namely, PIO and BBA are used.

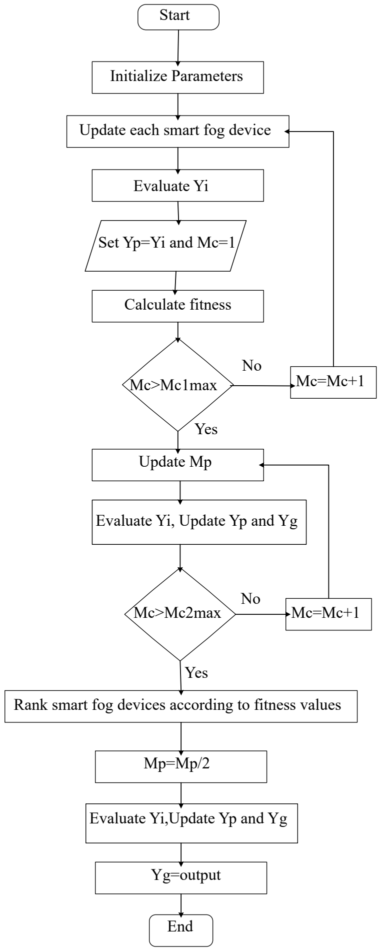


Fig. 3. Flowchart of PIO in the context of cloudlets.

A. *Working of Optimization Techniques*

The working of the two bio-inspired techniques used for determining the bills is given in this section. The description of which is also given in the form of flowcharts.

1) Pigeon Inspired Optimization (PIO): Considering the homing behaviour of pigeons, Duan and Qiao proposed PIO in 2014 [9]. The PIO has a better optimization performance and a high convergence speed [40]. Two major operators used in this technique are: (i) Map and Compass Operator; and (ii) Landmark Operator.

TABLE I. State-of-the-Art Work

Features					Limitations						Parameters					
Reference papers	Cost	Latency	Resource Management	Scheduling	Scalability	Efficiency	Performance	QoS	High Cost	Security	Servers	Energy Consumption	Memory	Throughput	Time	Experiments
[10]	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✓	✗	✓	✗	✓	✓
[11]	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✓	✗	✓
[12]	✓	✗	✗	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗	✓	✗	✓
[13]	✓	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✓	✓
[14]	✗	✗	✓	✗	✓	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✓
[15]	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓
[16]	✗	✗	✓	✗	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✓	✓
[17]	✓	✗	✓	✓	✗	✓	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗
[18]	✓	✓	✗	✗	✗	✗	✗	✓	✗	✗	✓	✓	✗	✓	✓	✓
[19]	✓	✓	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗	✓	✗	✓	✗
[20]	✓	✗	✓	✓	✗	✓	✓	✓	✗	✗	✗	✓	✗	✗	✓	✓
[21]	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓
[22]	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓
[23]	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗
[24]	✓	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗
[25]	✓	✗	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✓	✓
[26]	✓	✗	✓	✗	✗	✓	✓	✓	✓	✗	✗	✓	✓	✗	✓	✓
[27]	✓	✗	✓	✗	✓	✓	✗	✗	✓	✗	✗	✗	✓	✗	✗	✓
[28]	✗	✗	✓	✗	✓	✓	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗
[29]	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	✓	✗
[30]	✓	✗	✓	✗	✗	✓	✓	✗	✗	✗	✓	✗	✓	✗	✗	✓
[31]	✓	✗	✗	✓	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗
[32]	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓
[33]	✓	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	✗
[34]	✓	✗	✗	✓	✗	✗	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗
[35]	✓	✗	✓	✓	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✓	✓
[36]	✓	✗	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓
[37]	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✓	✓
[38]	✓	✗	✗	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
[39]	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✓	✓

i) Map and Compass Operator

Parameter initialization along with random generation of population is done. Then fitness is calculated. Pigeons use magneto-reception to judge the earths magnetic field. They depend on sun and magnetic particles while moving towards destination.

ii) Landmark Operator

The pigeons fly near to their target by leaning on the adjacent landmarks. If the landmarks are known to them, they will fly to their goal directly. On contrary, if the landmarks are unknown to them and pigeons are far apart from their destination then they will track the pigeons who are familiar with the landmarks. Using this operator, half of the population is stranded after sorting according to the fitness function. This is how pigeons find food and in this case, we will find an optimal solution i.e., the cost estimation. Fig. 3 shows the working of the PIO technique. Some of the steps are taken from [9] in order to find the global best solution using a fitness function.

2) Binary Bat Algorithm (BBA): Bat Algorithm (BA) was proposed by Xin-She Yang in 2010 [4] for solving the complex optimization problems. Inspired from the echolocative behavior of bats global optimization is performed. Binary version of BA i.e., BBA [5] uses artificial bats navigating and hunting in binary search spaces by switching their positions between 0 to 1 values. Bats use natural sonar in order to navigate and hunt. When finding the prey bats adopt two main characteristics. While chasing the prey bats tend to decrease the loudness and increase the emission rate of ultrasonic sound. It has been proven that BBA is capable of providing competitive results as compared to most well known algorithms such as Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) in terms of convergence speed and improved local optima avoidance [5], [6]. As stated by Mirjalili *et al.*, BBA has fastest convergence rate along with high performance in finding global solutions.

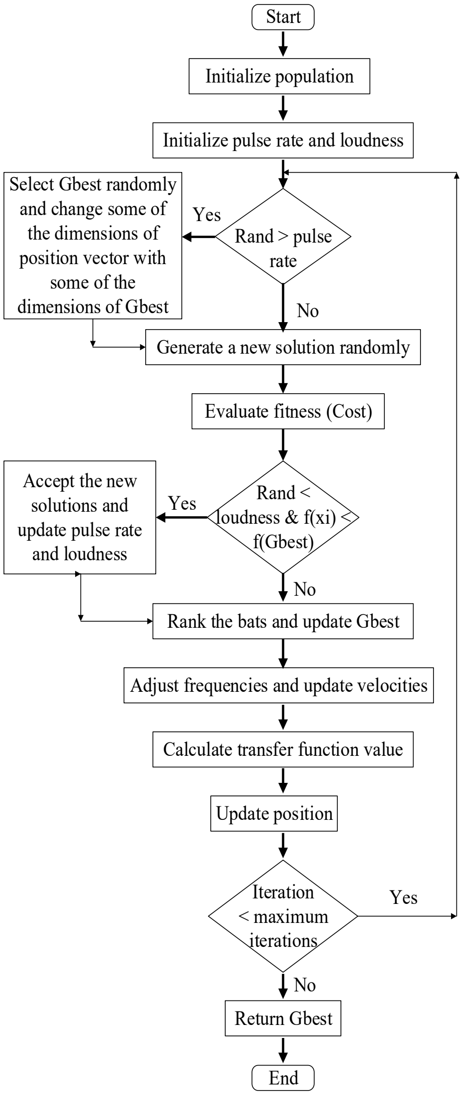

Start
Initialize population
Initialize pulse rate and loudness
Rand > pulse rate
Yes
Select Gbest randomly and change some of the dimensions of position vector with some of the dimensions of Gbest
No
Generate a new solution randomly
Evaluate fitness (Cost)
Rand < loudness & f(xi) < f(Gbest)
Yes
Accept the new solutions and update pulse rate and loudness
No
Rank the bats and update Gbest
Adjust frequencies and update velocities
Calculate transfer function value
Update position
Iteration < maximum iterations
Yes
No
Return Gbest
End

Fig. 4. Flowchart of BBA in the context of cloudlets.

The working of the BBA is given in Fig. 4, where fitness is considered as the cost and the Gbest is the one with the minimum cost.

IV. Simulations and Results

In order to evaluate the bio-inspired techniques for the purpose of estimation of bills, MATLAB is used. Simulations are held in MATLAB for ease of visualization by setting up a Fog computing environment, leveraging six fog nodes. Simulations are held by considering varying values of power consumption along with cloudlets operational time in which resources are being utilized. Experiments are conducted under similar conditions, i.e., under equal amount of load both PIO and BBA are compared and evaluated.

Along with the PIO and the BBA, ToU pricing signal is used to lessen the bill of the cloudlets. The scheduling of the cloudlets is done in order to utilize resources efficiently so that overall cost can be reduced.

Fig. 5, shows the resources utilized by users. With the increased resource utilization, cost will also be increased.

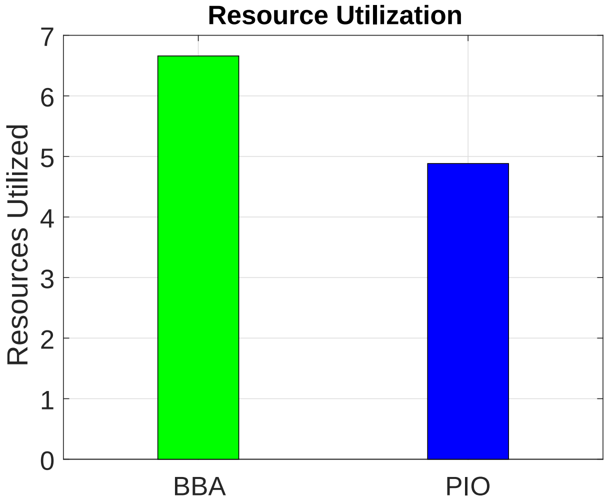


Fig. 5. Comparison of resource utilization for PIO and BBA.

Results clearly show that by using the BBA algorithm, more resources can be acquired as compared to the PIO. Around 25% more resources can be used in BBA in contrast to PIO.

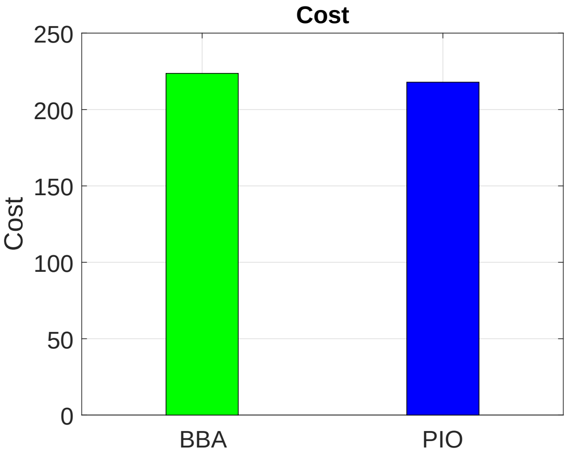


Fig. 6. Comparison of utilization cost for PIO and BBA.

In Fig. 6, cost of the resources which are utilized is given. This shows that the cost of BBA is little bit more than that of PIO. This is because through BBA more resources can be utilized in comparison with PIO. Therefore, consumers have to pay for the utilized resources. Although BBA requires a bit more cost as compared to PIO, still BBA is better as a lot more resources can be acquired using BBA with a minor increase in cost.

Fig. 7, represents the hourly cost of the resources utilized by the consumers. From figure, it can be seen that highest peaks are formed by the PIO throughout. At a certain point of the day BBA may result in high cost however, throughout the day cost of PIO is more i.e., highest peaks are formed by PIO. The maximum cost is increased till 54 cents while using BBA the cost is less throughout the day in contrast to PIO.

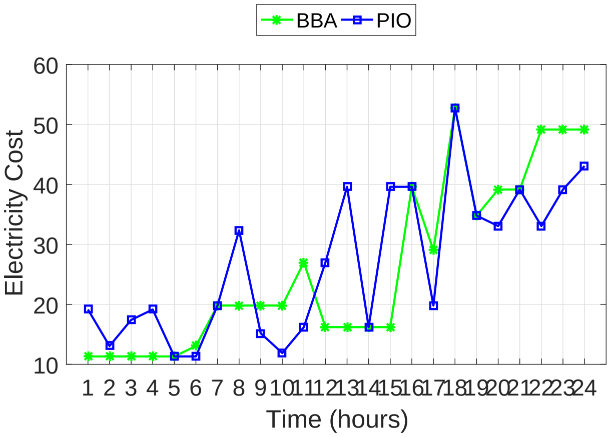


Fig. 7. Hourly cost for the simulation.

V. Conclusions

In this paper, two nature inspired algorithms were compared and evaluated regarding their performance in means of estimation of utility bills based upon the usage of cloudlets. The Binary Bat Algorithm (BBA) and Pigeon Inspired Optimization (PIO) along with Time of Use pricing signal were used for cost estimation. The evaluation is done while considering the resources utilized by the consumers and thus their hourly consumption cost. Simulation results show that the BBA can efficiently utilize large amount of resources i.e., 25% as compared to PIO, therefore, it has a minor increase in cost as compared to the PIO. The cost of BBA is approximately 5% more than that of the PIO. From the simulation results, we conclude that the overall performance of BBA is better than PIO, as with only 5% increase in cost 25% more resources can be utilized. These techniques lead to cost estimation based upon the resources utilized by the cloudlets. In future, we plan to perform experiments using other noteworthy bio-inspired optimization techniques, such as Earth Worm Optimization Algorithm (EWA), Bellman-Ford, Beetle Antennae Search Algorithm (BAS) for utility pricing where the simulations would be setup in a dominated environment fixed in devoted simulator such as iFogsim.

References

[1] Bonomi, Flavio, R. Milito, J. Zhu, and S. Addepalli. ”Fog computing and its role in the internet of things.” In Proceedings of the first edition of the MCC workshop on Mobile cloud computing, pp. 13-16. ACM, 2012.

[2] Rottondi, Cristina, A. Barbato, L. Chen, and G. Verticale. ”Enabling privacy in a distributed game-theoretical scheduling system for domestic appliances.” IEEE Transactions on Smart Grid 8, no. 3 (2017): 1220-1230.

[3] Alam, M. Raisul, M. B. Ibne Reaz, and M. A. Mohd Ali. ”A review of smart homesPast, present, and future.” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42, no. 6 (2012): 1190-1203.

[4] Yang, Xin-She. ”A new metaheuristic bat-inspired algorithm.” In Nature inspired cooperative strategies for optimization (NICSO 2010), pp. 65-74. Springer, Berlin, Heidelberg, 2010.

[5] Mirjalili, Seyedali, S. M. Mirjalili, and X. S. Yang. ”Binary bat algorithm.” Neural Computing and Applications 25, no. 3-4 (2014): 663-681.

[6] Yang, Xin-She, and X. He. ”Bat algorithm: literature review and applications.” International Journal of Bio-Inspired Computation 5, no. 3 (2013): 141-149.

[7] Arshad, Hafsa, S. Batool, Z. Amjad, M. Ali, S. Aimal, and N. Javaid. ”Pigeon Inspired Optimization and Enhanced Differential Evolution Using Time of Use Tariff in Smart Grid.” In International Conference on Intelligent Networking and Collaborative Systems, pp. 563-575. Springer, Cham, 2017.

[8] Amjad, Zunaira, S. Batool, H. Arshad, K. Parvez, M. Farooqi, and N. Javaid. ”Pigeon Inspired Optimization and Enhanced Differential Evolution in Smart Grid Using Critical Peak Pricing.” In International Conference on Intelligent Networking and Collaborative Systems, pp. 505-514. Springer, Cham, 2017.

[9] Duan, Haibin, and P. Qiao. ”Pigeon-inspired optimization: a new swarm intelligence optimizer for air robot path planning.” International Journal of Intelligent Computing and Cybernetics 7, no. 1 (2014): 24-37.

[10] Pham, X. Qui, N. D. Man, N. D. T. Tri, N. Q. Thai, and E. N. Huh. ”A cost-and performance-effective approach for task scheduling based on collaboration between cloud and fog computing.” International Journal of Distributed Sensor Networks 13, no. 11 (2017): 1550147717742073.

[11] Yang, Lei, J. Cao, G. Liang, and X. Han. ”Cost aware service placement and load dispatching in mobile cloud systems.” IEEE Transactions on Computers 65, no. 5 (2016): 1440-1452.

[12] Sidhu, H. Singh. ”Cost-deadline based task scheduling in cloud computing.” In Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference on, pp. 273-279. IEEE, 2015.

[13] Kang, D. Ki, S. H. Kim, C. H. Youn, and M. Chen. ”Cost adaptive workflow scheduling in cloud computing.” In Proceedings of the 8th International conference on ubiquitous information management and communication, p. 65. ACM, 2014.

[14] Aadkane, Trapti, and S. Monga. ”An Energy Efficient Cost Aware Virtual Machine Migration Approach for the Cloud Environment.”

[15] Aslanpour, M. Sadegh, M. G. Arani, and A. N. Toosi. ”Auto-scaling web applications in clouds: a cost-aware approach.” Journal of Network and Computer Applications 95 (2017): 26-41.

[16] Aazam, Mohammad, and E. N. Huh. ”Broker as a service (baas) pricing and resource estimation model.” In Cloud Computing Technology and Science (CloudCom), 2014 IEEE 6th International Conference on, pp. 463-468. IEEE, 2014.

[17] Kim, S. Hwan, D. K. Kang, W. J. Kim, M. Chen, and C. H. Youn. ”A science gateway cloud with cost-adaptive VM management for computational science and applications.” IEEE Systems Journal 11, no. 1 (2017): 173-185.

[18] Kim, Yeongjin, J. Kwak, and S. Chong. ”Dual-Side Optimization for Cost-Delay Tradeoff in Mobile Edge Computing.” IEEE Transactions on Vehicular Technology 67, no. 2 (2018): 1765-1781.

[19] Kim, Yeongjin, J. Kwak, and S. Chong. ”Dual-side dynamic controls for cost minimization in mobile cloud computing systems.” In Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2015 13th International Symposium on, pp. 443-450. IEEE, 2015.

[20] Cheng, Mingxi, J. Li, and S. Nazarian. ”DRL-cloud: Deep reinforcement learning-based resource provisioning and task scheduling for cloud service providers.” In Design Automation Conference (ASP-DAC), 2018 23rd Asia and South Pacific, pp. 129-134. IEEE, 2018.

[21] Kansal, Sahil, H. Kumar, S. Kaushal, and A. K. Sangaiah. ”Genetic algorithm-based cost minimization pricing model for on-demand IaaS cloud service.” The Journal of Supercomputing (2018): 1-26.

[22] Shahapure, Nagamani H., and P. Jayarekha. ”Load balancing with optimal cost scheduling algorithm.” In Computation of Power, Energy, Information and Communication (ICCPEIC), 2014 International Conference on, pp. 24-31. IEEE, 2014.

[23] Li, Ji, Y. Wang, T. Cui, S. Nazarian, and M. Pedram. ”Negotiation-based task scheduling to minimize users electricity bills under dynamic energy prices.” In Green Communications (OnlineGreencomm), 2014 IEEE Online Conference on, pp. 1-6. IEEE, 2014.

[24] Li, Ji, Y. Wang, X. Lin, S. Nazarian, and M. Pedram. ”Negotiation-based task scheduling and storage control algorithm to minimize user’s electric bills under dynamic prices.” In Design Automation Conference (ASP-DAC), 2015 20th Asia and South Pacific, pp. 261-266. IEEE, 2015.

[25] Patel, Suhradam, R. K. Bhujade, A. Sinhal, and S. Kathrotia. ”Resource optimization and cost reduction by dynamic virtual machine provisioning in cloud.” In Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on, pp. 857-861. IEEE, 2013.

[26] Xia, Qiufen, W. Liang, and W. Xu. ”Throughput maximization for online request admissions in mobile cloudlets.” In Local Computer Networks (LCN), 2013 IEEE 38th Conference on, pp. 589-596. IEEE, 2013.

[27] Sturm, Tobias, F. Jrad, and A. Streit. ”Storage CloudSim.” In Proceedings of the 4th International Conference on Cloud Computing and Services Science, pp. 186-192. SCITEPRESS-Science and Technology Publications, Lda, 2014.

[28] Youn, C. Hyun, M. Chen, and P. Dazzi. ”VM Placement via Resource Brokers in a Cloud Datacenter.” In Cloud Broker and Cloudlet for Workflow Scheduling, pp. 47-73. Springer, Singapore, 2017.

[29] Cao, Zijian, J. Lin, C. Wan, Y. Song, Y. Zhang, and X. Wang. ”Optimal cloud computing resource allocation for demand side management in smart grid.” IEEE Transactions on Smart Grid 8, no. 4 (2017): 1943-1955.

[30] Singh, Sukhpal, and I. Chana. ”Q-aware: Quality of service-based cloud resource provisioning.” Computers and Electrical Engineering 47 (2015): 138-160.

[31] Zhao, Jian, H. Li, C. Wu, Z. Li, Z. Zhang, and F. CM Lau. ”Dynamic pricing and profit maximization for the cloud with geo-distributed data centers.” In INFOCOM, 2014 Proceedings IEEE, pp. 118-126. IEEE, 2014.

[32] Sheeja, Y. S., and S. Jayalekshmi. ”Cost effective load balancing based on honey bee behaviour in cloud environment.” In Computational Systems and Communications (ICCSC), 2014 First International Conference on, pp. 214-219. IEEE, 2014.

[33] Ren, Shaolei, and M. V. D. Schaar. ”Dynamic scheduling and pricing in wireless cloud computing.” IEEE Transactions on Mobile Computing 13, no. 10 (2014): 2283-2292.

[34] Yaghmaee, M. Hossein, M. Moghaddassian, and A. L. Garcia. ”Power Consumption Scheduling for Future Connected Smart Homes Using Bi-Level Cost-Wise Optimization Approach.” In Smart City 360, pp. 326-338. Springer, Cham, 2016.

[35] Chopra, Nitish, and S. Singh. ”Deadline and cost based workflow scheduling in hybrid cloud.” In Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on, pp. 840-846. IEEE, 2013.

[36] Pandey, Manish, and S. K. Verma. ”Cost based resource allocation strategy for the cloud computing environment.” In Computing, Communication and Networking Technologies (ICCCNT), 2017 8th International Conference on, pp. 1-7. IEEE, 2017.

[37] Yang, Binxu, W. K. Chai, Z. Xu, K. V. Katsaros, and G. Pavlou. ”Cost-Efficient NFV-Enabled Mobile Edge-Cloud for Low Latency Mobile Applications.” IEEE Transactions on Network and Service Management (2018).

[38] Ibrahim, Rabha W., and A. Gani. ”A mathematical model of cloud computing in the economic fractional dynamic system.” Iranian Journal of Science and Technology, Transactions A: Science 42, no. 1 (2018): 65-72.

[39] Mehmi, Sandeep, H. K. Verma, and A. L. Sangal. ”Simulation modeling of cloud computing for smart grid using CloudSim.” Journal of Electrical Systems and Information Technology 4, no. 1 (2017): 159-172.

[40] Khan, Nasir, N. Javaid, M. Khan, A. Subhani, A. Mateen, and A. Iqbal. ”Harmony Pigeon Inspired Optimization for Appliance Scheduling in Smart Grid.”

Detection of Sentiment Polarity of Unstructured Multi-Language Text from Social Media

Saad Ahmed, Saman Hina, Raheela Asif
Department of Computer Science
NED University of
Engineering and Technology
Karachi, Pakistan

***Abstract*—In recent years, Twitter has caught the attention of many researchers because of the fact that it is growing very rapidly in terms of number of users and also all the data present as tweets on twitter is public in nature while other social media networks such as Facebook, data is not completely public as users can restrict their post to only users present in their friend list. In this research study, aspect based sentiment analysis (ABSA) was done on the data acquired from social media related to the major cellular network companies of Pakistan (Telenor Pakistan, Mobilink Jazz, Zong, Warid and Ufone). For this research, we have specifically selected all tweets which are not only in English and Roman Urdu but also mixture of above two languages. We have employed natural language processing (NLP) techniques for pre-processing the dataset and machine learning (ML) techniques to detect the sentiments present in the data. The results are interesting and informative specially for policy makers of cellular companies. These companies can utilize this information to increase the performance of their services. In comparison with the state of the art algorithms, the performance of bagging algorithm with this framework on the acquired dataset has produced F Score of 92.25, which is very encouraging outcome of this research work.**

***Keywords*—Social media; sentiment analysis; data mining; cellular networks**

I. Introduction

As the advancement in science and technology continues, the research plays a vital role in every science and technology related field. This work of research is done on the social media data associated with telecommunication domain. Twitter, a micro blogging website is one of the main stream social media website, which has seen tremendous growth in last few years. In a developing country like Pakistan, common people have now gained access to the Internet and are learning the advantages of social media as a source of information as well as using the same to express their views and ideas about politics, products and services. This makes social media a main source of user generated information which makes it a valuable source of data to perform opinion mining and sentiment analysis of general public.

In the last few years, researchers are working on social media data to extract information and then analyze it using different techniques. Some methods of sentiment analysis have been developed in areas of different domains but still a lot of research needs to be done.

The social media has become a vital part of everyday life where its users can express their ideas, views or comments about any product or service [1]. These views and comments about products and service are very important for companies which are the provider of those products and services. This information from social media can help these companies to refine their strategies for the improvement of their products and services.

Twitter, a micro-blogging real time social media network data is extracted from its website in this research. Twitter generates huge amount of data, this data is extremely valuable for data mining and analyzing sentiments of public. The simplicity of posting tweets in Twitter makes it a suitable data source for real-time sentiment analysis.

Twitter has about 300M+ active users who post about 500M tweets in a single day. This huge data which is generated by users is public and is easily available through APIs (Application Program Interface) to anyone who wants to use this data for analysis. That is why twitter is very popular among research scientists for research purposes. There are several features of twitter such as tweets are maximum of 140 characters, mentions (@) and hashtags (#) which are used by users to refers to any particular event or a company in their tweet. This can be used to collect tweets related to a particular event or company. Tweets have short length, use of local languages and local terms makes it more challenging to analyze and find out the sentiments and possible aspects present in it.

The Twitter is an important source of data acquisition, but it is very complex analyzing its content as large number of the tweets either use slang language or shorten words. Sentence level and word level polarity classification [2] was done using a method based on lexicons, namely, SentiCircles, which builds a dynamic depiction of words in order to determine their suitable semantics. Here, semantics refers to the co-occurrence patterns from each word in the dataset. A different method is feature engineering [3] which produces a result of seven dimensions. This feature engineering method was used to analyze aspects: frequency, affinity, valence, shifter, feature sentiment scoring and categorization. Different type of representations can be utilize, based on dictionaries and lexical aspects of sentences [4], word embedding [5], word and character n-gram [6] among others.

The extraction and classification of user opinion on the diverse topics is known as sentiment analysis which is also referred as opinion mining. Mostly, two forms of methods are used for sentiment Analysis, which are either based on machine learning or based on vocabulary. The machine learning method has two approaches, supervised learning and unsupervised learning. Supervised learning involves data which is labeled to train algorithms [7], whereas unsupervised learning does not need data to be labeled [8]. The mixture of labeled and unlabeled data makes semi-supervised learning [9].

We proposed a hybrid sentiment analysis framework. This method comprises of a customized dictionary of Roman Urdu words which are commonly used by social media users in Pakistan to express their views and share their comments on twitter. This has helped us extract more information from tweets and use this additional information [10] to detect the sentiments of the people. In addition to this, the proposed framework also includes the use of SENTIWORD dictionary which provides weights for each English word which appears in the tweet. By using these weights we were able to calculate more realistic sentiment polarity which improves overall performance of this framework.

Since the beginning of 21st century, Sentiment Analysis (SA) has become one of the main area of research in natural language processing (NLP) [11]. It is a complex problem with many exciting sub-problems, which includes sentence-level sentiment classification, which is the case in tweets. Research scientists have documented that different type of sentence require different handling for SA. Sentence can be of different types, which includes subjective sentences, comparative sentences, negation sentences, conditional sentences, target-dependent sentences and sarcastic sentences. The tweets extracted from twitter could carry all these types of sentences which present a more complex problem during analysis.

Aspects [12] are basically features; these are selected by using Information gain method. The sentiment of the feature is calculated by using the neighboring words of the aspect. These are acquired through N-gram methods. To calculate the effectiveness of this hybrid method, we obtained a corpus from Twitter, we collected data for the duration of six months from 15 Dec 2016 to 14 June 2017 and 2703 tweets were extracted which were then manually classify as positive, negative or neutral.

Our experimental results confirms that the good result was obtained through the N-gram around method [13] along with the use of customize Roman Urdu dictionary. In addition to this, documents such as customer reviews may contain fine-grained emotion for different features (e.g. a product or service) that are mentioned in the document. This information can be very valuable for understanding customers' opinion about a certain service or product on twitter.

Twitter has seen rapid growth in the last few years, where its registered users can post tweets related to events in real time [14]. Users of social media tend to tweet using highly unstructured language with many typographical errors and use local languages as well as slang words in tweets. A significant amount of tools and setup is required to work on social media data due to its speedy growth and to the difficulty of processing its data by using standard relational SQL databases [15].

Today, social media users share their views and opinions on internet, increasing the volume of information each day. Social networks like Twitter and Facebook sites are most popular. Facebook [16] reaches its 1 billion users in October 2012, while twitter had more than 500 million users on February 2012 [17], [18] and currently it has more than 690 million registered users and on average twitter recorded 9100 tweets/second.

According to www.Twitter.com, the most discussed topics [19] in 2016 are as follows:

1) Rio2016.
2) Election2016.
3) PokemonGo.
4) Euro2016.
5) Oscars.
6) Brexit.
7) BlackLiveMatter.
8) Trump.
9) RIP.
10) GameofThrones.

On 24 October 2015, more than 41m tweets related to "#ALDubEBTamangPanahon" of AlDub [20] were sent during special concert of the Kalyeserye segment of the show Eat Bulaga entitled Eat Bulaga: Sa TamangPanahon held at the Philippine Arena, the world's largest indoor arena in Bulacan, Philippines. This performance was attended by more than 55000 people. This was the most discussed topic on twitter.

The top most sports game ever discussed on twitter with over 35.6 million tweets was the 2014 FIFA World Cup semi final held on July 8, 2014 between Brazil and Germany.

Twitter is a very popular social media site for data mining research where there is significant amount of data available containing all type of information. The information regarding followers, followed, tweets, and posts can be used in Recommender system [21] and can also be used to mine valuable information like public mood, trends and sentiments.

ABSA emerges as excellent technique which enables us to find the best solution. Recently ABSA based on social network data is gaining importance in the field of data mining. The result of this proposed framework will reflect the mood of the general public.

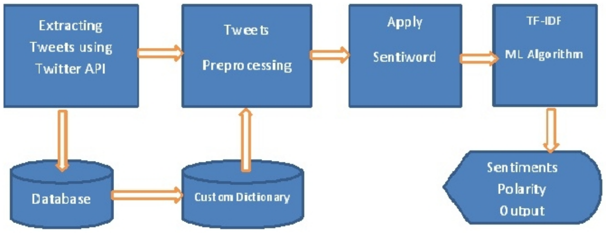
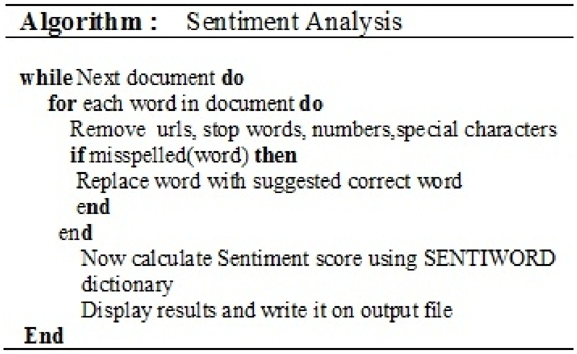


Fig. 1. Proposed system to detect sentiment polarity.

II. Methodology

The main focus of this research work was to develop a hybrid framework as shown in Fig. 1, which employs data mining and machine learning techniques using data from social media network and obtain results which are valuable for Pakistani cellular companies. Main steps involved in this framework are using data crawler to mine twitter website using twitter API. Unstructured data (tweets) is converted into structured data



```
Algorithm :   Sentiment Analysis

while Next document do
   for each word in document do
      Remove urls, stop words, numbers,special characters
      if misspelled(word) then
       Replace word with suggested correct word
       end
    end
       Now calculate Sentiment score using SENTIWORD
       dictionary
       Display results and write it on output file
 End
```

Fig. 2. Algorithm of sentiment analysis.

TABLE I. ANALYSIS OF TWEETS COLLECTED

NAME	Total	Negative	Positive	Neutral
Mobilink	204	34	120	50
Warid	487	35	149	303
Ufone	1000	84	764	152
Telenor	561	50	364	147
Zong	451	21	338	92

and is stored in a database. This data includes some irrelevant information which was cleaned by applying preprocessing steps on this data using Natural Language Processing (NLP) techniques and finally we use learning classifiers to find the sentiments of the tweets. We have used R programing language to perform statistical calculation on the dataset. This R platform has provided us all the tools, required for this research work. Algorithm of proposed sentiment analysis framework is shown in Fig. 2.

The algorithm uses SENTIWORDNET [23] for assigning weighted scores to determine the polarity of analyzed tweet.We then apply Machine learning Algorithm and use Term Frequency Inverse Document Frequency (TF-IDF) technique [18] to obtain the aspects and there weights. To achieve constant processing time the Twitter data corpus is divided into parts of equal size in the testing process. The block diagram of the framework is shown in Fig. 1. The process is discussed in depth in the following sections:

A. Data Preprocessing

The dataset which is downloaded from social website Twitter has 2703 tweets which is from 15 Dec 2016 to 14 June 2017 by using the API (application Program Interface) and the Data Crawler. Then on this data, preprocessing was done to clean data by removing the garbage like website address and links to images, which are of no use in sentiment analysis research project.

Total tweets collected were 2703 in dataset out of which 1000 tweets belongs to Ufone, 451 tweets belong to Zong, 561 belong to Telenor, 204 belongs to Mobilink and 487 tweets belongs to Warid as shown in Table I. These tweets were extracted from Timelines of official twitter account of these companies. This is depicted in Fig. 3.

The collected corpus has very detailed information about each tweet, it has 16 fields in it. We retain only the useful and essential fields of data and store it in the .csv files, the fields includes User ID, User Screen Name, Reference, Tweet ID, Date and Time and importantly Text field is stored in database and remaining data fields were filtered out. The Dataset is now structured and is in organized form to be tested.

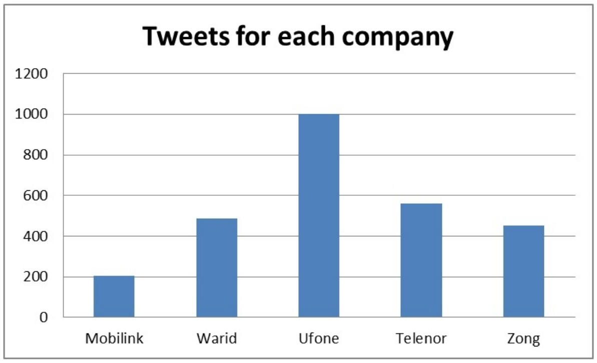


Fig. 3. No. of tweets extracted for each cellular company.

B. Text Preparation

Tweets (documents) are now parsed into a data corpus for text analysis. Text field of the tweet is considered. The text present in text field is prepared by cleaning for further analysis.

During text preparation, the numbers, URLs and links to images, videos and websites are removed from tweets as they do not serve any purpose. Stopwords such as “but”, “shall”, “by”, etc. are words which have no analytical importance but are commonly used, so these stopwords are removed from the text. After this Stemming process is done to reduce inflected words to their root form which makes system analyze words better, for this purpose suffix dropping algorithms are used in this step. Punctuation marks and whitespaces are removed as they also serve no purpose in sentiment analysis. Lemmatisation algorithms are applied finally to complete the data cleaning process.

In this preprocessing step of the tweets, a large level of noise is removed by using tokenizing which is a process of splitting text into a set of individual terms or tokens. Each tweet is tokenized into a sequence of terms. In NLP, the most commonly used words in a document are referred to as ‘stopwords’. All the tweets are checked against a standard stopword list to remove terms which carry little information. The token starting with ‘@’ (i.e. a reply or mention) will also be removed from the tweets in the filtration process. At the end of this process, each tweet is divided into a set of aspects which are in the vector space model.

C. TF-IDF Technique

In this research work we have utilized TF-IDF technique [14]. We have applied this technique to filter out tweets which have minimal or no information which helps in our analysis during this research.

This technique makes a sparse matrix (a matrix in which most elements are zero), this indicates that how many parameters are un-informative in the dataset. So we reduce sparsity by removing the terms that occur very regularly. This has shown to have the effect of reducing over fitting and improving the analytical capability of our system.

In this research we choose to set the sparsity to a maximum of 75% which has provide us improved results without bias to the context and perspective of sentiment of public. Tweets with no information to predict the sentiments were also filtered out to improve the performance of the ML algorithm.

Normally data mining frameworks use clustering methods, which groups similar items in a one subset. These subsets of items are called clusters. Different types of techniques such as Nearest Neighbor (NN) and K-mean [22] can perform clustering. The clusters which are created by these algorithms are used as polarities of the opinion or sentiments of people.

III. Experimental Results

The dataset is divided into two same size parts, i.e. training and testing data. The presented system was tested using different ML algorithms with TF-IDF, using the dataset acquired from the twitter.com. The other major feature of this work is that we have develop a list of words which are commonly used by users in their tweets which are not English words but are words written in Roman Urdu. With the help of this customize dictionary of Roman Urdu words, we were able to better understand the context of tweets written using words both from English and Roman Urdu in a single tweet and this results in better detection of sentiments at sentence level and increases the performance of this proposed system.

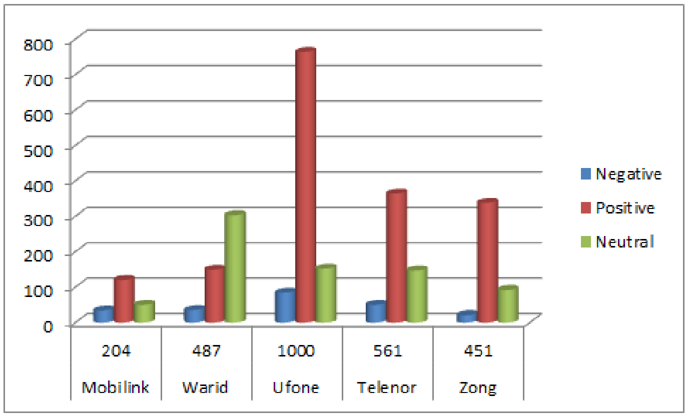


Fig. 4. Results of analysis showing polarity of tweets for each company.

The results obtained are illustrated in Fig. 4, which shows that the positive aspects in the tweets dataset are in abundance as compared to the negative or neutral with very few sarcastic tweets. This identifies that the telecommunication companies are giving better service to their customers in Pakistan. Ufone has the most number of positive sentiments towards their services which is in line with this common perception that Ufone is the most popular cellular network company in Pakistan.

Fig. 5 shows the performance of presented hybrid system to detect sentiment polarity. The services of cellular networks in Pakistan are becoming reliable and dependable, this statement was validated during this research work as the number of tweets predicted which carry positive sentiments are far more than the other sentiment combined.

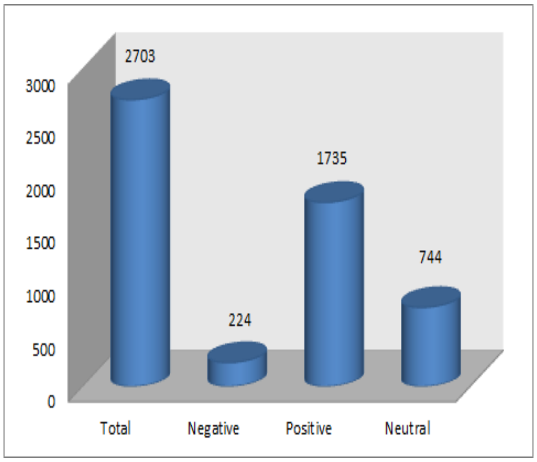


Fig. 5. Polarity of sentiments for all cellular companies.

TABLE II. Performance of Framework

ALGORITHM	PRECISION	RECALL	F SCORE
Boosting	0.9601	0.8925	0.9148
SVM	0.6811	0.5975	0.6311
Bagging	0.9651	0.8975	0.9225
Forest	0.4025	0.3799	0.3875
Tree	0.7001	0.6454	0.6601
Maxen	0.4755	0.6601	0.4925
Naive Bayes	0.8160	0.8150	0.8240

Predicted sentiments of all the tweets used in this research work are illustrated in Fig. 6. The closer look at these predicted polarity of tweets gives us the information that over the period of time during which tweets were collected there is an increase in positive sentiments towards cellular companies in general which is the indicator that services in telecommunication sector is improving very rapidly and customers are getting state of the art technology which is affordable.

The Precision, Recall and F Scores of algorithms which were used to compare the performance are shown in Table II.

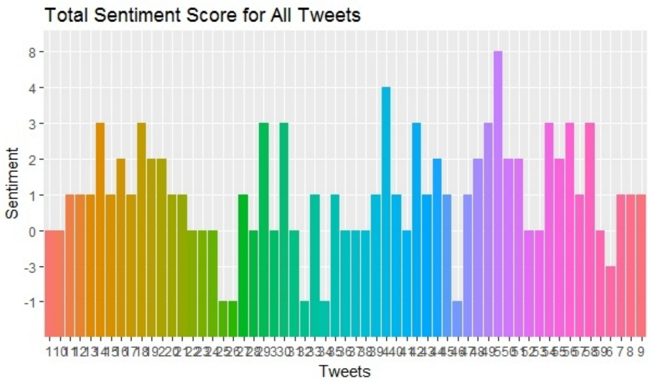


Fig. 6. Sentiments of tweets predicted.

The performance of bagging and boosting algorithm with this framework was superior as compared to other well-known algorithms, as depicted in Fig. 7. The Precision of Nave bayes is more than 81% while SVM and tree algorithms achieved above 60% precision and Forest algorithm achieved only 40%

precision on this dataset.

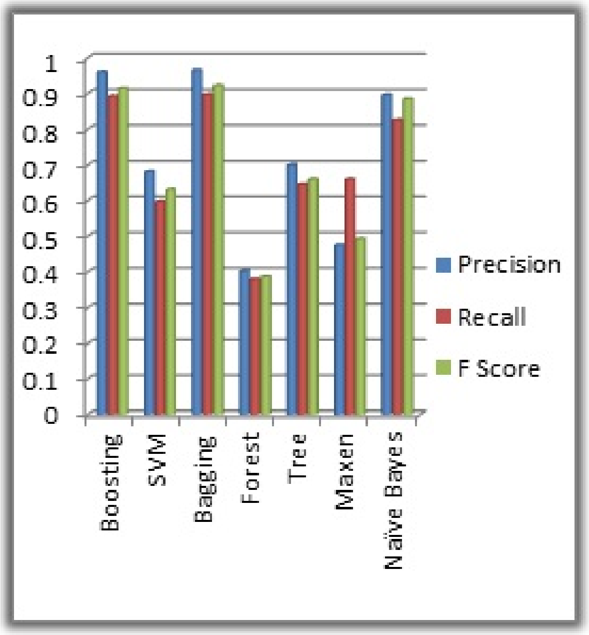


Fig. 7. Comparison of performance of proposed framework with different ML algorithms.

IV. Conclusion

Internet has open doors of information and social networks on internet have provided an important source of data regarding users and their sentiments towards a particular product, service or event. This is especially valuable in giving in depth knowledge about the current developments and attitudes of people who are using Internet. In this paper we have presented and applied a hybrid system of sentiment analysis to analyze tweets from twitter. This hybrid system has been weighed using twitter data related to cellular companies of Pakistan.It will be further evaluated by increasing data set to provide efficient/faster processing on big data.

We have collected real data from social media network Twitter which is related to cellular companies of Pakistan by using the Twitter streaming API. This API also provides detailed meta data for the tweets.

Data preprocessing has been done to improve the accuracy of this hybrid system. The addition of customize Roman Urdu dictionary has added a new dimension to this research work and produced motivating results.We were able to correctly detect the sentiment polarity of the tweets used in this research and these results were confirmed by human annotation of the same data. The framework which we have proposed uses TF-IDF technique with Bagging machine learning algorithm and is identifying the sentiments faster with improved F-scores, this proposed framework has also produced better results with boosting ML algorithm. We have also evaluated the accuracy of this system by comparing methods used by other researches in this area of research and found the performance of this framework is comparable with the other state of the art algorithms.

References

[1] Ravi, Kumar, and Vadlamani Ravi. “A survey on opinion mining and sentiment analysis: tasks, approaches and applications.” Knowledge-Based Systems 89 (2015): 14-46.

[2] Saif, Hassan, Yulan He, Miriam Fernandez, and Harith Alani. 2016. Contextual semantics for sentiment analysis of twitter. Information Processing and Management, 52(1):5 - 19

[3] Ghiassi, Manoochehr, David Zimbra, and Sean Lee. 2016. Targeted twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for articial neural networks. Journal of Management Information Systems, 33(4):1034-1058.

[4] Murillo, Edgar Casasola and Gabriela Marin Raventos. 2016. Evaluacion de modelos de representacion del texto con vectores de dimension reducida para analisis de sentimiento. In TASS@ SEPLN, pages 23 - 28.

[5] Quiros, Antonio, Isabel Segura-Bedmar, and Paloma Martnez. 2016. Labda at the 2016 tass challenge task: Using word embeddings for the sentiment analysis task. In TASS@ SEPLN, pages 29-33.

[6] Ceron-Guzman, Jhon Adrian and Santiago de Cali. 2016. Jacerong at tass 2016: An ensemble classifier for sentiment analysis of Spanish tweets at global level. In TASS@ SEPLN, pages 35-39

[7] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP, pages 7986, 2002.

[8] Peter D Turney. Thumbs up or thumbs down? Semantic Orientation applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), (July):417424, 2002.

[9] Andrew B Xiaojin. Introduction to Semi-Supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, pages 1130, 2009.

[10] Zhang, Wen, Taketoshi Yoshida, and Xijin Tang. “A comparative study of TF* IDF, LSI and multi-words for text classification.” Expert Systems with Applications 38.3 (2011): 2758-2765.

[11] Liu, Bing. Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press, 2015.

[12] Salas-Zrate, Mara del Pilar, et al. “Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach.” Computational and mathematical methods in medicine 2017 (2017).

[13] Cavnar, William B., and John M. Trenkle. “N-gram-based text categorization.” Ann Arbor MI 48113.2 (1994): 161-175.

[14] J. Lin and D. Ryaboy, “Scaling big data mining infrastructure: the twitter experience,” SIGKDD Explor.Newsl., vol. 14, pp. 6-19, apr, 2013.

[15] G. Mishne, J. Dalton, Z. Li, A. Sharma and J. Lin, “Fast data in the era of big data: Twitter’s real-time related query suggestion architecture,” in Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, New York, USA, 2013.

[16] Ashlee Vance (October 04, 2012) Facebook: The Making of 1 Billion Users. [Online]. Available: http://www.businessweek.com/articles/2012-10-04/facebook-the-making-of-1-billion-users

[17] Lauren Dugan (February 21, 2012) News, Statistics: Twitter to Surpass 500 Million Registered Users on Wednesday. (Online). Available : http://www.mediabistro.com/alltwitter/500-million-registered-usersb18842

[18] Twitter Inc. (2011) Year in Review: Tweets per second. [Online]. Available: http://yearinreview.twitter.com/en/tps.html

[19] https://blog.twitter.com/official/enus/ a/2016/thishappened-in-2016.html

[20] “Fans in the Philippines and around the world sent 41M Tweets mentioning #ALDubEBTamangPanahon”. Twitter Data Verified Account. October 27, 2015. Retrieved October 30, 2016.

[21] J. Bobadilla, F. Ortega, A. Hernando, A. Gutierrez, Recommender systems survey Universidad Politcnica de Madrid, Ctra. De Valencia, Km. 7, 28031 Madrid, Spain

[22] Hartigan, John A., and Manchek A. Wong. “Algorithm AS 136: A k-means clustering algorithm.” Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1 (1979): 100-108.

[23] sentiwordnet.isti.cnr.it/

An Algorithm that Prevents SPAM Attacks using Blockchain

Koichi Nakayama, Yutaka Moriyama
Faculty of Science and Engineering,
Saga University
Saga, Japan

Chika Oshima
Faculty of Medicine,
Saga University
Saga, Japan

***Abstract*—There are many systems and methods for preventing spam attacks. However, at present there is no specific tried-and-true method for preventing such attacks. In this paper, we propose an algorithm, "SAGA$_{BC}$" to prevent spam attacks using a blockchain technique and demonstrate its effectiveness by a simulation experiment. A person who sends an email using the "SAGA$_{BC}$" must pay the processing cost with cryptocurrency. If an e-mail sent using this algorithm is received normally at a destination e-mail account, this fee is refunded. However, a lot of spam e-mails are not received normally, because addresses of the spam e-mails are indiscriminate. If a spammer sends spam using the "SAGA$_{BC}$," he/she will lose the cryptocurrency fee for each such message. Thus, if using the "SAGA$_{BC}$" to send e-mail becomes a standard practice for the general public, receiving e-mail servers and/or mailers will be able to easily judge incoming messages without using the "SAGA$_{BC}$," because spammers cannot use the "SAGA$_{BC}$" without losing their cryptocurrency.**

***Keywords*—Cryptocurrency; wallet account; Mail Send Coin (MSC)**

I. Introduction

Unwanted electronic mail (e-mail), known as "spam" appear in many people's inboxes every day. People who send spam e-mail (called "spammer" in this paper) aim to spread advertisements and computer viruses and play tricks on their targets. Many spams impose a high load on a network and may affect the processing of legitimate e-mails.

Some technical methods to protect e-mail users from spam do exist. When a receiving server receives an e-mail, it authenticates the validity of the message using information from the sending DNS server. The technical methods of authenticating sender domains include "Sender Policy Framework (SPF) [1]", "Sender ID", "DomainKeys Identified Mail (DKIM [3])", "Domain Name System Blacklist (DNSBL [2])", among others. Whereas the SPF and the SenderID use IP addresses to authenticate a sender domain, the DKIM uses an electronic signature.

The receiving e-mail server refuses any e-mail from an IP address that does not have an SPF record. The SPF record is a text record verifying that a domain's administrator made and is registered to DNS. The SPF record includes the IP address of the server permitted to send e-mail using the domain name as an e-mail source. The receiving server checks the SPF record against the IP address of the sending server which forwarded the e-mail; if the IP address of the sending server does not match that of the SPF record, the receiving e-mail server considers the e-mail to be spam. Although the SPF refers to the sender's e-mail address as designated in the "From": field, the Sender ID refers to that of the header. The DKIM is a way of sending an electronic signature to the receiving e-mail server, which then acquires a public key from the sending DNS server and verifies the DKIM signature. The DNSBL is a software mechanism to stop spam. Lists consisting of the IP addresses of servers sending spam are then supplied to receiving e-mail servers.

These problems may be negligible on a daily basis, but over timeweeks or even dayshundreds and thousands of spam messages make it through flawed filters. We outline these problems below.

1) **Cannot verify the e-mail account unit.**
When a sending server is determined to be a spam server, all the e-mail accounts that use that server are prevented from sending and receiving e-mail. On the other hand, if a spammer switches to a different sending server, the attack can no longer be prevented.
2) **Cannot prevent a first spam attack.**
Most spam attacks go through one or more sending servers which are not the true server, and the number of these "zombie computers" can increase exponentially. This method also ensures that traditional protections cannot prevent a first spam attack.
3) **Sometimes a normal e-mail is erroneously identified as spam.**
Most filtering functions provided by Internet service providers and e-mail software identify spam according to contents of the message, and thus may erroneously mark normal messages as spam. Moreover, if the contents of an e-mail message contain graphics, the filtering function may not work because it scans the graphic but cannot determine what the graphic portrays.
4) **Infringing e-mail users' privacy.**
Because Internet service providers and e-mail software judges whether an e-mail is spam according to its contents, individual information may be disclosed.
5) **Receiving servers and e-mail software carry a heavy workload.**
Current spam filters screen all incoming e-mail to identify spam, and receiving servers and e-mail software need to renew their filtering function constantly to catch new types of spam. The workload for identifying spam is heavier than that of launching a spam attack. Receiving servers and e-mail software incur

considerable financial and processing costs in dealing with spam.

We will resolve these problems using a block-chain technology. In the next section, we offer precise definitions of the relevant terms in this paper before explaining "SAGA$_{BC}$". Then, in Section IV, we present the results of a simulation experiment using SAGA$_{BC}$. We also discuss the method's ability to prevent spam, based on the experiment's results. The paper's conclusion follows:

II. SAGA$_{BC}$

A. Concept

A genuine e-mail from a harmless person can be received normally. However, some e-mail addresses used in spam are fictitious or are rejected by receiving servers. Therefore, only a fraction of the spam sent out by a spammer reach their targets. In our proposed method, anyone who sends an e-mail message must pay a processing fee in cryptocurrency, but if the e-mail is received correctly, that fee will be refunded to the sender. Those sending genuine, harmless e-mail will pay a little, but spammers must spend much more to launch a spam attack. We expected that this will reduce spam attacks. We call this method the "SPAM Attack Guard Algorithm Using Block Chain (SAGA$_{BC}$)".

B. Definition of Terms

Because cryptocurrency is a relatively new concept, the definitions of relevant terms offered by various publications have been vague and sometimes contradictory. Therefore, we will offer precise definitions of the relevant terms in this paper.

Blockchain

Blockchain is a kind of a distributed database (Distributed Ledger Technology, or DLT). Data is accumulated per a unit "block". Each block records the Hash values of the unit immediately preceding it. Therefore, it is necessary to calculate Hash values for all the data leading to a falsified block to falsify the data on the way. In other words, it is very difficult to falsify the data for a blockchain. "Bitcoin [6]" and "Ethereum [7]" are well-known kinds of blockchain cryptocurrencies.

Cryptocurrency

"Electronic money is commonly defined as value stored electronically, issued on receipt of funds of an amount not less in value than the monetary value issued, and accepted as a means of payment by parties other than the issuer [8]". "Digital currency is a type of currency available only in digital form, not in physical. Examples include virtual currencies and crypto currencies or even central bank issued [9]". "Virtual currency is a digital payment mechanism for (and denominated in) fiat currency [10]". "Digital and virtual currencies can either be centralized or decentralized [4]". "Cryptocurrency refers to any electronic money created using a cryptographic technology [4]". "Cryptocurrency is a purely decentralized peer-to-peer electronic cash system for validating value transfers [5]". "BTC" and "ETH" are types of cryptocurrency comprised of blockchains, such as "Bitcoin [6]", and "Ethereum [7]".

Wallet

The "wallet" is a means of storing cryptocurrency. Anyone can freely create a wallet, and "users can send and receive bitcoins electronically using wallet software on a personal computer, mobile device, or web application [11]". We consider the wallet to be a mechanism for managing cryptocurrency.

Wallet account

A "wallet account" is an ID used to identify an individual wallet. Wallet users manage their cryptocurrencies using unique wallet accounts.

Transaction

A "transaction" is a record of sending cryptocurrency from one's own wallet account to another wallet account. The digital signature of the owner of the cryptocurrency as well as his/her private key are needed to issue the transaction.

Mining

"When the sending user transfers cryptocurrency to a recipient, the transaction is verified by a process called 'mining [13]'." There are public keys used to verify information and permit the execution of transactions requested by others (called "miners"). Once a transaction is verified and approved by a miner, it is executed and stored in a digital block [12]. The entire transaction, from issuance to verification, only takes a few minutes.

C. System Set-up

The SAGA$_{BC}$ cooperates with an e-mail account associated with a wallet account to prevent spam attacks. Generally, an e-mail client has one or more e-mail accounts. One or more wallet accounts are assigned to each e-mail account by the SAGA$_{BC}$. The e-mail client cooperates with one e-mail account associated with a wallet account.

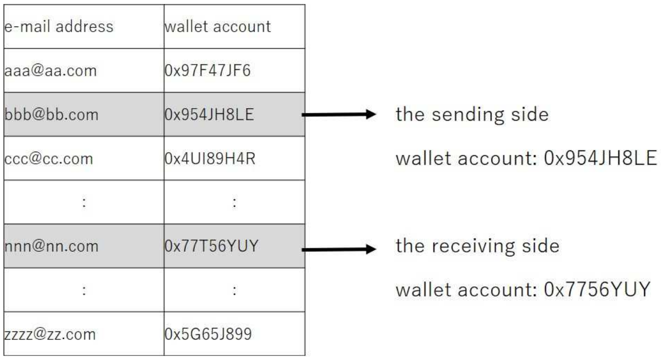
The SAGA$_{BC}$ system comprises the following components:

1). **Cryptocurrency: Mail Send Coin**
The Mail Send Coin (MSC) is one of the cryptocurrencies implemented by the SAGA$_{BC}$. The MSC is not a monetary token but a kind of utility token. Anyone using the SAGA$_{BC}$ can also use existing cryptocurrencies, e.g., Ethereum. However, in this paper, we will explain the SAGA$_{BC}$ with specifically in terms of using MSC.

2). **Mailers**
In the SAGA$_{BC}$, an expanded function (add-on) of the general mailer is implemented.

(2-1) The account management function

As shown in Fig. 1, the account management function extracts those wallet accounts that correspond not only to the owner's e-mail account but also to a destination e-mail account. This function then inquires of the blockchain whether the



e-mail address	wallet account
aaa@aa.com	0x97F47JF6
bbb@bb.com	0x954JH8LE
ccc@cc.com	0x4UI89H4R
:	:
nnn@nn.com	0x77T56YUY
:	:
zzzz@zz.com	0x5G65J899

Fig. 1. Extracting wallet accounts.

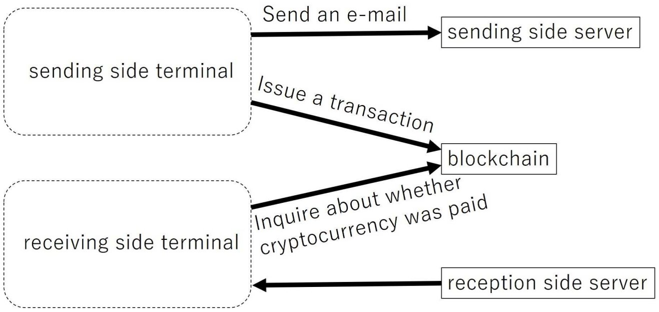


Fig. 2. Validating wallet accounts.

sending wallet account has paid the MSC into the receiving wallet account.

(2-2) Inquiring whether MSC was paid
This function inquires a blockchain about whether the MSC was paid from the sending side wallet. Any data gathered from such reference results are then stored in this function.

(2-3) The sorting function
This function assesses whether e-mails are spam according to the amount of MSC paid to send them and sorts them into a spam e-mail folder.

(2-4) The remittance function
The remittance function is an MSC payment function operating from the wallet account corresponding to the owner's e-mail account that contacts the wallet account corresponding to the destination (receiving) e-mail account.

(2-5) The validation function
The first time an e-mail is sent to a new recipient, this function validates the wallet account corresponding to the destination e-mail account. The sending mailer then checks the associated wallet account and determines whether it has already remitted the MSC fee to the receiving mailer. The receiving mailer connects its associated wallet account with the sending mailer depending on the identity of the sending wallet account and whether the appropriate amount of MSC has been paid. The sending mailer then sends the MSC fee to the receiving wallet account.

(2-6) The mining function
If the MSC paid is insufficient, a user can supplement it by mining MSC transactions that other users have issued. A spammer can also supplement his or her MSC in the same way, but it costs much more for an illegitimate user to do so.

D. Procedure to be Followed when Both the Sending and Receiving Mailers use the SAGA$_{BC}$

1) **Sending mailers**
As shown in Fig. 2, when a mailer sends an e-mail, it issues a certain number of transactions sending MSCs to the wallet account corresponding to the destination e-mail account.

2) **Receiving mailers**
The receiving mailer determines whether a received e-mail message is spam based on the amount of MSC attached and sorts the spam into the spam folder. The receiving mailer then automatically decides whether to refund the MSC fee paid according to how the e-mail message is processed. If the message is deleted or sorted into the spam folder, the MSCs paid for it is not refunded. However, if the message is not processed within a certain period of time, the amount of MSC paid can be refunded to the sending wallet account.

3) **Mining**
Issued transactions are recorded at the head of the blockchain by a miner. All dealings related to the transaction are then concluded.

E. Procedure to be Followed when Either the Sending or the Receiving Mailer does not use the SAGA$_{BC}$

1) **When only the sending mailer uses the SAGA$_{BC}$**
Sending mailers can determine whether the receiving mailer uses the SAGA$_{BC}$ with the validation function (see (2-5)). In this case, the sending mailer can send a regular e-mail without paying a transaction fee in MSC.

2) **When only the receiving mailer uses the SAGA$_{BC}$**
Receiving mailers can determine whether the sending mailer uses MSCs by the function for inquiring whether MSC was paid (see (2-2)). If the sending mailer does not use the SAGA$_{BC}$, the receiving mailer will know this because of the account management function (see (2-1)). In this case, the receiving mailer deals with incoming messages as normal e-mail that cannot be confirmed as not being spam.

3) **When neither the sending nor the receiving mailer uses the SAGA$_{BC}$**
In this case, e-mail is sent and received using a traditional method.

F. Anticipated Effects of a Spam Attack

When using the SAGA$_{BC}$, the e-mail sender must simultaneously send an MSC fee to the receiving wallet when sending an e-mail message. Because spammers send vast amounts of e-mail, they will lose MSCs doing this, which will eventually discourage them from sending e-mail. When the normal e-mails are received correctly, those sending such messages do not lose their MSCs: Even if the MSCs they sent disappears, they can restock by mining. We therefore expect that SAGA$_{BC}$ users will cease to receive spam.

III. Simulation Experiment

In this section, we verify whether the SAGA$_{BC}$ can prevent spam attacks using a simulation. The experimental simulation will not include e-mail senders who do not use the SAGA$_{BC}$.

A. Experiment Model

Fig. 3 shows the main routine of the simulation experiment.

1) **Initial setting**
The number of SAGA$_{BC}$ users is indicated by “N”. The initial value of the MSC that all users possess is indicated by “M”. Of N, the number of spammers and the number of genuine users are indicated by “S” and “(N-S)”, respectively.
2) **Sending e-mails and MSCs**
A genuine SAGA$_{BC}$ user sends an e-mail and 1 MSC to an address selected from those of the other users (except for the user’s own address and the spammers addresses). If a genuine user does not have any MSC, the e-mail cannot be sent to an address using the SAGA$_{BC}$.
3) **Refunds**
None of the e-mails sent by the (N-S) genuine users are spam. The 1 MSC fee sent to the receiving wallet is refunded to the wallet associated with the user’s e-mail account.
4) **Loop for genuine users**
The simulation repeats routines 1, 2, and 3 above for the (N-S) users. In general users do not perform mining.
5) **Sending e-mails and MSCs**
A spammer uses the SAGA$_{BC}$ to send a spam message and 1 MSC to an address selected from those of the other users (except for the spammer’s own address. If the wallet associated with the spammer’s account has no MSC, the spammer cannot send the spam message.
6) **Refunds**
Any e-mail sent by a spammer is considered spam, and thus the MSC that spammers send to receiving wallets are not refunded.
7) **Profit**
Spammers make a profit b via the probability p per spam message sent. Spammers acquire the same amount of MSC through the profit b they make from mining.
8) **Loop for sending spam**
The simulation repeats the above routines 5, 6, and 7 T times. T is selected as a uniform random number from natural numbers that satisfy 0T<N. Namely, each spammer will send T spam messages to an e-mail account except for his/her own e-mail account, without overlapping the unit time for each.

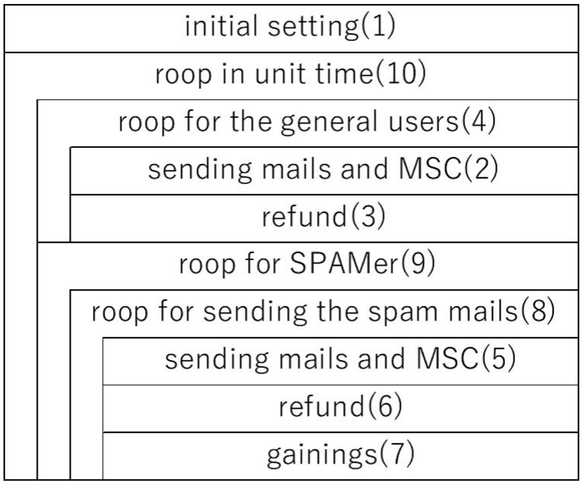


Fig. 3. Main routine of the simulation experiment.

9) **Loop for Spammers**
Routines $5 - 8$ are repeated for all spammers.
10) **Loop in unit time**
Routines $2 - 9$ are considered one unit time (t) and repeated.

B. Parameter

In this simulation, the parameters are set as follows: N=10,000 ($M \subset 980, 1000, 1020$), ($S \subset 300, 500, 700$). The probability distribution P of the profit G is calculated as follows:

$$P = 1000 \times (C)^{(-X)}, \qquad (1)$$

where the constant C is (27), and x is the uniform distribution of random numbers satisfying $0 < x < 330$.

C. Result

This simulation was performed 100 times for each of the three kinds of initial values of MSC, satisfying $S = 500$. Fig. 4 shows the shift of the average throughout each of the 100 runs for the three conditions ($M \subset 980, 1000, 1020$). The horizontal axis of the figure indicates the unit time t. The vertical axis of the figure indicates the ratio of spam to all e-mails sent.

This figure shows that the ratio of spam to all e-mails sent clearly decreases, although the speed of this decrease differs for each of the three kinds of initial values of MSC.

The next simulation was performed 100 times for each of the three numbers of spammers, satisfying ($M = 1,000$. Fig. 5) shows the shift of the average throughout each of the 100 simulations for the three conditions ($S \subset 300, 500, 700$).

This figure shows that the ratio of spam to all e-mail sent clearly decreases, although the speed of this decrease differs for each of the three kinds of profit that spammers make.

IV. Discussion

The results of the simulation show that the SAGA$_{BC}$ can prevent spam. The SAGA$_{BC}$ is more effective than traditional spam prevention methods. Because the spam prevention

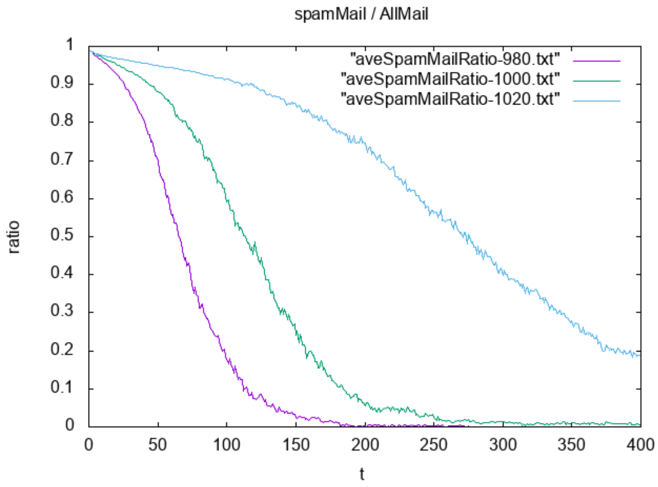


Fig. 4. Result of the simulation ($M \subset 980, 1000, 1020$).

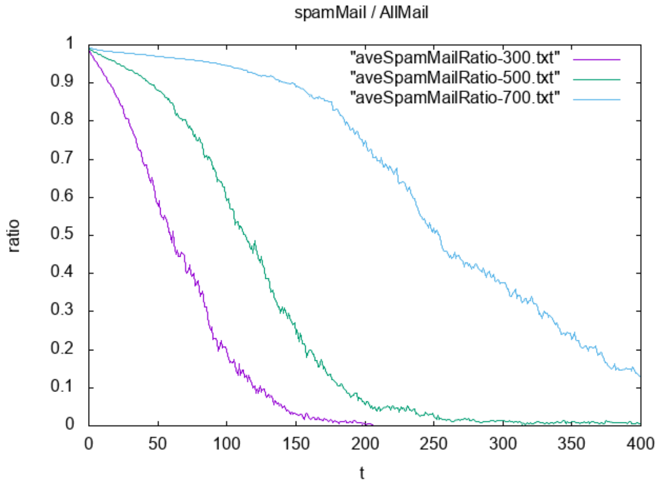


Fig. 5. Result of the simulation ($S \subset 300, 500, 700$).

takes place in both the sending server and the filter of the receiving side server, there are distinct advantages in using the SAGA$_{BC}$.

1) Even if the sending server of the user is same as that of the spammer, the SAGA$_{BC}$ can prevent spam attacks because the SAGA$_{BC}$ determines whether an e-mail is spam or legitimate in each e-mail account.
2) Even if the spammer switches to a different sending server, the SAGA$_{BC}$ will prevent him/her from sending spam unless he/she acquires MSC.
3) Because the receiving e-mail is paid for with MSC by the sender's wallet, the receiving server and mailer do not need to assess the contents of an e-mail and thus have a small workload.
4) Users sending e-mails have the assurance that their messages will not be classified as spam as long they pay the MSC fee.
5) If a user receives an e-mail for which the MSC has been paid that turns out to be spam, he or she accrues the MSC paid because the fee is not refunded to the spammer.
6) The results of the simulations indicate that spam attacks will decrease when e-mails are sent using the SAGA$_{BC}$.

Because the SAGA$_{BC}$ creates disadvantages for spammers, they will not use it. However, genuine users can be assured that e-mail they receive which have been paid for MSC are unlikely to be spam.

V. Conclusion

In this paper, we propose the SAGA$_{BC}$ algorithm to prevent spam attacks using a blockchain. SAGA$_{BC}$ cooperates with an e-mail account associated with a wallet account to achive this. Anyone who sends an e-mail message must pay a processing fee in cryptocurrency MSC. However, if the e-mail is received correctly, the fee will be refunded to the sender. We conducted a simulation experiment to demonstrate that spam attacks decreased when using SAGA$_{BC}$.

In a future experiment, we will add more effective functions to the algorithm, to ensure that it has a variety of uses.

Acknowledgment

This work was supported by JST-Mirai Program Grant Number JPMJMI17D3, Japan.

References

[1] W. Meng, and W. Schlitt, Sender policy framework (SPF) for authorizing use of domains in e-mail, version 1. No. RFC 4408. 2006.

[2] A. Ramachandran, N. Feamster, and D. Dagon, Revealing Botnet Membership Using DNSBL Counter-Intelligence. SRUTI 6. pp. 49-54. 2006.

[3] L. Barry and J. Fenton, DomainKeys Identified Mail (DKIM): Using Digital Signatures for Domain Verification. CEAS. 2007.

[4] G. C. Pieters, The Potential Impact of Decentralized Virtual Currency on Monetary Policy, Globalization and Monetary Policy Institute 2016 Annual Report. 2017.

[5] J. Herbert and A. Litchfield, A Novel Method for Decentralised Peer-to-peer Software License Validation Using Cryptocurrency Blockchain Technology, Proceedings of the 38th Australasian Computer Science Conference (ACSC 2015), Vol. 27, 2015.

[6] S. Nakamoto, Bitcoin: A Peer-to-peer Electronic Cash System. 2008.

[7] G. J. Wood, Ethereum: A Secure Decentralised Generalised Transaction Ledger, Ethereum project yellow paper, 151, pp. 1-32, 2014.

[8] P. L. Serge, Cryptocurrency and E-money Should not be Conflated. Medium. 2017.

[9] S. Yunus, What is the Difference between a Cryptocurrency, a Digital Currency, and a Virtual currency?. QUORA. 2018.

[10] D. He, K. Habermeier, R. Leckow, V. Haksar, Y. Almeida, M. Kashima, N. K. Saad, H. Oura, T. S. Sedik, N. Stetsenko, and C. V. Yepes, Virtual Currencies and Beyond: Initial Considerations, IMF STAFF DISCUSSION NOTE, SDN, 16 (3), 2016.

[11] A. Hayes, What factors give cryptocurrencies their value: An empirical analysis. 2015.

[12] A. K. M. Meera, Cryptocurrencies From Islamic Perspectives: The Case Of Bitcoin, Buletin Ekonomi Moneter Dan Perbankan, Vol.20, No.4, pp.443-460, 2018.

[13] M. Omri, A conceptual framework for the regulation of cryptocurrencies, U. Chi. L. Rev. Dialogue, Vol. 82, pp. 53-68, 2015.

Enhanced Textual Password Scheme for Better Security and Memorability

Hina Bhanbhro
Department of Computer Syst. Eng.
Faculty of Electrical, Electronics & Computer Systems Engineering
Shaheed Benazir Bhutto University

Syed Raheel Hassan
Department of Computer Science
Faculty of Computing and Information Technology
King Abdulaziz University

Shah Zaman Nizamani
Department of I.T.
Faculty of Science
Quaid-e-Awam University

Sheikh Tahir Bakhsh, Madini O.Alassafi
Department of Information Technology
Faculty of Computing and Information Technology
King Abdulaziz University

***Abstract*—Traditional textual password scheme provides a large number of password combinations but users generally use a small portion of available password space. Complex textual passwords are difficult to remember, therefore most users choose passwords with small length and contain dictionary words. Due to the use of small password length and dictionary words, textual passwords become easy to crack through offline guessability attacks. Traditional textual passwords scheme is also weak against keystroke logger attacks because alphanumeric characters are directly inserted into the password field. In this paper, enhancements are proposed in the registration and login screen of the traditional textual password scheme for improving security against offline guessability attacks and keystroke logger attacks. The proposed registration screen also improve memorability of traditional textual passwords through visual cues or pattern-based approach. In the proposed login screen, passwords are indirectly inserted into the password field, to resist keystroke logger attacks. A comparative analysis between the passwords created in traditional and proposed pattern-based approach is presented. The testing results show that users create strong and high entropy passwords in the proposed pattern-based approach as compared to the traditional textual passwords approach.**

***Keywords*—Security; usability; alphanumeric passwords; authentication**

I. Introduction

When users have the freedom to select any textual password, they select weak passwords or they select predictable password patterns [1]. Weak passwords can be guessed through dictionary attacks or brute force attacks [2]. The processing power of computing machines is increasing over time [3], due to which offline guessability attacks such dictionary and brute-force attacks become less effort taking. To resist offline guessability attacks, the users are encouraged to create strong textual passwords by using different password setting policies.

Password-based authentication techniques are called knowledge-based authentication techniques. In these techniques, passwords can be graphical or textual. Graphical passwords consist of some graphical elements such as pictures or drawings. Textual passwords consist of some combinations of alphanumeric characters. Graphical passwords contain visual cues for password memorization, therefore they are considered better than textual passwords in terms of password memorability. However, the graphical passwords are weak against shoulder surfing attacks because they can be easily viewed. Complex textual passwords are difficult to memorize, therefore users use dictionary words in their textual passwords. This approach makes textual passwords vulnerable to dictionary attacks.

A. Password Memorization Techniques

In the category of knowledge-based authentication, textual passwords are most widely used for authentication. However, users set easy to guess or weak textual passwords due to memorability limitations [2]. The textual passwords which contain mix of alphanumeric characters, and have high length are better for security but such passwords are difficult to memorize. Different techniques are suggested by researchers for memorizing hard to guess (strong) passwords, the techniques are listed here:

(a) Passphrase
(b) Cognitive passwords
(c) Associative passwords
(d) Mnemonic passwords

a) Passphrase: Passphrase is a set of words that together form a password. Passphrases are generally easy to remember and difficult to guess because they contain large number of alphanumeric characters. For example, the passphrase “clean the table at the corner” contains 29 characters and it is also easy to memorize due to logical meaning of the phrase.

b) Cognitive passwords: In cognitive passwords, a series of questions are asked and the users are authenticated when correct answers are given. The users select the set of questions and their answers. For example, “What is the name of your birth place?” Cognitive passwords are difficult to guess because the attackers have to correctly identify both questions and answers. These passwords are weak in terms of useability because selecting questions and writing answers take some time.

c) Associative passwords: In associative passwords, the users select some dictionary words provided by the system and their related textual response. For example, “wall=painting”. These passwords are difficult to guess because it is difficult to identify the words used as passwords. Associative passwords are easy to memorize for one or two user accounts but it becomes difficult when large number of associative passwords are required to be remembered.

d) Mnemonic passwords: In this technique of password memorization, a complete sentence is memorized by a user and the password consists of the first letter of each word of the sentence. For example, the sentence “Birth date of Aliza is on 1st May” can be memorized for the mnemonic password “BdoAio1M”. Through this approach random alphanumeric characters of a password can be easily remembered.

All the above password memorization techniques help in memorization of strong passwords but recalling multiple passwords for different accounts is a difficult task in all the password memorization techniques. Users feel difficulty in correctly recognizing which password belongs to which user account. These password memorization techniques also require some mental effort for recalling the passwords [4], therefore the techniques are not widely used by the users.

In this paper, some enhancements are suggested in password registration and login screens of the traditional textual password scheme for better memorization of passwords. In the proposed enhancements, visual cues can be added to the random alphanumeric characters of the password by using the pattern-based approach. The visual cues help in memorization of strong alphanumeric passwords and through the proposed pattern-based approach, the users’behavior of setting weak passwords can also be changed.

The remaining paper is divided into five sections. In Section 2 literature review is given about the problems in traditional textual passwords. Research methodology for pattern-based passwords is explained in Section 3. In Section 4 analysis of pattern-based approach is presented. Finally the conclusion is given in Section 5.

II. Related Work

Textual passwords were first analyzed by Morris and Thompson in 1979 [5]. They found that 86% of the passwords were weak e.g. passwords had small length, consisting of lowercase letters only, digits only or mixture of the two with dictionary words. After Morris and Thompson [5] a large number of studies have been done on understanding the characteristics of textual passwords set by the users. Researchers from Microsoft carried out a study [6] that involved half a million web-based passwords. From the study, the researchers found that users generally create weak passwords and they reuse the same password across multiple accounts. Same password weaknesses were also found by Smith [7] and Borges *et al.* [8]. These studies suggest that a large number of textual passwords can be cracked through dictionary attacks. klein [9] performed a dictionary attack on 15,000 passwords through the password dictionary of 3000000 alphanumeric strings. Klein [9] successfully cracked 25% passwords through the dictionary attack. The research studies on textual passwords highlight the need for motivating users for setting strong or secure passwords.

Liu *et al.* [10] analyzed the length of textual passwords set by the users, the researchers found that majority of users set passwords with a length of less than twelve alphanumeric characters. The result of Liu *et al.* [10] shows that small portion of textual password space is being used by the users, as a result different offline guessability attacks become easy to apply [11].

Viktor Taneski *et al.* [12] conducted a systematic literature review of articles and journals about password use and password security. The researchers [12] suggested different password setting policies and password checkers to guide users for setting strong passwords. Password meters [13] are used in some websites for informing users about the strength of the passwords. Egelman *et al.* [14] analyzed the effect of password meters. They found that password meters do not have a significant effect on changing behavior of users, towards setting strong passwords due to memorability issues. Strict password creation policies have a poor effect on memorability [15].

Strong textual passwords contain strings of alphanumeric characters which do not belong to dictionary words and they have larger lengths [16]. Different password setting policies are applied in applications for enforcing users to set strong passwords, such as minimum length and a mix of multiple categories of alphanumeric characters. However, research studies [17] suggest that users create weak passwords even after applying password setting policies.

Due to human limitations of information memorization, users reuse the same password across different accounts. Florencio and Herley [6] found that, on average users can easily remember 6.3 different textual passwords. However, users generally have more than six accounts, therefore they reuse same password in different accounts.

The advantage of strong textual passwords decreases when the same password is re-used in multiple user accounts [18]. The attacker after cracking a password from the less secure application, apply the cracked password on more secure applications [19]. Therefore, it is also important that users should create separate passwords for different accounts. Password managers are used to create strong and distinct password for each user account. However, password managers have some security and usability issues [20] i.e. a password manager can be hacked and it may not be available all the time. Privacy is also an issue with the password managers because all the passwords will be presented into a third party software.

Cognitive or visual cues are helpful for memorization of information. Therefore, in the proposed enhancements of traditional textual password scheme, the users can draw a visual pattern from the alphanumeric characters of a password on the registration screen. The password pattern serves as visual cue for better memorization of the textual password. Pattern-based passwords help users to easily set and remember strong or secure alphanumeric passwords.

III. Research Methodology

To analyze the effect of pattern-based passwords, a user study was conducted. The user study was divided into four phases. In the first phase, password setting trends in traditional textual passwords were analyzed through a pre-test survey. In the second phase, users were asked to register and login inside the specially designed application for testing usability, security, and memorability aspects of the pattern-based passwords. In the third phase, a post-test survey was conducted for understanding the feelings of the users about the pattern-based approach. In the fourth phase, the data of pre-test survey and application testing was analyzed for security, usability, and memorability aspects of the pattern-based approach.

For this research 110 users participated, out of which 43 were female and 67 were male. All the users belonged to different professions including students, teachers and administration staff. The users belonged to different institutions including the Quaid-e-Awam University of Engineering, Science & Technology, SZABIST Nawabshah and Shah Abdul Latif University.

A. *Pre-test survey*

The objective of the pre-test survey was to find out the password setting trends in traditional textual passwords. Due to privacy issues exact textual passwords were not collected from the users but they were asked to provide some information about the passwords. In the pre-test survey users were asked to provide information regarding size and type of alphanumeric characters used in their passwords. Results of the pre-test survey are shown in Table I.

TABLE I. PASSWORD TRENDS IN TRADITIONAL TEXTUAL PASSWORDS

Average password length	9.56
Lower-case letters used	87%
Upper-case letters used	19%
Numbers used	54%
Special characters used	21%

Table I shows that majority of users use lower-case letters in their passwords along with decimal numbers. Generally in computer applications, users are restricted to select at least two categories of alphanumeric characters. Decimal numbers are easy to remember along with some dictionary words. Therefore, comparably high percentage of decimal numbers are used in comparison with capital letters and special characters.

B. *Application Testing*

For analyzing usability and memorability aspects of pattern-based passwords, a web-based application was developed. In the application, users were asked to perform registration and login activities. Timings of registration and login activities were recorded through the testing application. Failed and successful login attempts were also saved in the database of the application.

1) Registration Activity: In the registration activity, the participants created their accounts in the testing application. Registration screen of the testing application is shown in Fig. 1. The alphanumeric characters are present on the registration screen along with profile and authentication fields as shown in Fig. 1. Different categories of alphanumeric characters are separately presented on the registration screen. For example, lower-case letters are present at the top of the screen while numbers are present at the bottom of the registration screen. This arrangement helps in recalling password characters from the password patterns.

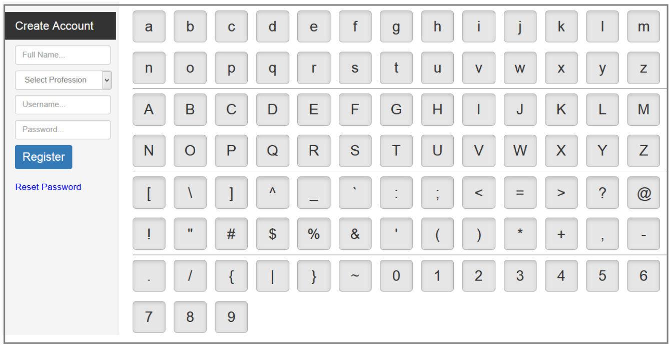


Fig. 1. Registration screen.

a) Password selection: Passwords can be entered through keyboard or mouse on the registration. Through keyboard, the passwords can be entered by pressing alphanumeric keys similar to the traditional textual password scheme. For mouse-based password entry, a user needs to drag or click over alphanumeric characters present on the registration screen. For example, if a user drags the mouse from “h” to “1”, “j” to “3” and “U” to “W” then some lines will be drawn over the alphanumeric characters as shown in Fig. 2. Visually the password will look like “H” but in the database, the password *huHU;(1jwJW=*3UVW* will be saved.

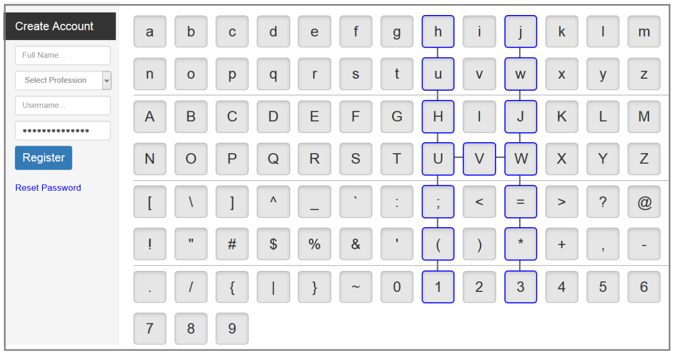
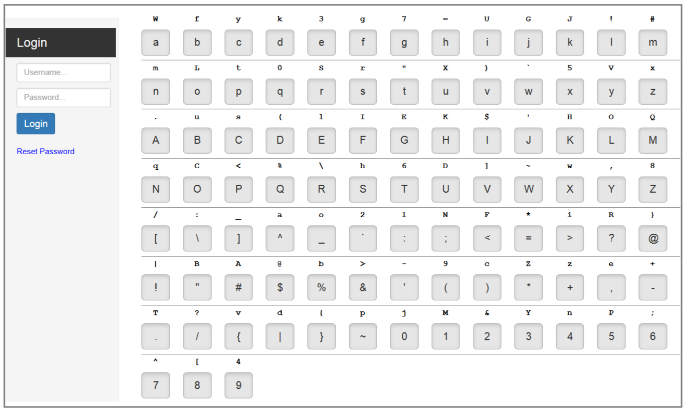


Fig. 2. Registration screen after password selection.

When the mouse is dragged multiple times over same alphanumeric characters then same password pattern will be created but alphanumeric characters will be repeated in the password. For example if a user drags the mouse two times from “h” to “1” then visually a straight line will be formed from “h” to “1” but in the database, the password *huHU;(1huHU;(1* will be saved.

Through pattern-based approach a strong textual password such as given in the example can be easily memorized by the visual cues. A user just needs to remember the visual shape of the password along with starting and ending alphanumeric characters of the password pattern.

2) Login activity: In this activity, users provide login credentials (username & password) for authentication. The login screen of the testing application is shown in Fig. 3. The login screen contains two sequences of alphanumeric characters, one for actual password characters and other for temporary representation of the password characters. Actual alphanumeric characters (present in the button shape) are sequentially present, while representing alphanumeric characters are randomly present in the login screen i.e. they change their position in every login session. For randomization of the representing characters, Algorithm-1 can be used.



W	f	y	k	3	g	7	=	U	G	J	!	#
a	b	c	d	e	f	g	h	i	j	k	l	m
m	L	t	0	S	r	"	X	)	`	5	V	x
n	o	p	q	r	s	t	u	v	w	x	y	z
.	u	s	(	1	I	E	K	$	'	H	O	Q
A	B	C	D	E	F	G	H	I	J	K	L	M
q	C	<	%	\	h	6	D	]	~	w	,	8
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
/	:	_	a	o	2	1	N	F	*	i	R	}
[	\	]	^	_	`	:	;	<	=	>	?	@
|	B	A	@	b	>	-	9	c	Z	z	e	+
!	"	#	$	%	&	'	(	)	*	+	,	-
T	?	v	d	{	p	j	M	&	Y	n	P	;
.	/	{	|	}	~	0	1	2	3	4	5	6
^	[	4										
7	8	9										

Fig. 3. Login screen.

Algorithm 1 Algorithm for Alphanumeric Characters Randomization

```
$TempElements \leftarrow list\ of\ alphanumeric\ characters$
$RepresentingElements \leftarrow NULL$
$ELength \leftarrow 95$
for $i = 0$ to 94 do
    $temp \leftarrow NULL$
    $ind \leftarrow random(0,\mathrm{ELength})$
    $RepresentingElements[\mathrm{i}] \leftarrow TempElements[ind]$
    $TempElements[\mathrm{i}] \leftarrow remove(\mathrm{ind},TempElements)$
    $ELength \leftarrow \mathrm{ELength - 1}$
end for
```

a) Password selection: On the login screen, a password can be entered through keyboard or mouse. Passwords are entered through the keyboard by typing characters representing the password of a user. For example, if the password of a user is *huHU;(1jwJW=*3UVW* and characters representation are same is shown in Fig. 4, then the user has to type *iUrE!)PVT6Jx}>J1E* characters in the password field for authentication. In this case, first password character "h" is represented by "i" and the password character "u" is represented by "U" and so on. This temporary representation of password characters helps in avoiding keystroke logger attacks.

Mouse-based password entry requires dragging or clicking the mouse over the password characters, similar to the registration screen. For example, in the current scenario when password of a user is *huHU;(1jwJW=*3UVW* then the password can be entered by dragging mouse from "h" to "1", "j" to "3" and "U" to "W". When passwords are entered through mouse then the passwords can be observed through shoulder surfing attacks. Therefore in public, it is better to enter passwords through keyboard. For improving security against shoulder surfing attacks in case of mouse-based password entry, the process of generating password lines should be removed from the login screen. In the login screen, a link is given for reset password. This option clears all the lines drawn from the login screen and the text written in the password field.

C. Post-test Survey

After completing registration and login activities, a post-test survey was conducted. In the post-test survey five questions were asked from the users about the proposed pattern-based approach. The questions and their answers are shown in Table II. The post-test survey was conducted for understanding how comfortable users are in using the proposed pattern-based approach. The results of the post-test survey are shown in Section IV-D.

IV. Results and Discussion

After completing pre-test, post-test surveys and application testing, the data was analyzed to know the performance of pattern-based approach. From the data, security, usability and memorability aspects of pattern-based passwords were analyzed.

A. Strength of passwords

Fig. 5 shows a different types of alphanumeric characters (lower-case, upper-case letters, numbers and special characters) used in traditional and pattern-based passwords. The data for traditional textual passwords were collected from pre-test survey and the data for pattern-based passwords were collected from the database of the testing application.

Fig. 5 shows that in pattern-based passwords, upper case letters and special characters are widely used by the users as comparison with traditional textual passwords. A high percentage of capital letters and special characters in pattern-based passwords show that the dictionary attacks will be difficult to apply in the passwords of pattern-based approach.

Password length and entropy of both the password setting approaches are given in Table III. Results show that password length is slightly higher in traditional textual passwords. However, password entropy of pattern-based passwords is higher than traditional textual passwords. The password entropy shows that users create strong passwords through pattern-based approach.

B. Timings

Password entry time for pattern-based passwords were analyzed through the log stored in the database of the testing application. Average password entry time when passwords were entered through keyboard was 18.24 seconds and when passwords were entered through mouse was 7.19 seconds. The reason for higher login time in keyboard based password entries is that, users have to identify alphanumeric characters which represent the password characters from the login screen.

C. Password Memorability

To test memorability of pattern-based passwords, users were asked to perform login activities in the testing application. Memorability tests were conducted in three different timings, which were immediately after registration, after one day and

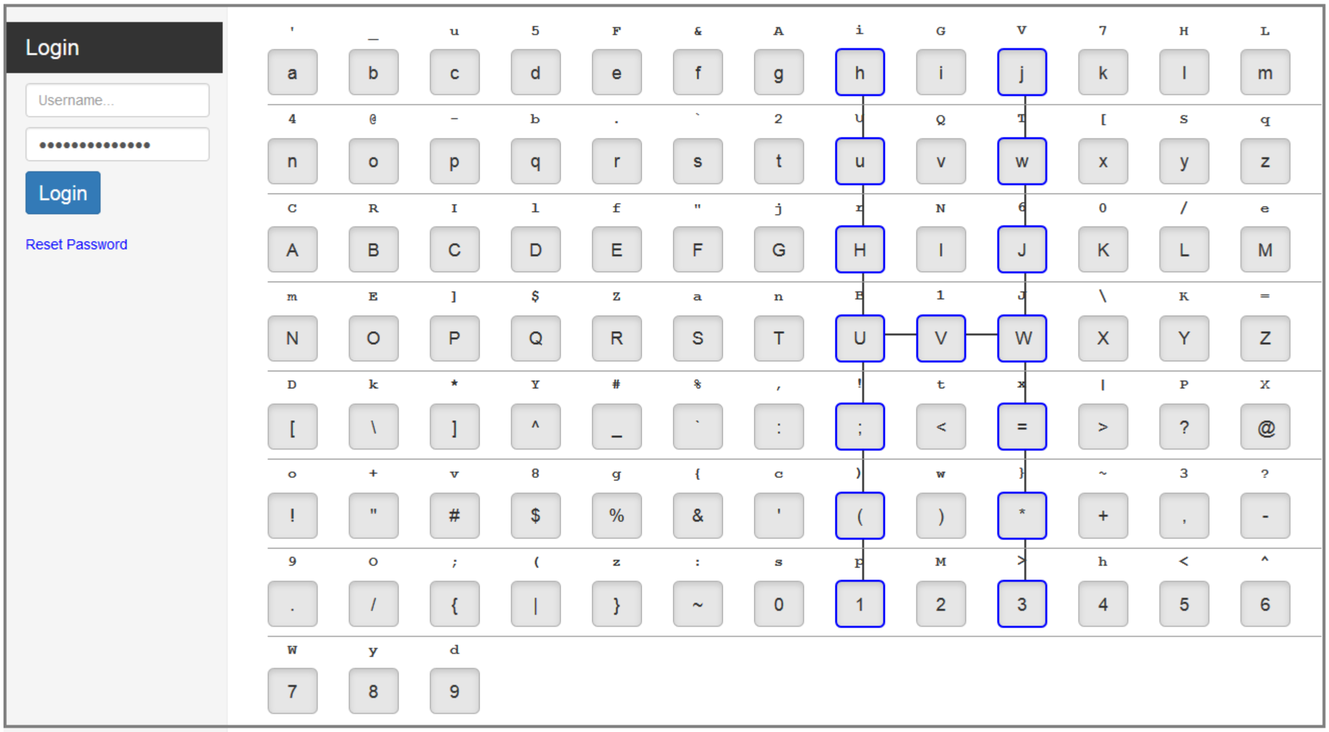


Fig. 4. Login screen after password selection.

TABLE II. POST-TEST QUESTIONNAIRE

No.	Question	Options
1	Do you think the pattern-based approach helps in password memorization?	(a) Yes (b) No (c) Little-bit
2	The pattern-based approach is most useful for?	(a) Registration page (b) Login page (c) Both of them (d) None
3	Pattern-based approach is suitable for?	(a) Desktop applications (b) Web-based applications (c) Mobile applications (d) All of them
4	How do you remember the password?	(a) Password patterns only (b) Alphanumeric characters (c) Both of them
5	How easy it is to use the pattern-based approach?	(a) Very Easy (b) Easy (c) Average (d) Difficult

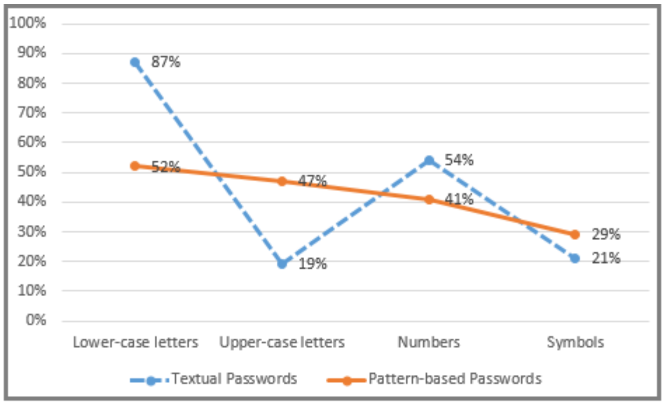


Fig. 5. Alphanumeric characters comparison.

TABLE III. PASSWORDS STRENGTH

Scheme	Password length	Password entropy
Textual passwords	9.56	53.88
Pattern-based passwords	9.21	57.62

after one week. Users were allowed maximum three attempts for a successful login.

Table IV shows password recall rate in the pattern-based approach. Results show that memorability of textual passwords is improved by the pattern-based approach. The results also show that with the passage of time password memorability decreases. In this experiment, users only interacted with the login page in specified timings (immediately after registration, after one day and one week), they were not allowed to login on 2nd to 6th day of registration. If the users were allowed to login inside the system on each day, then the memorability results would have been much better after one week duration because recalling information on short periods have positive effect on memorability.

TABLE IV. PASSWORDS MEMORABILITY

Time	1st attempt	Within 2 attempts	Within 3 attempts
Immediately after registration	83%	89%	98%
After 1 day	78%	82%	89%
After 1 week	69%	73%	76%

D. Qualitative Analysis

In the post-test survey, users were asked to share their views about the pattern-based password setting approach. In the survey five questions were asked from the users, the results are shown in Tables V to IX.

Pattern-based approach provides multiple ways (cognitive and visual) of password memorization. Therefore, users find this approach helpful for password memorization.

TABLE V. RESULTS OF QUESTION 1

Q.01: Do you think pattern-based approach helps in password memorization	
Option	**Answer**
Yes	81%
No	7%
Little-bit	12%

TABLE VI. RESULTS OF QUESTION 2

Q.02: Pattern-based approach is useful for	
Option	**Answer**
Registration page	17%
Login page	14%
Both of them	63%
None	6%

The results of Table VI show that majority of users found the proposed approach useful for both registration and login pages. The reason for this choice is that registration page helps in password memorization and login page improves password security.

TABLE VII. RESULTS OF QUESTION 3

Q.03: Pattern-based approach is suitable for	
Option	**Answer**
Desktop applications	11%
Web-based applications	68%
Mobile applications	5%
All of them	16%

Table VII shows that majority of users preferred web-based applications for the pattern-based. The web-based applications are more vulnerable to security attacks than desktop and mobile applications. The pattern-based approach improves password security, therefore this approach is most suitable for the web-based applications.

TABLE VIII. RESULTS OF QUESTION 4

Q.04: How do you remember the password	
Option	**Answer**
Password pattern only	73%
Alphanumeric characters	19%
Both of them	8%

Table VIII shows that majority of users remembered password patterns instead of alphanumeric characters of the passwords. Visual password shapes are easy to remember than random alphanumeric characters, therefore most of the users only remembered the password patterns.

Android unlock scheme is widely used by the users and the proposed pattern-based approach has similarities with the scheme. Therefore, users found easy to use the proposed pattern-based approach as shown in Table IX.

TABLE IX. RESULTS OF QUESTION 5

Q.05: How easy it is to use the pattern-based approach	
Option	**Answer**
Very easy	22%
Easy	59%
Average	14%
Difficult	5%

The qualitative analysis done through the post-test survey shows that users are satisfied with the performance of pattern-based approach i.e. they found pattern-based approach easy to use and helpful for password memorization. However, the majority of users prefer to use this approach in web based applications because chances of password hack are high in web based applications.

V. DISCUSSION

Many knowledge based authentication schemes are proposed which provide better security than textual password scheme. However, passwords of the secure knowledge based authentication schemes are difficult to memorize and password entry procedures are very complex. Due to memorability and usability issues, such schemes are not accepted by the software industry. Through the proposed scheme, security and memorability improvements are made in the traditional textual password scheme.

Although proposed scheme improves security of textual passwords against different security attacks such as dictionary attacks and keystroke logger attacks, but the proposed scheme is vulnerable to screen-scrapper attack. In this attack a password is captured by recording both password input and login screen. In the proposed scheme, one-to-one relationship exists between actual password characters and their representing alphanumeric characters for a particular session. Therefore, passwords can be captured by recording both representing password characters and screen-shot of the login screen. For further improving security of textual passwords, the proposed scheme needs to be enhanced to resist the screen-scrapper attacks.

VI. CONCLUSION

Strong textual passwords which contain mix of alphanumeric characters are difficult to memorize because they do not contain cognitive or visual cues for password memorization. Through the proposed pattern-based password setting approach, strong textual passwords become easy to memorize due to visual cues. Qualitative and quantitative results show that memorability of strong alphanumeric or textual passwords is improved through the pattern-based approach.

Users create strong textual passwords through the pattern-based approach, therefore brute-force and dictionary attacks will be difficult to apply when the pattern-based approach is used in the registration screen. Keystroke logger attacks are resisted in the proposed login screen by indirectly collecting password characters from the users.

ACKNOWLEDGMENT

The authors acknowledge the faculty members, students and administration staff of Quaid-e-Awam University, SZ-ABIST nawabshah and Shah Abdul Latif University for helping us in performing different surveys and application testing for the research work.

REFERENCES

[1] A. Adams and M. A. Sasse, “Users are not the enemy,” *Communications of the ACM*, vol. 42, no. 12, pp. 40–46, 1999.

[2] A. Forget, S. Chiasson, and R. Biddle, “Helping users create better passwords: Is this the right approach?” in *Proceedings of the 3rd Symposium on Usable Privacy and Security*. ACM, 2007, pp. 151–152.

[3] G. E. Moore, “Cramming more components onto integrated circuits, electronics magazine,” 1965.

[4] N. Ekstrom, “Password practice: The effect of training on password practice,” 2015.

[5] R. Morris and K. Thompson, “Password security: A case history,” *Communications of the ACM*, vol. 22, no. 11, pp. 594–597, 1979.

[6] D. Florencio and C. Herley, “A large-scale study of web password habits,” in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 657–666.

[7] S. W. Smith, “Humans in the loop: Human-computer interaction and security,” *IEEE Security & privacy*, vol. 99, no. 3, pp. 75–79, 2003.

[8] M. A. Borges, M. A. Stepnowsky, and L. H. Holt, “Recall and recognition of words and pictures by adults and children,” *Bulletin of the Psychonomic Society*, vol. 9, no. 2, pp. 113–114, 1977.

[9] D. V. Klein, “Foiling the cracker: A survey of, and improvements to, password security,” in *Proceedings of the 2nd USENIX Security Workshop*, 1990, pp. 5–14.

[10] Z. Liu, Y. Hong, and D. Pi, “A large-scale study of web password habits of chinese network users.” *JSW*, vol. 9, no. 2, pp. 293–297, 2014.

[11] S. Chiasson, A. Forget, R. Biddle, and P. C. van Oorschot, “Influencing users towards better passwords: persuasive cued click-points,” in *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction-Volume 1*. British Computer Society, 2008, pp. 121–130.

[12] V. Taneski, M. Hericko, and B. Brumen, “Password securityno change in 35 years?” in *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*. IEEE, 2014, pp. 1360–1365.

[13] M. Bishop and D. V. Klein, “Improving system security via proactive password checking,” *Computers & Security*, vol. 14, no. 3, pp. 233–249, 1995.

[14] S. Egelman, A. Sotirakopoulos, I. Muslukhov, K. Beznosov, and C. Herley, “Does my password go up to eleven?: the impact of password meters on password selection,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 2379–2388.

[15] R. Shay, S. Komanduri, P. G. Kelley, P. G. Leon, M. L. Mazurek, L. Bauer, N. Christin, and L. F. Cranor, “Encountering stronger password requirements: user attitudes and behaviors,” in *Proceedings of the Sixth Symposium on Usable Privacy and Security*. ACM, 2010, p. 2.

[16] P. C. Dean, J. Buck, and P. Dean, “Identity theft: A situation of worry,” *Journal of Academic and Business Ethics*, vol. 9, p. 1, 2014.

[17] D. Florencio, C. Herley, and B. Coskun, “Do strong web passwords accomplish anything?” *HotSec*, vol. 7, no. 6, 2007.

[18] B. Ives, K. R. Walsh, and H. Schneider, “The domino effect of password reuse,” *Communications of the ACM*, vol. 47, no. 4, pp. 75–78, 2004.

[19] B. Prince, “Twitter details phishing attacks behind password reset. eweek,” 2010.

[20] Z. Li, W. He, D. Akhawe, and D. Song, “The emperor’s new password manager: Security analysis of web-based password managers.” in *USENIX Security Symposium*, 2014, pp. 465–479.

Automated Quantification of Non-Calcified Coronary Plaques in Cardiac CT Angiographic Imagery

M Moazzam Jawaid[1]*, Sanam Narejo[1], Nasrullah Pirzada[1], Junaid Baloch[1], C.C. Reyes-Aldasoro[2],Greg Slabaugh[2]
[1]Faculty of Electrical & Electronic Engg.
Mehran University of Engineering & Technology, Jamshoro 76062
[2]Deptt. of Electrical & Electronic Engg.
City University London, EC1V 0HB

Abstract—**The high mortality rate associated with coronary heart disease (CHD) has driven intensive research in cardiac image analysis. The advent of computed tomography angiography (CTA) has turned non-invasive diagnosis of cardiovascular anomalies into reality as calcified coronary plaques can be easily identified due to high intensity values. However, detection and quantification of the non-calcified plaques in CTA is still a challenging problem because of their lower intensity values, which are often similar to the nearby blood and muscle tissues. In this work, we propose Bayesian posterior based model for precise quantification of the non-calcified plaques in CTA imagery. The only indicator of non-calcified plaques in CTA is relatively lower intensity. Hence, we exploited intensity variations to discriminate voxels into lumen and plaque classes. Based on the normal coronary segments, we computed the vessel-wall thickness in first step. In the subsequent step, we removed vessel wall from the segmented tree and employed Gaussian Mixture Model to compute optimal distribution parameters. In the final step, distribution parameters were employed in Bayesian posterior model to classify voxels into lumen or plaque. A total of 18 CTA volumes were analyzed in this work using two different approaches. According to the experimental results, mean Jaccard overlap is around 88% with respect to the manual expert. In terms of sensitivity, specificity and accuracy, the proposed method achieves 84.13% ,79.15% and 82.02% success, respectively. Conclusion: According to the experimental results, it is shown that the proposed plaque quantification method achieves accuracy equivalent to human experts.**

***Keywords*—Coronary segmentation; non-calcified plaques; vascular quantification; coronary wall analysis**

I. Introduction

Coronary heart disease (CHD) is related to the accumulation of fatty materials (also termed as coronary plaques) inside coronary arteries. The recent statistics of the National Health Services, United Kingdom [2] reveals that over 2.3 million people in the United Kingdom suffer from CHD where the annual death toll is approximately 73,000 (an average of one death every seven minutes). The substantial levels of growing morbidity and mortality have led to a intensified interest in new techniques for detecting coronary abnormalities to potentially avoid worst events [1], [2].

The recent advancements in non-invasive imaging have improved the diagnostic accuracy in terms of high temporal and spatial resolution [3]; however, detection and quantification of non-calcified plaques in CTA is still a challenging problem. Clinically, the non-calcified plaques have been established as the most important indicator of acute coronary syndromes due to their fragile nature [4]. The risk of sudden rupture has made soft plaques threatening in clinical context, i.e. for many individuals, sudden death becomes the first sign of soft plaque in contrast to the calcified plaques which often lead to disease symptoms at early stages. It should be noted that calcified plaques can be identified easily in a CTA image based on the high intensity value, consequently numerous methods have been reported with a reasonable quantification accuracy [5]–[8]; however, non-calcified plaque requires more sophisticated phenomenon. In context of the flow of paper, we start with relevant literature and CTA data specification. Subsequently, we explain the plaque quantification methodology which is followed with the Results section. The lumen - plaque quantification results are provided in the Results section using statistical metrics of sensitivity, specificity and accuracy, with respect to manual experts.

II. Related Work

Non-calcified plaque detection and quantification in CTA has been a challenging problem; hence, there is a little literature [9]–[12], [14] published addressing automatic segmentation, out of which the majority have been clinical pilot studies or generic anomaly detection techniques. The use of machine learning in soft plaque detection was first reported by Wei *et al.* [10] where a linear discriminant analysis (LDA) was used to reduce the false positives in a set of 120 preselected soft plaque candidates. Accordingly, the detection accuracies reported were 94% and 79%, respectively for the calcified and non-calcified plaques, along with a high number of false positives. Another interesting method for the automatic detection of vascular abnormalities was proposed by Zuluaga *et al.* [13]. In this work, an unsupervised SVM model trained on normal cross sections was used to detect the outliers i.e. the cross sections which violate the intensity pattern of normal class. The authors reported promising results for 9 clinical CTAs with NCP detection accuracy of 79.62%, however; the precise quantification was not performed in this work. Similarly, the detection methods were reported by Renard and Yang [14], Lankton *et al.* [11], Li *et al.* [15]; however, the precise quantification has not been reported frequently.

In context of the non-calcified plaque quantification, a number of algorithms [16], [17], [34], [35] have been proposed in recent years with a motive of correlating CTA based plaque quantification with intra-vascular ultrasound (IVUS) measurements. Athanasiou et al. [35] employed 4-class Gaussian Mixture Model to identify respective classes namely lumen, calcified plaque, non-calcified plaques and the background. Accordingly, the paper reported efficiency over the existing literature and a good correlation with IVUS measurements; however, the blooming effect of calcified plaque resulted in relatively low agreement for calcified plaque volume. Moreover, the vessel wall analysis was not performed explicitly, which is very crucial in context of the non-calcified plaques because of two-way vessel remodelling.

In addition, a number of studies [16], [18]–[23], [34], [36]–[39] have been reported in context of the non-calcified plaque quantification: however, the main focus in these studies was to demonstrate the capability of CTA imaging to reflect the non-calcified plaque rather than automated quantification of non-calcified plaque in CTA. Accordingly, non-calcified plaque lesions were manually selected in the first step, and plaque quantification results were compared with respect to intra-vascular ultrasound analysis to establish correlation between two imaging modalities.

Our contribution in this work is an efficient methodology for quantification of the non-calcified plaques with a human-equivalent accuracy. First, we present an efficient method for the vessel wall analysis in context of the non-calcified plaques. The proposed vessel-wall analysis can be used as a stand-alone plaque detection method as well it serves as important step towards plaque quantification. In addition, we formulate a posterior class based plaque quantification method for voxel-wise plaque quantification with a human-equivalent accuracy.

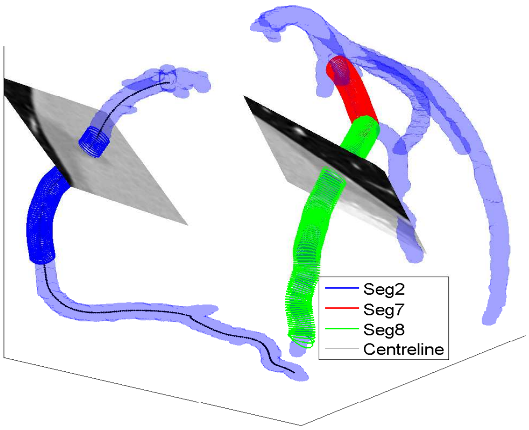
In this work, we employed clinical CTA data (a total of 16 CTA images) obtained from publicly available database of Rotterdam Coronary Artery Evaluation framework [24], [25]. The Rotterdam CTA data comes from different sources and is based on different vendors as described in [24]. The motive behind using Rotterdam data is the availability of the manual ground truth in terms of expert annotations i.e. segment- wise status (normal/abnormal) and the precise position of non-calcified plaque for the abnormal coronary segments. Based on the provided ground truth, we identified the individual coronary segments affected with non-calcified plaques as defined in Table I.

TABLE I. NON-CALCIFIED PLAQUE EFFECTED SEGMENTS IN ROTTERDAM CTA DATA

Segment ID	Plaque Specifications			
	Segment Type	Plaque Type	Plaque Grading	Stenosis(%)
DS1 seg6	Proximal	Non-calcified	mild	20
DS2 seg6	Proximal	Non-calcified	mild	25
DS4 seg1	Proximal	Non-calcified	Severe	65
DS4 seg2	Proximal	Non-calcified	Moderate	51
DS5 seg2	Proximal	Non-calcified	Moderate	57
DS5 seg8	Distal	Non-calcified	Moderate	45
DS7 seg2	Proximal	Non-calcified	Severe	71
DS7 seg3	Proximal	Non-calcified	Moderate	41
DS9 seg2	Proximal	Non-calcified	Moderate	51
DS11 seg7	Proximal	Non-calcified	Mild	22
DS15 seg2	Proximal	Non-calcified	Moderate	53
DS15 seg3	Proximal	Non-calcified	Mild	22
DS15 seg14	Distal	Non-calcified	Moderate	45

III. PROPOSED MODEL

Precise segmentation of the coronary vasculature serves as first step in plaque quantification. Accordingly, we employed hybrid energy model of [26] to extract the coronary tree as illustrated in Fig. 1. Subsequently, radial profile based plaque detection method [27] was applied to precisely localise the plaque in different coronary segments.



(a) DS04 seg1

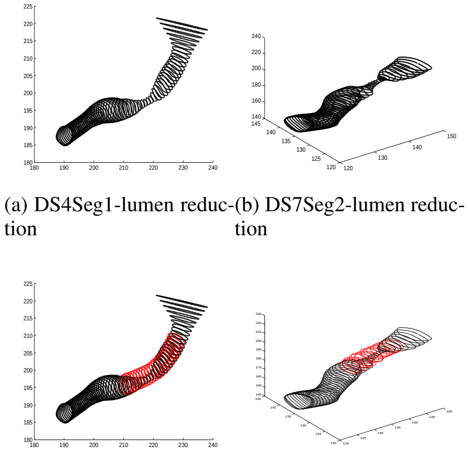
Fig. 1. Segmented coronary trees with overlaid centreline and two cross sectional planes. The centreline is overlaid in black colour for the right coronary artery, whereas blue, red and green represents the curved cylindrical approximations for coronary segments numbered 2, 7 and 8 respectively.

A. Ground Truth Construction

In context of plaque quantification, we started with the "reference" ground truth formulation using plaque position inside respective coronary segments. Because of the ambiguous appearance in CTA imagery, the non-calcified plaque is clinically estimated by evaluating lumen deformations. Accordingly, we used the annotated lumen boundary of Rotterdam experts to derive the voxel-wise plaque ground truth. The lumen diameter variations can be observed in the mid of the vessel as shown in Fig. 2a - 2b, indicating non-calcified plaque instance at respective locations. We approximated the ideal "plaque-free" vessel (red contours) for the plaque affected region using two "normal" cross sections (immediately before and after the plaque region) as shown in Fig. 2c - 2d. In the subsequent step, the annotated lumen (black contour) is subtracted from the ideal vessel (red contour) and the remaining voxels in the plaque free region are labelled as ground truth plaque voxels.

For mathematical formulation of the plaque estimation problem, we represent the coronary segment (lumen boundary annotations) using a tubular model $T_{model}\left[CP, \theta'_{cs}\right]$, where CP denotes the centreline of the segment and θ'_{cs} defines corresponding cross-sectional information. Accordingly, complete coronary segment is represented using an $[N_s]$ by $[m]$ array, where N_s represent the total number of points in segment centreline and m denotes cross-sections related parameters.

The elliptical model is used to represent vascular cross sections, as Vessels are elastic bodies which can accommodate local deformations of the lumen due to changes in the blood flow and intra-luminal pressure. Such deformations cannot



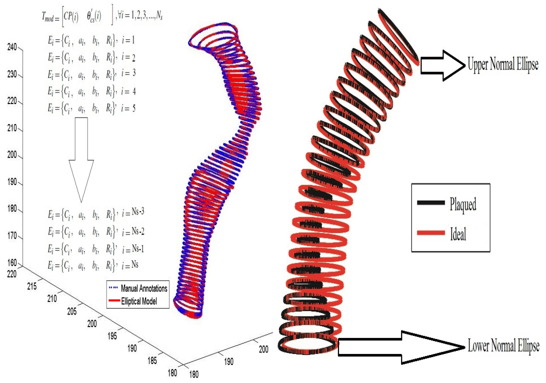
(a) DS4Seg1-lumen reduction (b) DS7Seg2-lumen reduction

(c) DS4Seg1-ideal vessel (d) DS7Seg2-ideal vessel

Fig. 2. Lumen boundary annotations for two non-calcified plaque effected coronary segments. Black contours represent manual annotations for lumen boundary in 3D space, (red) contours define "ideal" (plaque-free) vessel boundary for the plaque effected region of the coronary segment.

be accurately represented using circular cross section model proposed in [28]–[32].

The elliptical model based representation used in this work is illustrated in Fig. 3a. Accordingly, for the i^{th} point of the centreline CP, we define the parameter vector $\theta_{cs}^{'}(i)$ using ellipse as $[E_{xyz}(i) \approx \{a(i), b(i), C_{xyz}(i), R_{xyz}(i)\}]$, where $a(i)$ and $b(i)$ represent the semi axis length for major and minor axes of the current ellipse, $C_{xyz}(i)$ denotes the centre of the i^{th} ellipse of segment, $R_{xyz}(i)$ defines orientation information for i^{th} ellipse and $E_{xyz}(i)$ represents points on the ellipse circumference. Accordingly, the mathematical formulation (parametric representation) for a 3-dimensional ellipse is expressed by (1), where $t^{'}$ denotes the angular parameter varying between 0 to 2π.


$T_{model} = [CP(i) \quad \theta_{cs}^{'}(i)]$, $\forall i = 1,2,3,....,N_s$
$E_i = \{C_1, a_1, b_1, R_1\}$, $i = 1$
$E_i = \{C_2, a_2, b_2, R_2\}$, $i = 2$
$E_i = \{C_3, a_3, b_3, R_3\}$, $i = 3$
$E_i = \{C_4, a_4, b_4, R_4\}$, $i = 4$
$E_i = \{C_5, a_5, b_5, R_5\}$, $i = 5$
$i =$ Ns-3
$i =$ Ns-2
$i =$ Ns-1
$i =$ Ns
240
230
220
210
200
190
180
170
160
220 215 210 205 200 195 190 185 180
180 190 200
Manual Annotations
Elliptical Model
Upper Normal Ellipse
Plaqued
Ideal
Lower Normal Ellipse

(a) (b)

Fig. 3. Tubular model representation and estimation of ideal vessel boundary for plaque effected region of coronary segment. Black contours represent manually annotated lumen boundary in the plaque effected region, red shows the estimated ideal (plaque-free) vessel boundary based on two normal (upper and lower) cross sections.

$$E_{xyz} = \begin{bmatrix} Cx \\ Cy \\ Cz \end{bmatrix} + R_{xyz} \begin{bmatrix} a.cos(t\prime) \\ b.sin(t\prime) \\ 0 \end{bmatrix} \quad (1)$$

where, $R_{xyz} = R_1.R_2.R_3$, and individual rotation values are computed as follows:

$$R1 = \begin{bmatrix} cos(\alpha) & sin(\alpha) & 0 \\ -sin(\alpha) & cos(\alpha) & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$R2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(\beta) & sin(\beta) \\ 0 & -sin(\beta) & cos(\beta) \end{bmatrix},$$

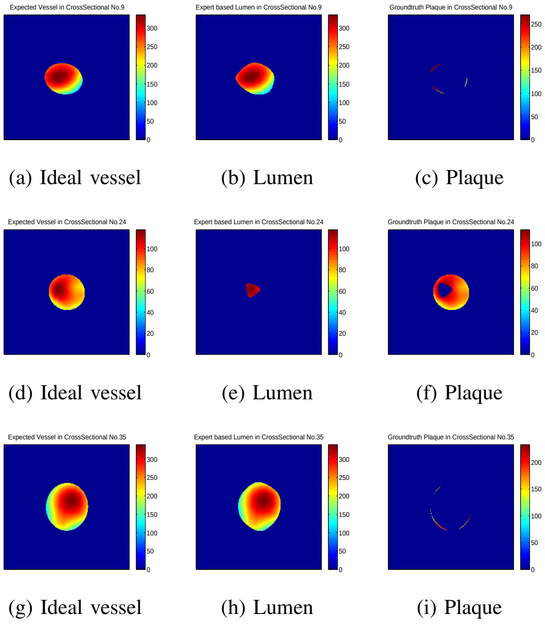
$$R3 = \begin{bmatrix} cos(\gamma) & sin(\gamma) & 0 \\ -sin(\gamma) & cos(\gamma) & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Accordingly, for an ellipse based modelling of the respective coronary segment, we approximated the manually annotated lumen boundaries (3D- contours) using best fitting ellipses on respective cross sections of the coronary segment using non-linear least square fitting. After obtaining the elliptical model $T_{model}\left[CP, \theta_{cs}^{'}\right]$ of the coronary segment, we used two "normal" ellipses adjacent to the lesion region i.e (immediately before and after the plaque region) to derive the parameters for ideal ellipse (plaque-free vessel) through the plaque affected region as illustrated in Fig. 3b. It should be noted that, in order to model the ideal plaque-free vessel at i^{th} point of the centreline, we employed the ellipse orientation information from the current fitted ellipse i.e. $R_{xyz}(i)$, whereas the major-minor axis lengths for ideal ellipse are derived from two "normal" ellipses E_s and E_e , which ensures that the 3D progression of vessel is tracked realistically.

$$E_{xyz}(i) = \{a(i) \quad b(i), \quad C_{xyz}(i), \quad R_{xyz}(i)\}$$

where $a(i)$ and $b(i)$ represent major-minor axes derived from two normal ellipses adjacent to the plaque region i.e. immediately before and after the plaque region.

After deriving the ideal ellipses for the plaque effected region, we subtracted the manually annotated lumen region which results in "reference" ground truth plaque voxels. The process of obtaining plaque ground truth is further illustrated in Fig. 4 where lumen boundary contours are used effectively in plaque identification process. It can be observed that due to the presence of a non-calcified plaque , the lumen shrinks in the proximal section and overcomes the diameter reduction as plaque region is passed. The left column of the figure represents the ideal vessel at respective cross sections of the segment, the middle column shows the manually annotated lumen and the right column represents the leftover to be interpreted as non-calcified plaque. It can be observed from the middle column that the lumen annotations are closely corroborating the plaque-free vessel for two normal contours (top and bottom row), whereas the lumen contour in middle row (plaque affected) appears significantly reduced. Likewise, the right column justifies that there exist a minimal plaque for two normal cross-sections, whereas the plaque effected cross-section results in a substantial amount of non-calcified plaque.

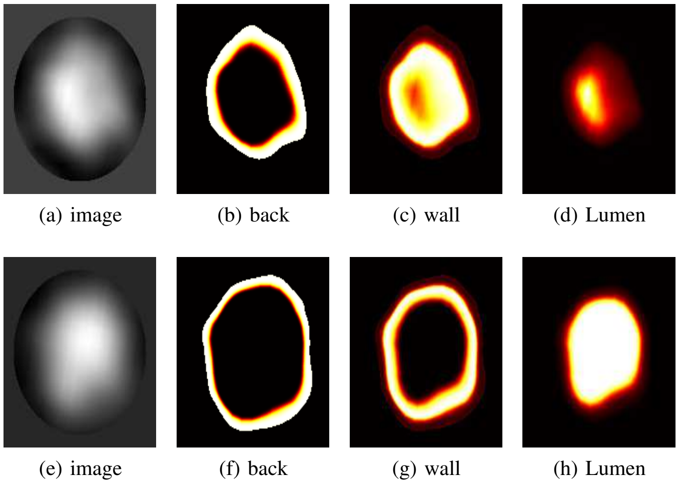

Expected Vessel in CrossSectional No.9
Expert based Lumen in CrossSectional No.9
Groundtruth Plaque in CrossSectional No.9
Expected Vessel in CrossSectional No.24
Expert based Lumen in CrossSectional No.24
Groundtruth Plaque in CrossSectional No.24
Expected Vessel in CrossSectional No.35
Expert based Lumen in CrossSectional No.35
Groundtruth Plaque in CrossSectional No.35

(a) Ideal vessel (b) Lumen (c) Plaque

(d) Ideal vessel (e) Lumen (f) Plaque

(g) Ideal vessel (h) Lumen (i) Plaque

Fig. 4. The ground truth estimation for plaque-effected cross-sections using lumen boundary annotations of manual expert. The Top and bottom rows show two normal slices at the start and end of the plaque region, whereas the middle row represents a severely effected plaque cross section. The first column shows ideal vessel, middle column shows the manually annotated lumen and right column shows derived plaque.



(a) image (b) back (c) wall (d) Lumen

(e) image (f) back (g) wall (h) Lumen

Fig. 5. Vessel wall analysis based on 3-class approximation of 6mm cylindrical model of DS4 seg1. First column shows 6mm region on the cross-sectional plane, second column represents the background of vessel that comes inside 6mm, next two columns shows the vessel wall and lumen respectively. First and third row represents two normal cross-sections, whereas the middle row represents an abnormal cross section.

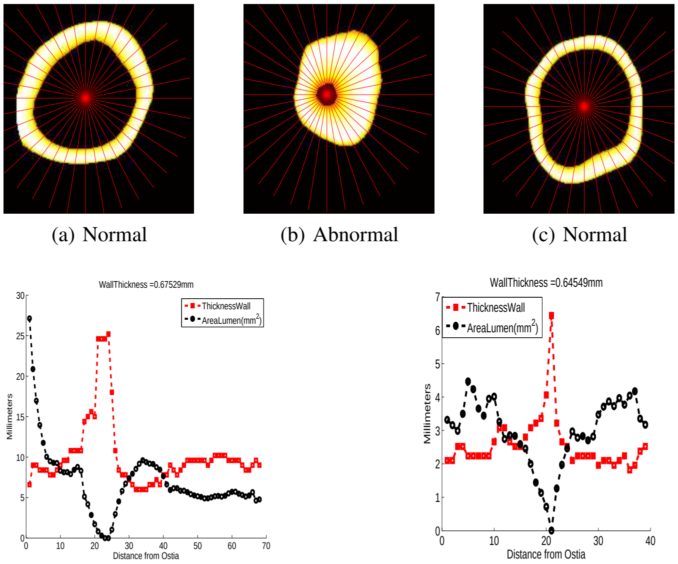
B. Vessel Wall Analysis

The non-calcified plaque quantification algorithm is based on the assumption that the input data (vessel) comprises of two components (i.e. blood lumen and the non-calcified plaque); however, the initial segmented tree violates this basic assumption. This is due to the fact that initially segmented tree includes the vessel wall, i.e. the interface of the lumen with the background in CTA imagery. Hence, the vessel wall must be identified and removed before applying the non-calcified plaque quantification algorithm. Accordingly, we started with the segmented coronary tree and computed the vessel wall thickness for normal segments in respective CTA_s. In the subsequent step, the vessel wall is removed using ray projection based thickness metric. In the final step, wall-removed coronary segments are evaluated for the lumen and non-calcified plaques.

The wall thickness computation process starts with the cylindrical model of Fig. 1, in which a coronary segment is approximated using 6 millimeters based cylindrical model for segment approximation. Based on the fact that 6-mm represents the maximum possible expansion of coronary vessel, the background data is often included in circular approximation. Accordingly, we used three class Gaussian Mixture Model (GMM), followed with the Bayesian Posterior's computation to classify the tubular segment voxels into three classes namely the background, vessel wall and the lumen as illustrated in Fig. 5.

Accordingly, it can be observed from the second column of the figure that background is generally well identified by "class-1" as first peak of the histogram corresponds to the low intensity regions that appears dark-black in the 6 mm circle of first column. Likewise, "class-2" defining vessel wall is represented in the middle column in which a ring pattern circumscribing the lumen can be clearly visualized. Class-3 representing lumen is shown in column 4 of the figure where a stable pattern can be observed for normal cross sections (top and bottom row) along the length of the segment. In case of plaque effected cross-section(middle row), the 3-class approximation reflects the abnormality in terms of violation of the normal patterns for both lumen and the vessel wall. The non-calcified plaque in general assumes intensity value comparatively lower than the blood lumen and close to the myocardial tissues. Hence, our 3-class approximation assigns the existing non-calcified plaque voxels to "class-2" i.e. the vessel wall. Consequently, the vessel wall shows unexpected increase in thickness for non-calcified plaque-effected sections with a significant reduction in lumen as illustrated in Fig. 5a - 5d.

After identifying the vessel wall, we employed ray-projection technique to compute the wall thickness for arterial cross section as illustrated in Fig. 6a - 6c. Based on the centre of the lumen, we projected a total of 36 rays outward with an angular interval of 10 degrees and computed ray-wise thickness of the vessel wall, which is averaged to obtain the wall thickness for respective cross section. This phenomena is further illustrated using wall thickness plots for two plaque affected segments as shown in Fig. 6d - 6e. It can be observed that for both segments, the lumen (black) starts with sharp decrement and becomes stable as we move away from the aorta. Similarly, the wall thickness (red) shows a stable thickness value for normal region of the segment. However, the plaque affected region shows unexpected reduction in lumen coupled with unexpected increase in the wall thickness. Once the mean wall thickness is computed for respective CTA volumes, the next step is to remove the wall of the segmented tree.



(a) Normal (b) Abnormal (c) Normal

(d) Lumen area versus wall thickness plot

(e) Lumen area versus wall thickness plot

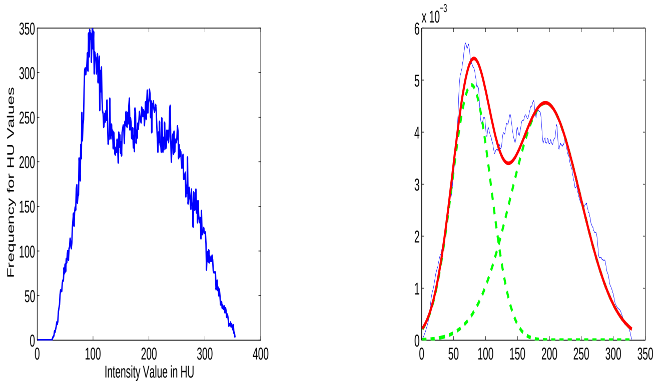
Fig. 6. Computation of the Vessel Wall thickness for coronary segment DS4 seg1. (a-c) shows the ray-projection to compute the mean thickness of the vessel wall, (d) represents the graphical comparison between lumen area and the normalized vessel wall thickness to reflect the anomalous lesion area. Cross sectional representing normal segment (a and c) leads to stable vessel wall, whereas abnormal cross section leads to expansion of the vessel wall based on low density soft plaques.

IV. Pixel-Based Segmentation

After removing the vessel wall from the segmented tree, it is expected that the leftover is true lumen and the non-calcified plaque (if any). Accordingly, we derive hand crafted discriminative features capable of differentiating voxels into lumen or non-calcified plaque. For voxel-wise discriminative features, we employed the spatial neighbourhood information, optimized 2-class GMM based posteriors, signed distance function, distance from the arterial orifice, pixel distance from the medial axis and histogram based fuzzy label as explained in this section.

V. 2-Class Posteriors

It is notable that the non-calcified plaques present inside coronary vasculature do not follow any particular shape or structure; hence, the use of shape-prior information is not very effective in the problem domain. Consequently, the extensively investigated feature in context of non-calcified plaque segmentation is the intensity distribution in the vessel, as the plaque region undergoes an unexpected intensity drop relative to the normal blood HU distribution. Accordingly, we computed the intensity histogram for the plaque affected region with an expectation of two peaks representing the plaque and lumen respectively, as illustrated in Fig. 7a.

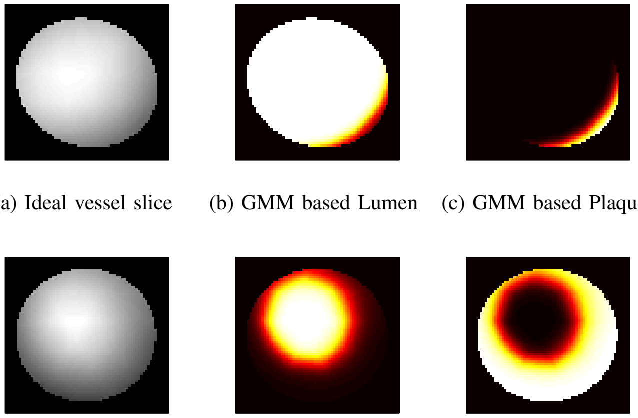


(a) Intensity histogram (b) 2-class GMM

Fig. 7. 2-class approximation for the plaque effected section of the coronary segment DS4 seg1. (a-b) shows the plaque effected boundary and respective bi-modal intensity histogram, (c) represents 2-class Gaussian Mixture Model and respective HU intensity peaks.

Next, the bi-modal intensity histogram of the plaque affected section is approximated using 2-class Gaussian Mixture Model, followed with the application of expectation maximization (EM) algorithm for optimal representation of two classes. Fig. 7b shows GMM approximation, with first class defining low density non-calcified plaque and the second class representing high intensity blood lumen.

After obtaining EM based optimal distribution parameters, we used Bayesian modelling approach to compute the posterior probabilities for two classes respectively as represented in Fig. 8.



(a) Ideal vessel slice (b) GMM based Lumen (c) GMM based Plaque

(d) Ideal vessel slice (e) GMM based lumen (f) GMM based plaque

Fig. 8. 2-class approximation based initial estimation for lumen and plaque. Top row represents a normal cross-section, i.e. at the immediate start of the non-calcified plaque region, and second row represents cross-section in the mid of plaque region. Left column shows a 2D intensity based cross-section of the coronary vessel, whereas middle and right columns respectively shows the 2-class GMM based lumen and non-calcified plaque.

The left column represents the cross sectional view for an ideal vessel (plaque free vessel), the middle and right columns represents 2-class GMM based lumen and the plaque voxels, respectively. It can be observed that top row (start of the plaque) shows the 2-class lumen much close to the ideal vessel with a minimal plaque, however second row reflecting the mid of the plaque region shows significantly reduced lumen along with an expended plaque. Moreover, the relative position of the lumen and plaque validates the clinical fact that non-calcified plaque generally sticks with the vessel walls leading to Napkin ring signs [33].

From a statistical point of view, the computed plaque is compared against expert-based manual demarcations.Accordingly,we employed metrics true positive (TP), true negative (TN), false positive (FP), false negative(FN), respectively. True positive refers the case when expert-based lumen voxel is identified as lumen by posterior method. Similarly, True negative refers the case when expert-based plaque voxel is identified as plaque by the posterior method. In contrast, false positive and false negative defines two cases for mismatch among the expert-based lumen/plaque demarcation and the output of the posterior method. Using individual metrics, we computed Jaccard similarity index as follows.

$$Jaccardindex = \frac{TP}{(TP + FP + FN)}$$

It is important to mention that for an ideal plaque quantification, the Jaccard index approaches to one, whereas two dissimilar annotations result in Jaccard score of $zero$. According to the experimental results, mean Jaccard overlap is around 88% with respect to the manual expert. In terms of sensitivity, specificity and accuracy, the proposed methods achieves 84.13% ,79.15% and 82.02% success.

VI. Conclusion

In this work, we proposed a method for voxel-wise quantification of coronary non-calcified plaque using Bayesian Posterior probability model. Based on the normal coronary segments, we computed the vessel-wall thickness in first step. In the subsequent step, we removed vessel wall from the segmented tree and employed Gaussian Mixture Model to compute intensity based clusters. According to the experimental results, it is shown that the automated plaque segmentation method achieves accuracy equivalent to human experts. We aim to extend this work in future in context of deep learning based solutions for the said problem. Application of convolutional neural network (CNN) shows promising results in recent years; however, this requires a bulk amount of data for adequate training.

Acknowledgment

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

References

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

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

[3] T. Flohr, B. Ohnesorge, Multi-slice ct technology, in: Multi-slice and Dualsource CT in Cardiac Imaging, Springer, 2007, pp. 41-69.

[4] R. Virmani, A. P. Burke, A. Farb, F. D. Kolodgie, Pathology of the vulnerable plaque, Journal of the American College of Cardiology 47 (8s1) (2006) C13-C18.

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

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

[7] I. Isgum, A. Rutten, M. Prokop, B. van Ginneken, Detection of coronary calcifications from computed tomography scans for automated risk assessment of coronary artery disease, Medical physics 34 (4) (2007) 1450-1461.

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

[9] M. E. Clouse, A. Sabir, C.-S. Yam, N. Yoshimura, S. Lin, F. Welty, P. Martinez-Clark, V. Raptopoulos, Measuring noncalcified coronary atherosclerotic plaque using voxel analysis with mdct angiography: a pilot clinical study, American Journal of Roentgenology 190 (6) (2008) 1553- 1560.

[10] J. Wei, C. Zhou, H.-P. Chan, A. Chughtai, P. Agarwal, J. Kuriakose, L. Hadjiiski, S. Patel, E. Kazerooni, Computerized detection of noncalcified plaques in coronary ct angiography: Evaluation of topological soft gradient prescreening method and luminal analysis, Medical physics 41 (8) (2014) 081901.

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

[12]] M. Tessmann, F. Vega-Higuera, D. Fritz, M. Scheuering, G. Greiner, Multiscale feature extraction for learning-based classification of coronary artery stenosis, in: SPIE Medical Imaging, International Society for Optics and Photonics, 2009, pp. 726002-726002.

[13] M. A. Zuluaga, I. E. Magnin, M. H. Hoyos, E. J. D. Leyton, F. Lozano, M. Orkisz, Automatic detection of abnormal vascular cross-sections based on density level detection and support vector machines, International Journal of Computer Assisted Radiology and Surgery 6 (2) (2011) 163-174.

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

[15] Y. Li, W. Chen, K. Liu, Y.Wu, Y. Chen, C. Chu, B. Fang, L. Tan, S. Zhang, A voxel-map quantitative analysis approach for atherosclerotic noncalcified plaques of the coronary artery tree, Computational and mathematical methods in medicine 2013.

[16] S. Achenbach, F. Moselewski, D. Ropers, M. Ferencik, U. Homann, B. MacNeill, K. Pohle, U. Baum, K. Anders, I.-k. Jang, et al., Detection of calcified and noncalcified coronary atherosclerotic plaque by contrastenhanced, submillimeter multidetector spiral computed tomography, Circulation, 109 (1) (2004) 14-17.

[17] A. W. Leber, A. Becker, A. Knez, F. von Ziegler, M. Sirol, K. Nikolaou, B. Ohnesorge, Z. A. Fayad, C. R. Becker, M. Reiser, et al., Accuracy of 64-slice computed tomography to classify and quantify plaque volumes in the proximal coronary system: a comparative study using intravascular ultrasound, Journal of the American College of Cardiology 47 (3) (2006) 672-677.

[18] T. Schepis, M. Marwan, T. Pederer, M. Seltmann, D. Ropers, W. G. Daniel, S. Achenbach, Quantification of non-calcified coronary atherosclerotic plaques with dual-source computed tomography: comparison with intravascular ultrasound, Heart 96 (8) (2010) 610-615.

[19] Z. Sun, L. Xu, Coronary ct angiography in the quantitative assessment of coronary plaques, BioMed research international 2014.

[20] P. Schoenhagen, E. M. Tuzcu, A. E. Stillman, D. J. Moliterno, S. S. Hal liburton, S. A. Kuzmiak, J. M. Kasper, W. A. Magyar, M. L. Lieber, S. E. Nissen, et al., Non-invasive assessment of plaque morphology and remodeling in mildly stenotic coronary segments: comparison of 16-slice computed tomography and intravascular ultrasound, Coronary artery disease 14 (6) (2003) 459-462.

[21] S. Schroeder, A. F. Kopp, A. Baumbach, C. Meisner, A. Kuettner, C. Georg, B. Ohnesorge, C. Herdeg, C. D. Claussen, K. R. Karsch, Noninvasive detection and evaluation of atherosclerotic coronary plaques with multislice computed tomography, Journal of the American College of Cardiology 37 (5) (2001) 1430-1435.

[22]] H. Brodoefel, C. Burgstahler, M. Heuschmid, A. Reimann, F. Khosa, A. Kopp, S. Schroeder, C. Claussen, M. Clouse, Accuracy of dual-source ct in the characterisation of non-calcified plaque: use of a colour-coded analysis compared with virtual histology intravascular ultrasound, The British journal of radiology 82 (982) (2009) 805-812.

[23] D. Dey, T. Schepis, M. Marwan, P. J. Slomka, D. S. Berman, S. Achenbach, Automated three-dimensional quantification of noncalcified coronary plaque from coronary CT angiography: comparison with intravascular us, Radiology 257 (2) (2010) 516-522.

[24] W. Theo, The Great Challenge,, coronary artery stenoses detection and quantification evaluation framework (2016). URL http://www.http://coronary.bigr.nl/stenoses/

[25] H. Kirisli, M. Schaap, C. Metz, A. Dharampal, W. B. Meijboom, S.-L. Papadopoulou, A. Dedic, K. Nieman, M. De Graaf, M. Meijs, et al., Standard ized evaluation framework for evaluating coronary artery stenosis detection, stenosis quantification and lumen segmentation algorithms in computed tomography angiography, Medical image analysis 17 (8) (2013) 859-876.

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

[27]] M. M. Jawaid, A. Riaz, R. Rajani, C. C. Reyes-Aldasoro, G. Slabaugh, Framework for detection and localization of coronary non-calcified plaques in cardiac cta using mean radial profiles, Computers in Biology and 385 Medicine 89 (C) (2017) 84-95.

[28]] W. C. Wong, A. C. Chung, Augmented vessels for quantitative analysis of vascular abnormalities and endovascular treatment planning, IEEE transactions on medical imaging 25 (6) (2006) 665-684.

[29] D.G. Kang, D. C. Suh, J. B. Ra, Three-dimensional blood vessel quantifi cation via centerline deformation, IEEE Transactions on Medical Imaging 28 (3) (2009) 405-414.

[30] A. F. Frangi, W. J. Niessen, R. M. Hoogeveen, T. Van Walsum, M. A. Viergever, Model-based quantitation of 3-d magnetic resonance angiographic images, IEEE Transactions on medical imaging 18 (10) (1999) 946-956.

[31] A. F. Frangi, W. J. Niessen, P. J. Nederkoorn, J. Bakker, W. P. T. M. Mali, M. A. Viergever, Quantitative analysis of vascular morphology from 3d MR angiograms: in vitro and in vivo results, Magnetic Resonance in Medicine 45 (2) (2001) 311-322.

[32]] P. J. Yim, J. J. Cebral, R. Mullick, H. B. Marcos, P. L. Choyke, Vessel surface reconstruction with a tubular deformable model, IEEE transactions on medical imaging 20 (12) (2001) 1411-1421.

[33] N. R. Pal, S. K. Pal, A review on image segmentation techniques, Pattern recognition 26 (9) (1993) 1277-1294.

[34] O. MastO, B.Nico *et al.*,Quantification of coronary plaque by 64-slice computed tomography: a comparison with quantitative intracoronary ultrasound, Investigative radiology, 43(9), (2008), 314-321.

[35] L. Athanasiou, G. Rigas, A.I. Sakellarios, T.P.Exarchos *et al.*, Three-dimensional reconstruction of coronary arteries and plaque morphology using CT angiographycomparison and registration with IVUS, BMC medical imaging, 16(1), (2016), p.9.

[36] T. Pflederer, M. Schmid, D. Ropers, U. Ropers, S. Komatsu, W. G. Daniel, and S. Achenbach, Interobserver variability of 64-slice computed tomography for the quantification of non-calcified coronary atherosclerotic plaque. RFo-Fortschritte auf dem Gebiet der Rontgenstrahlen und der bildgebenden Verfahren 179(9), 953-957.

[37] C. L. Christopher *et al.*, How to assess non-calcified plaque in CT angiography: delineation methods affect diagnostic accuracy of low-attenuation plaque by CT for lipid-core plaque in histology, European Heart JournalCardiovascular Imaging 14(11), (2013): 1099-1105.

[38] D. Damini, *et al.*, Automated 3-dimensional quantification of noncalcified and calcified coronary plaque from coronary CT angiography, Journal of cardiovascular computed tomography 3(6), (2009), 372-382.

[39] H. Brodoefel, C. Burgstahler, A. Sabir, C.S.Yam, F. Khosa, *et al*, Coronary plaque quantification by voxel analysis: dual-source MDCT angiography versus intravascular sonography. American Journal of Roentgenology, 192(3), (2009), W84-W89.

Wakes-Ship Removal on High-Resolution Optical Images based on Histograms in HSV Color Space

Fidel Indalecio Mamani Maquera, Eveling Gloria Castro Gutierrez
Universidad Nacional de San Agustín de Arequipa
CiTeSoft
Arequipa, Perú

***Abstract*—Ship detection on optical remote sensing images is getting great attention; however, some images called wakes-ship have not been taken into account yet. Current works in ship detection are focusing on in-shore detection where ships are in calm; furthermore, their methods get high Intersection Over Union (IoU), above 70%, but when computing IoU using only wakes-ship images the ratio is 22%. In this paper, it is presented a new framework to improve ship segmentation on wakes-ship images. In order to achieve this goal, it was considered HSV color space and histograms. First, ship detection was done using state-of-the-art ship detection methods. Second, bin histograms in HSV color space was used to get the colors that rely on wakes. Finally, the removal of wakes from ships was done using some discriminative properties. In this way, the increase of the IoU performance at wake-ship segmentation goes from 22% to 63%, which is an improvement of 186%.**

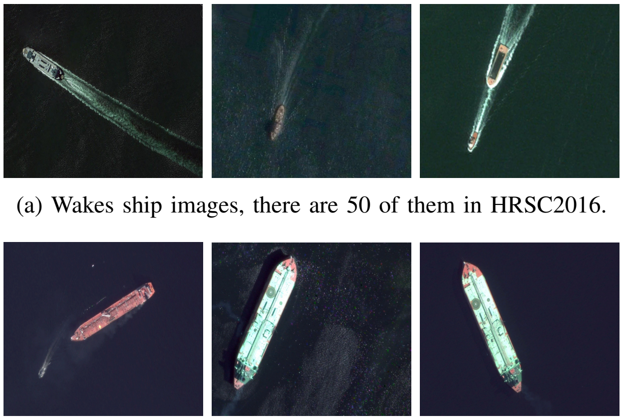
***Keywords*—Wakes-ship removal; optical remote sensing; ship detection; HSV color space; histograms; intersection over union**

I. Introduction

Ship detection on optical satellite images is attracting great interest with the growing use of optical remote sensing images in recent years [1] and because of its importance in maritime security, fishery management and other applications [1]. Most of the state-of-the-art works in ship detection are using Convolutional Neural Network (CNN) as the benchmark [2]–[5] and they are getting high detection ratio, results are above 85% in precision and recall.

There are two kinds of images used in ship detection task, one is off-shore as shown in Fig. 1, where ship detection task appears to be easy and the other is in-shore, Fig. 2, which shows great difficulty when it comes to ship detection because of the land, harbor, and other issues that may happen in optical remote sensing images [1], [6]. Most of the works are focused on in-shore detection since it represents a big challenge.

The tests were run on HRSC2016[1] dataset released by (Lui's et al.,2016) [6], it contains 1680 high-resolution optical images of ships in in-shore and off-shore collected from different sources. Due to the difficult task in in-shore ship detection, most of the images used on current works are in-shore, where many ships are in calm and lying next to the harbor as shown in Fig. 2.



(a) Wakes ship images, there are 50 of them in HRSC2016.

(b) Static-ship images, there are 192 of them in HRSC2016.

Fig. 1: There are 242 images in off-shore in HRSC2016. (a) Wakes-ship images, there are 50 of them. (b) Static-ship images, there are 192 of them.

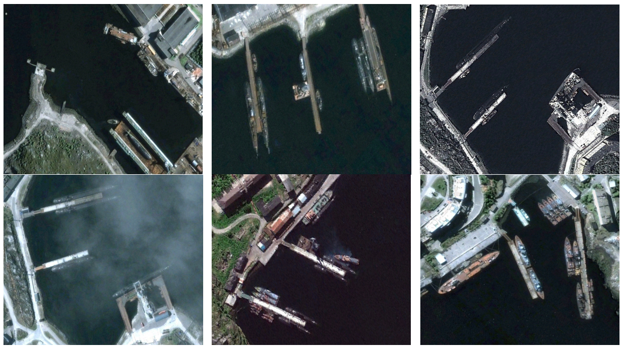


Fig. 2: 1438 in-shore images out of 1680 from HRSC2016. current ship detection works focus on these images since it represents a big challenge.

In this paper, it is wanted to address the problem with off-shore images, Fig. 1. There are two kinds of images in off-shore, images where ships are moving and generating wakes around them, Fig. 1a, in this paper they are called as wakes-ship images and the others are static-ship images, Fig. 1b, where ships are not moving, thus, they are not generating wakes. These two kinds of images could be seen in Fig. 1.

While ship detection in static-ship images, Fig. 1b, appears to be easy. However, it is difficult to detect ships when the sea part has noise. On the other hand, there are two problems that have not been taken into account yet on wakes-ship images,

[1] HRSC2016 is a dataset that contains images about ships, it can be downloaded from http://www.escience.cn/people/liuzikun/DataSet.html

Fig. 1a. First, most of the methods on these images detect the ship, but including the wakes, thus, that means that they get low Intersection Over Union (IoU[2]), and because of the vast use of in-shore images where static-ships presence is dominant, they get high IoU. To show this behavior, tests were run using current ship detection methods on static-ship images and the results have shown that they got high IoU. Getting high IoU is important because it allows to get the right measurements from ships such as length and width and consequently it helps to ship classification task [5], [7], [8].

A research [9] approached this problem using clustering k-means and watershed segmentation, they tested their method in a small dataset of 54 images, also they point it out how important is the segmentation in this kind of images. Another recent research [10] remarks the importance of segmentation on ship-wakes images, they used CIELAB color space to segment ship from wakes as a preliminary study. Both works were done using remote sensing images, but with low resolution.

To tackle the problem with wakes-ship images, in this paper, it was used bin histograms generation in HSV[3] color space to remove the wakes from ships, but before that, a ship detection method was needed because it allowed knowing whether there are ships with wakes on the image or not. Two ship-detection methods and Otsu method were used to deal with ship detection as described in Section II-B, consequently, they were compared each other at IoU on wakes-ship images. So, there are two stages in the proposed framework in this paper, at first stage, it was used a ship detection method, it allowed knowing that there exist a ship on the image. The second stage, once the segmented image has been taken with only wakes and ship, the removal of wakes was done by changing the image into HSV color space and by generating HSV histograms on this color space to evaluate the colors that rely on the wakes. Fig. 3 shows the images that were taken in the entire process of our framework.

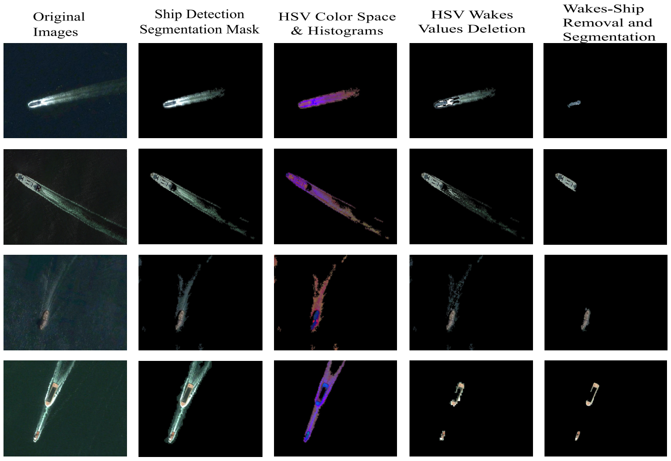

Original Images
Ship Detection Segmentation Mask
HSV Color Space & Histograms
HSV Wakes Values Deletion
Wakes-Ship Removal and Segmentation

Fig. 3: Images obtained at every step of the entire process of the framework, from original wakes-ship images until the wakes-ship removal and ship segmentation.

The remainder of this paper is organized as follows. Section 2 explains the data-set and ship detection methods evaluated and used in this paper. The proposed method presented in this paper is described in detail in Section 3. Section 4, discussion of the results and finally, in Section 5 conclusions and future works are explained.

II. HRSC2016 Dataset and Ship Detection Methods

This section is covered. First, the HRSC2016 dataset is described, furthermore, the kind of images that have been taken into account in this paper, then it is described two ship detection methods used in this paper and the evaluation of them when applied to wakes-ship images.

A. HRSC2016 Dataset

The HRSC2016 dataset [6] contains images from two scenarios, ships far from the harbor and ships next to the harbor, they are typically called off-shore and in-shore images, respectively. All of the images were collected from famous harbors. The resolution of the images is between 2-m and 0.4-m.

From these two groups of images; most of the images are in-shore, a small number of images are off-shore, (Liu et al.) explains it is because of the big challenge that in-shore ship detection demands. From the total 242 off-shore found images, it has been selected 50 wakes-ship images, which means, ships are moving and generating wakes around them, like in Fig. 1, the rest of the images the ships are next to the harbor in calm and without wakes. In summary, from the 242 off-shore images, there have been taken 50 wakes-ship images and 192 static-ship images. They were used both off-shore images to test the accuracy of IoU of ship detection methods.

B. Ship Detection Methods

It was used OTSU method as a baseline and two ship detection methods [11], [12] to compare, these methods were implemented and tested in Liu et al. [6], [13], both of them focuses on off-shore ship detection, so they were not prepared for wakes-ship images.

Tang's method [12] is based on sea-land segmentation with some improvements, they include ship location, feature representation, and classification, but only ship location criteria was used in this paper. The other method, Liu's ship extraction method [11] is based on "V" shape ship-head detection and it is well suited for high-resolution optical images.

To know the accuracy of these segmentation methods, the metric Intersection Over Union (IoU) used in PASCAL VOC 2007 challenge was applied, which is described as follows, S represents the bounding-box area of the original image and S' represents the bounding-box area of the proposed method. When this ratio is above 50%, it means a good detection rate.

$$IoU = \frac{S \cap S'}{S \cup S'} \tag{1}$$

After running tests on wakes-ship images using these ship detection methods presented previously, it has been seen that these methods performed with a high Intersection over

[2] IoU: Intersection over Union, it measures the accuracy for object detection.
[3] HSV: Color space with three parameters Hue, Saturation and Value.

Union(IoU) ratio on the 192 static-ship images, between 62% and 73% as shown in Table I. However, when applied only using wakes-ships images, the result of the Intersection over Union (IoU) ratio was between 22% and 32% as shown in Table II. These results show that segmentation on static-ship images appears to be easy; however when static-ship images present noises, false alarm removal is required. While in wakes-ship images the methods do not perform well since the wakes generated by ships have too much noise, furthermore, IoU ratio is lower because of the presence of wakes around the ship.

TABLE I: Comparison Intersection Over Union (IoU) among Ship-Detection Methods on the 192 Static-Ship Images in Off-Shore

Ship Detection Method	Intersection over Union(IoU)
Liu's method	0.731
Tang's method	0.691
Otsu's method	0.646

TABLE II: Comparison Intersection Over Union (IoU) among Ship-Detection Methods on the 50 Wakes-Ship Images in Off-Shore

Ship Detection Method	Intersection over Union(IoU)
Liu's method	0.259
Tang's method	0.222
Otsu's method	0.323

III. Methodology

In this paper, it is proposed a new framework for wakes-ship removal of off-shore images on high-resolution optical images as shown in Fig. 4. By going throughout this framework, the removal of wakes from ships can be done, thus, to get the right measurements from ships, such as length and width could be used for ship classification task. Classification of ships is beyond the needs and scope of this paper.

The proposed framework is divided into two stages. First, ship-detection segmentation is the stage where the ship with wakes is taken from original images; current ship detection methods aforementioned were used at this stage, during the test results applying Tang's method [14] excelled, it could be seen in Fig. 5 along with other methods used. The second stage, the wakes-ship removal task is done, where the use of the segmented image performed previously takes place, by converting such image in HSV color space. To deal with colors that rely on wakes, the generation of histograms in HSV color is performed, such histograms clearly indicate the wakes color values, next the separation of wakes from ship is done by deleting the values obtained from histograms and consequently generating a small group of pixels. Finally, false alarms are removed to get the ships. The next subsections explain in detail every single step of the second stage of our framework, Fig. 4.

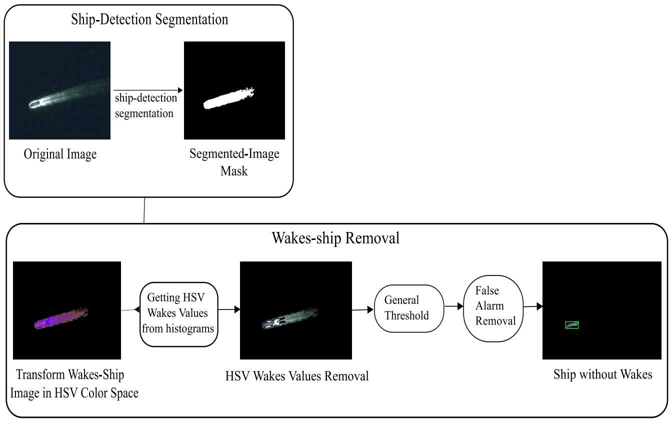

Ship-Detection Segmentation
ship-detection segmentation
Original Image
Segmented-Image Mask
Wakes-ship Removal
Getting HSV Wakes Values from histograms
General Threshold
False Alarm Removal
Transform Wakes-Ship Image in HSV Color Space
HSV Wakes Values Removal
Ship without Wakes

Fig. 4: The methodology presented in this paper with two stages. Ship-detection Segmentation, where the ship detection task is done and Wakes-ship Removal, where the removal of wakes is performed.

A. *Wakes-Ship Segmentation Image and HSV Color Space Conversion*

First, getting the wakes-ship segmented image was needed; most of the methods for ship detection provide one, they detect the ships, but including the wakes as shown in Fig. 5. It is necessary to remark, these methods are not prepared for wakes-ship images since they wanted to detect ships in in-shore, but according to the tests carried out they still detect the ship in off-shore. Liu's method [11], Tang's method [12] and Otsu's method as a baseline were used. From these methods, the proposed framework outperformed with Tang's method since it provides the mask of the complete ship and wakes, in addition, it allows getting a better representation of HSV histograms of HSV wakes colors.

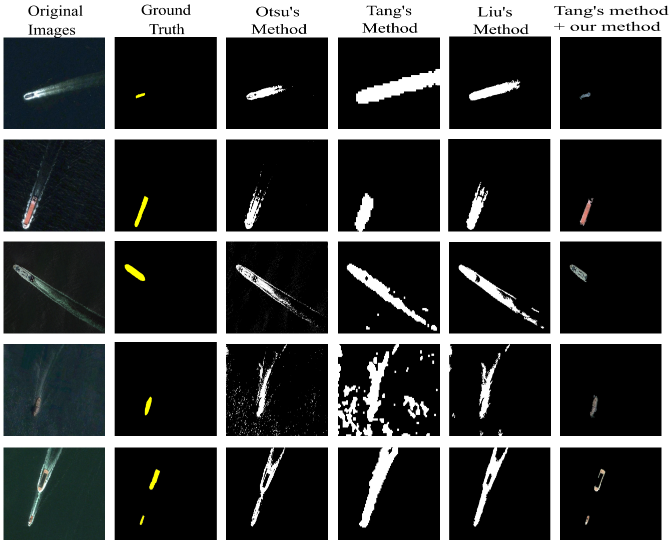


Fig. 5: Several ship detection methods detect the ship, but including the wakes around them. So, a removal of them is needed.

Once the mask for ship and wakes were obtained, conversion of the image in HSV color space using the method proposed in [14], [15] is performed. HSV color space has been largely used for image segmentation and image retrieval [14], [15]. H(Hue) is defined as the angle in the range [0, 2π] starting at red axis with red at 0, green at $2\pi/3$, blue at $4\pi/3$ and red at 2π again, the transformation of these angles to values between 0-1 was done. S(Saturation) is measured as a radial distance from the central axis with a value between 0 and 1, it defines the brilliance and intensity of a color and V(value) defines the lightness or darkness of a color, it goes from 0 to 1.

From these three values, after running some tests, H(hue) and V(value) were the only values that clearly indicated the presence of the colors that rely on wakes. Unlike S(saturation), even its value was ranged from 0 to ∞ it did not indicate big changes and did not cause many troubles. So H(hue) and V(value) values were used to generate histograms as seen in Fig. 6 and 7.

B. Getting HSV Wakes Color Values from Histograms and Wakes Removal

Second, to get the colors that rely on wakes, the differentiation of the wakes colors from ship colors was done by generating histograms in HSV color space for H(hue) and V(value) as shown in Fig. 6. Tang's method excelled combined with the proposed framework because it segments both ships and wakes without damaging the integrity of the ship and taking a good proportion of wakes. Although Liu's method has better detection ratio and performs well in static-ship images, it deletes most of the wakes around the ship and part of the ship in wakes-ship images, so the HSV values obtained are hard to normalize since most of the wakes pixels were removed from the image.

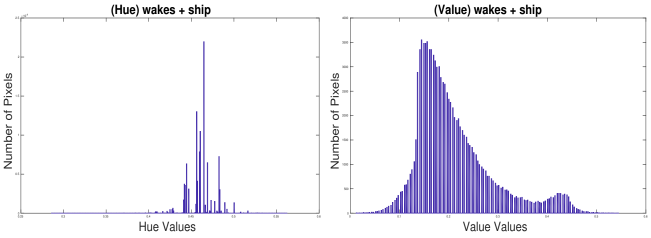


Fig. 6: Histograms in Hue and Value color space without normalization.

In order to get the right H(hue) and V(value) wakes values, a normalization of these values was needed and also adding an extra tolerance for each value when deleting the pixels from wakes-ship images. So the chosen number of bins for H(hue) was set 16 as suggested in [14]. Saturation values did not cause many problems on the tests as explained before, even when its value was changed between 0 and ∞, thus, it was not taken into account. V(value) value was set to 18 bins empirically, Fig. 7 shows these histograms.

Once the values were normalized, the calculation of the HSV wakes values to delete them from the image was carried out. In order to accomplish that, setting the HSV values $I_{from(h,s,v)}$, from which starting off deleting the pixels until reaching $I_{to(h,s,v)}$ HSV values. This range is calculated for each image as it is written below.

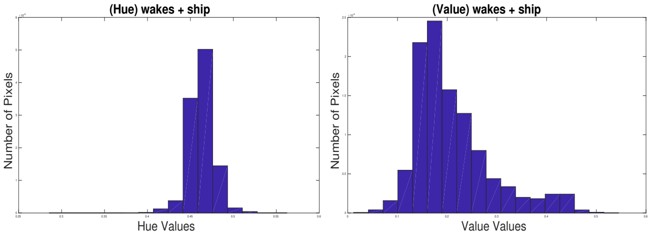


Fig. 7: Histograms in Hue and Value for Hue with 16 bins, and Value with 18 bins. The mode in each figure represents the Hue and Value for wakes values.

$$I = I_{from(h,s,v)}, I_{to(h,s,v)} \tag{2}$$

$$I_{from(h,s,v)} = \begin{cases} Mode(H)_{binIndex} - 1 & \text{for} \quad I_{from(h)} \\ 0 & \text{for} \quad I_{from(s)} \\ Mode(V)_{binIndex} - 1 & \text{for} \quad I_{from(v)} \end{cases} \tag{3}$$

$$I_{to(h,s,v)} = \begin{cases} Mode(H)_{binIndex} + 2 & \text{for} \quad I_{to(h)} \\ \infty & \text{for} \quad I_{to(s)} \\ Mode(V)_{binIndex} + 3 & \text{for} \quad I_{to(v)} \end{cases} \tag{4}$$

C. General Thresholding and False Alarm Removal

Third, after the removal of the HSV wakes pixels, the pixels left form multiple small connecting regions. General thresholding was applied to the entire image, then 4-connected neighborhood components were used to get all the regions left. Finally to remove the false alarms that may exist it has been used three discriminative properties, area, compactness and length-width as in [11], [16].

1) Area: on the tests, it has been noticed that just calculating the largest area among all the regions on the 50 wakes-ship images it has been gotten 96% accuracy, which means the elimination of almost all the wakes were done by breaking them down.
2) Compactness : It measures the circular similarity degree, and it is as follows.

$$Compactness = \frac{P^2}{A} \tag{5}$$

Where P represents the perimeter and A represents the Area.
3) Length-Width : Due to the most of the ships are long and thin, this simple ratio can eliminate false alarms.

$$LengthWidth = \frac{L}{W} \tag{6}$$

Where L is the length of the bounding-box and W is the width.

IV. Evaluation

The proposed method was tested on 50 wakes-ship images from HRSC2016 dataset, the results are shown in Table III, which shows that ship detection methods on wakes-ship images get better performance when combined with our method. Table III shows the importance of taking into account the wakes around ships since the more wakes pixels it has, the more HSV representative wakes values it gets and it is easy to normalize their values. Liu's method did not excel in Table III as it did in Table II because it was prepared to work on high-resolution images, furthermore, it deletes part of the wakes from images, which difficulties the task of normalizing since the HSV wakes values are almost the same as HSV ship values.

TABLE III: Improvement for Ship Detection Methods, Getting Higher IoU by Removing Wakes from Images

Ship Detection Method + Our Method	Intersection over Union(IoU)
Liu + our method	0.563
Tang + our method	0.626
Otsu + our method	0.456

V. Conclusion and Future Works

In this paper, it has been proposed a new framework for wakes-ship removal by generating histograms in HSV color space. It has been used current ship detection methods. By doing this, the only task that remains is the removal of wakes from ships. In order to do that a complete analysis of the colors that rely on wakes was carried out by generating histograms of HSV values to get representative wakes HSV value colors. The elimination of the wakes was done by eliminating pixels values near the mode(Mo) in histograms, next the elimination of false alarms and the extraction of the ship was done by using some discriminative properties.

The results show that the average success rate of Intersection over Union (IoU) goes from 22% to 63% which is an improvement of 186%, IoU above 50% means a good detection rate. It is also important to mention that on some images where the integrity of the ship has been affected by using ship detection methods, it was impossible to recover the pixels. Another observation, when applied our framework to the image without a ship detection method, the HSV values represent almost all the sea values, so it is hard to get the right wakes values, thus, it is important to know whether the image contains a ship with wakes or not.

In order to get a better representative HSV values from wakes, it is necessary to collect more wakes-ship images, in that way, a supervised learning model for wakes-ship removal in HSV color space could improve the results for others ship detection methods.

Acknowledgment

This research is part of the Project "Detección de Embarcaciones Pesqueras Industriales dentro de las 05 Millas en la región Arequipa Utilizando Computación de Alto Desempeño e Imágenes Satelitales IBA-018737-2016, Resolución Vicerrectoral N°030-2016-V.R.INV.", which is supported by Universidad Nacional de San Agustín de Arequipa. We would like to thank CiTeSoft for providing us materials, books, tools and computers to test the proposed framework for wakes-ship removal elaborated in this paper.

References

[1] U. Kanjir, H. Greidanus, and K. Oštir, "Vessel detection and classification from spaceborne optical images: A literature survey," *Remote Sensing of Environment*, vol. 207, pp. 1–26, 2018.

[2] R. Zhang, J. Yao, K. Zhang, C. Feng, and J. Zhang, "S-cnn-based ship detection from high-resolution remote sensing images." *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 41, 2016.

[3] Z. Zou and Z. Shi, "Ship detection in spaceborne optical image with svd networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 5832–5845, 2016.

[4] Z. Liu, J. Hu, L. Weng, and Y. Yang, "Rotated region based cnn for ship detection," in *2017 IEEE International Conference on Image Processing. Piscataway, NJ: IEEE*, 2017.

[5] K. Raineya, J. D. Reedera, and A. G. Corellia, "Convolutional neural networks for ship type recognition," in *Proc. of SPIE Vol*, vol. 9844, 2016, pp. 984 409–1.

[6] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines." in *ICPRAM*, 2017, pp. 324–331.

[7] Y. Liu, H.-Y. Cui, Z. Kuang, and G.-Q. Li, "Ship detection and classification on optical remote sensing images using deep learning," in *ITM Web of Conferences*, vol. 12. EDP Sciences, 2017, p. 05012.

[8] S. Li, Z. Zhou, B. Wang, and F. Wu, "A novel inshore ship detection via ship head classification and body boundary determination," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1920–1924, 2016.

[9] H. Bouma, R. J. Dekker, R. M. Schoemaker, and A. A. Mohamoud, "Segmentation and wake removal of seafaring vessels in optical satellite images," in *Electro-Optical Remote Sensing, Photonic Technologies, and Applications VII; and Military Applications in Hyperspectral Imaging and High Spatial Resolution Sensing*, vol. 8897. International Society for Optics and Photonics, 2013, p. 88970B.

[10] A. Klimkowska and I. Lee, "a prealiminary study of ship detection from uav images based on color space conversion and image segmentation," *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 189–193, 2017.

[11] G. Liu, Y. Zhang, X. Zheng, X. Sun, K. Fu, and H. Wang, "A new method on inshore ship detection in high-resolution satellite images using shape and context information," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 3, pp. 617–621, 2014.

[12] J. Tang, C. Deng, G.-B. Huang, and B. Zhao, "Compressed-domain ship detection on spaceborne optical image using deep neural network and extreme learning machine," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 3, pp. 1174–1185, 2015.

[13] Z. Liu, H. Wang, L. Weng, and Y. Yang, "Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 8, pp. 1074–1078, 2016.

[14] S. Sural, G. Qian, and S. Pramanik, "Segmentation and histogram generation using the hsv color space for image retrieval," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 2. IEEE, 2002, pp. II–II.

[15] R. Tamilkodi, R. Karthika, G. RoslineNesaKumari, and S. Maruthuperumal, "Segment based image retrieval using hsv color space and moment," in *Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1*. Springer, 2015, pp. 239–247.

[16] F. Yang, Q. Xu, and B. Li, "Ship detection from optical satellite images based on saliency segmentation and structure-lbp feature," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 602–606, 2017.

Automatic Cyberbullying Detection in Spanish-language Social Networks using Sentiment Analysis Techniques

Rolfy Nixon Montufar Mercado, Hernan Faustino Chacca Chuctaya, Eveling Gloria Castro Gutierrez
National University of San Agustin
CiTeSoft
Arequipa-Peru

***Abstract*—Cyberbullying is a growing problem in our society that can bring fatal consequences and can be presented in digital text for example at online social networks. Nowadays there is a wide variety of works focused on the detection of digital texts in the English language, however in the Spanish language there are few studies that address this issue. This paper aims to detect this cybernetic harassment in social networks, in Spanish language. Sentiment analysis techniques are used, such as bag of words, elimination of signs and numbers, tokenization and stemming, as well as a Bayesian classifier. The data used for the training of the Bayesian classifier were obtained from the Spanish Dictionary of Affect in Language (SDAL), which is a database formed by more than 2500 words manually evaluated in three affective dimensions: Pleasantness, activation and imagery, as well as same 595 words obtained following the same procedure of SDAL was used with the help of the members of the Research Center, Technology Transfer and Software Development. As a result, the software developed has 93% success in the validation tests carried out.**

***Keywords*—Cyberbullying; social media analytics; sentiment analysis; tokenization; stemming; bag of words**

I. Introduction

As online social networks (OSN) have grown in popularity, instances of cyberbullying at OSN have become a growing concern. The prevalence of Cyberbullying in Peru is 20 to 40% in the last 10 years, according to the report "Cyberbullying: Approach to a comparative study: Latin America and Spain", by Albert Clemente, professor at the International University of Valencia (VIU) [1].

The VIU expert explains that in general, prevalence is understood as the set of individuals involved in the phenomenon of harassment or cyberbullying, that is, both victims, perpetrators and spectators. And stresses that "cyberbullying has not stopped growing and has become a problem in all cultures and regions of the world, both in its traditional and online" [1]. In addition, research has been conducted between technical performance tests and negative results, such as decreased school performance, absenteeism, school absenteeism, school dropout and violent behavior [2], and potentially psychological effects. devastating, such as depression, low self-esteem, suicidal ideation, and even suicide, which may have long-term effects on the future life of victims [3], [4]. Incidents of cyberbullying with extreme consequences, such as suicide, are reported routinely in the popular press.

Given the seriousness of the consequences that cyberbullying has on its victims and its rapid spread among college and university students, there is an immediate and compelling need for the research to understand how cyberbullying occurs in today's OSN. So things can be done to detect with cyberbullying.

The sentiment analysis, also called opinion mining [5], is the field of study that analyzes opinions, feelings, evaluations, attitudes and emotions of people towards entities such as products, services, organizations, individuals, problems, events, themes and their attributes. While most papers address it as a simple categorization problem, the sentiment analysis is actually a research problem [6] that requires addressing many natural language processing (NLP) tasks, including recognition of entities named [6], [7], the disambiguation of the polarity of the word [8], the personality recognition [9], the detection of sarcasm [10] and the extraction of the aspect [11]. In particular, the subtask is an extremely important subtask that, if ignored, the accuracy of the sentiment analysis in the presence of multiple points of opinion can be reduced consistently.

Therefore, the aspect-based sentiment analysis (ABSA) [6], [10], [12], [13], extends the feeling analysis section with a more realistic assumption that the polarity is associated with specific aspects (or characteristics of the product) instead of the whole text unit. For example, in the sentence "Food is delicious but service is horrible", the feeling expressed towards the two aspects is completely opposite. Through the aggregation of the analysis of feelings with the aspects, ABSA allows the model to produce a detailed opinion of the opinion of the people towards a particular product.

The objective sentiment classification (or objective-dependent) [14], [15], [16], instead, solves the polarity of the feeling of a given goal in its context, assuming that a prayer could express different opinions towards different specific entities. For example, in the sentence "I just logged into my Facebook and found an ugly picture of Anastacia", the sentiment expressed towards Anastacia is negative, while there is no clear feeling for Facebook.

Recently, Saeidi et al. [17], have tried to address the challenges of ABSA and the analysis of specific feelings. The task is to jointly detect the aspect category and resolve the polarity of the aspects with respect to a given objective. The deep learning methods [18], [19], [20], [21] have achieved great accuracy when applied to ABSA and analysis of specific feelings. Especially, sequential neural models, such as short-term long memory networks (LSTM) [22], are of increasing interest for their ability to represent sequential information. In addition, most of these sequence-based methods incorporate the attention mechanism, which is rooted in the alignment model of machine translation [23]. Such a mechanism takes an external memory and representations of a sequence as input and produces a probability distribution that quantifies the concerns in each position of the sequence.

Currently the industry around the sentiment analysis, increased its popularity due to the proliferation of commercial applications, offering many challenging problems becoming a very active research area with a broad domains offering a strong motivation for research and offering many challenging problems, which had not been studied before, such as processing information from social networks Facebook, Twitter, Instagram, blogs, wikis and other mass media online [24], [25], [26], which speed up the way of sharing private and/or intimate information through its platforms facilitating users to get in close contact with others without taking into account the dangers that these involve [5], [24], [27].

This type of communication can be dangerous and have serious consequences, because the post messages can contain some types of abusive or offensive content through which threats such as cyberbullying may emerge [24], [27]. In general, adults may be able to establish a secure line of communication and are often more aware of curiosity to explore new fields without the capacity of the dangers existing in social networks. Conversely, children or teenagers [24], [27], often have a misperception of threats and must weigh the potential risks of this communication.

The remaining part of the paper is organized as follows. A related works in this paper is explained in Sections 2, 3 and 4. Materials and Methods are described in Section 5. The results with the experiment settings is introduced in Section 6. Conclusions are presented in Section 7 and some future works are provided in Section 8.

II. Related Work

Dan Olweus [28], one of the leading specialists in the world in bullying, developed the first criterion to identify the specific form of bullying, when it was discovered that the phenomenon is associated with a high rate of suicide attempts among adolescents and defined a harassment situation as one in which ”a student is assaulted or becomes a victim if he is exposed, repeatedly and for a time, to negative actions carried out by another student or several of them”. For this author in the harassment there is a clear intention to harm the other, either physically or morally, in such a way that the intimidation is constant and persists over time. It is very remarkable the imbalance of forces between the aggressor and the victim, especially because the latter has difficulty overcoming mockery or aggression and decides to remain silent.

Vilares David [29], describes a system of opinion mining that classifies the polarity of texts in Spanish. He proposed an approach based on natural language processing that led to a segmentation, tokenization and labeling of the texts to then obtain the syntactic structure of the sentences by algorithms of dependency analysis.

The syntactic structure is then used to deal with three of the most significant linguistic constructions in the field we are dealing with: intensification, adversative subordinate clauses and denial. The experimental results show an improvement of the performance with respect to the purely lexical systems and reinforce the idea that the syntactic analysis is necessary to achieve a robust and reliable sentiment analysis.

Hernandez Li [30], carried out an investigation on the sentiment analysis in texts based on semantic approaches with linguistic rules for the classification of polarity of texts in Spanish, the classification was made according to a dictionary of semantic orientation where each The term is marked with a use value and emotional value, along with linguistic rules to solve several constructions that could affect the polarity of the text. For this evaluation a sample of 60,798 Twitter messages was used, each tweet is labeled with a global polarity, indicating whether the text expresses a Strongly Positive, Positive, Neutral, Negative, Strongly Negative feeling and no feeling. Among the results, it was found that 35.22% do not express any feelings, the 34.12% company positive feelings and 18.56% express negative feelings.

Martnez et. al and Alonso [31], [32], carried out a research approach to the study of the analysis of opinions in Spanish, where a survey of the researchs on the analysis of feelings is made and the small number of researches is expressed in Spanish, being the majority in English; also highlights the research in Spanish conducted by the group ITALICA of the University of Seville. Similarly, it tells us about the unbridled growth of the use of social networks where users give opinions of any type and topic, encouraging the use of these data for future research.

Baquero Abel [33], designed an instrument to detect cyberbullying in a school context and analyzed its psychometric properties. As participants, it had 299 adolescents (54.2% women and 45.8% men) with an average age of 15 years, belonging to the low stratum (22.1%) and middle stratum (78%). A quantitative study was carried out with a non experimental design of instrumental type and cross section. Under the classical theory of the tests, an adequate internal consistency was obtained, as well as convergent validity with the other measures.

The exploratory factor analysis was carried out in SPSS version 21, which yielded three factors. From the item response theory, it was found that the INFIT of the items ranged between 0.73 and 1.23 and the OUTFIT between 0.74 and 1.24. Based on the favorable results of the psychometric analysis, it is concluded that the instrument can be used for the detection of cyberbullying in a school context.

As instruments, the bullying prevention and dismantling project was used, which included bullying and cyberbullying questionnaires and workshops conducted by the school guidance team. Among the results revealed for the research, 58.32% have more of 200 Facebook contacts, a 21% share their password with pairs, and five students of the course answered having been bothered by this page.

Becerra Martn [34], analyzed the large volumes of data generated in social networks about public opinion and proposed to analyze a set of data using a sentiment classifier to tag publications made by Twitter users, in conjunction with clustering algorithms for to be able to detect which are the topics on which opinions are expressed. He used a base of 2000 reviews of films labeled as positive and negative to then train an SVM classifier of feelings, then the K-Means clustering algorithm to get a general overview of the topics and an approximation of the feeling associated with them.

III. Sentiment Analysis

The sentiment analysis [30], [34], [35], seeks to extract opinions, about a certain entity and its different aspects from the natural language of texts. This is done automatically using algorithms for classification. Opinions are classified according to the feeling they transmit, that is, as positive, negative or neutral. Its importance is that our perception of reality, and thus also the decisions we make, is conditioned in a certain way by how other people see and perceive the world. That is why, from a point of view of utility, we want to know the opinions of other people on topics of interest, since they have various applications such as recommending products and services, determining which political candidate to vote in the next elections or even measuring public opinion before the measure taken by a company or a government.

A. Types of Sentiment Analysis

At the time of extracting this information, there is a great variety of methods and algorithms depending on the level of granularity of the analysis that we want to carry out. The levels [34], [36], [37], document, sentence or aspect are distinguished. The analysis at the document level determines the general feeling expressed in a text, while the analysis at the sentence level specifies it for each of the sentences in the text. However, these two types of analysis do not delve into in detail the element that people like or dislike. They do not specify what is the opinion, since considering the general opinion of an object as positive or negative does not mean that the author has a positive or negative opinion of all aspects of that object. For this work we focus on conducting a document level analysis as a first instance, due to the limit in the messages, the authors are usually concise and go straight to the point without having the possibility of including several different aspects in a single post. For this reason, using the post as a unit of analysis seems to provide an adequate level of granularity to make a broken down analysis of sentiment.

1) The sentiment analysis at the document level: The document-level analysis [34], [36], aims to classify the opinion of a document, in this case a post. This task does not consider the details regarding entities or aspects, but considers the document as a whole, which will be labeled as positive or negative. This can be considered as a traditional text classification task, where classes are different orientations in terms of feelings. However, to ensure that this type of analysis makes sense, we assume that each document expresses a single opinion on a single entity. Although this may seem a limitation, because in a post one could express more than one opinion towards different entities, in practice it produces positive results, since users tend to focus on only one aspect in each post.

IV. Cyberbullying

Cyberbullying, [24], [27], [38], [39], is the use of digital media to harass a person or group of people, through personal attacks, disclosure of confidential or false information among other means. It may constitute a criminal offense. Cyberbullying involves recurrent and repetitive damage inflicted through digital media.

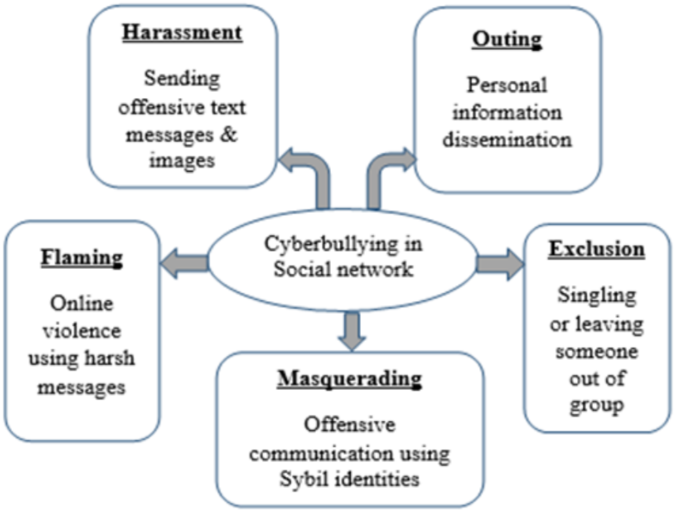

Harassment
Sending offensive text messages & images
Outing
Personal information dissemination
Flaming
Online violence using harsh messages
Cyberbullying in Social network
Exclusion
Singling or leaving someone out of group
Masquerading
Offensive communication using Sybil identities

Fig. 1. Types of cyberbullying (Source: Hosseinmardi [27]).

According to Karthik Dinakar, [40], [3], cyberbullying is a more persistent version of traditional forms of intimidation, which extend beyond the physical confines of a school, sports field or workplace, with the victim often does not experience any respite from it. Cyberbullying gives an individual the power to embarrass or hurt a victim before an entire online community [41], especially in the realm of social networking websites. This is widely recognized as a serious social problem, [38], [40], [3], [42], [43], especially for teenagers.

The mitigation of cyberbullying involves two key components, robust techniques for effective detection and reflective user interfaces that encourage users to reflect on their behavior and choices. The types of cyberbullying usually occurs in the social network that shows in Fig. 1 are Harassment (sending offensive text messages and images), Flaming (Online violence using harsh messages), Masquerading (Someone might create fake email addresses or instant messaging names or someone might use someone else's email or mobile phone to bully another person), Outing (personal information dissemination) and Exclusion (Singling or leaving someone out of group) [27].

V. Materials and Methods

A. Database

As a first step for the detection of cyberbullying through the analysis of feelings, it is necessary to have a database for the training of the Bayesian network. The database of Agustn Gravano (SDAL) [25], from the Faculty of Exact and Natural Sciences of the University in Buenos Aires, Argentina, was used.

The database SDAL [25] is a lexicon of 2880 words in Spanish, which have been annotated manually with respect to three affective dimensions:

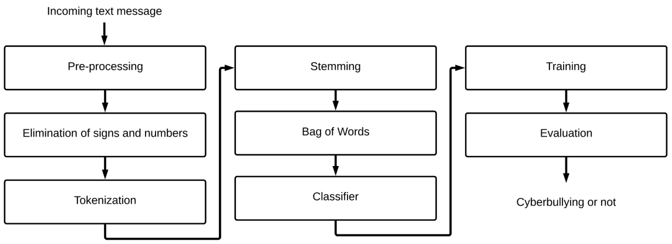


Fig. 2. Cyberbullying detection process.

- Pleasant (pleasant, neutral, unpleasant)
- Activation (active, neutral, passive)
- Imaginability (easy to imagine, neutral, hard to imagine)

Likewise, 595 words obtained following the same SDAL procedure were used, collected with the help of the members of the Research Center, Technology Transfer and Development of Software (CiTeSoft), mostly Peruvian and Spanish slang in order to improve the results.

B. Cyberbullying Detection

For cyberbullying detection, the developed software follows the procedure shown in Fig. 2. Each process will be explained in greater detail in the following subsections.

C. Preprocessing

The message or post must be preprocessed because it contains unstructured text. The purpose of preprocessing is to transform messages into a uniform format that can be understood by the learning algorithm. In preprocessing, the process of tokenization, stemming, elimination and stoppage of meaningless words, elimination of numbers and blank spaces is carried out.

D. Bag of words

One of the most important subtasks in the text classification with bullying is the extraction of characteristics. Through the use of machine learning algorithms to train the classifier, the representation of the text as a feature vector is required. For that, a model commonly used in the processing of natural language is the Bag of Words (BoW) model. The main stage of this model is the creation of a vocabulary of words that, in our approach, indicates the vocabulary or the collection of abusive words. Among the reference approaches for text classification, the BoW approach has the highest recovery rate of 66% [44]. In the BoW model, each word is associated with a count of occurrences. This vocabulary can be understood as a set of non-redundant words where order does not matter. The BoW approach ignores grammar and detects offensive sentences by checking whether or not they contain offensive or offensive words.

E. Natural Language Processing

The stage of Processing of Natural Language [45], is very important for the implementation of models of analysis of feelings. It is necessary to carry out some processes both to the text that we are going to analyze, and to the text that the classifying algorithm will train. The processes that they applied are the following.

1) Elimination of signs and numbers: It is necessary to eliminate signs and numbers from the text, signs like "!", "?", "+", etc., since the existence within the text could affect the recognition of the expressions by the classifier. Table I shows two examples.

TABLE I. EXAMPLE OF ELIMINATION OF SIGNS AND NUMBERS

Original text	Transformed Text
Que buena pelcula!	Que buena pelcula
Eres una mala persona :8	Eres una mala persona

2) Tokenization: It consists in breaking up the text in the different words of the ones that appear, naming these resulting elements tokens [46]. Each document in our corpus is transformed into a list of terms called tokens. This representation of data is also known as a bag of words. Tokens are strings of characters between spaces or punctuation, but this is not always the case, as for example in the case of the abbreviations [34]. The total set of words used, distinct and unique, is the vocabulary of the corpus. Table II shows two tokenization examples.

TABLE II. TOKENIZATION EXAMPLE

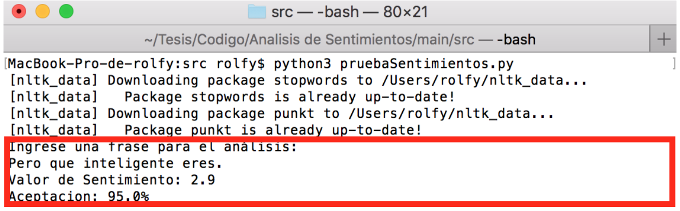
Original text	Transformed Text
Que buena pelcula	["Que" "buena" "pelcula"]
Eres una mala persona	["Eres" "una" "mala" "persona"]

3) Stemming: It consists of extracting stems of the tokens obtained in the previous process. So the different forms, such as diminutives, superlatives, gender, etc. do not affect the result [47].

Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form-generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.

4) Naive Bayes classifier: Naive Bayes classifier, [26], assign probabilities to the data entered, building a tree of probabilities according to the data entered, within the NLTK tool set we find the nltk. Naive Bayes Classifier class that allows us to use this type of classifier and train it according to our needs.

5) Training: To train the Naive Bayes classifier we need known data, so it is a supervised learning algorithm. This is where the need for a lexicon arises because the analysis was based on them. The lexicon is a file that can vary in its structure, but it must contain a list of words, with its respective subjectivity value in order to be processed in order to train the classifying algorithm. In this case we are using the database SDAL [25], which we saw in detail in the database section.



```
MacBook-Pro-de-rolfy:src rolfy$ python3 pruebaSentimientos.py
[nltk_data] Downloading package stopwords to /Users/rolfy/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /Users/rolfy/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
Ingrese una frase para el análisis:
Pero que inteligente eres.
Valor de Sentimiento: 2.9
Aceptacion: 95.0%
```

Fig. 3. Example of text with positive polarity (low probability bullying).

6) *Implementation:* It was imported and used:

- NLTK [48] as a Python library for natural language processing.
- Pickle to save the classifier instance as a binary file.
- OS to be able to interact with the system.

The classifiers require to receive a dictionary that recognizes them as features that are those that will describe conditions for a result to be given. In the particular case of our analysis we simply send the expression or word as a characteristic "word" the dictionary.

Then declare the global variables that correspond to the classifier and the list of words known by the classifier.

VI. Results and Validation

The software provides: the feeling value of the text or phrase which is in the range of 1 to 3, where 1 is negative, 3 is positive and 2 is neutral. In addition to acceptance, which is in the range of 0% to 100%, where 0% is a text or phrase has a high probability of containing bullying and 100% a low probability of containing bullying.

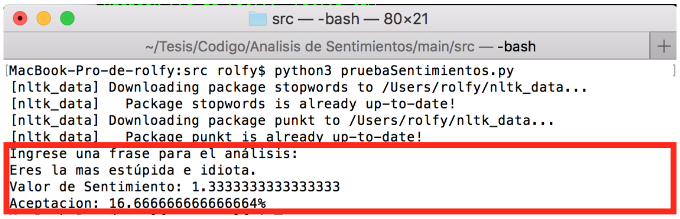
To validate the operation of the software a list of phrases was made, classified by three types: phrases and texts with positive polarity (high probability of bullying), negative (low probability of bullying) and neutral, which were evaluated by the software and confronted with the manual evaluation by the members of the CiTeSoft [49] of The National University of San Agustn [50]. Below are three types of phrases.

A. Without Bullying

The software successfully responded to the tests that were carried out with phrases and texts of positive polarity, as seen in Fig. 3 the phrase "But how intelligent you are" obtains an acceptance of 95 % indicating that there is a low probability of bullying in addition to indicating that it is a very positive phrase.

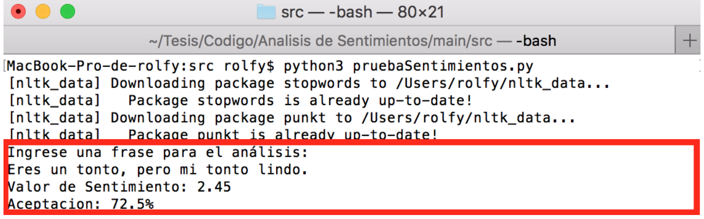
B. With Bullying

Tests were carried out with simple and complex negative polarity text, as we can see in Fig. 4, which is a container text of bullying. The software successfully responded to bullying text tests, as shown in the figure "You are the stupidest and idiot person." obtains an acceptance of 16% indicating a high probability of the existence of bullying and validating the operation of the software for the detection of text containing bullying.



```
MacBook-Pro-de-rolfy:src rolfy$ python3 pruebaSentimientos.py
[nltk_data] Downloading package stopwords to /Users/rolfy/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /Users/rolfy/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
Ingrese una frase para el análisis:
Eres la mas estúpida e idiota.
Valor de Sentimiento: 1.3333333333333333
Aceptacion: 16.666666666666664%
```

Fig. 4. Example of text with negative polarity (high probability of nullying).



```
MacBook-Pro-de-rolfy:src rolfy$ python3 pruebaSentimientos.py
[nltk_data] Downloading package stopwords to /Users/rolfy/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /Users/rolfy/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
Ingrese una frase para el análisis:
Eres un tonto, pero mi tonto lindo.
Valor de Sentimiento: 2.45
Aceptacion: 72.5%
```

Fig. 5. Example of ambiguous phrase (neutral).

C. Ambiguous (Neutral)

Tests were performed with neutral polarity text, neutral polarity occurs when a text is ambiguous because in the first part the sentence contains a high probability of containing bullying but in the second part a low probability or vice versa.

The software successfully responded to tests with neutral text, as seen in Fig. 5 the phrase "You're a fool but my cute fool." The first part of the sentence has an insult, but in the second it is clarified that it is an expression of affection; obtaining an acceptance of 62% indicating a low probability of the existence of bullying and validating the operation of the software for the detection of ambiguous text.

D. Validation

To evaluate the effectiveness of the software, a collection of phrases from social networks (Facebook, Twitter, Instagram and Youtube) of diverse topics was done and a manual score was made between 0 and 10, where 0 represents a hurtful, offensive or bullying phrase and 10 a pleasant phrase or without bullying. This evaluation was carried out by the members of the CiTeSoft [49] (Center for Research, Technology Transfer and Software Development), then an arithmetic average was made between the evaluations of these members to be able to compare with the evaluation of the software developed. On the other hand, the evaluation of the same sentences by the software was made, then the range of acceptation that the software gives us from 0-100 to 1-10 was made to make a confrontation and see the effectiveness of it. As we observe in Table III, it shows the results of 100 sentences evaluated by members of CiTeSoft and the resulting average of each sentence. Likewise, the comparison between the average of the evaluations and the evaluation of the developed software, in version 1 and version 2, is shown in Table IV.

Finally, a comparative graph was drawn up as shown in Fig. 6 to see better the difference between the results obtained and the error percentage of the software. As can be seen in test 27, there was a very high error margin. This occurs because the software does not know the words that were used in the evaluation phrase.

TABLE III. RESULTS OF THE EVALUATION OF THE SENTENCES BY MEMBERS OF CITESOFT [49]

PHRASE	T 1	T 2		T 8	T 9	T 10	AVERAGE
1	1	5		6	3	5	3.8
2	1	3		1	1	1	1.8
3	7	5		7	6	5	6
4	9	7		9	8	9	7.8
5	1	5		5	1	5	3.1
6	9	6		9	8	8	8.1
7	1	5		3	2	3	2.4
8	1	2		1	2	1	2
9	1	4		2	2	1	1.8
10	1	3		3	2	2	2.3
11	9	6		9	7	9	7.6
12	1	2		3.5	3	4	2.5
13	9	7		8	8	9	7.9
14	1	2		2	2	2	1.7
15	9	6		9	9	9	8.4
16	1	2		1	1	1	1.6
17	1	2		1	1	1	1.7
18	1	3		6	3	6	3.4
19	1	3		3	2	2	2.4
20	8	6		9	9	9	8.5
21	1	3		1	2	1	1.6
22	9	7		8	8	8	7.4
23	9	6		9	7	9	7.7
24	1	4		3	4	3	2.6
25	1	3		1	1	1	1.6
26	1	5		3	1	3	2.3
27	1	5		1	2	1	2.6
100	1	4		3	4	3	3

TABLE IV. COMPARISON BETWEEN THE EVALUATION OF THE SOFTWARE (VERSION 1 AND 2) AND THE EVALUATION OF THE CITESOFT [49] MEMBERS OF THE TEST SENTENCES

PHRASE	AVERAGE	SOFTWARE V1	SOFTWARE V2
1	3.8	3.9	3.9
2	1.8	5	1.6
3	6	6.5	6.5
4	7.8	7.3	7.3
5	3.1	4	2.6
6	8.1	7.5	7.5
7	2.4	6.5	3
8	2	7.6	2.6
9	1.8	2.6	2.6
10	2.3	3.2	3.2
11	7.6	7.2	6.6
12	2.5	4	4
13	7.9	9.5	9.5
14	1.7	7.5	3.3
15	8.4	7.4	6.8
16	1.6	4.6	3.3
17	1.7	3.5	3.5
18	3.4	5.2	5.2
19	2.4	6	4.3
20	8.5	7.3	6.6
21	1.6	3.7	3.7
22	7.4	6.4	5.3
23	7.7	6.2	5.5
24	2.6	6.6	5
25	1.6	6.6	4
26	2.3	7.7	6.1
27	2.6	6.5	6.5
...	...	..	..
100	3	7.6	7.6

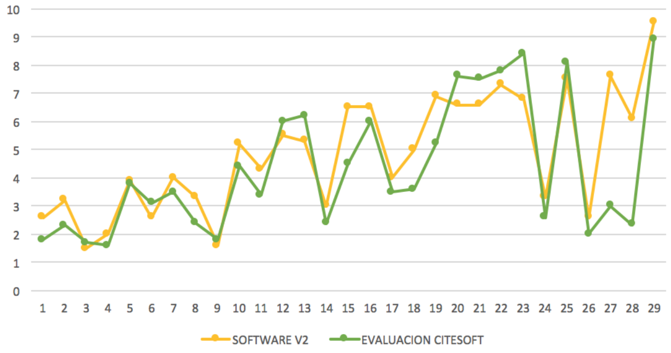


Fig. 6. Comparison between the evaluation of the test phrases of the software and the manual evaluation by the CiTeSoft members.

VII. CONCLUSION

In the validation of the software, three types of tests were carried out, without bullying, with bullying and ambiguous (neutral), then they were confronted with the manual evaluation of the members of the CiTeSoft [49], as shown in Table III, passing successfully the same in 93% of the cases, demonstrating its correct functioning.

The performance of the software developed depends directly on the number of words used in the sentences to be evaluated and if they are found in the word bag. In this work we worked with Peruvian and Spanish words and slang if we want to use software to evaluate phrases from other countries, we recommend adding words from these countries to the word exchange for better performance.

VIII. FUTURE WORK

As future work it is proposed to optimize the detection of cyberbullying by: Replacement of emoticons: A bag of emoticons and their respective meaning will be created, then this chain of characters will be replaced by a string that can be searched in the semantic orientation dictionary.

Correction of abbreviations: some of the most common abbreviated words will be replaced by their recognized grammatical form (Example:“q” → “que”, “xq” → “$porque$”).

Spelling correction: The Levenshtein algorithm with its notion of distance will be used. To correct the words, a dictionary of words will be used, which is made up of the complete list of forms of the Corpus of Reference of Actual Spanish (CREA) of the Royal Spanish Academy, with frequencies of use and with the conjugated forms most used , approximately 128 thousand forms. If a word is not found in the dictionary, the algorithm will take the nearest word with distance 1, and replace it with.

Correction of repeated characters: especially in the case of vowels, the repetition of the same concurrence will be replaced by a single one, with the exception of cc, rr, ll. Once the clean text is obtained, we proceed to carry out the lemmatization of the words to obtain their motto without conjugation, together with the tokenization and the segmentation of the sentences in order to classify the polarity (for example: “$largoooooo$” → “$largo$”).

Acknowledgment

We would like to thank Citesoft UNSA Research Center for providing us materials, books, tools and computers to test algorithms proposed in this paper.

References

[1] I. Viu, "Ciberacoso. Aproximación a un estudio comparado: Latinoamérica y España 1," pp. 1–28, 2015.

[2] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth," *Psychological Bulletin*, vol. 140, no. 4, pp. 1073–1137, 2014.

[3] K. Dinakar, R. Picard, and H. Lieberman, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2015-Janua, no. 3, pp. 4168–4172, 2015.

[4] M. van Geel, P. Vedder, and J. Tanilon, "Relationship Between Peer Victimization, Cyberbullying, and Suicide in Children and Adolescents," *JAMA Pediatrics*, vol. 168, p. 435, 5 2014.

[5] B. Liu, *Sentiment Analysis and Opinion Mining*. 2012.

[6] E. Cambria, D. Das, S. Bandyopadhyay, and A. F. Editors, "Socio-Affective Computing 5 A Practical Guide to Sentiment Analysis," 2017.

[7] Y. Ma, H. Peng, and E. Cambria, "Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM," 2014.

[8] Y. Xia, E. Cambria, b. Amir, H. @bullet, and H. Zhao, "Word Polarity Disambiguation Using Bayesian Model and Opinion-Level Features," 2015.

[9] Y. Ma, E. Cambria, and S. Gao, "Label Embedding for Zero-shot Fine-grained Named Entity Typing," pp. 171–180, 2016.

[10] E. Cambria, S. Poria, A. Gelbukh, I. P. Nacional, and M. Thelwall, "AFFECTIVE COMPUTING AND SENTIMENT ANALYSIS Sentiment Analysis Is a Big Suitcase," 2017.

[11] A. Mukherjee and B. Liu, "Aspect Extraction through Semi-Supervised Modeling," *Jeju, Republic of Korea*, pp. 339–348, 2012.

[12] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. María Jiménez-Zafra, and G. Eryiğit, "SemEval-2016 Task 5: Aspect Based Sentiment Analysis," pp. 19–30, 2016.

[13] E. Cambria, J. Fu, F. Bisio, and S. Poria, "AffectiveSpace 2: Enabling Affective Intuition for Concept-Level Sentiment Analysis," 2015.

[14] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for Target-Dependent Sentiment Classification," 2015.

[15] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, "Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification," pp. 49–54, 2014.

[16] B. Wang, M. Liakata, A. Zubiaga, and R. Procter, "TDParse: Multi-target-specific sentiment recognition on Twitter," vol. 1, pp. 483–493, 2017.

[17] M. Saeidi, G. Bouchard, M. Liakata, and S. Riedel, "SentiHood: Targeted Aspect Based Sentiment Analysis Dataset for Urban Neighbourhoods," 2016.

[18] T. H. Nguyen and K. Shirai, "PhraseRNN: Phrase Recursive Neural Network for Aspect-based Sentiment Analysis," pp. 2509–2514, 2015.

[19] Y. Wang, D. Zeng, B. Zhu, X. Zheng, and F. Wang, "Patterns of news dissemination through online news media: A case study in China," *Information Systems Frontiers*, vol. 16, no. 4, pp. 557–570, 2014.

[20] D. Tang, B. Qin, and T. Liu, "Aspect Level Sentiment Classification with Deep Memory Network," 2016.

[21] Y. Wang, M. Huang, L. Zhao, and X. Zhu, "Attention-based LSTM for Aspect-level Sentiment Classification," pp. 606–615, 2016.

[22] Hochreiter Sepp and Schmidhuber Jurgen, "Long Short-Term Memory," 2001.

[23] D. Bahdanau, K. Cho, and Y. Bengio, "NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE," 2015.

[24] K. B. Kansara and N. M. Shekokar, "A Framework for Cyberbullying Detection in Social Network," *International Journal of Current Engineering and Technology*, vol. 5, no. 1, pp. 494–498, 2015.

[25] D. A. Ríos and G. Matías, "A Spanish Dictionary of Affect in Language Spanish DAL : A Spanish Dictionary of Affect in Language," 2015.

[26] J. H. Xue and D. M. Titterington, "Comment on "on discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes"," *Neural Processing Letters*, vol. 28, no. 3, pp. 169–187, 2008.

[27] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Detection of Cyberbullying Incidents on the Instagram Social Network," 2014.

[28] D. Olweus, "Bullying at School," pp. 97–130, 1994.

[29] D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez, "Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias Polarity classification of opinionated Spanish texts using dependency parsing," 2013.

[30] R. Hernández Petlachi and X. Li, "Análisis de sentimiento sobre textos en Español basado en aproximaciones semánticas con reglas linguísticas," *Tass 2014*, 2014.

[31] L. F. Hurtado and F. Pla, "ELiRF-UPV en TASS 2016: Análisis de sentimientos en twitter," *CEUR Workshop Proceedings*, vol. 1702, no. September, pp. 47–51, 2016.

[32] L.-F. Hurtado and F. Pla, "ELiRF-UPV en TASS 2014: Análisis de Sentimientos, Detección de Tópicos y Análisis de Sentimientos de Aspectos en Twitter," *Procesamiento del Lenguaje Natural*, pp. 1–7, 2014.

[33] C. A. Baquero and P. B. L. Avendaño, "Diseño y análisis psicométrico de un instrumento para detectar presencia de ciberbullying en un contexto escolar," *Psychology, Society, & Education*, vol. 7, no. 2, pp. 213–226, 2015.

[34] C. D. T. Estudiantiles, "Análisis de sentimientos en Twitter : El bueno , el malo y el ¿ :(," pp. 184–209, 2017.

[35] T. Wilson, J. Wiebe, and P. Hoffman, "Recognizing contextual polarity in phrase level sentiment analysis," *Acl*, vol. 7, no. 5, pp. 12–21, 2005.

[36] M. Hu and B. Liu, "Mining and summarizing customer reviews," *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, p. 168, 2004.

[37] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," *Proceedings of the 20th international conference on Computational Linguistics*, p. 1367, 2004.

[38] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. De Jong, "Improving cyberbullying detection with user context," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7814 LNCS, pp. 693–696, 2013.

[39] T. Nasukawa, T. Nasukawa, J. Yi, and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," *Proceedings of the 2nd international conference on Knowledge capture*, p. 70–77, 2003.

[40] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," *Association for the Advancement of Artificial Intelligence*, pp. 11–17, 2011.

[41] N. Tsirakis, V. Poulopoulos, P. Tsantilas, and I. Varlamis, "Large scale opinion mining for social, news and blog data," *Journal of Systems and Software*, vol. 0, pp. –, 2016.

[42] M. Dadvar and F. D. Jong, "Cyberbullying detection: a step toward a safer Internet yard," *Proceedings of the 21st international conference . . .*, pp. 121–125, 2012.

[43] M. Dadvar, F. M. G. de Jong, R. J. F. Ordelman, and R. B. Trieschnigg, "Improved cyberbullying detection using gender information," *12th Dutch-Belgian information retrieval workshop (DIR 2012)*, pp. 23–25, 2012.

[44] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, pp. 71–80, 2012.

[45] K. Dinakar, R. Reichart, H. L. T. S. M. Web, and U. 2011, "Modeling the detection of Textual Cyberbullying.," *aaai.org*, 2011.

[46] C. Za'in, M. Pratama, E. Lughofer, and S. G. Anavatti, "Evolving Type 2 Web News Mining," *Applied Soft Computing*, pp. –, 2017.

[47] J. B. Lovins, "Development of a stemming algorithm," *Mechanical Translation and Computational Linguistics*, vol. 11, no. June, pp. 22–31, 1968.

[48] E. Loper and S. Bird, "NLTK: The Natural Language Toolkit," no. July, pp. 69–72, 2002.

[49] CiTeSoft, "CiTeSoft – Centro De Investigación, Transferencia De Tecnologías Y Desarrollo De Software I+D+I - UNSA."

[50] UNSA, "Universidad Nacional de San Agustin de Arequipa."

A Resource Recommendation Approach based on Co-Working History

Nada Mohammed Abdulhameed, Iman M. A. Helal, Ahmed Awad, Ehab Ezat
Information Systems Department
Faculty of Computers and Information
Cairo University, Egypt

Abstract—**Recommending the right resource to execute the next activity of a running process instance is of utmost importance for the overall performance of the business process, as well as the resource and for the whole organization. Several approaches have recommended a resource based on the task requirements and the resource capabilities. Moreover, the process execution history and the logs have been used to better recommend a resource based on different human-resource recommender criteria like frequency and speed of execution, etc. These approaches considered the recommendation based on the individual's execution history of the task that will be allocated to the resource. In this paper, a novel approach based on the co-working history of resources has been proposed. This approach considers the resources that had executed the previous tasks in the current running process instances. Then, it recommends a resource that has the best harmony with the rest of the resources.**

Keywords—Business process; process instance; co-working history; human-resource recommender criteria; harmony

I. Introduction

Resource allocation is highly relevant to process mining and its applications. This relevance has been an important issue in business process management (BPM) [1]. Some researchers have discussed ways to optimize allocating the resources in an organization, to improve its business process [2], [3]. They have studied the business process structural features and the way to optimize the available resources to reach the perfect fit for the business needs. Other researchers have described the resource patterns and the correlation between different activities and the available resources [4]. In order to allocate resources, a clear set of rules need to be specified at the beginning of the process lifetime, though this can be challenging. In order to better allocate resources, some researchers have provided different resource patterns, e.g. creation, push, pull, detour [4]. Some of these patterns, such as *push* (from system to worker) and *pull* (from worker to system) patterns, do not rank the process performance [3].

A *resource* is an important indicator of a business process performance. Resources can be machines, manpower, money, software, etc. The process of allocating the human resources can be optimized by analyzing their behavior and mining the event logs to find the rules and the different resource patterns. These resources need to be allocated dynamically to improve the efficiency of the process performance in BPM through a resource recommendation approach.

The main contribution is a resource recommendation approach based on the *co-working history* from the event log. This approach considers the resources executed in the previous tasks at the current running process instances. In order to recommend a resource that has the best harmony with the rest of the resources, the proposed approach considers the *frequency* and the *duration* criteria.

The remainder of this paper is organized as follows: an overview of our approach that briefly discusses most of the background ideas, techniques and tools used to cover this paper in Section II. Section III covers a discussion of previous work. The contribution in resource recommendation based on the co-working history is discussed in Section IV. Implementation details and evaluation are discussed in Section V. Finally, the paper concludes with an outlook for the future work in Section VI .

II. Background

This section starts with some basic concepts about business process management, as well as describing some of the basic definitions used in the resource recommendation approach (the proposed approach). It starts with a brief overview about the business process and its components in section II-A. Then, Section II-B introduces the event log concept as the main input to the proposed approach. Finally in section II-C, the raw performance measure [5] is explained to be used later in extracting the co-working history.

A. Business Process Management

Business processes are used to organize the tasks performed in an organization by different resources [6]. The concept of business process has expanded in the domain of BPM [1], where a business process is represented as a set of activities and tasks performed in an organization or cross-organizations. Each business process serves a set of business goals in an organization or in cross-organizations [7].

BPM has several definitions in the community, one of them states that it is composed of a set of concepts, methods and techniques; each of which support the whole business life cycle (i.e. analysis/design, configuration, runtime, and mining) [1]. These methods and techniques manage the execution of business processes in a Business Process Management System.

Business processes are modeled as a set of activities that transit from activity to another using a control flow. Fig. 1 illustrates a simple example for a travel agency [7], [8], where a customer sends a travel request which will be processed for further actions. The request can be either accepted or rejected by a travel agent, (i.e., a resource in the travel

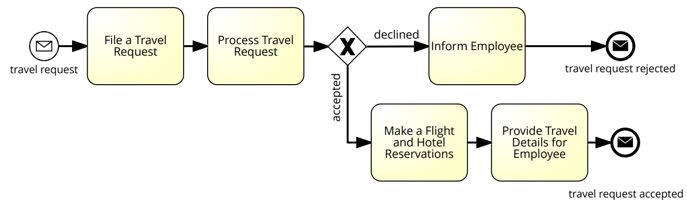


Fig. 1. Travel agency process model [7], [8].

agency). If the request is accepted, the travel agent will reserve both the flight and the hotel for the customer. However, if the travel agent rejected the travel request, the agency will inform the employee about declining the request. This example is graphically presented as a process model using Business Process Model and Notations (BPMN) [**?**].

B. Event Logs

In information systems, an event log contains all saved events that are related to the implemented activities by the specified resources. An event log is composed of a set of events that are correlated to a set of cases. An event is composed of a set of attributes, which includes the activity name (i.e. task), the resource responsible, and the timestamp of event occurrence, etc (cf. Definition 1). The series of registered events in a case is given the term *trace* (cf. Definition 2).

Definition 1 (Event): Let C be the set of all case identifiers, T the set of all task identifiers, R the set of all resource identifiers, S the set of all states, and M the set of all timestamps; So, the event $e \in (C \times T \times R \times S \times M)$ represents an occurrence of a state change in a process instance. We can access properties of an event by the dot notation. $e.case$ refers to the case identifier of e, analogously, $e.task$, $e.resource$, $e.state$, $e.timestamp$.

An event represents an evolution in the execution of a process instance (case), cf. Definition 2. This evolution occurs when one of the task instances (work items) within the case experience has a change of state. For instance, when a work item starts execution, or shows completes, fails, skipped, etc., an event should at least contain information about the case, the task instance, the resource, the type of state change, and the timestamp indicating the time of the event. The resources here refers to human performers involved in the execution of task instances.

Definition 2 (Execution Trace (Case)): An execution trace σ, case, is a sequence of events, $\sigma =<< e_1, e_2, \ldots, e_n >>$, where $e_i, 1 \leq i \leq n$, is an event as per Definition 1. The event can be $e_x < e_y$ if $e_x.timestamp < e_y.timestamp$. If an event e occurs within a trace σ, it is denoted as $e \in \sigma$. Also, the dot notation is used $\sigma.e$ to access event e of the trace. $|\sigma|$ is used to denote the length of the trace. Finally, an event can be accessed by its position in a trace, $\sigma[i],\ where\ 0 \leq i \leq |\sigma|$.

An event log contains different attributes. It is a set of cases each of which contains a set of events (cf. Definition 3). In this paper, an event log contains (task instance, resource, state, and timestamp) attributes. All the event logs provide information about the implementation of a single process by the process model [10].

Definition 3 (Event Log): An event log W is a set of traces. $W = \{\sigma_1, \sigma_2, \ldots, \sigma_k\}$, where $\sigma_i, 1 \leq i \leq k$, is a case as per Definition 2.

C. Co-working History

In order to determine the significance of the co-working history of a set of resources in an event log, these resources must have clear measures for their performance. ***Raw performance measure*** (RW) is concerned with deciding different performance measures for each resource in the execution log [5]. It is stated as a tuple for which the measure is calculated. It contains *process model instance*, *case instance*, *task instance* with the *resource* responsible, number of *occurrence* for the task performed by this resource within a *start* and *end* timestamps, and finally the *value* of the performance measure. RW can measure a resource with respect to its effective time, waiting time, service time, etc. In this paper, only the effective time is considered from RW as presented in [5].

Definition 4 (Trace History): For a work item $t \in T$, which has an event e, i.e. $e.task = t$, in a case σ, we define $\sigma_{<e} = \ll e_i | e_i < e \gg$ to be a sub-sequence of σ including all events that occurred before e.

The co-working history is based mainly on the event logs. It is one of the key notations that is defined as a task over an event log W (cf. Definition 5).

Definition 5 (Co-working History): For a work item $t \in T$ and an event log W, we define $W_{<t} \subseteq \mathcal{P}(W) = \{\{\sigma_v | \exists e \in \sigma_v \wedge e.task = t \wedge \sigma_{v<e} is\ a\ trace\ history\ for\ e\}\}$. Moreover, $\forall \sigma_x, \sigma_y \in W_{<t} : \sigma_x[i].task = \sigma_y[i].task \wedge \sigma_x[i].resource = \sigma_y[i].resource,\ where\ 0 \leq i \leq |\sigma_x|$.

Definition 5 finds $t \in T$ the different sets of trace histories that have common tasks and common resources performing them to recommend the resource who will execute the task.

Many studies have identified and classified the main criteria used in resource allocation approaches. These studies aimed at improving tasks performance within the process which are related to the properties of human resources (a taxonomy of resource allocation criteria) [11]. These specified criteria are *Amount, Experience, Expertise, Preference, Previous performance, Role, Social context, Trustworthiness,* and *workload.*

This paper presents a criteria for resources recommendation based on the co-working history, which considers the resources that performed the previous tasks in the current running process instances. This criteria is used to recommend a resource having the best harmony with the rest of the resources. This aspect of co-working history, *frequency* and *duration*, as a *harmony*-based aspect can be inserted among standards of social context, which includes *Collaboration, Compatibility, Influence* and *Social position.* This consideration works for the improvement of tasks performance in the process. Fig. 2 illustrates the proposed criteria and their inclusion in the classifications for resource allocation [11].

III. Related Work

In [4], the authors proposed 43 workflow resource patterns, classified into six categories. These categories cover, among

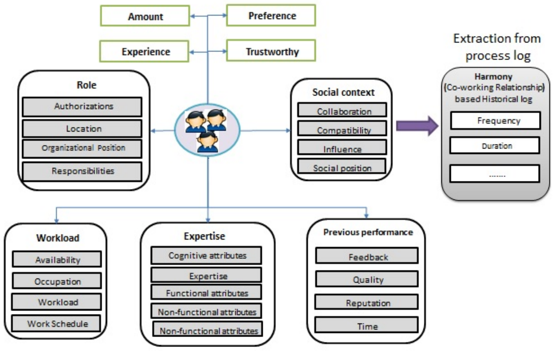


Fig. 2. Taxonomy of resource allocation criteria [11].

other aspects, how the resources can be assigned at the process-design time and how they can be allocated at runtime. Those patterns provide different approaches to identify the eligible resources, as in creation patterns, but they do not provide any means to recommend one candidate over the others. In [12], the authors provide a life-cycle support for the rules of staff assignment based on an organizational model and an event log to discover the allocation rules by learning the decision tree.

In [13], the problem of resource allocation optimization is modeled as Markov decision processes and solved using reinforcement learning. The optimization was fixed to either cost or flow-time based queuing for ordering the queue of work items allocated to each resource. A similar approach is presented in [14]. It is worth noting that both approaches considered role assignment at process *design time* in contrast to resource allocation at *runtime* which this research aims for.

In [15], the authors applied machine learning techniques on the process logs and process models to extract classifiers about the resource who can execute the activity. These classification models are then used to recommend resources for running instances. In [16], the authors used data mining techniques to support and identify resource allocation decisions by extracting information about the process context and process performance from past process executions histories.

In [17], [18], data mining techniques are used to extract resource allocation rules from process logs. The approach recognizes the so-called dependent resource assignment, e.g., if activity a_1 is executed by resource r_1, activity a_2 is executed by resource r_2, then activity a_3 should be executed by resource r_3. However, it is unclear how the approach would deal with other runtime aspects like workload or the unavailability of the resource. The proposed approach in this research recommends the resource based on the co-working history among other aspects, cf. [19]. Then, it adapts to the actual allocation that will take place on an instance level for the recommendation of the next task. Thus, the aim in this research is to introduce flexibility at runtime compared to the rigidness of the extracted association rules.

In [20], the authors have specified the preferences for different resources using expressions based on a Resource Assignment Language (RAL). These preferences are then used at runtime to rank the potential performers of an activity. The concept of providing a list of performers along with their rank is interesting and is of practical relevance. At runtime, it is not helpful to recommend just one resource as he might be engaged in other work or not present. Thus, it is important to provide several alternatives to do not block the process instance waiting for a free resource. In [19], a framework for recommending resource allocation based on process mining is defined. It introduces six dimensions to compare between potential resources and the user who can change the weight of each dimension to control the final recommendation. This approach can be seen as an extension of the work in [20]. Compared to what this research aims to achieve, this proposed approach targets going beyond the one-to-one relation between a task to allocate for a resource by studying the $n-ary$ relationships between groups of tasks and their respective potential performers.

Cooperation correlation among pairs of resources have been introduced and measured in [21]. Compared to the proposed approach perspective, the cooperation is applied only on pairs of resources whereas $n-ary$ sets of resources are considered based on the completed tasks within a case. Moreover, the authors have identified resource recommendation as a use case for calculated measures. However, for that specific use case they do not consider cooperation correlation as a criterion for resource recommendation. Rather, they consider resource preferences and competency. A similar approach about resource cooperation is presented in [22].

In [23], the authors have provided the resource allocation method under constraints of preference, availability and the total cost constraints. Then they analyzed the influence of collaboration between resources on process performances. In [24], the authors introduced a method to compute the social relation between two resources; then, they computed the influence of the previous resources on the candidate resources by using a Q-learning algorithm for dynamic task allocation. In [25], the authors presented a model which measures the compatibility among resources when assigning work items to the collaborative groups by using compatibility matrix. They have also developed an allocation algorithm to maximize team cooperation, the needs for inquiring the effect of cooperation on throughput and other process results.

IV. A Resource Recommendation Approach

This section presents the proposed approach for resources recommendation based on co-working history by specifying and applying frequency and duration criteria.

A. Recommendation Criteria based on Co-working History

As in Section II, the proposed approach is based on the co-working history for resource recommendation. It determines the criteria and metrics from the event log and uses them as a new dimension for resource recommendation which has been termed as *co-working* history. These criteria are as follows:

- Frequency Criterion (FC): It recommends the appropriate resource to perform the target task based on the number of times in which the resource works with the previous resources in the same cases at the event log. It is suitable to recommend a resource that works more times with the previous resources in the same cases for the event log to perform the target task; i.e.

the number of times during which the resources work with each other in the same cases in the event log.

- Duration Criterion (DC): It recommends the appropriate resource with less average time to perform the target task based on the previous resources. It is suitable to recommend a resource to perform the target task based on previous resources that has less average time in the execution of the tasks.

Based on these criteria, a resource recommendation approach is needed to find the most appropriate resources to work with previous resources based on co-working history.

B. Calculating Co-working History

The resource allocation based on co-working history approach mainly includes two parts:

Part 1, The *preprocessing steps*: The raw performance measures (RW) are generated from the implementation of the approach presented in [5]. Both of the event log and the raw performance measures (RW) are inputs to a set of preprocessing steps to obtain the co-working relationships. After extracting the event log and RW from [5], data preprocessing has been conducted as follows: (1) filter out cases in which sum METRIC-VALUE are less than 0, and cases that contain less than three activities, (2) choose the effective time as a measure for the proposed approach, (3) find the latest resource who performed each activity in each case within the event log and RW, and finally (4) detect whether the resource executes the same activity more than once in the same case. In the last step, the average is calculated as the effective time for the activity. As an output, the event log is ready to be used as an input to the proposed approach. Fig. 3 illustrates how to obtain and calculate *co-working* relationships.

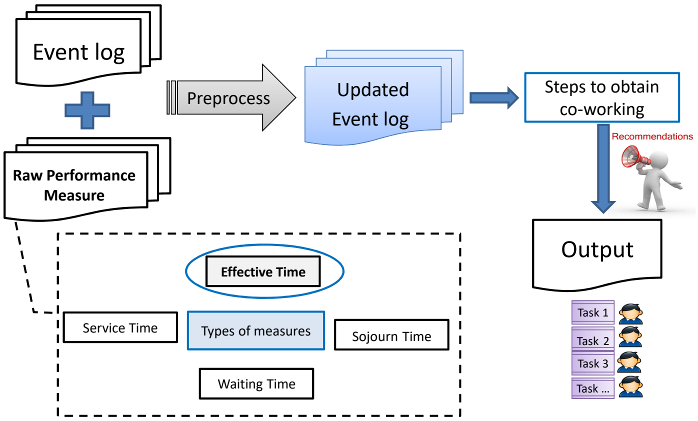


Fig. 3. An overview of the proposed approach.

Part 2, *Recommending a resource*: There are two steps to recommend a suitable resource for working with the previous resources for each activity based on the co-working history. They are as follows: (1) divide the event log into training and testing sets, (2) recommend resources for each activity according to the proposed criteria (*Frequency* and *Duration*).

V. Evaluation

The evaluation of the proposed approach is applied on two logs:

1) Synthesized Log: For the evaluation, a synthesized Log that had been generated from the ProM [26] plugin "Perform a simple simulation of (stochastic) Petri net" [27] was used. This log was taken from [5]. It contains 100 cases with a total of 4677 events, 10 activities, and 9 resources. For further references clarification in this paper, this log is referred to as $W1$.
2) Real Log: The approach has been applied on a real log from the Business Process Intelligence (BPI) Challenges. The log was taken from a Dutch Financial institute (referred to as $W2$), and it contains data that that represent the process of personal loans applications [1]. the log contains 13087 cases with a total of 262200 events, 25 activities and 69 resources.

The proposed approach has been implemented using Java and a relational database. And it has been tested on Windows 8 with 4G RAM and a Core i5 processor.

A. Co-working Effect on Process Performance

The aim of this section is to statistically prove the significance of the co-working relationships (team harmony) on resource recommendation and process performance. Definition 5: The co-working history works on finding, for a given task, $t \in T$, the different sets of trace histories that have common tasks and common resources performing them. Suppose that $W = \{\sigma_1 = \ll e_1(t_1, r_1, complete, tm_1), e_2(t_2, r_2, complete, tm_2), e_3(t_3, r_3, complete, tm_3), \cdots \gg, \sigma_2 = \ll e_{10}(t_1, r_1, complete, tm_{10}), e_{11}(t_2, r_2, complete, tm_{11}), e_{12}(t_3, r_3, complete, tm_{12}), \cdots \gg, \sigma_3 = \ll e_{100}(t_1, r_7, complete, tm_{100}), e_{101}(t_2, r_5, complete, tm_{101}), e_{102}(t_3, r_3, complete, tm_{102}), \cdots \gg\}$. If we consider task t_3, then the resulting co-working history will be $W_{<t_3} = \{\{\ll e_1(t_1, r_1, complete, tm_1), e_2(t_2, r_2, complete, tm_2) \gg, \ll e_{10}(t_1, r_1, complete, tm_{10}), e_{11}(t_2, r_2, complete, tm_{11}) \gg\}, \{\ll e_{100}(t_1, r_7, complete, tm_{100}), e_{101}(t_2, r_5, complete, tm_{101}) \gg\}\}$.

Here, traces σ_1 and σ_2 are grouped together because they have a common trace history for task t_3, the same tasks executed before t_3 with the same human resources. Trace σ_3 is in another set because it deviates from the other two traces with respect to the human resources.

In order to check whether the co-working history affects human resource performance of the task-in-hand, a statistical test was formulated where the statistical significance of such hypothesis was tested. The null hypothesis H_0 is that the harmony (common co-working history) is ineffective and has no influence on the performance of human resources. The alternative hypothesis states otherwise. That is, the co-working history has an influence on human performance. To test the hypothesis, a paired T-Test using unequal variance was applied. For this test, there is need to prepare two sets. The first set contains the time taken by the different human resources who executed the target task t with all cases (traces) in which t was executed. The second set contains human resources who executed t but within cases that have common co-working history.

[1] https://tinyurl.com/FinacialLog

To explain how the testing works to prove the hypothesis assumption, a set of steps on the event log has been applied: (1) Obtain the process model for the event log, (2) Select cases or traces that contain at least three activities and more, (3) Select the target activity; the chosen target activity is not the first one in the trace but must be preceded by a number of activities to have a co-working history, (4) Find all the resources who execute the target activity, (5) Find the effective time for all resources who perform the target activity for each case in the event log (situation 1), (6) Identify specific resources for the target activity, and finding all possibilities for co-working history to all activities that precede the target activity, c.f. Definition 5, (7) Compile similar groups in co-working history of all activities that precede the target activity (situation 2), (8) Run the paired T-Test with un-equal variance.

In this paper, the approach in [5] has been used to extract the event log and the performance indicators. To give an example about how the test works, trace from event log for process model was chosen as shown in Fig. 4. Table I illustrates a sample from the event log.

TABLE I. A SAMPLE EVENT LOG WITH SEQUENCES W-C,W-N,W-V

EVENTID	CASEID	RESOURCES	ACTIVITY	Effective-Time
1	173688	Dummy	W-C	8
2	173688	11049	W-N	0
3	173688	10629	W-V	32
4	173844	11201	W-C	6
5	173844	11049	W-N	0
6	173844	10629	W-V	15
.	.	.	.	.
511	173691	Dummy	W-C	6
512	173691	11049	W-N	1
513	173691	10809	W-V	7.33333
514	173913	Dummy	W-C	16
515	173913	10899	W-N	0
516	173913	10809	W-V	19.5
517	174511	10909	W-C	5
518	174511	11259	W-N	0
519	174511	10809	W-V	17
.	.	.	.	.
1468	173715	10912	W-C	11
1469	173715	10899	W-N	0
1470	173715	10138	W-V	4
1471	173751	Dummy	W-C	11
1472	173751	10899	W-N	0
1473	173751	10138	W-V	10
.	.	.	.	.

After choosing trace, the target activity $W-V$ was selected from the trace shown in Table I. Then, all the resources that perform this activity along with their effective time (18 resources) were provided. This step is referred to as (Situation 1). Table II shows all resources and their effective time in performing the target activity in all cases of the log. Then, $W_{<W-V}$,cf. Definition 5 was constructed, which contains the sets of activities that precede the target activity to get the co-working history of the target activity for each resource. This step is referred to as (Situation 2). Table III shows the special groups for each resource in each case according to the co-working history.

The paired T-Test used contains several tests for data analysis. Two tests were chosen. These two tests are Equal-Variance-Test and Unequal-Variance T-Test. The focus was on Unequal-Variance T-Test to prove the assumption, as it is the most common type of T-tests and the most used tests that cover large part in statistical test or hypothetical tests.

TABLE II. SAMPLE SITUATION 1 FROM TABLE I AND ACTIVITY W-V

EVENTID	CASEID	RESOURCES	ACTIVITY	Effective-Time
516	173688	10629	W-V	32
517	173844	10629	W-V	15
518	174105	10629	W-V	7
519	174141	10629	W-V	23
.	.	.	.	.
362	173694	10609	W-V	40
363	173790	10609	W-V	17
364	173817	10609	W-V	14
365	173868	10609	W-V	24
366	174009	10609	W-V	20
.	.	.	.	.
686	173691	10809	W-V	7.33333
687	173913	10809	W-V	19.5
688	174511	10809	W-V	17
689	174707	10809	W-V	16
.	.	.	.	.

TABLE III. SAMPLE SITUATION-2 FROM TABLE I AND ACTIVITY W-V

CASEID	RESOURCE	RESOURCE	RESOURCE	ACTIVITY	Effective
	T1	T2	Target	Target	Time
173688	Dummy	11049	10629	W-V	32
175798	Dummy	11049	10629	W-V	29
176000	Dummy	11049	10629	W-V	6
177317	Dummy	11049	10629	W-V	35
182101	Dummy	11049	10629	W-V	32
198107	Dummy	11049	10629	W-V	10.5
203648	Dummy	11049	10629	W-V	25
203702	Dummy	11049	10629	W-V	32
.	.	.	.	.	.
173844	11201	11049	10629	W-V	15
179456	11201	11049	10629	W-V	26
183910	11201	11049	10629	W-V	11
.	.	.	.	.	.
180187	11169	11259	10809	W-V	14
185771	11169	11259	10809	W-V	12.5
188317	11169	11259	10809	W-V	19.5
196228	11169	11259	10809	W-V	2.66667
201710	11169	11259	10809	W-V	6
205803	11169	11259	10809	W-V	11.25
.	.	.	.	.	.

To give an example, in situation 1, without co-working history for the resource 10629 the effective time according to the target activity $W.V$ in all cases is 32 min in case 173688, 15 min in case 173844, etc., cf. Table II. In situation 2, with co-working history for resources $Dummy, 11049\ and\ 10629$, the effective time in this group for resource 10629 is $32, 29, 6, 35, 32, 10.5, 25, 32$ min in all cases respectively. For the other group in situation 2, The effective time for the resource 10629 in co-working history with $11201\ and\ 11049$ is $15, 26, 11$; cf. Table III. Each group from situation 2 will be tested against situation 1 individually. Note that the size of situation 1 is not equal to the size of situation 2. The size of situation 1 is larger than the size of situation 2.

B. Process Performance Results

The results stated that there is a certain percentage of each resource confirming the assumption of this research which says that "the harmony among resources with co-working history has an influence on the human resource performance". The percentage, which confirmed the assumption for each resource that has performed the target activity, is calculated using the following equation:

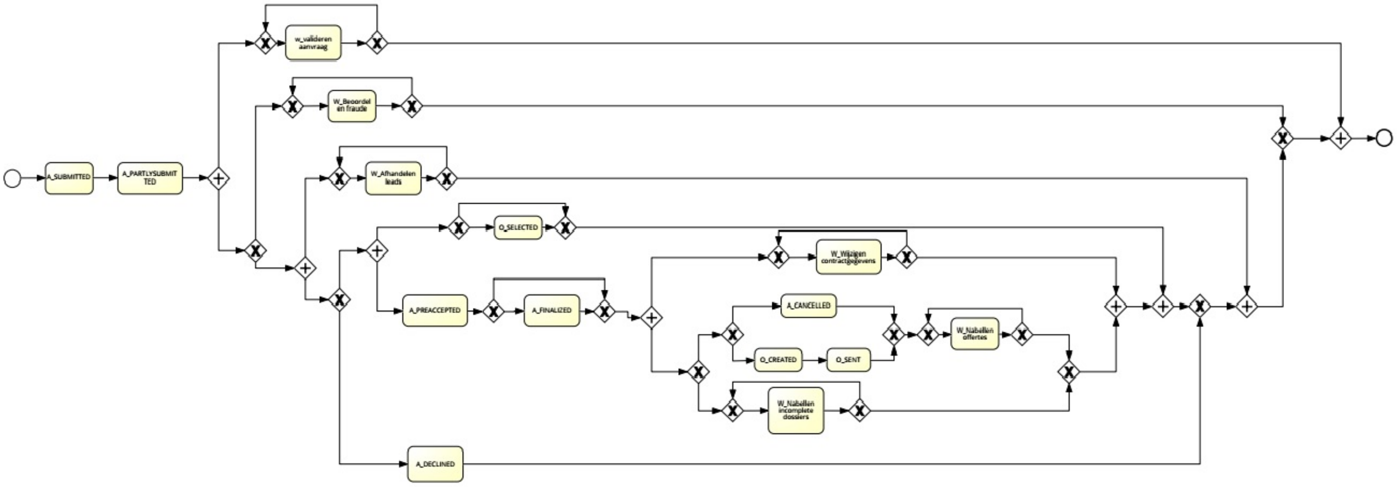


Fig. 4. Personal loans business process.

Co-working-Hypothesis =

$$\frac{\sum_{i=1}^{n} \text{ Cases When Reject } H_0\text{=Yes}}{\text{Count of Cases for all groups}}, \ \forall \ R \qquad (1)$$

n is the number of cases where H_0 is rejected and the count of cases for all groups is the summation of the cases for each resource with co-working history (situation 2). And the confidence level (CL) was 95 % when was the default value α = 0.05. Table IV shows the results of the statistical tests for real log and the percentages obtained by each resource in the provided example. These results prove that the test result for all the groups have proven the hypothesis for each resource.

TABLE IV. THE RESULTS OF THE STATISTICAL TESTS FOR REAL LOG BY UNEQUAL VARIANCE T-TEST

#	RESOURCES	Count of Group(cases)	Count Group Reject H_0 =Yes(cases)	Significant Different(%)
1	10629	80 (170 case)	45 (70 Casa)	0.41
2	10809	82 (165 case)	50 (58 case)	0.35
3	10609	81 (154 case)	51 (57 case)	0.37
4	10138	120 (361 case)	49 (63 case)	0.17
.	.	. .	.	.

As an example, when resource 10629 performed the target activity $W-V$, 80 groups with 170 cases which have common co-working history were formed. There are 45 groups among the 80 groups that confirmed the hypothesis (i.e., count of groups where H_0 was rejected is 45 groups out of the 80 groups). The number of cases in the 45 groups which confirmed the hypothesis is 70 cases. When (1) was applied on resources, the result of resource 10629, was $70/170 = 41\%$, 35% for resource 10809, 37% for resource 10609, etc.

C. Implementation and Experimental Evaluation

In this section, the proposed approach is described and applied along with details in order to verify the influence and the effectiveness of the harmony among resources. In order to obtain co-working relationships, the event log is preprocessed and split into training and testing sets (80% for training set, 20% for test set). In this part, the real-life event log (i.e. $W2$) was used to test the proposed approach. The approach was implemented in [5] which calculates the RW out of the input event log (Table VI). Table V shows a sample of the event log and Table VI shows a sample of the raw performance measures for cases 205715 and 205721 as an example.

TABLE V. A SAMPLE OF THE EVENT LOG

EVENTID	CASEID	RESOURCES	ACTIVITY	EVENT TYPE	TIMESTAMP
.	.	.	.	.	.
209686	205715	112	A-SUBMITTED	complete	2/1/2012 20:04
209687	205715	112	A-PARTLYSUBMITTED	complete	2/1/2012 20:04
209688	205715	112	W-Afhandelen leads	allocate	2/1/2012 20:04
209689	205715	10933	W-Afhandelen leads	start	2/1/2012 20:06
209690	205715	10933	A-PREACCEPTED	complete	2/1/2012 20:10
209691	205715	10933	W-Completeren aanvraag	allocate	2/1/2012 20:10
209692	205715	10933	W-Afhandelen leads	complete	2/1/2012 20:10
209693	205715	10933	W-Completeren aanvraag	start	2/1/2012 20:10
209694	205715	10933	A-ACCEPTED	complete	2/1/2012 20:17
.	.	.	.	.	.
209737	205721	10933	W-Completeren aanvraag	start	2/1/2012 20:26
209738	205721	10933	W-Completeren aanvraag	complete	2/1/2012 20:26
209739	205721	11119	W-Completeren aanvraag	start	2/1/2012 20:27
209740	205721	11119	W-Completeren aanvraag	complete	2/1/2012 20:27
.	.	.	.	.	.

The proposed approach needs preprocessing for the current event log in Table V. The event log is filtered using Table VI to remove all cases where the sum metric value for all resources is less than or equal to zero, and using Table V to remove all cases that contain less than three activities. Hence, only the cases that contain three or more activities are considered. Moreover, effective time is used as the main performance metric which is one of the measures extracted from [5].

After filtering both the event log and the raw performance measure (RW) tables, the event log was scanned to find the latest resource who performed each activity in each case within the event log. Then, these resources are linked with the performance measures when the event type = $complete$. For example, in Table V, the resource 10933 has *allocated* activity "$W - Completeren\ aanvraag$" for case 205721 at time tc = $2 - 1 - 201220 : 26$, and the resource 11119 has *allocated*

TABLE VI. A SAMPLE OF RAW PERFORMANCE MEASURE(RW)

CASEID	ACTIVITY	RESOURCES	OCCURRENCE	MEASURE TYPE	METRIC VALUE
.	.	.	.	.	.
205715	W-Afhandelen leads	10933	1	Effective	3
205715	W-Completeren aanvraag	10933	1	Effective	8
205715	W-Nabellen offertes	10933	1	Effective	1
205715	W-Nabellen offertes	11179	2	Effective	2
205715	W-Nabellen offertes	11181	3	Effective	0
205715	W-Nabellen offertes	10629	4	Effective	0
205715	W-Valideren aanvraag	10629	1	Effective	29
205715	W-Nabellen incomplete dossiers	10982	1	Effective	4
205715	W-Nabellen incomplete dossiers	11003	2	Effective	20
205715	W-Nabellen incomplete dossiers	11003	3	Effective	0
205715	W-Nabellen incomplete dossiers	11169	4	Effective	4
205715	W-Nabellen incomplete dossiers	10889	5	Effective	3
.	.	.	.	.	.
205721	W-Completeren aanvraag	10933	1	Effective	0
205721	W-Completeren aanvraag	11119	2	Effective	0
205721	W-Completeren aanvraag	11119	3	Effective	19
205721	W-Nabellen offertes	11119	1	Effective	2
205721	W-Nabellen offertes	11119	2	Effective	0
205721	W-Nabellen offertes	Dummy	3	Effective	1
205721	W-Nabellen offertes	11259	4	Effective	0
205721	W-Valideren aanvraag	10629	1	Effective	15
.	.	.	.	.	.

activity “$W - Completeren\ aanvraag$” for case 205721 at time $tc = 2 - 1 - 201220 : 27$, in this scenario, the resource 11119 is chosen as the latest one. This strategy is applied to all cases in the event log. Then, these recent resources are connected with the performance measures from Table VI.

TABLE VII. A SAMPLE OF THE HISTORY FOR EVENT LOG EXTRACTED FROM TABLE V

EVENTID	CASEID	RESOURCES	ACTIVITY	EVENT TYPE	TIMESTAMP
.	.	.	.	.	.
209686	205715	112	A-SUBMITTED	complete	2/1/2012 20:04
209687	205715	112	A-PARTLYSUBMITTED	complete	2/1/2012 20:04
209690	205715	10933	A-PREACCEPTED	complete	2/1/2012 20:10
209692	205715	10933	W-Afhandelen leads	complete	2/1/2012 20:10
209694	205715	10933	A-ACCEPTED	complete	2/1/2012 20:17
209695	205715	10933	A-FINALIZED	complete	2/1/2012 20:19
209696	205715	10933	O-SELECTED	complete	2/1/2012 20:19
209697	205715	10933	O-CREATED	complete	2/1/2012 20:19
209698	205715	10933	O-SENT	complete	2/1/2012 20:19
209700	205715	10933	W-Completeren aanvraag	complete	2/1/2012 20:19
209708	205715	10629	O-SENT BACK	complete	2/16/2012 15:36
209710	205715	10629	W-Nabellen offertes	complete	2/16/2012 15:36
209713	205715	10629	W-Valideren aanvraag	complete	2/16/2012 16:10
209723	205715	10889	O-DECLINED	complete	2/18/2012 13:26
209724	205715	10889	A-DECLINED	complete	2/18/2012 13:26
.	.	.	.	.	.
209743	205721	11119	A-FINALIZED	complete	2/1/2012 20:47
209748	205721	11119	W-Completeren aanvraag	complete	2/1/2012 20:47
209754	205721	11202	O-SELECTED	complete	2/6/2012 12:27
209755	205721	11202	O-CANCELLED	complete	2/6/2012 12:27
209756	205721	11202	O-CREATED	complete	2/6/2012 12:27
209757	205721	11202	O-SENT	complete	2/6/2012 12:27
.	.	.	.	.	.

TABLE VIII. A SAMPLE OF THE RESULT OF JOINING TABLES VI AND VII

EVENTID	CASEID	RESOURCES	ACTIVITY	EVENT TYPE	MEASURE	METRIC VALUE
.	.	.	.	.	.	.
15591	205715	10933	W-Afhandelen leads	complete	Effective	3
15592	205715	10933	W-Completeren aanvraag	complete	Effective	8
15593	205715	10629	W-Nabellen offertes	complete	Effective	0
15594	205715	10629	W-Valideren aanvraag	complete	Effective	29
15595	205715	10889	W-Nabellen incomplete dossiers	complete	Effective	3
15596	205721	11119	W-Completeren aanvraag	complete	Effective	0
15597	205721	11119	W-Completeren aanvraag	complete	Effective	19
15598	205721	11259	W-Nabellen offertes	complete	Effective	0
15599	205721	10629	W-Valideren aanvraag	complete	Effective	15
.	.	.	.	.	.	.

Next, if any activity which is executed by the same resource more than once is found in the same case, the average is calculated as the effective time for the activity. For example, in Table VIII, activity “$W - Completeren\ aanvraag$” is executed by the resource 11119 more than once in the case 205721, and the effective time for the activity “$W - Completeren\ aanvraag$” by the resource 11119 is $(0, 19)$ respectively. The average time is $(0 + 19/2 = 9.5)$. The same is applied for all the cases in the event log. Finally, Table IX illustrates the final result of the preprocessing steps.

After the preprocessing steps, the new event log is used as input for the proposed approach. This event log contains information about 3718 cases, 13704 events, 58 resources and 9 activities. The attributes for each case include EVENTID, CASEID, RESOURCE, ACTIVITY, and METRIC VALUE (Effective Time), cf. Table IX. This table is used to calculate the *co-working* relationships based on applying (Frequency and Duration Criteria) for recommending the resources based on co-working history. This co-working history verifies the influence of the harmony among resources on the performance of resources, where a significant difference has emerged.

TABLE IX. SAMPLE OF THE FINAL EVENT LOG FOR OUR APPROACH

EVENTID	CASEID	RESOURCE	ACTIVITY	METRIC-VALUE
.	.	.	.	.
10966	205715	10933	W-Afhandelen leads	3
10967	205715	10933	W-Completeren aanvraag	8
10968	205715	10629	W-Nabellen offertes	0
10969	205715	10629	W-Valideren aanvraag	29
10970	205715	10889	W-Nabellen incomplete dossiers	3
.	.	.	.	.
10971	205721	11119	W-Completeren aanvraag	9.5
10972	205721	11259	W-Nabellen offertes	0
10973	205721	10629	W-Valideren aanvraag	15
.	.	.	.	.

The event log data (cf. Table IX) is split into training and testing sets to obtain ***co-working*** relationships. The *training* set is used to extract the co-working relationships using SQL queries, which generate a *co-working relationship* table. This table is used to recommend the resource based on both (Frequency and Duration Criteria) after applying some SQL queries. On the other hand, the *test* set is used to compare the results before and after applying the proposed approach.

Table X presents some comparative examples before and after implementing the approach. It compares the original log (i.e., test set) and the output of the proposed resource rec-

TABLE X. Some Comparative Examples before and after Implementing of our Approach

		original log		Co-working Relationships			
				Frequency Criterion		Duration Criterion	
CASEID	ACTIVITY	RESOURCES	METRIC VALUE	RESOURCES	METRIC VALUE	RESOURCES	METRIC VALUE
205733	W-C	10932	20.00	10932	9.80	10932	9.80
205733	W-N	10789	0.00	11259	0.00	10138	0.00
205733	W-V	10138	8.00	10138	14.86	10629	8.84
.	.	.	.	.	.	.	.
205745	W-Afhandelen	11169	2.00	11169	4.53	11169	4.53
205745	W-C	11119	19.00	11189	14.00	11203	8.03
205745	W-N	10629	0.00	11259	0.28	10899	0.00
205745	W-V	10629	20.00	10138	13.98	10138	13.50
.	.	.	.	.	.	.	.
205766	W-C	11201	11.00	11201	9.07	11201	9.07
205766	W-N	11259	0.00	11049	0.67	10609	0.00
205766	W-V	11289	50.50	10138	15.99	10629	8.05
205766	W-N incomplete dossiers	11289	0.00	10899	0.27	10899	0.15
.	.	.	.	.	.	.	.

ommendation approach after applying frequency and duration criteria. For example, there are cases where each case records the resource which performs the task. In the case 205733 of the original log, resource10932 executes task $W-C$, resource 10789 executes task $W-N$, resource 10138 executes task $W-V$, and so on. Each resource has an effective time for its corresponding activity $(20.00, 0.00, 8.00$ min), respectively.

According to *frequency* criterion for resources recommendation, when the resource 10932 executes task $W-C$, the appropriate resource to execute task $W-N$ is 11259 with average time (0.00 minute). Hence, when the resource 10932 executes task $W-C$ and the resource 11259 executes task $W-N$, then the appropriate resource to execute task $W-V$ is 10138 with average time (14.86 min). While according to *duration* criterion, different resource recommendations are as follows: when the resource 10932 executes task $W-C$, the appropriate resource to execute task $W-N$ is 10138 with average time (0.00 min). Moreover, when the resource 10932 executes task $W-C$ and the resource 10138 executes task $W-N$, the appropriate resource to execute task $W-V$ is 10629 with average time (8.84 min).

Another example, in the case 205745 of the original log, the resource 11169 executes task $W-Afhandelen$, the resource 11119 executes task $W-C$, the resource 10629 executes task $W-N$, and the resource 10629 executes task $W-V$. Each resource takes an effective time for an activity $(2.00, 19.00, 0.00, 20.00$ min) respectively. Table X shows the different variations on both frequency and duration criteria after applying the proposed approach.

D. Evaluation Results

The evaluation of the results is based on synthesized and real life logs. In order to investigate whether the proposed approach contributes to get better results and improve the performance of tasks, (2) was used to calculate the overall result for applying the approach for the duration criterion.

$$\text{Overall}=\Sigma_{i=1}^{n}\text{ETB approach} - \text{ETA approach} \tag{2}$$

where n is the number of test case, and the overall represents sum of the total difference between before and after the proposed approach application according to the criterion of duration. The Effective Time Before (ETB) applying the approach represents the effective times of activities that resources have performed in the original log. While, the Effective Time After (ETA) applying the approach represents the effective times of activities that resources have performed after applying the criteria. Table XI summarizes the results of applying (2) on synthesized and real life logs. It shows resources recommendation based on the average time, the minimum time, and the maximum time to execute each activity in each case over all the log.

TABLE XI. Overall Co-Working Relationships

	Co-working Relationships (Duration Criterion)		
	Overall		
Logs	Avg	Min	Max
W_1	23279.16	62893.99	25998.2
W_2	10496.77366	6123.24469	-2091.40666

The total of the effective time after and before applying the proposed recommendation approach is computed using the following equation:

$$\text{Total Effective Time}=\Sigma_{i=1}^{n}\text{Effective time(A} \setminus \text{B)} \tag{3}$$

where n is the number of test cases, and the effective time for each test case (after(A) and before(B)) the recommendation is the summation of the effective time (after and before) applying the proposed recommendation approach.

In (4), the average of the effective time for test set (20%) before applying the proposed recommendation approach was computed.

$$\mathbf{Avg}_{BR} = \frac{\sum_{i=1}^{n}\text{Effective time (BR)}}{n} \tag{4}$$

where n is the number of test case, and the effective time for each test case before the recommendation (BR) is the summation of the effective time before applying the recommendation approach.

In (5), the average of the effective time for test set (20%) after applying the recommendation approach is computed.

$$\mathbf{Avg}_{AR} = \frac{\sum_{i=1}^{n}\text{Effective time(AR)}}{n} \tag{5}$$

TABLE XII. RESULTS OF APPLYING THE PROPOSED APPROACH ON REAL AND SYNTHESIZED EVENT LOGS

	Co-working Relationships			
	Frequency Criterion		Duration Criterion	
Logs	W_1	W_2	W_1	W_2
Sum Effective Time (min)	187384.85	17555.064	159901.054	12804.2593
Sum Effective Time (min) original	183180	22488.88615	183180	22488.88615
$\mathbf{Avg}_{BR}$	9641.052632	33.02332767	9641.052632	33.02332767
$\mathbf{Avg}_{AR}$	9862.360526	25.778361	8415.844947	18.8021429
Improvement Rate	-11.64778393 (min)	0.2193894 (min)	0.127082356 (min)	0.43064057 (min)

where n is the number of test case, and the effective time for each test case after the recommendation (AR) is the summation of the effective time after applying the proposed recommendation approach.

Table XII shows the results of applying the proposed approach on real and synthesized event logs. It uses (3), (4) and (5) on test set (20% of the event log). For the order fulfillment log ($W1$), the total of the effective time after applying the approach based on the criteria (Frequency, Duration) is 187384.85 min, 159901.054 min, respectively. The total effective time of the original log before applying the approach is 9641.052632 min. The $\boldsymbol{Avg}_{AR}$ after applying the approach recommendation based on the criteria (Frequency, Duration) is 9862.360526 min, 8415.844947 min respectively. On the other hand, $\boldsymbol{Avg}_{BR}$ of original log before applying the recommendation approach is 9641.052632 min.

For the Financial log ($W2$), the total of the effective time after applying the recommendation approach based on the criteria (Frequency, Duration) is 17555.064 min, 12804.2593 min respectively. The total effective time of original log before applying the approach is 22488.88615 min. The $\boldsymbol{Avg}_{AR}$ after applying the proposed recommendation approach based on the criteria (Frequency, Duration) is 25.778361 min, 18.8021429 min, respectively. On the other hand, $\boldsymbol{Avg}_{BR}$ of original log before applying the recommendation approach is 33.02332767 min.

The improvement rate of the proposed approach was calculated and evaluated by using the following equation:

$$\text{Improvement Rate} = (\mathbf{Avg}_{BR} - \mathbf{Avg}_{AR})/\mathbf{Avg}_{BR} \quad (6)$$

The results of the proposed approach have an improvement of the real data set and synthesized data set. The results show that the time is minimized to 0.2350476 min with frequency criterion and 0.43064057 min with duration criterion of the real data set. For synthesized logs, the results show that the time is minimized to 0.127082356 min with duration criterion, while the results state that the time is maximized to -11.64778393 min with frequency criterion. The negative value implies that the resources recommendation approach gives bad results.

The real data set has a bigger improvement because it contains a greater number of cases, activities and resources. In other words, considering co-working history for task allocation and resource recommendation is efficient. It also reduces process execution time significantly by taking resource harmony into account. Note that, more satisfactory results can be obtained as the number of process instances increases. The reason is that more event logs can generate more accurate harmony measurement which in turn provides more effective allocation recommendation.

VI. CONCLUSION AND FUTURE WORK

This paper has proposed a resource recommendation approach. This approach is built upon the co-working history from an event log. It considers the resources that had performed the previous tasks in the current running process instances, in order to recommend a resource that has the best harmony with the rest of the resources. This proposed approach focuses on the organizational perspective. It depends on a procedure-approach to extract time-related key performance indicators from process execution logs. This procedure-approach supports four measures: effective, service, waiting and sojourn time. The effective time measured was used in the proposed approach.

The proposed approach works to determine the criteria and the metrics from event log for resource recommendation. These criteria are (frequency and duration) based on the *co-working history*. The approach has been implemented and tested on both real and synthesized logs. The results show that it is possible to obtain the appropriate resources recommendation based on the criteria of co-working history. This approach has contributed to reducing the tasks time and to improving both the process and the resources performance.

As a future work, the researcher aims to add the co-working history approach as a new dimension and extend the related approaches for the resource recommendation with other algorithms.

REFERENCES

[1] M. Weske, *Business Process Management - Concepts, Languages, Architectures, 2nd Edition*. Springer, 2012.

[2] J. Xu, C. Liu, and X. Zhao, "Resource allocation vs. business process improvement: How they impact on each other," *Lecture Notes in Computer Science*, vol. 5240 LNCS, pp. 228–243, 2008.

[3] W. Zhao and X. Zhao, *Process Mining from the Organizational Perspective*, pp. 701–708. Springer, 2014.

[4] N. Russell, W. M. P. van der Aalst, A. H. M. ter Hofstede, and D. Edmond, "Workflow Resource Patterns: Identification, Representation and Tool Support," in *17th International Conference, CAiSE 2005, Porto, Portugal, June 13-17, 2005. Proceedings* (O. Pastor and J. and Falcão e Cunha, eds.), vol. 3520, pp. 216–232, Springer, 2005.

[5] N. M. Zaki, A. Awad, and E. Ezat, "Extracting accurate performance indicators from execution logs using process models," in *AICCSA*, pp. 1–8, IEEE, 2015.

[6] Y.-C. Chen, "Business Process Reengineering," *Encyclopedia of library and information science*, vol. 67, no. 1, p. 23, 2001.

[7] I. M. A. Helal, *Towards Monitoring Compliance of Business Processes*. PhD thesis, Cairo University, Giza, 2017.

[8] A. Awad, E. Pascalau, and M. Weske, "Towards Instant Monitoring of Business Process Compliance," *EMISA Forum*, vol. 30, no. 2, pp. 10 – 24, 2010.

[9] OMG, "Business Process Model and Notation (BPMN), Version 2.0," tech. rep., Object Management Group, January 2011.

[10] W. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer, 1st ed., 2011.

[11] M. Arias, J. Munoz-Gama, and M. Sepúlveda, "Introducing a taxonomy of human resource allocation criteria," tech. rep., Pontificia Universidad Católica, Santiago, Chile, 2017.

[12] S. Rinderle-Ma and M. Wil, "Life-cycle support for staff assignment rules in process-aware information systems," tech. rep., TU Eindhoven, 2007.

[13] Z. Huang, W. M. P. van der Aalst, X. Lu, and H. Duan, "Reinforcement learning based resource allocation in business process management," *Data Knowl. Eng.*, vol. 70, no. 1, pp. 127–145, 2011.

[14] A. Koschmider, Y. Liu, and T. Schuster, "Role assignment in business process models," in *BPM Workshops*, vol. 99 of *LNBIP*, pp. 37–49, Springer, 2011.

[15] Y. Liu, J. Wang, Y. Yang, and J. Sun, "A semi-automatic approach for workflow staff assignment," *Computers in Industry*, vol. 59, no. 5, pp. 463–476, 2008.

[16] R. Sindhgatta, A. Ghose, and H. K. Dam, "Context-aware analysis of past process executions to aid resource allocation decisions," in *Int'l Conf. on Adv. Info. Syst. Eng.*, pp. 575–589, Springer, 2016.

[17] Z. Huang, X. Lu, and H. Duan, "Mining association rules to support resource allocation in business process management," *Expert Syst. Appl.*, vol. 38, no. 8, pp. 9483–9490, 2011.

[18] T. Liu, Y. Cheng, and Z. Ni, "Mining event logs to support workflow resource allocation," *Know.-Based Syst.*, vol. 35, pp. 320–331, Nov. 2012.

[19] M. Arias, E. Rojas, J. Munoz-Gama, and M. Sepúlveda, *A Framework for Recommending Resource Allocation Based on Process Mining*, pp. 458–470. Springer, 2016.

[20] C. Cabanillas, J. M. García, M. Resinas, D. Ruiz, J. Mendling, and A. Ruiz-Cortés, *Priority-Based Human Resource Allocation in Business Processes*, pp. 374–388. Springer, 2013.

[21] Z. Huang, X. Lu, and H. Duan, "Resource behavior measure and application in business process management," *Expert Sys. with Apps.*, vol. 39, no. 7, pp. 6458 – 6468, 2012.

[22] W. Zhao, H. Liu, W. Dai, and J. Ma, "An entropy-based clustering ensemble method to support resource allocation in business process management," *Know. Apps. Info. Syst.*, vol. 48, no. 2, pp. 305–330, 2016.

[23] W. Zhao, L. Yang, H. Liu, and R. Wu, "The optimization of resource allocation based on process mining," in *Int'l Conf. on Intel. Comp.*, pp. 341–353, Springer, 2015.

[24] X. Liu, J. Chen, Y. Ji, and Y. Yu, "Q-learning algorithm for task allocation based on social relation," in *Int'l Workshop on Process-Aware Sys.*, pp. 49–58, Springer, 2014.

[25] A. Kumar, R. Dijkman, and M. Song, "Optimal resource assignment in workflows for maximizing cooperation," in *BPM*, pp. 235–250, Springer, 2013.

[26] H. Verbeek, J. Buijs, B. van Dongen, and W. van der Aalst, "ProM6: The process mining toolkit," in *BPM 2010 Demonstration Track*, CEUR Workshop Proceedings, (USA), pp. 34–39, CEUR-WS.org, 2010.

[27] A. Rogge-Solti, W. van der Aalst, and M. Weske, "Discovering stochastic Petri nets with arbitrary delay distributions from event logs," *BPM Workshops*, vol. 171, pp. 15–27, 2013.

Global Citation Impact rather than Citation Count

Gohar Rehman Chughtai, Jia Lee,
Muhammad Mehran Arshad khan
College of Computer Science, Chongqing University,
Chongqing 400044, China

Rashid Abbasi
College of Computer Science and Technology,
Anhui University, 230601,
Hefei, China

Asif Kabir
College of Communication Engineering, Chongqing
University, 400044, China

Muhammad Arshad Shehzad Hassan
State Key Laboratory of Power Transmission Equipment &
System Security and New Technology, School of Electrical
Engineering, Chongqing University, Chongqing 400044,
China

***Abstract*—The progressing bloom in the tome of scientific literature available today debars researchers from efficiently shrewd the relevant from irrelevant content. Researchers are persistently engrossed in impactful papers, authors, and venues in their respective fields. Impact of an article depends on the citation received but just a citation count can't give readers in-depth information about the article. That is the reason some articles are quantified unfairly on the basis of a citation count. In this paper, Global Citation Impact (GCI) is proposed which addresses the issue of considering citations of papers equally. Intuitively, the papers citing a paper are not of the same worth. The proposed index not only considers the number of citations (popularity) like many existing methods did but also considers the worth of citations (prestige). Results and discussions show that researcher whose work is cited by other prestigious papers gets higher rank which is quite fair crediting for research impact.**

***Keywords*—Citation weighting; popular; global citations; prestigious; Global Citation Impact (GCI); research impact**

I. Introduction

Currently, scientific work evaluation and quantification is an important research area for researchers, for the sake of unbiased and fair crediting of their research work. It is a challenging task for an index to be the best fit and acceptable to the whole scientific community. Scientific work evaluation is necessary to decide whether or not a researcher is promoted, is suitable as a principal investigator for a project, should get a PhD degree, should be given tenure or should be awarded an important research funding.

Traditionally, research work contribution crediting is done by involving number of papers and number of citations by the state of the art H-Index [1], G-Index [2] q2-index [3] and all the variant of H-index such as A-index [4], R-index and AR-index [5], w-index a significant improvement of H-index [6], fractional counting of authorship [7], Weighted Citation [8] and E-index [9], [10]. Reviews various indicators that can possibly be used to measure the performance of an author. These variations of H-index have considered the citations as their target to quantify the scientific work, but they punish the new emerging scientists because of lesser citations received. To address this issue career length is considered in the quantification of research to address the ignorance of new scientists in the field by m-quotient [4]. Later, co-authorship contributions [11], such as the order of authors contributing to paper and a number of authors in a paper is considered by Kth-rank [12] and fractional H-index. The self-citations should not be given the same weight as citations by others issue is investigated and F-Index [13] was proposed. Finding the rising star in academia [14] where the star is the authors who have not enough citation at the start of their career but predicted as a rising star in the future. Their contribution actually highlights the new researchers irrespective of traditional indexing methods to give credits to researchers on the basis of citations and number of publications. Topic-based ranking of authors [15] and consistent annual citation based index [16], identifying authorities of a given topic within a particular domain [17] which is a great contribution to expert finding. Impact of hot paper on individual's research contribution demonstrated by [18] which is an encouragement for new emerging researchers. The IF of a journal, therefore, is not representative of the number of citations of an individual article in the journal [19].

Existing indexing methods have majorly considered the number of citations (popularity), researcher age factor, the order of authors (co-author contribution) and self-citations, while prestige (worth of citing papers) was ignored. The difference between popularity and prestige is provided in the following example. Suppose a researcher has two papers A and B and both are cited by 10 papers. Eight of the papers that cited paper A are written by the prestigious researchers while only three of the papers cited paper B are written by prestigious researchers. Though the number of citations is the same for both paper A and B, it seems that paper A is more useful for researcher A, due to getting the attention of eight prestigious researchers. In this work, considering the worth of papers citing a paper is referred to as the impact of global citations. Global Citation Index (GC-Index) is proposed which considers the worth of citing papers for indexing researchers. The more the worth of the citing papers of a paper the more the prestige it has and is highly important for the scientific community.

The major contributions of this paper towards researcher indexing are (1) differentiating between popularity (citation count) and global impact of a paper (2) highlighting positive impact of prestigious paper citing a paper and (3) proposal of PageRank based method to calculate the global impact of a paper which shows significance of paper. To the best of our knowledge, this work will give more insight into the significance of an article which could be helpful for readers to find the quality article.

II. Global Citation Impact (GCI)

The popularity of an article depends on the number of citation received by an article without knowing the worth of citing the paper. In past publication count and citation count is extensively used in author ranking in most of the bibliometric indicators, but this way we cannot find the worth of the article if we don’t know the citing article and author. H-index is generally acknowledged by major online databases, albeit, one blemish of H-index is that its value never goes down over time.

Scientific impact measurement achieves a move from popularity to the prestige of scientific productions. Since PageRank is acquainted with the scholastic assessment. In spite of these researchers have advanced the improvement of scholarly impact assessment, the most eminent constraint of the PageRank-based assessment tool is estimating the prestige of citation-publication network. The PageRank algorithm and its variants were used for the assessment of various types of citation-publication networks. The question has been raised whether better assessment results were depended directly on an author network or on a publication-citation network [20].

PageRank calculation without anyone else is initially intended to use web pages ranking, web pages propagations are considered as equivalent significance, accentuating the restrictive measurement. Specifically using PageRank to assess the academic impact regards all citations as equivalent weights, which adequately ignores the impact from citing authors. HR-PageRank, incorporate weighted PageRank according to individual’s H-index, and pertinence between citing and cited papers [21]. We contend that the PageRank-index is an impartial and more nuanced metric to evaluate the publication records of researchers contrasted with existing measures [22].

Basically, a global citation impact is the weighted impact of the citation on the basis of the global weight of its citing article which shows how worthy article is, not just popularity. In other words, if a paper is cited by impactful paper then it is more worthy because it attracts a more impactful audience. In this paper, the impact of global citations is considered. It is quite reasonable to think that a paper cited by many prestigious papers is likely to be more worthy as compared to a paper cited by not prestigious papers. Fig. 1 provides an example of four nodes where vertices represent papers and edges represent citations. It demonstrates that how the global citation impact is effective in the provision of more insight into paper significance.

A directed paper-citation graph in Fig. 1 is a pair $(V;E)$, where edges point toward and away from vertices. For a vertex $v_i \in V; In(v_i)$ is the set of vertices that point to it and $Out(v_j)$ is the set of vertices that v_i point to, such that:

$$In(v_i) = \{v_j \in V(v_j, v_i) \in \varepsilon\} \quad (1)$$

$$Out(v_j) = \{v_j \in V(v_i, v_j) \in \varepsilon\} \quad (2)$$

$$V(v_i) = In(v_i) \cup Out(v_i) \quad (3)$$

In Table I, rows show the *in*-degree of nodes while columns show the *out*-degree of nodes calculating prestige of paper using PageRank.

In Table II, P1 got the highest global weight because it is cited by P2 which has the highest local weight and P4. Every citation has different weight then it should be considered in the researcher indexing. The local weight is calculated by using PageRank’s (4) and global weight is the summation of the paper’s local weights citing a paper.

$$S(v_i) = (1-\sigma) + \sum_{j \in V(Vj)} \frac{S(v_j)}{\left|V(v_j)\right|} \quad (0 \le \sigma \le 1) \quad (4)$$

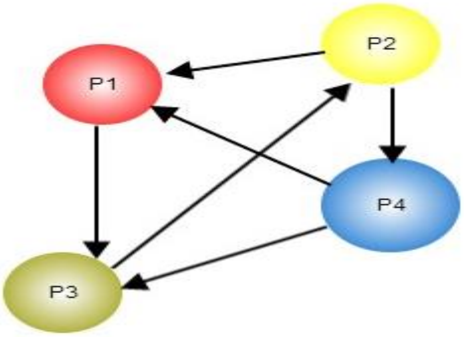


Fig. 1. Paper-citation graph.

TABLE I. Matrix of Degrees of Nodes

Paper	P1	P2	P3	P4
P1	0	½	0	1/2
P2	0	0	1	0
P3	1	0	0	1/2
P4	0	½	0	0

TABLE II. Global and Local Weight

Paper	Local weight	Global weight	W_{cit} 1	W_{cit} 2	W_{cit} 3	W_{cit} 4
P1	1.25	3.25	0	2.50	0	0.75
P2	2.50	2	0	0	2	0
P3	2	2	1.25	0	0	0.75
P4	0.75	2.50	0	2.50	0	0

$S(v_i)$ and $S(v_j)$ denote the score of vertex v_i and v_j respectively, $V(v_i)$ and $V(v_j)$ denote the set of vertices connecting with v_i and v_j respectively, and σ is a damping factor that integrates into the computation the probability of jumping from a given vertex to another random vertex in the graph.

Originally, Google PageRank algorithm assumed the parameter d to be 0.15. This value was incited by the informal observation that a person surfing the web will usually follow the order of 6 hyperlinks, equivalent to a leakage probability d = 1/6 ≃ 0.15, before becoming either bored or frustrated with this line of search and begin a new search. In the context of citations, it is common hypothesis that entries in the reference list of a typical paper are collected following somewhat shorter paths of average length 2, making the choice d = 0.5 more appropriate for a similar algorithm applied to the citation network. is that approximately 50% of the articles [23].

In our scenario, GCI is computed as

$$W_{local}(n) = \sigma(\frac{1}{G}) + (1-\sigma)\sum_{m\in L(n)} \frac{W_{local}(m)}{C(m)} \qquad (5)$$

Where G is the total number of nodes (papers) in the graph, σ is the random damping factor (n) is the set of papers that link (cite) n and $C(n)$ is the out-degree of the node m.

Therefore,

Global citation weight of the citation is as follow:

$$W_{Global}(n) = \sum_{m\in L(n)} W_{local} \qquad (6)$$

Whereas $W_{Global}(n)$ is the global weight of the node (paper). To index the author on the basis of global citation weight as H-index does use citation count we have arranged the documents in descending order on the basis of global citation weight. The index which satisfies, $r_0(n) \le W_{Global}(n)$, where r_0 is the highest rank, will be the GCI-rank of an author.

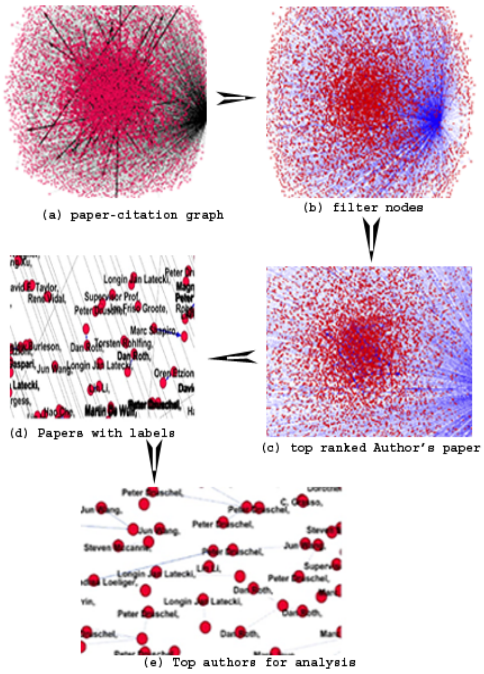
III. Experiments

Authors have carried out a series of experiments which is discussed in detail in this section in which analysis for the changes in the ranks of researchers is provided on the large real bibliographic dataset.

A. Dataset

Initially, we have taken 1000 researchers from Cite seer then completed their graph. In this process of getting the citing papers of each paper of a researcher, a total number of authors becomes 24567 and the total number of publications is 140883. Our data set consists of data variables PID (Publication ID), Authors, in links, out links and citations. Authors have removed inconsistencies and duplications in the dataset in order to get accurate results and completed the data set by extracting each publication of an author, its citing publications and completed the matrix of the dataset. Authors have chosen 30 authors having a difference in their rank from 50, 40, 30, 20, 10 ranges of H-index rank so that to analyze the variation with respect to proposed GCI-Impact ranks.

Fig. 2 represents the paper-citation graph and the process of selection of top authors for analysis. In graph red nodes represents the papers of an author and blue lines are the edges which represent the *in* degree and the *out*-degree of an article. In the following Table III, authors have given some statistics of data set used for simulations.



(a) paper-citation graph

(b) filter nodes

(c) top ranked Author's paper

(d) Papers with labels

(e) Top authors for analysis

Fig. 2. Paper-citation Directed Graph.

TABLE III. Cite Seer Dataset

No. of Authors	**No. of publication**	**Average No. of Citation**	**Average No. of Global Citation**
24567	140883	9.80	55.11

B. Parameter Settings

In H-index square root of total citations of all publications of an author are divided by Proportionality constant "*a*" whose values range from 3 to 5. Authors have set the value of "*an*" equal to 4 because it's a minimum number of publications and number of citations per paper per year of an author. Higher the value of "*a*" higher is the consideration of highly cited publications into H-index core which ignore the real participating articles in the *h* core.

Authors have tested GC-Index on two different values of “damping factor”, 0.85 which is used in PageRank calculation and 0.50 to represent paper citation network according to [23]. They have shown that average citation link in the citation network for an academic researcher is ½ that’s why 0.50 suits more for paper citation network.

C. Results and Discussions

The results of the proposed GCI are compared with H-Index. Authors have chosen 30 researchers randomly for our analysis. Table IV shows papers, citations, the rank of authors according to H-Index and GCI-rank, average citations, average global citations, and variation in the rank. Average G_{cit} shows that how much the citing paper is important and how much it serves the audience? If a paper attracts well-known researchers then it definitely means that work has more worth. Knowing that how much worth the citing articles have is very important to analyze the quality of one’s research. GC-Index shows the variation in rank along with variation in damping factor on 0.50 and 0.85, which also affect the results.

1) Comparison of H-Index and GCI-rank with damping factor 0.85

Table IV has shown the impact of global citations on author rank with damping factor 0.85. Their scenarios are discussed, that is; position earned, a position lost and position stable by authors with respect to H-Index. In Table IV, variation in rank column + symbol means position earned, -symbol means position lost and 0 means position stable.

TABLE IV. COMPARISON OF RANKS ON DAMPING FACTOR 0.85

S No	Authors	Publications	Citations	Average cit	Average G_{cit}	H-Index	GC-Index	H-rank	GC-rank	Variation in rank
1	Peter Druschel	124	10185	82.137	1374.121	50	52	1	1	0
2	Oren Etzioni	128	5516	43.093	622.125	44	30	2	4	+2
3	Dan Roth	157	4022	25.617	438.051	35	32	3	3	0
4	Lin Li	292	4661	15.962	178.537	32	21	4	7	-3
5	Steven Mccanne	53	4248	80.150	1689.264	30	40	5	2	+3
6	Jun Wang	268	3872	14.447	152.947	26	15	6	10	-4
7	Judea Pearl	92	2048	22.260	353.434	25	27	7	5	+2
8	Hao Che	160	2353	14.706	126.618	24	14	8	11	-3
9	Marc Shapiro	69	1775	25.724	365.144	22	19	9	8	+1
10	Eyal Amir	60	1173	19.55	299.733	21	15	10	10	0
11	Ira Cohen	57	1252	21.964	262.368	21	13	10	12	-2
12	Giedrius Slivinskas	60	1265	21.083	158.216	20	11	11	14	-3
13	Robert Nieuwenhuis	59	1058	17.932	259.186	19	14	12	11	+1
14	Dorothea Wagner	110	1140	10.363	108.863	19	11	12	14	-2
15	Longin Jan Latecki	86	1188	13.813	169.976	19	19	12	8	+4
16	Jan Friso Groote	75	1006	13.413	117.906	19	16	12	9	+3
17	Marc Joye	77	915	11.883	101.103	17	16	13	16	-3
18	Shigang Chen	45	1223	27.177	320.755	17	13	13	12	+1
19	Kristian Torp	46	923	20.065	142.782	16	11	14	14	0
20	S. Keshav	38	805	21.184	254.263	16	12	14	13	+1
21	Jian Chen	83	831	10.012	91.831	16	14	14	11	+3
22	Yang Yu	53	688	12.981	111.207	15	12	15	13	+2
23	Judith Donath	39	554	14.205	161.743	14	11	16	14	+2
24	Wei Sun	76	588	7.7368	95.210	13	7	17	16	+1
25	Michal Feldman	25	608	24.32	273.52	13	11	17	14	+3
26	Hans-rea Loeliger	56	1920	34.285	588.75	12	23	18	6	+12
27	Y. Rich	21	553	26.333	331.047	12	10	18	15	+3

28	Abhishek Kumar	12	322	26.833	392	10	7	19	16	+3
29	A. Bondavalli	22	173	7.863	67.090	7	7	20	16	+4
30	Gerrit Muller	24	66	2.75	12.333	4	2	21	17	+4

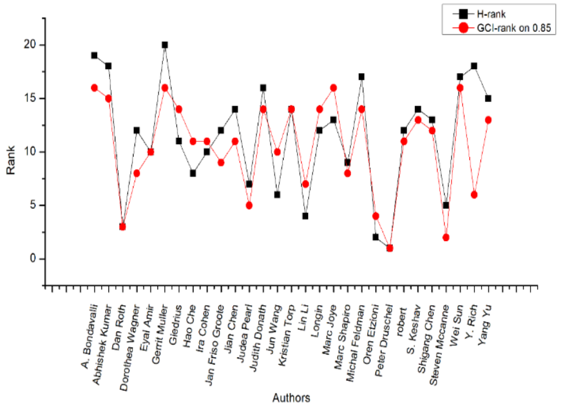


Fig. 3. Comparison of H-index and CGH with damping factor 0.85.

2) Scenario 1: Relocation with respect to H-Index: position up

Fig. 3 depicts the variation of ranks of researchers on damping factor 0.85 and more detailed elaboration of their results is explained in Table IV Judea Pearl who's H-Index was 25 and rank at position 7, with GC-Index his rank is increased by 2. This is because of his two publications having citations from worthy nodes which enhanced its rank. Another case in which the author gets a higher rank is the highest variation in selected authors for analysis. Hans-rea Loeliger having 56 and H-index is 12 but according to our proposed GC-Index, its index increased to 23. His index increased by 11 because its average global citation is quite higher, which is 588.75. He should be given higher rank because his work stimulated many other worthy papers in that domain. S. Keshav and Judith Donath, have almost same number of publications and their H-Index is 16 and 14, respectively. S. Keshav has average global citations 254.263 and Judith Donath has average global citations 161.743 on the basis of this S. Keshav has more worthy citations which so he should have a higher rank than Judith Donath. With respect to our proposed GC-Index, S. Keshav earned 7 positions higher than in H-Index rank while Judith Donath got just 2 positions high in with GC-Index.

3) Scenario 2: Relocation with respect to H-index rank: position down

Some authors have enough citations of their publications and they have higher h index but beside citations, according to their global weight, these citations are not so important because these are not weighted heavily enough to include it into its effective rank calculation. That's why these authors have lost their position with respect to CGH-index. As in h index, every citation of a publication matters a lot. In case of CGH-index not only the citation is considered but it must be an important publication to be the part of CGH-index core.

Authors have lost their rank in CGH-index even though some of them have a higher average global citation. It is just because they have few publications which have a more weighted citation but few are not so important that could include them in the CGH-index core. On average they have a high global citation but when computing CGH-index individual publication should have more weighted citations which can include them in CGH core. As in Table V two authors Ira chohen and GiedriusSlivinskas have almost an equal number of publication 57 and 60 and 1252, 1265 citations respectively. Both have lost the rank in CGH-index. Irachohen whose h index is 21 and rank 10 he lost rank by 2 and index by 8 while GiedriusSlivinskas has,h index 20 and rank 11 has lost his h rank by 3 and CGH-index by 9. GiedriusSlivinskas has lost more than Irachohen even though he has more publication and citation but he has average global citation 158.216lesser than average global citation of Irachohen whose average global citation is 262.368 which effects its rank and index more than Irachohen. Oren Etzioni has the highest global citation weight in Table VI that's 622.125 but still, he lost his rank by 2 because it is replaced by Steven Mccanne, whose G_{cit} is 1689.264 which is high enough than Etzioni which ranked him by 3.

4) Scenario 3: Position Stable With Respect to H-index

Peter Druschel has H-index 50 and his CGH-index is 52 which was already at first rank in h index rank other authors in Table VII has lost CGH-index but their rank remains stable. Damping factor is also important in graph-based ranking as we varied the damping factor rank of authors is also a varied little bit.

5) Comparison of H-Index and GC-Index on damping factor 0.50

Authors have discussed the comparison of CGH-index in following three different scenarios. Following scenarios clearly, describe the variation in results by varying the damping factor.

As in Table V first author, Steven Mccanne got higher rank which increases by 3 and he replaced Oren Etzioni because Steven has 53 publication 4248 citations and 1689.264 average global citation count while Etzioni has 128publication 5516 citation and 622.125 global citation count. Although he has more publications and citations, his global citation count is very less than Steven that's why he has replaced him and got position 2. In another case author Hans-rea Loeliger whose publication is 56, citations are 1920 and average global citation count is 588.75 earned position by 13. No other author has a high average global citation but they have a higher citation which gives them higher rank in H-index. Fig. 5 shows the gain in rank with respect to CGH-index.

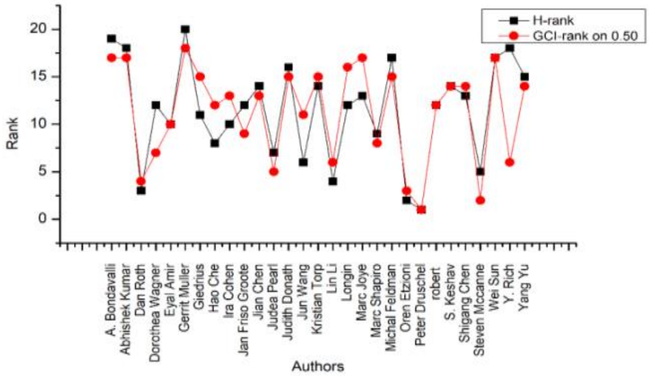


Fig. 4. Comparison of H-Index and CGH with damping factor 0.50.

Fig. 4 shows the variation of ranks among different authors on damping factor 0.50 which are explained in Table VI, first author Oren Etzioni has lost rank by 1 and he has been replacing by Steven whose average global citation count is quite higher than Oren which is 1689.264 and Oren has just 622.125. Dan Roth has,h-index rank 3 but here he lost rank by 1 and replaced by Oren Etzioni whose average global citation count is 622.125 while Dan Roth has 438.051 which is lesser than him that’s why he could not maintain his rank.

In Table VII first author Peter Druschel has maintained his position on H-index rank on both damping factor values because he has highest average Global citations. Some of the authors have lost their CGH-index but when ranked according to it, they have maintained their position with respect to H-index because variations occur in other’s rank.

TABLE V. SCENARIO 1 “RELOCATION WITH RESPECT CGH-INDEX RANK: POSITION UP”

S No	Authors	publications	citations	Average G_{cit}	H-index rank	Earned Position on 0.50
1	Steven Mccanne	53	4248	1689.264	5	+3
2	Judea Pearl	92	2048	353.434	7	+2
3	Longin Jan Latecki	86	1188	169.976	12	+3
4	Hans-rea Loeliger	56	1920	588.75	18	+13
5	Robert Nieuwenhuis	59	1058	259.186	12	+1
6	Kristian Torp	46	923	142.782	14	+1
7	Marc Shapiro	69	1775	365.144	9	+3
8	Judith Donath	39	554	161.743	16	+2
9	Michal Feldman	25	608	273.52	17	+2
10	Y. Rich	21	553	331.047	18	+2
11	Abhishek Kumar	12	322	392	19	+1
12	A. Bondavalli	22	173	67.090	20	+3
13	Gerrit Muller	24	66	12.333	21	+2

TABLE VI. “SCENARIO 2: RELOCATION WITH RESPECT TO CGH-INDEX RANK ON 0.50: POSITION DOWN”

Authors	Publications	Citations	Average G_{cit}	H-index rank	Position Lost
Oren Etzioni	128	5516	622.125	2	-1
Dan Roth	157	4022	438.051	3	-1
Lin Li	292	4661	178.537	4	-4
Jun Wang	268	3872	152.947	6	-1
HaoChe	160	2353	126.618	8	-5
Eyal Amir	60	1173	299.733	10	-2
GiedriusSlivinskas	60	1265	158.216	11	-4
Dorothea Wagner	110	1140	108.863	12	-3
Marc Joye	77	915	101.103	13	-4
Ira Cohen	57	1252	262.368	10	-3

TABLE VII. "SCENARIO 3: POSITION STABLE"

Authors	**Publication**	**Citations**	**Average G_{cit}**	**H-index rank**	**CGH-index rank**
Peter Druschel	124	10185	1374.121	1	1
Jan Friso Groote	75	1006	117.906	12	12
Shigang Chen	45	1223	320.755	13	13
S. Keshav	38	805	254.263	14	14
Jian Chen	83	831	91.831	14	14
Yang Yu	53	688	111.207	15	15
Wei Sun	76	588	95.210	17	17

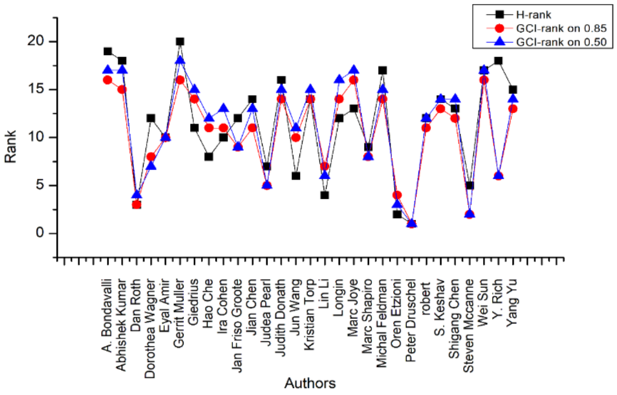


Fig. 5. Comparison of H-Index and CGH with varies damping values.

6) *Varied damping factor study*

This is the comparison of the ranks of authors which is clearly depicted in Fig. 5. When the value of the damping factor increases the global citation weight also increases, which ultimately increases the rank of authors.

Authors who have maintained their rank have more influence in the citation network because they got citations from the important nodes in this citation network which excels their weights. We have seen that as the damping factor increase rank of scientist also increases. Results on 0.50 damping factor are very close to the baseline method's result. Authors who have maintained their index with lower damping factor mean they have influence in the citation network.

Fig. 5 depicts the variations of rank with respect to H-rank on varied damping factor. The scholastic approach is adopted to see the variation in rank when considering the impact of global citation weight rather than just counting the citations into the ranking. This variation in rank gives us more insight into the worth of an article. Traditionally, the worth of an article is quantified on the basis of citation count which can give in detail information about the prestige of an article.

IV. CONCLUSIONS AND FUTURE WORK

Evaluating scientific research production is a challenging task. Authors have proposed GCI which calculates the global citation weight of each publication and index them. It provides more credit to the researchers whose work penetrates many well-known researchers which is quite fair and acceptable quantification. It is concluded that not only the number of citations could affect his/her index but also worth of citing papers is important. Results and discussion show the useful impact of global citations for researcher indexing. It is clear that if citing paper has received more citation then it means it serves more audience and more important for the scientific community. The analysis shows that even if a paper received lesser citations than any other paper but if its citing papers are worthy then it is better.

F-index considers the co-terminal citations into indexing and shows that how the scientific works penetrate into different scientific communities. As a future work, the global citation impact of unique citing authors in addition to simply considering their count can be added in F-Index for improved researcher indexing.

ACKNOWLEDGMENTS

I am thankful to my respected professor for his support and encouragement. I am also grateful to all my co-authors for their valuable contributions.

REFERENCES

[1] J. E. Hirsch, "An index to quantify an individual's s scientific research output," Proc. Natl. Acad. Sci. U.S.A., vol. 102, no. 46, pp. 16569–16572, 2005.

[2] L. Egghe, "Theory and practice of the g -index," Scientometrics, vol. 69, no. 1, pp. 131–152, 2006.

[3] F. J. Cabrerizo, S. Alonso, E. Herrera-Viedma, and F. Herrera, "q 2 -Index : Quantitative and qualitative evaluation based on the number and impact of papers in the Hirsch core," vol. 4, pp. 23–28, 2010.

[4] Q. L. Burrell, "On the h-index, the size of the Hirsch core and Jin's A-index," J. Informetr., 2007.

[5] B. Jin, L. Liang, R. Rousseau, and L. Egghe, "The R-index and AR-indices: Complementing the h-index," Chinese Sci. Bull., vol. 52, no. 6, pp. 855–863, 2007.

[6] Q. Wu, "The w-index: A significant improvement of the h-index," J. Am. Soc. Inf. Sci. Technol., vol. 10, no. May, pp. 609–614, 2008.

[7] L. Egghe, "Mathematical theory of the h-index and g-index in case of fractional counting of authorship," J. Am. Soc. Inf. Sci. Technol., 2008.

[8] E. Yan and Y. Ding, "Weighted citation: An indicator of an article's prestige," J. Am. Soc. Inf. Sci. Technol., vol. 61, no. 8, pp. 1635–1643, 2010.

[9] C. T. Zhang, "The e-Index, Complementing the h-Index for Excess Citations," PLoS One, vol. 4, no. 5, 2009.

[10] L. Wildgaard, J. W. Schneider, and B. Larsen, "A review of the characteristics of 108 author - level bibliometric indicators," Scientometrics, vol. 101, no. 1, pp. 125–158, 2014.

[11] M. Ausloos, "A scientometrics law about co-authors and their ranking: The co-author core," Scientometrics, vol. 95, no. 3, pp. 895–909, 2013.

[12] Q. C. Contributions, "edited by Jennifer Sills Biofuels : Clarifying," October, pp. 4–7, 2006.

[13] D. Katsaros, L. Akritidis, and P. Bozanis, "The f index: quantifying the impact of coterminal citations on scientists' ranking," J. Am. Soc. Inf. Sci. Technol., vol. 60, no. 5, pp. 1051–1056, 2009.

[14] A. Daud, R. Abbasi, and F. Muhammad, "Finding rising stars in social networks," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2013, vol. 7825 LNCS, no. PART 1, pp. 13–24.

[15] T. Amjad, Y. Ding, A. Daud, J. Xu, and V. Malic, "Topic-based heterogeneous rank," Scientometrics, vol. 104, no. 1, pp. 313–334, 2015.

[16] A. Daud and F. Muhammad, "Consistent Annual Citations based Researcher Index," Collnet J. Sci. Inf. Manag., vol. 8, no. 2, pp. 209–216, 2014.

[17] T. Amjad and A. Daud, "Indexing of authors according to their domain

of expertise," vol. 22, no. 1, pp. 69–82, 2017.

[18] G. Rehman and J. Lee, "Quantifying the Impact of Hot-paper on new Researchers," Proc. 2017 2nd Int. Conf. Commun. Inf. Syst. （ICCIS 2017）, pp. 329–334, 2017.

[19] L. Waltman and V. A. Traag, "Use of the journal impact factor for assessing individual articles need not be wrong," pp. 1–32, 2017.

[20] M. Nykl, K. Ježek, D. Fiala, and M. Dostal, "PageRank variants in the evaluation of citation networks," J. Informetr., vol. 8, no. 3, pp. 683–692, 2014.

[21] F. Zhang, "Evaluating journal impact based on weighted citations," Scientometrics, vol. 113, no. 2, pp. 1155–1169, 2017.

[22] U. Senanayake, M. Piraveenan, and A. Zomaya, "The PageRank-index: Going beyond citation counts in quantifying the scientific impact of researchers," PLoS One, vol. 10, no. 8, 2015.

[23] P. Chen, H. Xie, S. Maslov, and S. Redner, "Finding scientific gems with Google's PageRank algorithm," J. Informetr., vol. 1, no. 1, pp. 8–15, 2007.

Scientific Articles Exploration System Model based in Immersive Virtual Reality and Natural Language Processing Techniques

Luis Alfaro, Ricardo Linares, José Herrera
National University of San Agustin
Santa Catalina 107 - Arequipa, Perú

Abstract—**After having carried out a historical review and identifying the state of the art in relation to the interfaces for the exploration of scientific articles, the authors propose a model based in an immersive virtual environment, natural user interfaces and natural language processing, which provides an excellent experience for the user and allows for better use of some of its capabilities, for example, intuition and cognition in 3-dimensional environments. In this work, the Oculus Rift and Leap Motion Hardware devices are used. This work aims to contribute to the proposal of a tool which would facilitate and optimize the arduous task of reviewing literature in scientific databases. The case study is the exploration and information retrieval of scientific articles using ALICIA (Scientific database of Perú). Finally, conclusions and recommendations for future work are laid out and discussed.**

Keywords*—*Immersive virtual environment; human computer interaction; natural user interfaces; natural language processing; Oculus Rift

I. Introduction

The production of scientific articles, which are the result of lines of research in various fields and which must be made available to research communities, are currently indexed in databases or scientific repositories, such as Scopus, Web of Science, Latindex, etc., which allow one to perform searches, explore the database, and obtain information. These computer systems generally use traditional interfaces composed of forms, windows, and menus which can even utilize hypertext or hypermedia resources. In addition to using keywords, the name of the author, or the year of publication, among other search parameters, the traditional interface presents the results generally as a textual list [1]. However, the information indexed in the databases is growing rapidly, thus creating difficulties for the exploration and adequate selection of scientific articles when using traditional interfaces and making the researchers searches tedious without any guarantee that they will actually find what they are looking for.

Furthermore, the traditional interaction conguration is limited. Common devices like the mouse and keyboard are used, whereas in the virtual environment, interaction possibilities are broader and more intuitive to the user due to elements existing in the users context in the immersive environment. Virtual Reality can provide solutions to the problems mentioned above [2], [3], as well as new opportunities to search for articles, for example, creating more intuitive and conversational interfaces for the user.

Therefore, it is important to develop new ways of representing and interacting with the information in order to facilitate the efficient selection of scientific articles stored in the extensive databases or repositories that index them with available techniques of knowledge extraction, such as Data Mining, Information Recovery, etc. Despite being important alternatives and becoming emergent and very robust computational techniques, in the end the user only has the graphical interfaces to narrow down information, and these traditional interfaces do not generally take advantage of human capacities (perception, cognition, intuition, etc.).

The aforementioned problems make it necessary to consider the opportunities that other approaches, methodologies, and technologies (e.g., Virtual Reality) can offer for the search for articles, as well as for the design of intuitive and even conversational interfaces for the user.

In this work, we propose a model for the exploration of scientific articles with Virtual Reality and Natural Language Processing [4], which aims to facilitate and simplify the work of researchers in the process of searching for and selecting scientific articles.

This project focuses on more efficient ways to navigate the database ALICIA, using immersive environment (i.e., virtual reality) technology for example, Oculus Rift, Leap Motion, headsets, computers, and speech recognition engines.

The sections regarding the theoretical framework (Section II), system model development (Section III), and conclusions (Section IV), will subsequently be developed.

II. Theoretical Framework

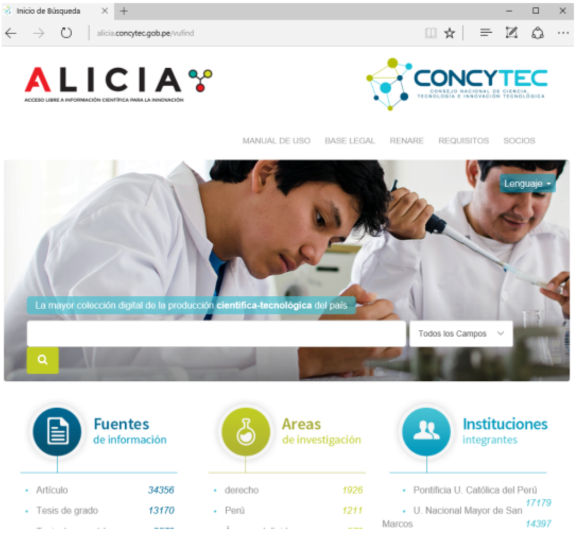
In this section, the authors conduct a historical review, identify the state of the art, and explore the theoretical foundation of the research project.

A. Scientific Databases

Scientific databases contain information about books and other materials in a library or a bibliographic index, for example, the content of a set of journals and other scientific publications such as research articles, conference proceedings, book chapters, etc. These databases usually have an electronic format and are consulted via the Internet. They contain bibliographic citations, references, abstracts, and the full text of the indexed contents or links to the full text. Many scientific databases are bibliographic databases, but others are not. Even outside the scientific sphere, the same thing occurs. There are databases that contain citations for art-history, journal articles, or databases containing only artistic images [5].

For Concytec[1], there are 20 main scientific databases, thirteen of which are freely accessible and available to the scientific/academic community as well as the general population of the Republic of Perú.

On the other hand, the database ALICIA[2] (Fig. 1), provides open access to this intellectual heritage resulting from advances in science, technology, and innovation achieved by public sector entities or with state funding.


Inicio de Búsqueda
alicia.concytec.gob.pe/vufind
ALICIA
ACCESO LIBRE A INFORMACIÓN CIENTÍFICA PARA LA INNOVACIÓN
CONCYTEC
CONSEJO NACIONAL DE CIENCIA, TECNOLOGÍA E INNOVACIÓN TECNOLÓGICA
MANUAL DE USO
BASE LEGAL
RENARE
REQUISITOS
SOCIOS
Lenguaje
La mayor colección digital de la producción científica-tecnológica del país
Todos los Campos
Fuentes de información
Areas de investigación
Instituciones integrantes

Fuentes de información		Areas de investigación		Instituciones integrantes	
Artículo	34356	derecho	1926	Pontificia U. Católica del Perú	17179
Tesis de grado	13170	Perú	1211	U. Nacional Mayor de San Marcos	14397

Fig. 1. ALICIA web interface.

B. Virtual Reality

The term virtual, its modern use was popularized by Jaron Lanier in the 1980s [6]. There are several definitions for Virtual Reality. According to one, Virtual Reality is a term used to describe a computer generated virtual Environment that may be moved through and manipulated by a user in real time [7]. Virtual Reality can also be defined as the way humans can visualize, manipulate, and interact with computers and extremely complex data [8]. Virtual Reality is only obtained when you are in a network, and several people share their realities amongst each other [9].

The main advantage of virtual environments is that they can be modeled and controlled exactly to an experiment's requirements, without having to build something similar in the real world [10]. Nowadays, virtual environments can be constructed with relative ease and it requires few resources to be able to run the necessary software [11]. Finally, These terms are the most popular and most often used [12], Virtual Environments (VE) and Virtual Reality (VR), are used in computer community interchangeably, but there are many others: Virtual Worlds, Artificial Worlds, Artificial Reality or Synthetic Experience.

For this research, the following themes are considered and developed:

1) Virtual Reality technology: Virtual Reality and immersive environments, as part of the emerging technological evolution involving our senses and cultural, symbolic and representative factors, may present interdisciplinary approaches [12]. Virtual reality refers to immersive, interactive, multisensory, viewer-centered, three dimensional computer generated environments and the combination of technologies required to build these environments [7]. In immersive VR, simulated objects appear solid and have an egocentric location much like real objects in the real world [13].

In terms of psychology, immersion refers to being completely involved in something while in action. In immersive virtual reality, it is possible to give people the illusion that they have a different body. They wear a head-tracked head-mounted display, and when they look down towards themselves they see a virtual body that is spatially coincident with their real body [14]. Mental immersion has different levels in a Virtual Reality experience. Such an experience can consist of partial or complete mental immersion, although it is worthwhile to note that reaching complete mental immersion in a Virtual Reality experience is still an ongoing challenge for research [15].

The unique characteristics of immersive virtual reality can be summarized as follow [7]: head-referenced viewing, stereoscopic viewing, the virtual world is presented in full scale and relates properly to the human size, realistic interactions with virtual objects, the convincing illusion of being fully immersed in an artificial world.

Steuer [16], proposed three different factors involved with an individuals presence in Virtual Reality: vividness, interactivity, the influence of a users personal characteristics. This derives from individual differences in the sense of presence when subjects are confronted with the same virtual environments. The individuals personal experiences and history contribute to this factor.

Virtual Reality possesses the characteristic of immersion, allowing users to immerse themselves in the simulation while stimulating intense emotions in the process. During the initial stages of development, Virtual Reality could only be used by scientific laboratories funded by the military at excessive costs. However, today virtual-reality glasses can be obtained at an affordable cost, for example, the Oculus Rift [17].

2) Display modes (2D vs 3D vs VR): There is currently an extensive discussion regarding the use of 2D, 3D, and VR modes in order to represent information [2]. The following briefly describes each.

- 2D is the rendering mode that uses only two dimensions. The graphics are flat, providing simplicity, clarity, and precision in the display of information.
- The 3D vision can be simulated by stereoscopy, meaning de display of two or more images perceived by the brain through the eyes and recompounded by it in an spatial image similar to what we naturally perceive in the everyday reality. The normal human vision is stereoscopic [18].
- VR is a technique which enables immersion in a multimodal, visualization environment, also utilizing

[1] https://portal.concytec.gob.pe/index.php/informacion-cti/biblioteca-virtual
[2] http://alicia.concytec.gob.pe/vufind/

stereoscopic images to improve depth perception. This type of VR system encases the audio and visual perception of the user in the virtual world and cuts out all outside information so that the experience is fully immersive [19].

- Sensory displays are used to display the simulated virtual worlds to the user. The most common sensory displays are the computer visual display unit, the head-mounted display (HMD) for 3D visual and headphones for 3D audio [20].

3) Virtual Reality Devices: The Oculus Rift uses stereoscopic vision technology, rendering a slightly different perspective of the 3D image for each eye. This enables the natural interaction of looking around while exploring a virtual three dimensional world [21].

It uses see-through lenses and 2 light engines to project the augmented content. It automatically calibrates pupillary distance, has a “holographic resolution” of 2.3M total light points and a “holographic density” of more than 2.5k light points per radian [22]. Samsung Gear VR is wireless, it only requires a smartphone. Finally, Oculus Go is an all-new standalone headset that doesn’t require you to snap in a phone, attach a cable or power up a PC. Oculus Go is hands-down the easiest way for people to get into VR. It’s perfect for those who haven’t experienced VR to play games, watch movies and connect with friends and family in new ways

4) Natural User Interfaces (NUI): NUI refers to both sensory inputs such as touch, speech and gesture [23]. NUI interfaces are those in which the interaction is natural, common, and familiar to the user [3]. This involves the use of new devices which enable new forms of interaction beyond the keyboard and mouse. This new generation of devices allow the use of gestures, tactile contact, body movement, and voice communication. An example of NUI devices is Kinect [24], a device which is able to record the movements of the body and transfer the data to a virtual avatar [17].

5) Evolution of Interfaces (CLI vs WIMP vs NUI): The user interface is the method or means by which the user communicates with the software. With the Natural User Interface, the interaction is natural; talking is prioritized over writing, listening over reading, and touch over clicking [25]. For this reason, the best choice to interact with the software is to use natural user interfaces [26].

C. Natural Language Processing

Natural Language Processing (NLP) is the function of software and/or hardware components in a computer system that analyzes or synthesizes spoken or written language [27].

The processing of Natural Language contributes to the retrieval of texts ranging from the length of a paragraph to that of a book. Technological advances have now made it possible to store, search for, and retrieve all or part of the full-text document online [28].

With a web search conducted with traditional approaches, several techniques are applied. These include building an index of the web content, building a database of the indices, and doing a search utilizing keywords which match the database contents. The problem with this approach is that it is not conducive to the acquisition of intelligence information [29].

On the other hand, the display of historical metadata in a virtual 3D environment implies several issues related to both user interface and data visualization and manipulation. A lot of work has been done in the field of user interfaces for Virtual Reality, even if the development of new devices is constantly challenging this research area [30].

Natural Language Processing possesses many applications, because on the Internet individuals communicate via natural language. The applications which are being developed range from the translation of texts to the implementation of talking robots. Some of them are mentioned below [31], [32]: automatic translation, text classification, sentiment analysis, information retrieval systems, search engine development, documentary databases, generation of automatic summaries, generation of natural language, etc. In this work, some approaches, methodologies, and techniques that use these applications will be used.

D. Related Work

As for the exploration of scientific articles, there are many applications which were developed for traditional environments, such as the case of scientific databases, which were designed for common interaction [33], where the majority of difficulties are due to the traditional configuration of the environment.

The literature mentions some frequent problems inherent in the traditional interfaces of scientific databases, such as Scopus, Web of Science, Scielo Google Scholar, etc. [34][35], which limit and complicate the search for scientific articles. These problems are detailed below:

Reduced Display Space: This is the weakness that most of the devices used in traditional visualization environments are prone to. LCD screens are the most commonly used display devices today, they come in different sizes, ranging from as small as cell phone screens, to medium, like those used in personal computers, to large screens such as those used in smart televisions. Obviously, the larger the visualization device, the more elements that can be visualized. Therefore, the size of the visualization device establishes a limit to the capacity of elements which can be displayed [36].

For example, in Google Scholar when one executes a search through its traditional interface, millions of results are obtained, but of the documents found, only about ten or even less are actually shown on the computer screen. There are alternative scanning and visualization interfaces, such as VR-Miner software which shows thousands of documents at the same time using graphical objects, but the disadvantage of this visualization technique is that occlusion or the overlapping of elements occurs, generating a disordered field of view for the user.

Faced with this problem, several researchers have proposed prototypes of visualization devices larger than the traditional ones. This problem of reduced space, limiting the visualization of search results is directly related to the device generating the image, which also has the characteristic of displaying the information in a flat way. Regardless of whether one is using two-dimensional or three-dimensional images, the final representation is without depth.

Interface overload: This generally occurs in the traditional interfaces of scientific databases of the WIMP-based interface type [37]. These interfaces are currently the most common, where the excessive use of menus, buttons, lists of options, icons, scroll bars, dialog boxes, text boxes, tabs, etc., cause the interface to be saturated with elements, reducing its versatility and ease of use [38].

While the tasks of retrieving user information are the same in each of the bibliographic query systems, this saturation of elements, commands, controls, parameters, etc., can confuse the user. What is more, most of these options are never used by the researcher upon using these tools to conduct a literature review due to the great complexity they usually represent.

Limited Interaction: Interaction is the set of actions through which the user communicates with the interface. These actions, in the context of exploration of scientific articles, include selecting elements, manipulating them, moving within the visualization environment, and reading the content of each document [39]. Generally, interaction with traditional scientific databases, whose interfaces are based on windows, icons, menus, and pointers, is deficient because only the mouse and keyboard are used as interaction elements, leaving the user's voice completely aside, not to mention the potential of the natural movement of his hands. In order to overcome these interactional limitations, several hardware prototypes have been developed and studied.

DocuWorld is an immersive virtual environment for document organization, where each file is displayed as a three-dimensional icon [40]. In Fig. 2, an example of visualization can be seen.

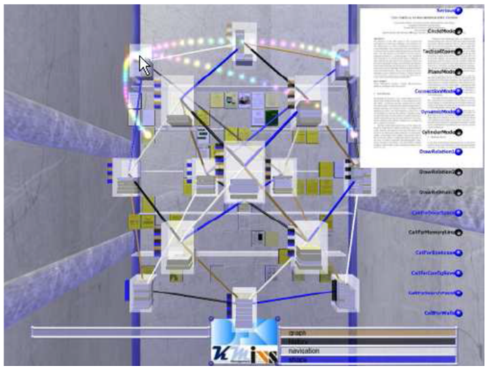


Fig. 2. DocuWorld - document visualizer.

@Visor is a software program designed for scanning and reading documents, using an electronic glove which is used instead of a mouse [17]. It is noteworthy due to its use of natural hand movements.

3D Spatial Data Mining is a visualization tool which differentiates itself by using a metaphor based on the real world for the exploration of automobile fault reports. It is a 3D model in the form of a car engine [41]. The great advantage of using this metaphor is that the user easily becomes familiar with it.

III. Model System Development

Next, we will describe the different elements that were developed as well as the procedural steps that were followed in order to build the model which the authors call AliciaVR.

A. Physical Configuration Immersive Environment

The first step consisted of configuring the equipment (Fig. 3). The Leap Motion device was attached to the front of Oculus Rift. Then both devices were connected to a laptop via two USB ports and an HDMI port. Regarding the configuration of the software, version 5.02 of the Unity3D graphics engine was installed.

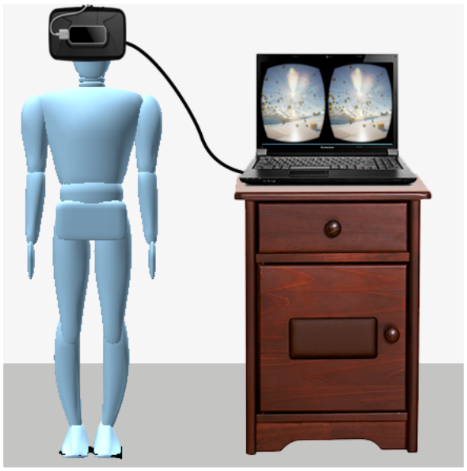


Fig. 3. Configuration scheme of the immersive environment.

B. Development of Virtual Environment

For the development of a virtual, immersive 3D environment (Fig. 4), which is suitable for the exploration of scientific articles, the following considerations and recommendations are taken into account:

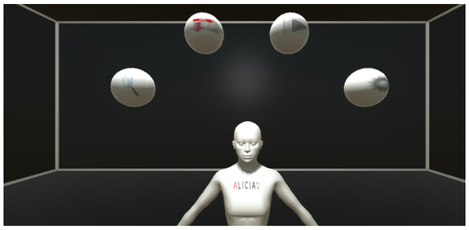


Fig. 4. Construction of virtual environment in unity3D.

- Simplicity: This is to avoid overloading the interface, using only the necessary elements. In this way, perception of the interface is smoother, thus maximizing the user's cognitive activities.
- Dark background: This is to mitigate dizziness and decrease the frequency of update snapshots required for the rotation of the observer's point of view.
- White lines: These delimit the virtual environment, allowing the user to perceive depth with greater precision.

- Color range close to the real: this is to avoid dizziness or discomfort if the images are not perceived as real.
- Rounded elements: These accommodate themselves to the shape of the hand better and allow degraded shadows to be generated, which facilitates their visualization.
- Visualization of an Avatar: This is in order to optimize the user's sense of presence in the environment [42].

Taking into consideration the last recommendation, an avatar was developed (Fig. 5). This avatar is defined as a digital character, which may resemble a real person, an animal, or any other form that could be given to it, according to system requirements and user preferences. The purpose is for the user to imagine having an assistant, which would allow them to intuit that it is possible to interact through voice commands. It is important to point out that its basic functions are defined within the context of the exploration of scientific articles.

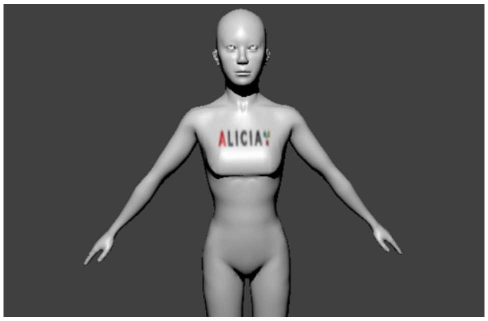


Fig. 5. AliciaVR model system avatar.

C. Natural Interaction Model System Implementation

After the immersive environment was configured, the interaction techniques for immersive, virtual environments were investigated and explored, according to the following parameters:

- Selection Technique by Direct Manipulation: this was done through the use of a virtual hand, which copies exactly the movements of the real hand. In Fig. 6, the selection of a node by direct manipulation is shown. As can be seen, the process is identical to grabbing an object in the physical world.
- Handling by Gestures: gesture manipulation consists of selecting elements in the immersive environment, utilizing gestures with the virtual hand. Specifically, in order to grasp an object, the 'pinch' gesture must be made, based on the metaphor of a pinch, and in order to release the element, the 'stop' gesture must be performed, which consists of extending the fingers so that the palm is exposed (i.e., the opposite of a 'grab' gesture). In Fig. 7a, the node is selected with the first gesture described, then in Fig. 7b, the node is released by the second gesture. This technique is not entirely intuitive because both gestures can have alternative meanings depending on the user or when performing them, they might expect a different action to be instantiated.
- Technique of Selection by Artificial Manipulation: This technique was proposed within a framework emphasizing the concepts of 'dimension', 'transduction', and 'reification' [43], which allow for prior-configuration experiences such as crossing objects. Therefore, in order to select an object, it is only necessary to cross it using the virtual hand. In order to release the element, it is only necessary to cross it again. In Fig. 8a the cube is selected by changing color to green and in Fig. 8b the cube returns to its previous state.

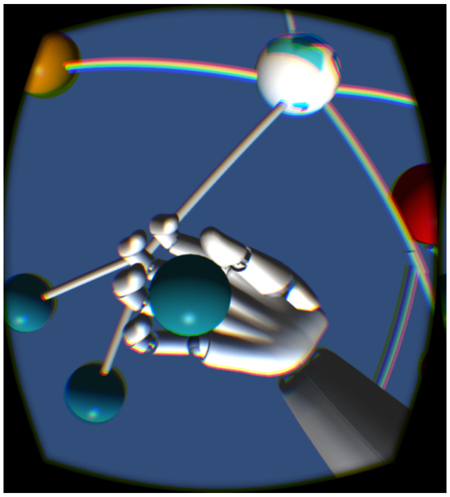
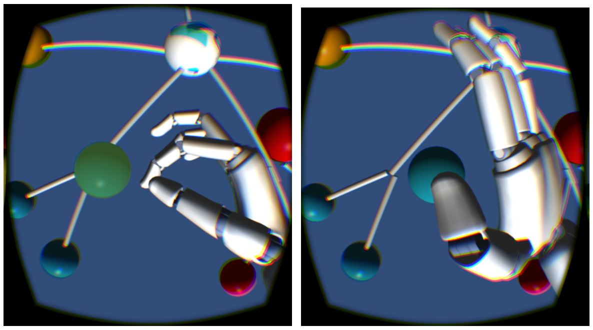


Fig. 6. Example of selection by direct manipulation of a node.

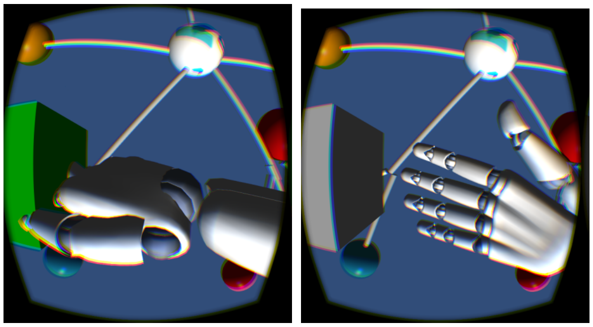


(a) Selecting the item (b) Releasing element

Fig. 7. Selection example gesture of a node.

D. Implementation of Web Crawler for Information Retrieval from the Model

The proposed interface model (AliciaVR), is connected to the ALICIA scientific database. In order to retrieve information from this system, a program of the crawler type was developed and implemented, which automatically communicates with ALICIA and extracts, through code analysis (source of the visual field), the information required for visualization within the interface based on Virtual Reality. This implementation involved studying the proper pattern in order to perform the tracking and recovery extraction in an automated way.



(a) Selecting the item (b) Releasing element

Fig. 8. Selection example by articial manipulation of a node.

E. Development of Speech Recognition Engine and Conversion to Text

For the recognition engine, the Intel Perceptual Computing Library was used, which incorporates audio-to-plain-text conversion functions. For more sensitive speech recognition, a wireless microphone certified by Dragon Natural Speaking has been added to the prototype design. A library has been developed in C# .Net, in order to capture the user's audio at all times, subsequently sending the data via sockets to Unity3D using network protocols, where a listening port was implemented through which all the user's voice commands are received.

Fig. 9 shows a representation of the process by which the sound waves produced by the user when speaking are captured by the microphone and through the Intel Perceptual Computing Library, they are interpreted and converted into plain text.

Fig. 9. Voice capture and conversion to text.

Using the previously-mentioned library, the desktop application was developed and configured using Visual Studio Community 2013. In order to start the speech recognition engine, it is necessary to start this application directly after initiating the Virtual Reality application developed with Unity3D. This is due to fact that within the Unity3D application, a listening server is implemented, and in the application for voice recognition the client is implemented for sending, in plain text, the voice of the user to the server. In Fig. 10, the speech-recognition engine interface is shown.

F. Implementation of the Syntatic Analyzer

In order to interpret the syntactic analysis of each of the voice commands, the Freeling library was used, because of the functions it possesses for the processing and conversion of the Spanish Language. This process of syntactic analysis first involves separating each of the words, parsing their content, and then adding part-of-speech labels, such as 'verb', 'adjective', 'noun', 'adverb', 'article', 'pronoun', etc. In Fig. 11, an example of the phrase analysis is shown, where the user says, "Alicia, save me this article please", generating its respective syntactic, phrase-structure tree.

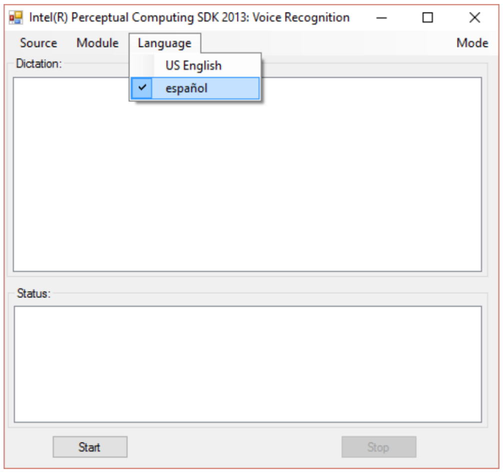


Fig. 10. Interface of the speech recognition engine.

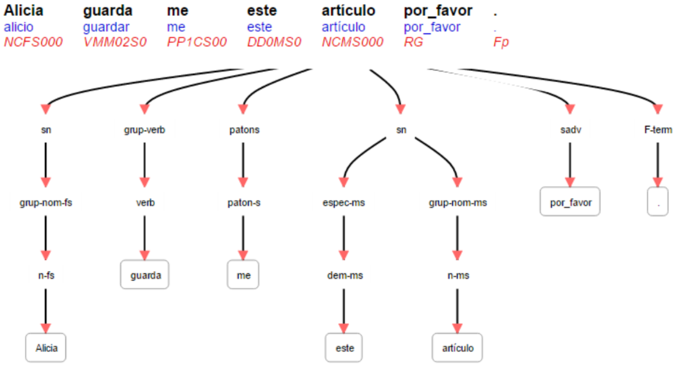


Fig. 11. Generation of the syntactic tree for a voice consultation.

The proposal, integrating all that is mentioned above, is described in the interface-model scheme for the exploration of scientific articles with Virtual Reality and Natural Language Processing, referred to in the research as AliciaVR and graphically described in Fig. 12.

Basically, the proposed model begins with the consultation of a user who wants to search for a scientific article in AliciaVR. The consultations are made by voice, which the speech recognition engine detects and converts into plain text. Then this information is processed by the semantic analyzer, where the respective syntactic tree is generated. This analyzer differentiates the interaction commands from the query, and then sends the latter to the tracker, which will be in charge of connecting with ALICIA and retrieving the information requested by the user. Finally, the results obtained are plotted in the immersive environment, where the user begins to interact with it. If the user is satisfied, then the process ends. Otherwise, a new search is initiated.

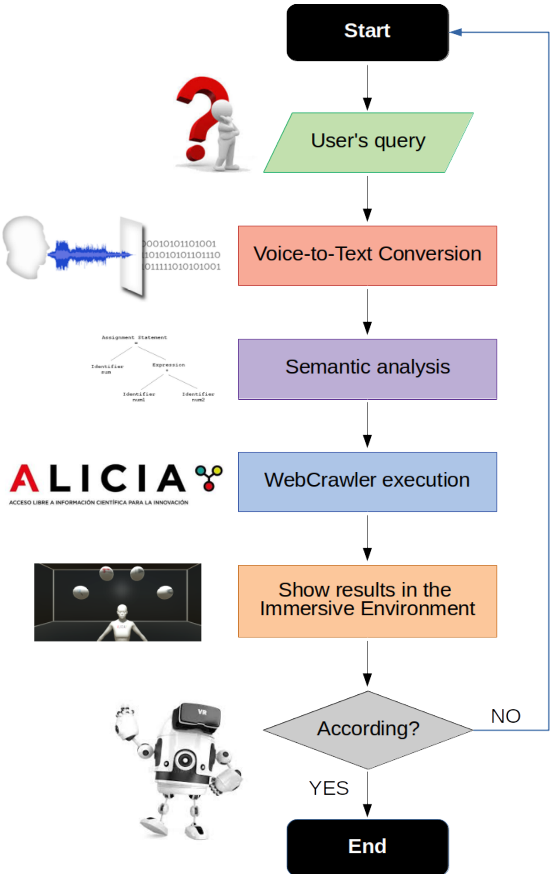


Fig. 12. Scheme of the proposed model AliciaVR.

G. Evaluation of proposed Model regarding Traditional

To evaluate the proposed interface model (AliciaVR) with respect to the traditional model (Alicia-Web), the ISO 9241 Interface Usability Standard will be used, which specifically evaluates three dimensions of user interaction with the interface: efficiency, efficacy, and satisfaction.

According to ISO 9241 Standard defines 'usability' as, "the degree to which a product can be used by specific users to achieve specific objectives with effectiveness, efficiency, and satisfaction in a specific context of use". The variables and their indicators are described below:

- 'Efficacy' is defined as the accuracy and integrity with which users reach their specified objectives, therefore implying ease of learning, absence of errors in the system, or ease of it to be remembered. The metrics used in the investigation are shown in Table I.
- 'Efficiency' is the relation of the resources used (effort, time, etc.) to the accuracy and integrity with which users reach their specified objectives. The metrics used in the investigation are shown in Table II.
- 'Satisfaction' is a subjective factor which implies a positive attitude in regards to the use of the product. The satisfaction parameters used in the research are shown in Table III.

H. Preparation of the Experiment with Users

After having set up the apparatus of interconnected devices (Oculus Rift, Leap Motion, a wireless headset and computer) and initiated all the software modules (i.e., immersion environment, speech recognition engine, web server for the tracker, etc.), the investigation then proceeds to the evaluation of the scientific-article exploration model with Virtual Reality and Natural Language Processing (AliciaVR) and also that of the traditional ALICIA model (Alicia-Web), with the assistance of five experimental users, this being the number of subjects that Nielsen recommends for the evaluation of interfaces [44]. A high speed internet connection of 8 Mbps is also be utilized.

Each of them is assigned the task of exploring scientific articles for thirty minutes using each interface within both interface models. During the experiment it is thought that the user will ask for help from an expert, who at the same time will be responsible for the observation and documentation of the experiment.

The first model to be evaluated is the exploration of scientific articles with Virtual Reality and Natural Language Processing. In this model the user is immersed in the software, being able to communicate by voice with the search engine and touch the elements selected. In Fig. 13 a user is shown using the proposed AliciaVR interface. What the user actually observes is shown in Fig. 14.

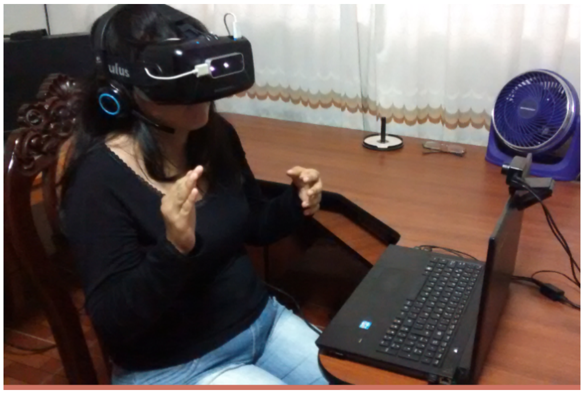
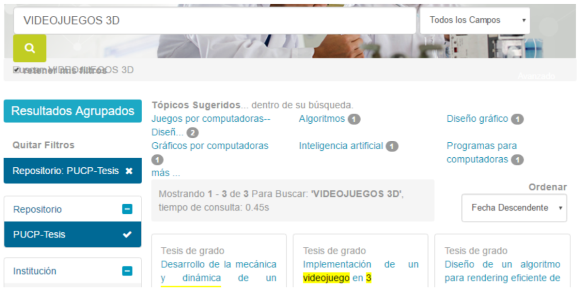


Fig. 13. Evaluation of the first interface model - AliciaVR.



VIDEOJUEGOS 3D

Todos los Campos

retener mis filtros

Resultados Agrupados

Tópicos Sugeridos... dentro de su búsqueda.

Juegos por computadoras--Diseñ... 2

Algoritmos 1

Diseño gráfico 1

Gráficos por computadoras 1

Inteligencia artificial 1

Programas para computadoras 1

más ...

Quitar Filtros

Repositorio: PUCP-Tesis ✖

Mostrando 1 - 3 de 3 Para Buscar: 'VIDEOJUEGOS 3D', tiempo de consulta: 0.45s

Ordenar

Fecha Descendente

Repositorio

PUCP-Tesis ✓

Institución

Tesis de grado
Desarrollo de la mecánica y dinámica de un

Tesis de grado
Implementación de un videojuego en 3

Tesis de grado
Diseño de un algoritmo para rendering eficiente de

Fig. 14. User's view - AliciaVR.

The second model to be evaluated is the exploration of scientific articles in the traditional way, which is the default interface of the ALICIA scientific database. In this interface model, the user has only the monitor with which to view the queries and only uses the mouse and keyboard to interact.

I. Evaluation of Results

This section presents the measurements made during the evaluation of and comparison between the traditional ALICIA (Alicia-Web) interface and the proposed interface model based on Virtual Reality (AliciaVR).

1) Tabulation of Collected Data: Table I, presents the data obtained by observing users using the two different interfaces, measuring the efficiency of each interface with the metrics defined by the ISO 9241 Standard. It should be recalled that 'efficiency' refers to the achievement of objectives or fulfillment of tasks. Therefore, it is represented by a number which indicates the number of objectives met. In the different metrics, the number of objectives fulfilled is counted, these being the tasks performed, functions used, etc.

TABLE I. COMPARISON OF INTERFACE EFFICACY

Metrics ISO 9241 Measurement of Efficacy	Alicia-Web (Half Quantity)	AliciaVR (Half Quantity)
Number of important tasks performed.	4.2	2.0
Number of relevant functions used.	7.8	3.8
Number of tasks completed successfully at the first attempt.	2.2	1.2
Number of calls for support.	0.6	3.4
Number of accesses to the aid.	0.2	0.4

Table II, presents the data obtained by observing users using the two different interfaces, measuring the efficiency of each interface with the metrics defined by the ISO 9241 Standard. Recall that 'efficiency' is the optimal use of resources required for the fulfillment of an objective, where time is the main resource used in the use of a software interface. The longer it takes for the user to perform the tasks, the less effective the interface will be considered. On the other hand, if this time is shorter, then the interface will be considered to be more efficient.

TABLE II. COMPARISON OF INTERFACE EFFICIENCY

Metrics ISO 9241 Measurement of Efficiency	Alicia-Web (Half time in seconds)	AliciaVR (Half time in seconds)
Time used in the first attempt.	47.3	35.6
Time used to relearn functions.	31.2.8	16.6
Productive time.	78.6	54.0
Time to learn characteristics.	48.6	34.2
Time to relearn characteristics.	29.4	10.2
Time used in the correction of errors.	54.6	55.8

TABLE III. COMPARISON OF INTERFACE SATISFACTION

Metrics ISO 9241 Measurement of Satisfaction	Alicia-Web (Half Percent)	AliciaVR (Half Percent)
Calibration of satisfaction with important characteristics.	0.32	0.72
Calibration of the learning facility.	0.44	0.64
Calculation of error handling.	0.28	0.24
Rate of voluntary use of the product.	0.64	0.92

2) Analysis by Inferential Statistics: To determine whether, judging by the data obtained, it can be concluded that efficiency is greater using one of the interfaces over the other, a statistical test to make the global inference is applied. The T-Student statistical test is used in order to differentiate between two paired means. In Table IV, the values obtained when applying the statistical test are shown.

The p value being 0.019, which is less than 0.05 ($p < \alpha$), is interpreted as sufficient, statistically-significant evidence, affirming that the groups are indeed different. Moreover, the difference being positive (0.81) indicates that the effectiveness of the first interface (Alicia-Web) is greater than the efficiency of the second interface (AliciaVR).

TABLE IV. T-STUDENT TEST FOR EFFICACY

	Difference by pairs				t	df	Sig.
	Dif.Average	Error D.E.	95% Confidence Int.				
			Low	High			
Efficacy	0.81	1.25	-1.989	3.589	0.639	10	0.019

To determine whether, with the data obtained, it can be concluded that efficiency is greater using one interface over other, the same "T-Student" statistical test used in order to differentiate between two paired means. In Table V, the values obtained upon applying the statistical test are shown.

The p value being 0.021, which is less than 0.05 ($p < \alpha$) is interpreted as sufficient evidence, confirming that the groups are significantly different, and, what is more, the difference being positive (13.87) indicates that the time spent on each task with the first interface (Alicia-Web) is greater than the time used using the second interface (AliciaVR). Therefore, since less time is spent on the second interface, it can be concluded that efficiency is greater using the AliciaVR interface than it is using the traditional ALICIA interface.

TABLE V. T-STUDENT TEST FOR EFFICIENCY

	Difference by pairs				t	df	Sig.
	Dif.Average	Error D.E.	95% Confidence Int.				
			Low	High			
Efficiency	13.87	10.57	-9.68	37.43	1.31	10	0.021

To determine whether, with the obtained data, it can be concluded that satisfaction is greater using one interface rather than the other, the "T-Student" statistical test is once again applied to make the global inference, measuring the difference between two paired means.

In Table VI, the values obtained upon applying the statistical test are shown. Here, it can be seen that the p-value is 0.024, which, being less than 0.05 ($p < \alpha$) is interpreted as sufficient evidence, confirming that the groups are significantly different. The difference being negative (-0.21) indicates that satisfaction with respect to the first interface (Alicia-Web) is less than that of the second interface (AliciaVR).

TABLE VI. T-STUDENT TEST FOR SATISFACTION

	Difference by pairs				t	df	Sig.
	Dif.Average	Error D.E.	95% Confidence Int.				
			Low	High			
Satisfaction	-0,21	0.16	-0.61	0.19	-1.28	6	0.024

J. Results Empirical Analysis

This section presents the results of empirical analyses based on the data obtained from the users' experiences [25][26]. The proposed model has attracted considerable interest from users, considering that it is a novel form of digital interaction which is full of possibilities. With every user a preliminary difficulty in both models was noticed.

Specifically, users who have never used ALICIA only use the basic functions, and not all possible tools which have been incorporated into the system. In contrast, the same users utilizing the proposed interface based on Virtual Reality were observed to have delayed initiation in order to accommodate themselves to the apparatus, at first failing to distinguish the three-dimensional character of the Oculus Rift device. However, after a few minutes of adapting to the equipment and watching their hands, they really wanted to explore all options available to perform.

For recovery-task items, users utilizing the AliciaVR interface managed to perform these tasks in less time than users of traditional, web-based interfaces were able to. Nevertheless, the big problem that arises is that the Virtual-Reality glasses are a bit uncomfortable and cumbersome for the user to say the least. This is to say that having to support a large, heavy apparatus on the head becomes annoying and tiresome after a while, especially during the last half hour, prompting the user to want to return to the real world.

IV. Conclusion

We have presented the whole process of creating a new interface for the exploration of scientific articles in scientific databases, based on Virtual Reality and Natural Language Processing techniques.

Users working with the traditional, web-based interface interact only with a mouse and keyboard. However, with the proposed new interface, the user interacts with the system in a much richer and more robust way, using Virtual Reality, full-immersion techniques. The proposed interface model and the traditional model of the ALICIA scientific database were evaluated using the ISO 9241 Standard, taking into consideration the three attributes of efficiency, efficacy, and satisfaction.

The statistical T-Student test was applied to evaluate the data obtained by means of inferential statistics. The results indicate that the proposed interface model, AliciaVR, is superior to the traditional interface of ALICIA (Alicia-Web) in the dimensions of efficiency and satisfaction, but it is inferior in the effectiveness dimension. Therefore, it is concluded that the model of exploration of scientific articles with Virtual Reality and Natural Language Processing is superior to the traditional exploration model.

An interesting future research project would be the inclusion in the model of the option to choose from among Oculus Go and Oculus Rift or to switch from one to the other.

Another future project would consist of extending the model that is currently focused on scientific databases, generalizing it to the exploration of business and management databases which are relative to the explicit knowledge defined in the discipline of Knowledge Management, for example, reports, contracts, operation manuals, schemes, and plans, among others. In this way, of particular importance would be the extension and development of the model for the exploration of tacit knowledge adequately linked to that of explicit knowledge utilizing, for example, a hybrid e-learning system that incorporates "case-based reasoning".

This would make it possible to treat and examine processes and procedures, industrial manuals, and other types of tacit knowledge, whose explication and transference process is complex and could be facilitated enormously utilizing immersive, virtual reality and natural language processing techniques.

Virtual Reality is an emerging technology in the process of constant evolution and diffusion, at the level of both hardware and software, having already witnessed the advent of revolutionary advances, such as the Holographic Processor (HPU). The current tendency is in the direction of everyone possessing access to this technology at a low cost, and the first specialized operating system is already in development, executing applications developed in Virtual Reality.

Acknowledgment

The authors would like to thank the UNSA and the CONCYTEC for supporting this research.

References

[1] T. Sakai, B. Flanagan, J. Zeng, T. Nakatoh, and S. Hirokawa, "Search engine focused on multiple features of scientific articles," in *2012 IIAI International Conference on Advanced Applied Informatics*. IEEE, 2012.

[2] M.-S. Yoh, "The reality of virtual reality," in *Proceedings of the Seventh International Conference on Virtual Systems and Multimedia (VSMM'01)*. IEEE Computer Society, 2001, pp. 666–. [Online]. Available: http://dl.acm.org/citation.cfm?id=882502.884873

[3] B. S. Zohra, G. Fabrice, R. Paul, B. Julien, and P. Fabien, "An overview of interaction techniques and 3d representations for data mining," in *Applications of Virtual Reality*. Intech, 2012.

[4] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *Journal of the American Medical Informatics Association: JAMIA*, vol. 18, no. 5, pp. 544–551, 2011.

[5] A. Clark, C. Fox, and S. Lappin, *The handbook of computational linguistics and natural language processing*. John Wiley & Sons, 2013.

[6] S. L. Valle, *Virtual Reality*. Cambridge University Press. London, 2017.

[7] S. Mandal, "Brief introduction of virtual reality & its challenges," *International Journal of Scientific & Engineering Research*, vol. 4, no. 4, 2013.

[8] L. A. A. Casas, V. Bridi, and F. Fialho, "Virtual reality full immersion techniques for enhancing workers performance," in *Re-engineering for Sustainable Industrial Production*. Springer, 1997, pp. 399–411.

[9] L. A. A. Casas, "Contribuições para a modelagem de um ambiente inteligente de educação baseado em realidade virtual," Ph.D. dissertation, Tese (Doutorado em Engenharia de Produção)–Programa de Pós-graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 1999.

[10] H. Sauzéon, P. A. Pala, F. Larrue, G. Wallet, M. Déjos, X. Zheng, P. Guitton, and B. N'Kaoua, "The use of virtual reality for episodic memory assessment: effects of active navigation," *Exp Psychol*, vol. 59, no. 2, pp. 99–108, 2011.

[11] S. Mills and J. Noyes, "Virtual reality: an overview of user-related design issues," *Special Issue on Virtual reality: user Issues*, pp. 375–386, 1999.

[12] J. Rubio-Tamayo, M. Gertrudix, and F. García, "Immersive environments and virtual reality: Systematic review and advances in communication, interaction and simulation," *Multimodal Technologies and Interact.*, vol. 1, no. 21, 2017.

[13] J. Psotka, "Immersive training systems: Virtual reality and education and training," *Instructional Science*, vol. 23, pp. 405–431, 1995.

[14] M. Slater and M. Sanchez-Vives, "Transcending the self in immersive virtual reality," *Computer*, vol. 47, no. 7, 2014.

[15] M. A. Muhanna, "Virtual reality and the CAVE: Taxonomy, interaction challenges and research directions," *Journal of King Saud University - Computer and Information Sciences*, vol. 27, no. 3, pp. 344–361, 2015.

[16] J. Steuer, “Defining virtual reality: Dimensions determining telepresence,” *Journal of Communication*, vol. 42, no. 4, pp. 73–93, 1992.

[17] S. Baumgärtner, A. Ebert, M. Deller, and S. Agne, “2D meets 3D,” in *CHI’07 extended abstracts on Human factors in computing systems - CHI’07*. ACM Press, 2007.

[18] D. Lacrama and D. Fera, “Virtual reality,” *Annals in Computer Science Series*, vol. 5, no. 1, 2007.

[19] A. Saeed, L. Foaud, and K. Abdulaziz, “Environments and system types of virtual reality technology in stem: A survey,” *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, 2017.

[20] M. Okechukwu and F. Udoka, “Understanding virtual reality technology: Advances and applications,” *Advances in Computer Science and Engineering*, 2011.

[21] M. Ruyg, C. Teunisse, and S. Verhage, “Virtual reality for the web: Oculus rift,” 2014.

[22] A. Coppens, “Merging real and virtual worlds: An analysis of the state of the art and practical evaluation of microsoft hololens,” 2017.

[23] M. Kaushik and R. Jain, “Natural user interfaces: Trend in virtual interaction,” 2010.

[24] W. D. Ra, “Brave NUI world,” *ACM SIGSOFT Software Engineering Notes*, vol. 36, no. 6, p. 29, nov 2011.

[25] R. Linares, J. Herrera, and L. A. A. Casas, “AliciaVR: Exploration of scientific articles in an immersive virtual environment with natural user interfaces,” in *2016 IEEE Ecuador Technical Chapters Meeting (ETCM)*. IEEE, oct 2016.

[26] R. Linares, “Exploración de artículos científicos con realidad virtual y procesamiento del lenguaje natural,” Ph.D. dissertation, Tesis de Maestría en Ciencias Informáticas, Universidad Nacional de San Agustín, 2017.

[27] P. Jackson and I. Moulinier, *Natural language processing for online applications: Text retrieval, extraction and categorization*. John Benjamins Publishing, 2007, vol. 5.

[28] D. D. Lewis and K. S. Jones, “Natural language processing for information retrieval,” *Communications of the ACM*, vol. 39, no. 1, pp. 92–101, jan 1996.

[29] D. K. M. Alhawiti, “Natural language processing and its use in education,” *Computer Science Department, Faculty of Computers and Information technology, Tabuk University, Tabuk, Saudi Arabia*, 2014.

[30] D. Bowman, E. Kruijff, J. LaViola, and I. P. Poupyrev, *3D User interfaces: theory and practice, CourseSmart eTextbook*. Addison-Wesley, 2004.

[31] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: State of the art, current trends and challenges,” *arXiv preprint arXiv:1708.05148*, 2017.

[32] A. E. Thessen, H. Cui, and D. Mozzherin, “Applications of natural language processing in biodiversity science,” *Advances in Bioinformatics*, vol. 2012, pp. 1–17, 2012.

[33] C. J. Noblejas and A. P. Rodríguez, “Recuperación y visualización de información en web of science y scopus: una aproximación práctica,” *Investigación Bibliotecológica: archivonomía, bibliotecología e información*, vol. 28, no. 64, pp. 15–31, 2014.

[34] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas, “Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses,” *The FASEB journal*, vol. 22, no. 2, pp. 338–342, 2008.

[35] I. Rafols, L. Leydesdorff, A. O’Hare, P. Nightingale, and A. Stirling, “How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management,” *Research Policy*, vol. 41, no. 7, pp. 1262–1282, 2012.

[36] E. Bertini, L. Dell’Aquila, and G. Santucci, “Reducing infovis cluttering through non uniform sampling, displacement, and user perception,” in *Visualization and Data Analysis 2006*, vol. 6060. International Society for Optics and Photonics, 2006.

[37] S. M. Redwan, “Instant-access start menu for an imaginary wimp based future operating system,” in *10th international conference on Computer and information technology, ICCIT 2007*. IEEE, 2007, pp. 1–5.

[38] B. W. Pickering, V. Herasevich, A. Ahmed, and O. Gajic, “Novel representation of clinical information in the icu: developing user interfaces which reduce information overload,” *Applied Clinical Informatics*, vol. 1, no. 2, p. 116, 2010.

[39] A. Kerren, A. Ebert, and J. Meyer, *Human-centered visualization environments*. Springer-Verlag Berlin Heidelberg, 2007.

[40] K. Einsfeld, S. Agne, M. Deller, A. Ebert, B. Klein, and C. Reuschling, “Dynamic visualization and navigation of semantic virtual environments,” in *Tenth International Conference on Information Visualisation (IV 06)*. IEEE, 2006.

[41] T. Götzelmann, K. Hartmann, A. Nürnberger, and T. Strothotte, “3d spatial data mining on document sets for the discovery of failure causes in complex technical devices,” in *GRAPP (AS/IE)*, 2007, pp. 137–145.

[42] R. Klevjer, “What is the avatar? Fiction and embodiment in avatar-based singleplayer computer games,” 2006.

[43] W. Winn, “A conceptual basis for educational applications of virtual reality,” *Technical Publication R-93-9, Human Interface Technology Laboratory of the Washington Technology Center, Seattle: University of Washington*, 1993.

[44] J. Nielsen, “Why you only need to test with 5 users,” 2000.

An Efficient Fault Tolerance Technique for Through-Silicon-Vias in 3-D ICs

Mohamed BENABDELADHIM
Department of Physics
Faculty of Sciences of Monastir
Monastir, Tunisia

Fakhreddine ZAYER
Department of Electrical Engineering
National Engineering School of Monastir
Monastir, Tunisia

Wael DGHAIS
Department of Electronics
ISSAT Sousse
Sousse, Tunisia

Belgacem HAMDI
Department of Electronics
ISSAT Sousse
Sousse, Tunisia

***Abstract*—Three-dimensional integrated circuits (3D-ICs) based on Through-Silicon-Vias (TSVs) interconnection technology appeared as a viable solution to overcome problems of cost, reliability, yield and stacking area. In order to be commercially feasible, the 3D-IC yield must be as high as possible, which requires a tested and reparable TSVs. To overpass this challenge, an integration of interconnect built-in self-test (IBIST) methodology for 3D-IC is given in the aims to test the defectives TSVs. Once the interconnection has been tested, the result extracted from IBIST initiate the repairing defectives TSVs based on the built-in self-repair (BISR) structure. This paper superposed two fault tolerance techniques in particularly the redundant TSV and the time division multiplexing access (TDMA) in case of multi defectives TSV. This novel repair architecture increase the yield of 3D-ICs in a high failure rate. It leads to 100% reparability for 30% failure rate. A parallel processing approach of the proposed structure is presented to accelerate the test and repair operations. Achieved experimental results with the proposed methodology scheme show a good performance in terms of repair rate and yield.**

***Keywords*—Fault tolerance; 3D-IC; TSV; IBIST; BISR; TDMA**

I. Introduction

The 3D-ICs with TSVs provides smaller interconnect delay and higher device density [1]. These characteristics allow for the fastest IC design that offers flexible and low-cost packaging solutions. Interconnection of the various tiers troughs the TSVs pledges to increase the interconnect bandwidth and the performance of 3D-ICs while lowering power dissipation and manufacturing cost [2], [3]. Nevertheless, there are several challenges that can affect and decrease 3D ICs yield, i.e., technological challenges, test challenges, thermal and power challenges and infrastructure challenges [4]. In other terms, to guarantee the quality of the 3D chips, the TSV must be tested to locate manufacturing defect, a lack of precision in every fabrication step or integration process can produce a several kind of defects as revealed in Fig. 1 [5].

All the kind of defect decrease the performance of TSVs channels in terms of electrical characteristics and reduces the ability to withstand physical and thermal stresses during and after the fabrication process [6], [7]. The defect mechanisms must be deeply assessed to decrease the failure in terminal product. Thus, a test of TSVs is very challenging because current wafer-level testing cannot deal with a large number of small TSVs at an affordable cost. Hence, an IBIST technique with low hardware overhead becomes a potentially good solution to achieve both test and cost-effectiveness [8].

Testing and repairing the 3D-ICs was considered as a key challenge to improve yield and reliability [9]. In the last few years, several researchers on 3D IC testing and TSVs repairing have been proposed [9]–[20]. The plurality of proposed solution in the literature to repair defective TSV based on adding redundant TSVs method by several design technique [9]–[15]. However, the redundant TSV made to repair one faulty [10]. One spare TSV for each defectives TSVs must been used in the repair process.

From area distribution and performance enhancement, it can be concluded that is not practicable to fabricate a redundant TSV for each one. Thus, the adopted technique to minimize number of spars TSVs was implemented to divide TSVs into sets with redundant TSV. After test process was achieved, a spares is used by shift operations [11], [12], to replace faulty TSVs. In [11] and [16] a novel amelioration of repairing methodology was proposed. It enables to exchange faulty TSVs by distant redundant TSVs. In [20] a new fault-tolerant technique was addressed without using of redundant TSV. The basic idea is to localize defective TSVs and rerouting the signal through faulty free one using time division multiplexing access (TDMA) technique. Although this evaluations in repairing.

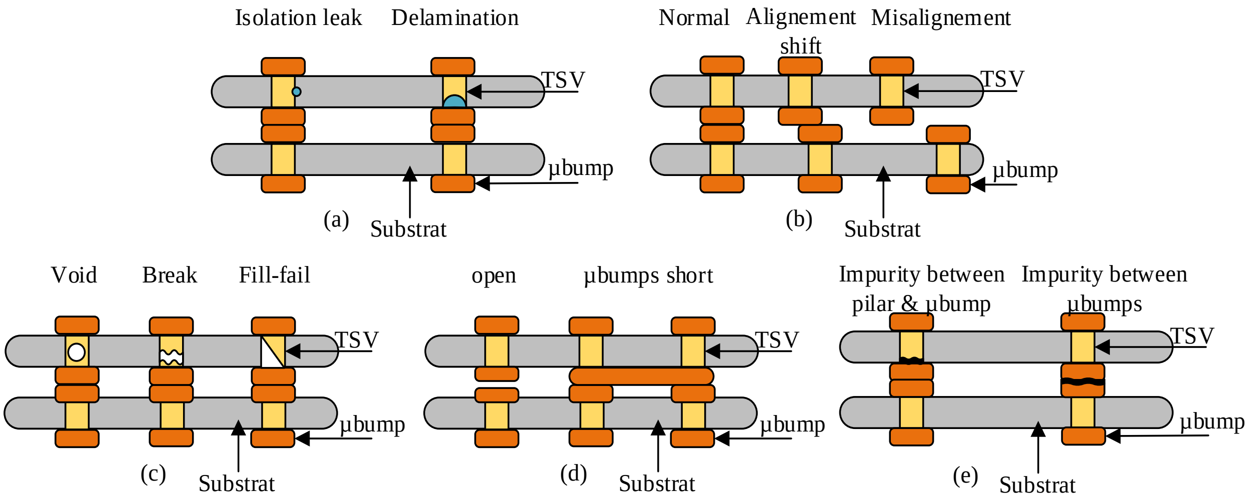


Fig. 1. Schematic presentations of TSV defects: (a) Defects due to substrate; (b) Defects due to misalignment; (c) Defects due to Cu pillar; (d) Defects in micro-bump; (e) Defects due to impurities.

This paper proposes a novel test and repair methodology of TSVs defect in 3D-ICs that integrates IBIST and BISR techniques. The proposed TSV's test architecture use the IBIST methodology in which a parallel test is applied in dies (layer) to speed the test process. The outcomes of the test results are investigated with a view to repair the identified faulty TSVs. The classification of TSVs into clusters with spare TSV aims to increase reparability rate. This work extends the repairing capabilities of BISR by enabling the correction of multi defectives TSVs based on TDMA technique. In fact, the TSV cluster was portioned in two bundles in which the shifting and replace process were properly implemented to transmit data. Moreover, the data is sequentially conveyed in two packets based on TDMA technique, which increase the number of spare TSV in each cluster.

The remains of this paper are organized as follows. Section II demonstrates the proposed combination between the new IBIST structure and the repairing architecture BISR. Section III presents the proposed repair process. Experimental results are illustrated in Section IV. Finally, conclusions are drawn in SectionV.

II. Proposed 3D IC based IBIST Methodology Test with TSV Repair

A. *Testing TSV based IBIST Methodology*

Fig. 2 presents the IBIST architecture, which is composed of detecting circuit, which is designed to be included between, transmitter and receiver components of each layer. The controller manages the different test tasks by means of an appropriate algorithm. The main component to detect the defectives TSVs is the test pattern generator (TPG) that includes a linear feedback shift register (LFSR).

The pattern source generates the test vectors, which are employed to test the TSVs. It comprises a sub-circuit that provides the LFSR functionality [18]. LFSR is a shift register used to generate pseudo-random test vectors. In the proposed test architecture, pattern generator in layer k-1 sends a test vector to layer k to test a k group of TSV in order to accelerate the test process for a big number of TSV groups. The test vectors are generated at layer k are simultaneously sent to the layer k+1 and the XOR logic port. In which the pattern coming from the layer k-1 is analyzed with XOR gate.

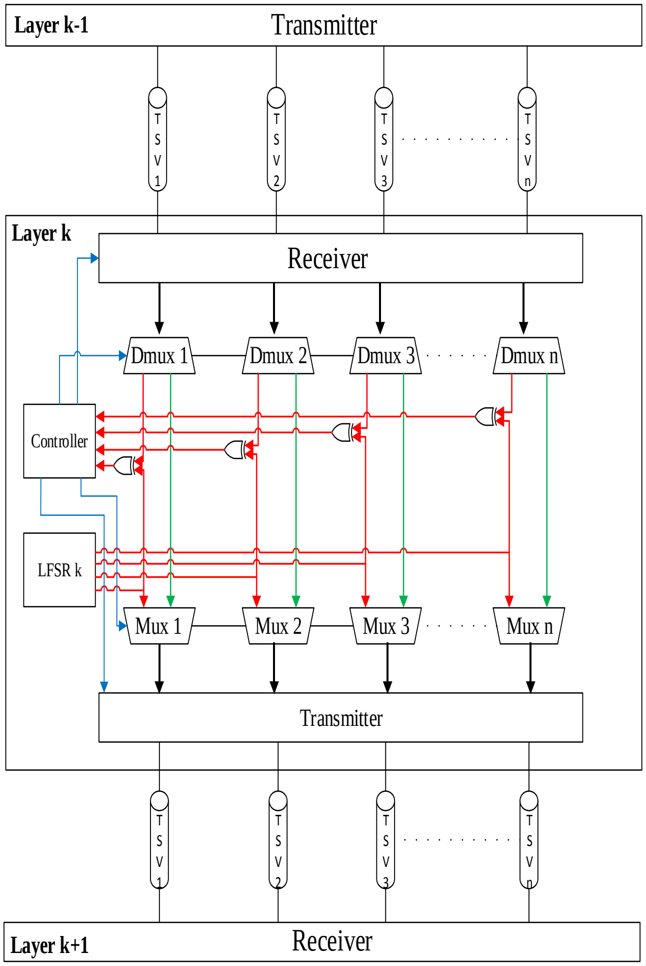


Fig. 2. The proposed architecture of the IBIST technique.

When the IBIST is in the test mode, the test vectors will be transmitted to the TSVs under test. In the end of test process, the test result well introduce to the analyzer. This last gives out an error signature, based on comparison between test data resaved from layer k-1 and data generated in layer k, which localize the defectives TSVs. This last step in test process verified the state of TSV, whether is correct or has another impact. While the analyzer is composed of XOR gate, according to the error signature (0, means that the TSV under test is defect free) or (1, means that the TSV is defective).

The controller in Fig. 2 is the management of the data traffic and getting synchronization between several blocks of circuits. It starts by producing a test pattern with LFSR bloc and continue the computation path by triggering the generated pattern from the layer k-1 to layer *k* via TSVs groups. Thereafter, the received test pattern at layer k will be analyzed according to same vectors generated in this last. Thus, the faulty TSVs are labeled and localized. Then, the result given by the analyzer stocked in the signature register. Moreover, the controller's backup register can be used to reduce the allocated area of BISR and IBIST .The algorithm controlling the IBIST process test procedure in all 3D-IC rows. The labeled localized signal information will be transmitted to the BISR component to start repair step.

B. *Architecture for TSV Repair*

To enhance the yield of TSVs, it is crucial to extend the test architecture shown in Fig. 2 to support the TSV repair step. Fig. 3 shows the conceptual scheme of the enhanced repair TSV architecture. The repair circuits are inserted at the two terminals the transmitter and the receiver. In addition, a repair register designed in the bottom die for storing the repair information (error signature and selected signal of MUXs) of all dies (each die have a repair register to store repair information of TSV between dies). Each terminal of a TSV has one MUX to switch the signal path of TSV and exchange defective ones by redundant TSV.

The error signature is extracted from IBIST component to the repair register in first step of repair process. In this architecture, a spare TSVs (STSV) is used to repair in case of one defective TSV. The MUXs is serves to switch signals that give more flexibility in the substitution of the faulty TSV and repair the crosstalk noise. Crosstalk defect is the most widely mentioned hindrance of 3-D. It emerges due to undesirable interaction between TSVs and active components like the MOSFET and the FinFET or among TSVs [21].

In the second step of repair process, two classes are proposed based on the number of defectives TSVs. The premier one is launched at the detection of unique defective TSV as illustrated in Fig. 4(a). The defective TSV is insulated and switched by redundant TSV. Fig. 4(b) and (c) describe the second-class, which start in case of detecting multi defectives TSVs based on TDMA fault tolerance technique.

In [22] a high-speed time division multiplexing (HSTDM) methodology was instituted to bridge problem of high frequency of traffic in FPGA communication while not altering the delay. Furthermore, according to this new added HSTDM technique, no additional delay was acquired with strict time budgeting and user-constraints [22]. The introducing of the TDM technique in the aim to reduce the complexity of the global interconnect hardware and to reduce the global area and power consumption, lead to the efficiency very high inter-FPGA communication. The range of transferring data on the TSVs at very high clock frequency is in the scale of Giga hertz [23]. By this ability to transmit data in high-frequency TSV used as high-speed interconnects in [24] and they are served as inter-die communication interface in 3-D NoC [25].

Therefore, the high-speed characteristics of TSV can be exploited to design TDMA-based 3-D IC. The high-data-rate transfer were achieved by the design operating at high clock frequency for inter-die communication this lead to no extra delay is induced on signal path.

The principal idea in repair using TDMA technique is to subdivide input bits in tow bundle. The first faulty TSV is switched by using the spare then the three bits of available TSVs is shipped from bottom to upper layer and stored in a FIFO.

After that the bit of second defective TSVs are transmitted on fault-free ones (Fig. 4(c)). If all bits are correctly saved in FIFO, then it will be triggered to the output. Input data bits are organized in two bundles in the case of multi defectives TSVs that lead to increase the number of spare TSVs for each bundle. For example, in Fig. 4(b) there are four TSVs and one spare.

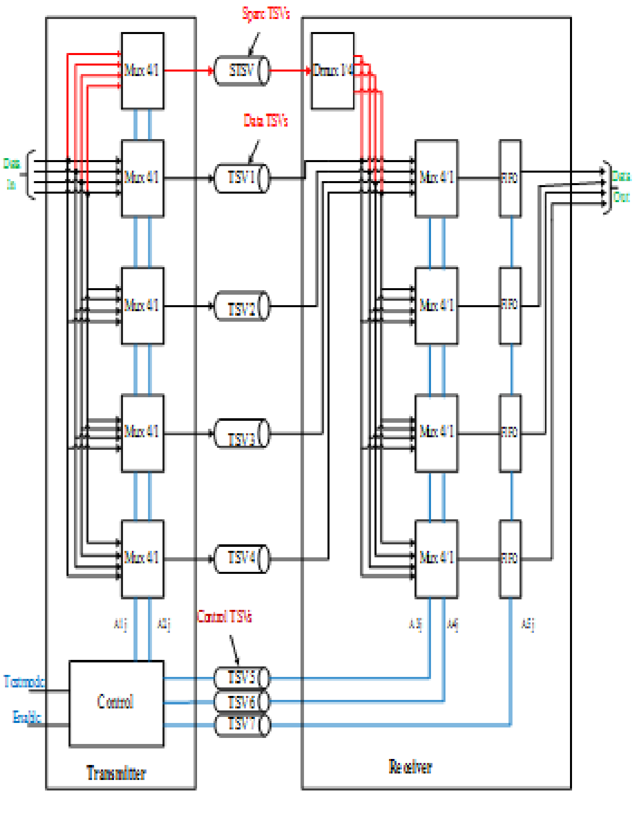


Fig. 3. Built-in Self-Repair (BISR) architecture for TSV defect.

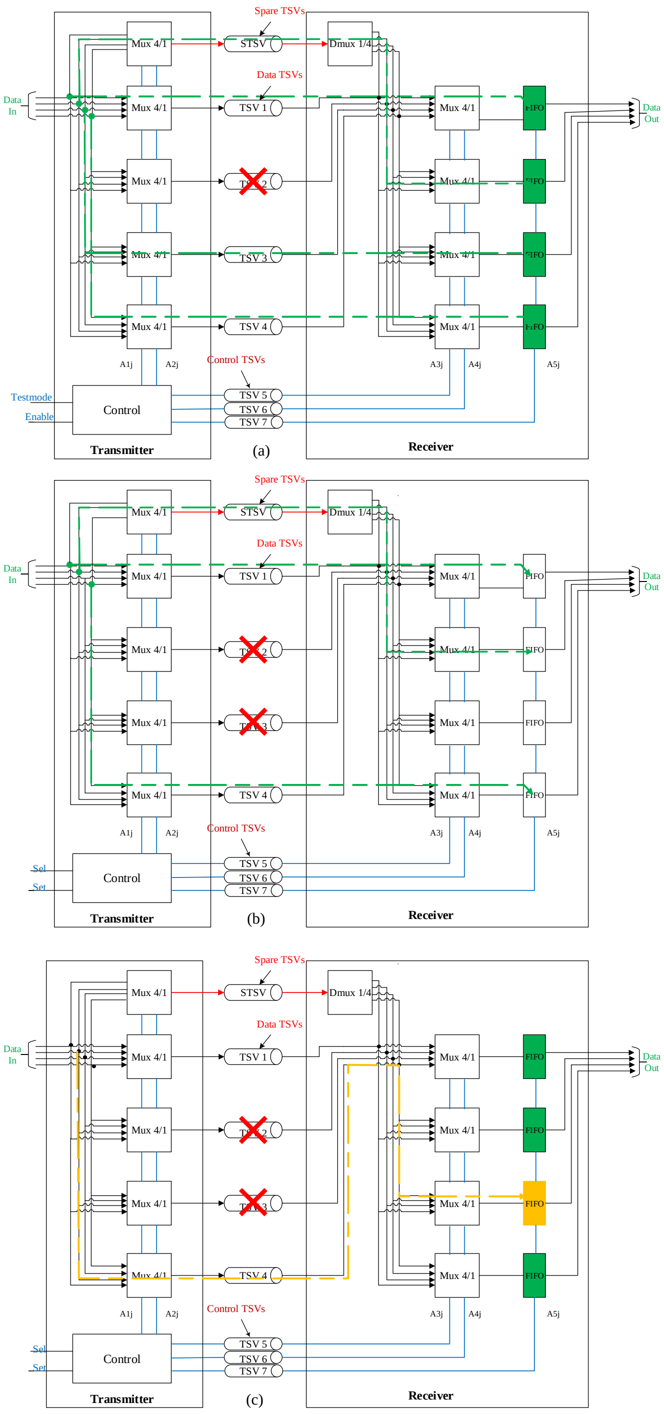
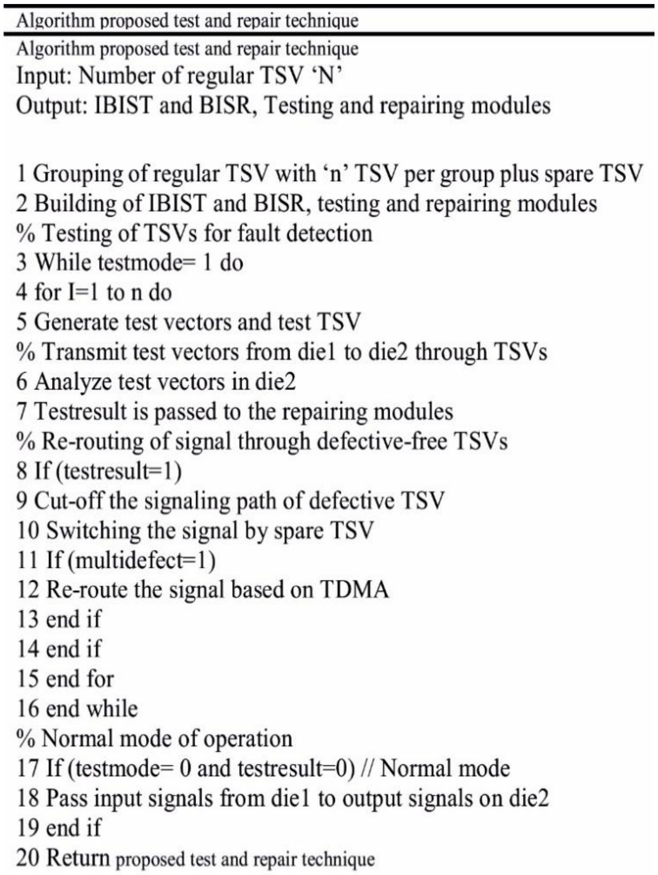


Fig. 4. Type of repair process: (a) Repair in case of one faulty TSV; (b)-(c) Steps of repair process in case of defectives TSVs.

Spare TSVs for each bundle. For example, in Fig. 4(b) there are four TSVs and one spare. In state of two defectives TSVs, TDMA is exploited and then the first bundle that comprise three bits are send on the two fault-free and spare TSVs in the repair process (Fig. 5). In Fig. 4(c), three TSVs are valid to send the last bit, as three spare instead of one (i.e. the other available TSV are counted as spares).



```
Algorithm proposed test and repair technique
Algorithm proposed test and repair technique
Input: Number of regular TSV 'N'
Output: IBIST and BISR, Testing and repairing modules

Grouping of regular TSV with 'n' TSV per group plus spare TSV
Building of IBIST and BISR, testing and repairing modules
% Testing of TSVs for fault detection
While testmode= 1 do
    for I=1 to n do
        Generate test vectors and test TSV
        % Transmit test vectors from die1 to die2 through TSVs
        Analyze test vectors in die2
        Testresult is passed to the repairing modules
        % Re-routing of signal through defective-free TSVs
        If (testresult=1)
            Cut-off the signaling path of defective TSV
            Switching the signal by spare TSV
            If (multidefect=1)
                Re-route the signal based on TDMA
            end if
        end if
    end for
end while
% Normal mode of operation
If (testmode= 0 and testresult=0) // Normal mode
    Pass input signals from die1 to output signals on die2
end if
Return proposed test and repair technique
```

Fig. 5. Algorithm of the proposed test and repair technique.

C. Enhanced Test Architecture for TSV Repair

A new test structure proposed for the TSV repair in the aim to reduce the area distribution and the cost. As described in previous section, to stock the control signals of repair architecture a new repair register is needed. The reconfigurable architecture has a new register for each TSV. As Fig. 3 shows, repair register aims to control the switching of repair MUXs. This proposition is to get the control signals of repair MUXs in the test process by overall enable signal (A_{ij}).

Fig. 6 presents a normal test operation of 3D-IC after repairing component receives error signature. In this simulation, When error signature (error in Fig. 6) is "0000000" that means there isn't a defect in the TSVs that leads to the normal operation. Then data_OUT receives data_IN without passing in repair component.

The new enunciated TSV self-repair architecture is auto performed in chip, based on the error signature delivering by IBIST, which identifies faulty TSVs. The BISR include a control, match register/MUXs, TSV mapping unit and TSV spared. When the control in BISR receives the trigger signal or address of faulty TSVs from IBIST, the repair process will be started immediately. After match register/MUX receives the labeled localized information from signature register via a controller in IBIST, the information should be decoded at first to acquire the reallocation of defective TSVs.

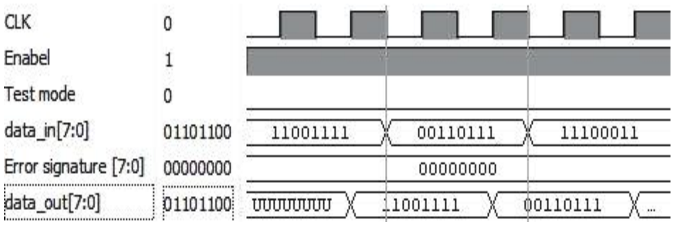


Fig. 6. Simulation of IBISR architecture in normal operation.

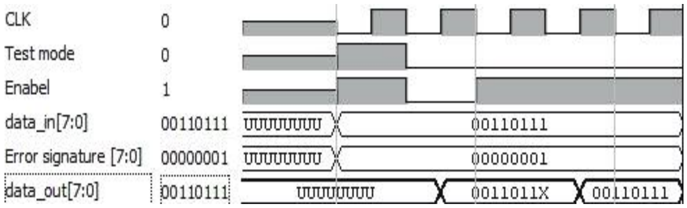


Fig. 7. Simulation of IBISR architecture in repairing operation.

TSV mapping unit will repair the defective TSVs by the method of shifting and replacing as shown in Fig. 7 that presents the simulation results of the repairing method. If there is any defective TSV, the MUX will isolate the defective TSV by labeled positional information, and switch the signal to the neighboring TSV. In Fig. 7, the test pattern is sent in 8 bits from layer1 to layer 2 of 3D-IC to test TSVs. After analyzing data in layer 2 data test, error signature stored in repair register and it used to repair TSV. In this simulation, there is “00000001”, the bit in high-level indicates the defect location. For instance, the BISR uses the first type of repair process; the signal through the defective TSV originally will be switched to the neighboring TSV. The count of STSVs (spare TSVs) has an effect on the performance of the overall BISR scheme and increases the reparability in the proposed architecture. Moreover, the yield of 3D IC will be positively influenced. Using the parallel processing of IBIST and BISR will reduce the processing time of BISR structure.

To compare the proposed repair architecture of TSV and those of existing methodology, the difference is in using the multiplexer (MUX) for repairing. In fact, the MUX increases the number of register and signal in the circuit but it has a good advantage. Moreover, the use of the coder makes the solution more flexible to move the defective TSV to another TSV and overcome all problem like the crosstalk defect.

III. Experimental Results

The 3D-ICs based on TSV technique have been proposed to overcome current limitations in manufacturing process of IC by stacking multiple dies via hundreds to thousands inter-die interconnection (TSVs). New proposed technique includes failure affect the system reliability and yield. To improve the repair process, a bundle of N TSVs is divided into set with spare ones. This leads to a much higher number of repairable faulty TSVs [11]. Furthermore, in the proposed approach, the spare TSVs are shared among all TSV clusters, the reparability rate R (the probability of successful repair) for data bundles in the multi defective TSVs that includes the uncorrelated TSV faults is:

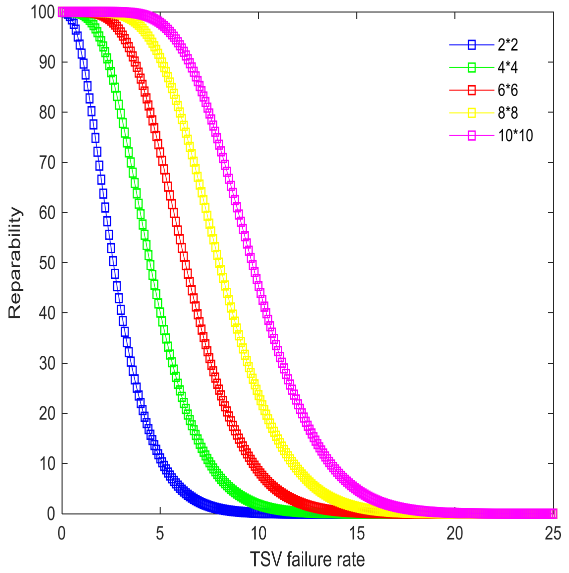


Fig. 8. Impact of spare TSV number on the reparability rate.

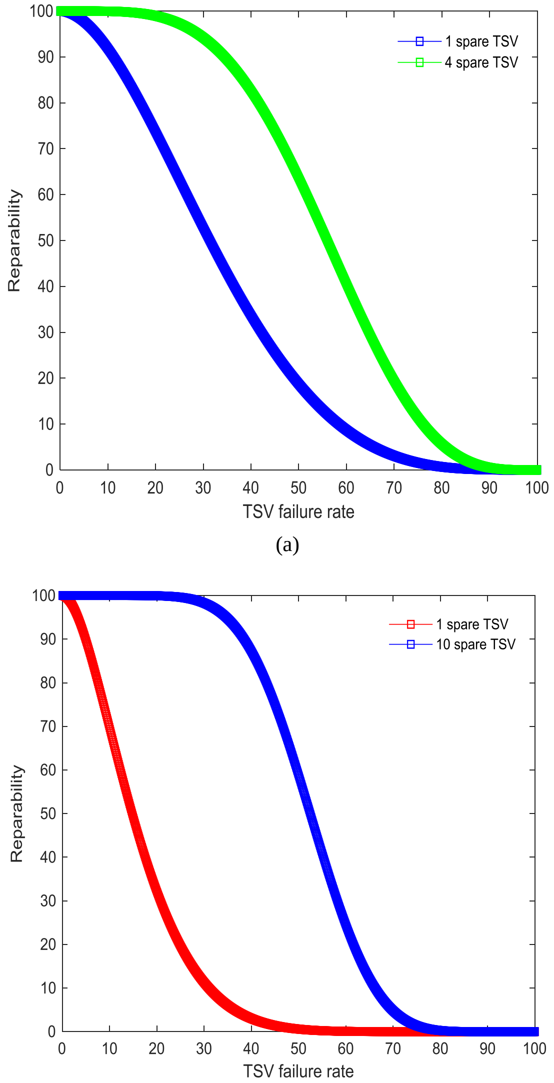
$$R=\sum_{k=N}^{N+r}\binom{N+r}{k}\left(1-D_{TSV}\right)^{k} D_{TSV}{}^{N+r-k} \qquad (1)$$

Where D_{TSV} is the failure rate for single TSV.

Fig. 8 presents the reparability and repair group for growing number of spares and rates of failure for N equal to 100 regular TSVs. For each repair group, there is one redundant TSV, while for two groups there are two spare TSVs and for ten groups there is ten spare. It can be concluded that the number of spare TSVs has a high impact to achieve higher reparability in- repair group.

In this repair architecture of multi defectives TSVs, data is transmitted in two-bundles whose can increase the reparability of circuit. For example, a group that consist of a set of four TSVs with one spare and in case of two defective TSVs, (Fig. 9(a)) two faulty free TSVs and one spare. With this solution, the total of spare TSVs is four. This is because partition of data in two- bundle give us in the first bundle one spare TSVs and in the second we can use all existing TSVs (two faulty free ones plus the redundant TSV) as spare then total of spare becomes four. In addition, the same for Fig. 9(b) in case of set of teen TSV.

The proposed test architecture was implemented in FPGA Virtex 5 ML 507 to extract the synthesis results. Tables I and II summarize the synthesis results of test UNIT architecture and the BISR, respectively. The tables show the number of slice registers, number of occupied Slice Lookup tables (LUTs), number of occupied slices and number of LUT Flip Flop pairs used in a various number of data bits of both proposed architecture.



(a)

(b)

Fig. 9. Reparability rate based on proposed repair methodology of group TSVs.

TABLE I. SYNTHESIS RESULTS OF TEST UNIT (IBIST)

Data Bits	Number of slices registers	Number of slices LUTs	Number of occupied slices	Number of LUT Flip Flop pairs used
4	46	23	17	54
8	82	39	29	96
16	154	64	50	176
32	298	115	100	339
64	586	177	204	680

TABLE II. SYNTHESIS RESULTS OF BISR

Data Bits	Number of slices registers	Number of slices LUTs	Number of occupied slices	Number of LUT Flip Flop pairs used
4	5	5	3	5
8	9	9	5	10
16	17	22	17	23
32	33	47	30	48
64	65	78	61	79

IV. CONCLUSION

In this paper, a novel repair process of a defective TSV in 3D-IC is presented. The proposed technique is based on two complimentary methods: IBIST for test, localization of faulty TSV and identify a kind the defects based on error signature, the second technique is BISR methodology for repair process. This last superpose to fault tolerance techniques, redundant TSV, and TDMA. In addition, the defective TSVs effectively isolated and repaired by a neighboring TSV of the proposed BISR structure. In case of multi defective TSVs, existing technique by dividing and transmitting data in two bundles to increase the reparability rate is improved based on TDMA technique. Experimental results and discussion show that the great yield improvement can be achieved (100% reparability for 30% of failure rate) with little area overhead penalty by using the proposed BISR structure.

REFERENCES

[1] L. W. Schaper, S. L. Burkett, S. Spiesshoefer, G. V. Vangara, Z. Rahman, and S. Polamreddy, “Architectural implications and process development of 3-D VLSI Z-axis interconnects using through-silicon vias,” IEEE Trans. Adv. Packag., vol. 28, no. 3, pp. 356–366, 2005.

[2] P. S. Andry et al., “Fabrication and characterization of robust through-silicon vias for silicon-carrier applications,” IBM J. Res. Dev., vol. 52, no. 6, pp. 571–581, 2008.

[3] R. S. Patti, “Three-dimensional integrated circuits and the future of system-on-chip designs,” Proc. IEEE, vol. 94, no. 6, pp. 1214–1224, 2006.

[4] J.-Q. Lu, “3-D hyperintegration and packaging technologies for micro-nano systems,” Proc. IEEE, vol. 97, no. 1, pp. 18–30, 2009.

[5] D. H. Jung et al., “Through silicon via (TSV) defect modeling, measurement, and analysis,” IEEE Trans. Compon. Packag. Manuf. Technol., vol. 7, no. 1, pp. 138–152, 2017.

[6] J. Li, X. Zhang, C. Zhou, J. Zheng, D. Ge, and W. Zhu, “New applications of an automated system for high-power LEDs,” IEEEASME Trans. Mechatron., vol. 21, no. 2, pp. 1035–1042, 2016.

[7] J. Li, L. Liu, L. Deng, B. Ma, F. Wang, and L. Han, “Interfacial microstructures and thermodynamics of thermosonic Cu-wire bonding,” IEEE Electron Device Lett., vol. 32, no. 10, pp. 1433–1435, 2011.

[8] C. Wang et al., “BIST methodology, architecture and circuits for pre-bond TSV testing in 3D stacking IC systems,” IEEE Trans. Circuits Syst. Regul. Pap., vol. 62, no. 1, pp. 139–148, 2015.

[9] Q. Xu, S. Chen, X. Xu, and B. Yu, “Clustered Fault Tolerance TSV Planning for 3-D Integrated Circuits,” IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 36, no. 8, pp. 1287–1300, Aug. 2017.

[10] A.-C. Hsieh and T. Hwang, “TSV redundancy: Architecture and design issues in 3-D IC,” IEEE Trans. Very Large Scale Integr. VLSI Syst., vol. 20, no. 4, pp. 711–722, 2012.

[11] M. Nicolaidis, V. Pasca, and L. Anghel, “Through-silicon-via built-in self-repair for aggressive 3D integration,” in On-Line Testing Symposium (IOLTS), 2012 IEEE 18th International, 2012, pp. 91–96.

[12] L. Jiang, Q. Xu, and B. Eklow, “On effective through-silicon via repair for 3-D-stacked ICs,” IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 32, no. 4, pp. 559–571, 2013.

[13] Y. Zhao, S. Khursheed, and B. M. Al-Hashimi, “Online fault tolerance technique for TSV-based 3-D-IC,” IEEE Trans. Very Large Scale Integr. VLSI Syst., vol. 23, no. 8, pp. 1567–1571, 2015.

[14] Y.-G. Chen, W.-Y. Wen, Y. Shi, W.-K. Hon, and S.-C. Chang, “Novel spare TSV deployment for 3-D ICs considering yield and timing constraints,” IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 34, no. 4, pp. 577–588, 2015.

[15] C.-W. Chou, J.-F. Li, Y.-C. Yu, C.-Y. Lo, D.-M. Kwai, and Y.-F. Chou, “Hierarchical test integration methodology for 3-D ICs,” IEEE Des. Test, vol. 32, no. 4, pp. 59–70, 2015.

[16] D. Arumí, R. Rodríguez-Montañés, and J. Figueras, “Prebond testing of weak defects in TSVs,” IEEE Trans. Very Large Scale Integr. VLSI Syst., vol. 24, no. 4, pp. 1503–1514, 2016.

[17] J. Park, M. Cheong, and S. Kang, “R 2-TSV: A Repairable and Reliable TSV Set Structure Reutilizing Redundancies,” IEEE Trans. Reliab., vol. 66, no. 2, pp. 458–466, 2017.

[18] W.-H. Hsu, M. A. Kochte, and K.-J. Lee, “Built-In Test and Diagnosis for TSVs With Different Placement Topologies and Crosstalk Impact Ranges,” IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., 2017.

[19] J. Park, H. Lim, and S. Kang, “FRESH: A New Test Result Extraction Scheme for Fast TSV Tests,” IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 36, no. 2, pp. 336–345, 2017.

[20] R. P. Reddy, A. Acharyya, and S. Khursheed, “A Cost-Effective Fault Tolerance Technique for Functional TSV in 3-D ICs,” IEEE Trans. Very Large Scale Integr. VLSI Syst., 2017.

[21] S. Mondal, S.-B. Cho, and B. C. Kim, “Modeling and Crosstalk Evaluation of 3-D TSV-Based Inductor With Ground TSV Shielding,” IEEE Trans. Very Large Scale Integr. VLSI Syst., vol. 25, no. 1, pp. 308–318, 2017.

[22] SNUG Silicon Valley. (2015). High Speed TDMA. [Online].Available: https://www.synopsys.com/news/pubs/snug/2015/siliconvalley/ta07_maheshwari_pres_snps.pdf

[23] I. Ndip et al., “High-frequency modeling of TSVs for 3-D chip integration and silicon interposers considering skin-effect, dielectric quasi-TEM and slow-wave modes,” IEEE Trans. Compon. Packag. Manuf. Technol., vol. 1, no. 10, pp. 1627–1641, 2011.

[24] A. Sheibanyrad and F. Pétrot, “Asynchronous 3D-NoCs Making Use of Serialized Vertical Links,” in 3D Integration for NoC-based SoC Architectures, Springer, 2011, pp. 149–165.

[25] F. Sun, A. Cevrero, P. Athanasopoulos, and Y. Leblebici, “Design and feasibility of multi-Gb/s quasi-serial vertical interconnects based on TSVs for 3D ICs,” in VLSI System on Chip Conference (VLSI-SoC), 2010 18th IEEE/IFIP, 2010, pp. 149–154.University Science, 1989.

Intrusion Detection and Prevention Systems as a Service in Could-based Environment

Khalid Alsubhi
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

Hani Moaiteq AlJahdali
Faculty of Computing and Information Technology Rabig
King Abdulaziz University, Jeddah Saudi Arabia

Abstract—**Intrusion Detection and Prevention Systems (IDPSs) are standalone complex hardware, expensive to purchase, change and manage. The emergence of Network Function Virtualization (NFV) and Software Defined Networking (SDN) mitigates these challenges and delivers middlebox functions as virtual instances. Moreover, cloud computing has become a very cost-effective model for sharing large-scale services in recent years. Features such as portability, isolation, live migration, and customizability of virtual machines for high-performance computing have attracted enterprise customers to move their in-house IT data center to the cloud. In this paper, we formulate the placement of *Intrusion Detection and Prevention Systems (IDPS)* and introduce a model called *Incremental Mobile Facility Location Problem (IMFLP)* to study the IDPP problem. Moreover, we propose a novel and efficient solution called *Adaptive Facility Location (AFL)* to efficiently solve the optimization problem introduced in the IMFLP model. The effectiveness of our solution is evaluated through realistic simulation studies compared with other popular online facility location algorithms.**

Keywords*—*Facility Location Problem; Intrusion detection and Prevention Systems; Cloud Computing

I. Introduction

Cloud computing has become a cost-effective model for sharing large-scale services in recent years. Its success is due to the attractive features offered by the underlying virtualization concept, including portability, isolation, live migration, and customizability of virtual machines. Popular examples of cloud-based services are Microsoft Azure, Google AppEngine, and Amazon Elastic Computing Cloud (EC2). Cloud services are generally categorized into three areas: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). In SaaS, a third-party provider host customer's application over the Internet (i.e., Rackspace and SAP Business ByDesign). In PaaS model, both hardware and software are provided and hosted by third-party (i.e., Google App Engine and Microsoft Windows Azure). Finally, IaaS refer to providing virtualized computing resources, usually in terms of VMs (i.e., Amazon EC2, GoGrid and Flexiscale).

Intrusion Detection and Prevention Systems are an essential defensive measure against a range of attacks [44, 47]. In enterprise networked system, IDPSs examine packets sent over networks and trigger alerts when malicious content is discovered and defend against attacks when prevention mode is active. Most issues regarding security in cloud systems are inherited by the current enterprise network [34]. Traditional distributed IDPSs are best practice in providing security for large scale networks. However, the deployment of distributed IDPSs in cloud systems raise many challenges due to the diversity of its services and the complexity of its infrastructure [43].

Network Functions Virtualization (NFV) [1] [2] promises a reprive from the vertically integrated hardware middlebox model followed for decades, by advocating the use of software Network Functions (NFs) running on commodity hardware. This means a reduced acquisition and operational costs, flexible programability, and easier management [31] [42]. Another orthogonal idea is the Software Defined Networking (SDN) that advocates flexible programability in the network. This is done by the separation of the control-plane from the data plane and centralized logical control of the network. SDN simplifies the overall management of the network by allowing deeper programability of the networking devices. Leveraging SDN in environments where NFV are used can leads to several interesting use cases. The high precision control of forwarding elements (switches) provided by SDN can be used to orchestrate traffic patterns between various appliances and NFVs across a data center [22]. In recent years, the cloud has become a mature platform for deploying scalable and cost effective services. With huge growth forecasts, the public cloud industry has grown to become a multi-billion dollar industry [6]. Combining the agility of the cloud with the flexibility of Virtualized Network Functions (VNFs) and the fine-grained control of SDN can bring about a new class of cloud based services for IDPSs [13].

In this paper, we introduce a model in which infrastructure providers support Vritual Intrusion Detection and Prevention Systems (IDPSs) as a Service (IDPSaaS) by leveraging NFV, SDN, and cloud. IDPSaaS services can be enabled or disabled for tenant's Virtual Machines (VMs) on their demands and can be scaled up or down to cope with their service workloads. Moreover, the deployment of multiple IDPS instances of a network functions motivates an interesting challenge, which we call Intrusion Detection and Prevention Systems Placement problem (IDPSP). In order to study the IDPSP problem, we propose Incremental Mobile Facility Location Problem (IMFLP) based on the online facility location problem. IMFLP takes into account the online actions, such as live migrations in cloud, which are ignored in almost all of the existing models [21]. To the best of our knowledge, it is the first time that the online version of facility location problem has been used to study placement of IDPS. Furthermore, we present an efficient solution for the optimization problem defined in this model called Adaptive Facility Location (AFL). This solution by employing online actions, such as migrations and switches, adjusts the placement of IDPS instances to efficiently adapt to changes in service demands. The effectiveness of our solution is evaluated though realistic simulation studies and empirically compared with several popular online facility location algorithms.

The remainder of this paper is organized as follows. In section II, we formulate the IDPSP problem and present the IMFLP model for studying this problem. We present AFL in section III and conduct experiments to evaluate this algorithm in section IV. The related works are discussed in section V. Finally, we conclude and discuss about future works in section VI.

II. Problem Formulation

As mentioned before, the placement module receives an event of an arrival or leaving of a demand, and by information and functions supported by the management module, adjusts the placement of facilities. In this section, we introduce the Intrusion Detection and Prevention System Placement problem (IDPSP) in section II-A. In section II-B we formally define our model of facility location problem that can be used for modeling the IDPSP problem.

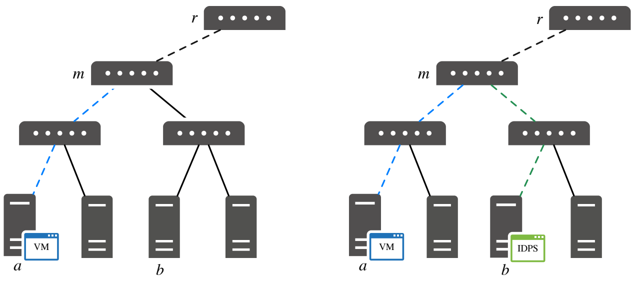
A. Intrusion Detection and Prevention System Placement Problem (IDPSP)

Without loss of generality, we introduce this problem through an example. Suppose that an infrastructure provider offers a IDPSaaS service. From the client's point of view, her VMs can be installed any time, and the IDPSaaS service can be requested and enabled for her VMs at any moment. Moreover, VMs are different and have various service workload on the IDPS instances (*IDPSInst*). Let call each unit of VM's workload as a *demand*. Thus, we can view the problem as dynamic demands that should be served by multiple IDPSInsts.

From the view point of the infrastructure provider, enabling this service incurs certain amount of the installation, operational, and management costs. The installation cost includes the cost of resource consumption of a host machine on which a IDPSInst is installed, and the cost of certain messages between the controller and the host. In our system, all IDPSInsts are same, and therefore the installation cost is same for all IDPSInsts. The operational cost consists of the traffic processing delay cost, and the cost of steering the traffic to the IDPSInst and then to the destination VM. It can be shown that the cost of steering the traffic is related to the distance between IDPSInst and the VM. Finally, the management cost includes the cost of certain statistics collection and syncronization messages between the controller and the IDPSInsts. The management cost is related to the cost of shortest path between the controller(s) and the IDPSInst. Optimizing the management cost is similar to the placement of SDN controllers [8] [29], and is outside of the scope of the current paper.

Considering Figure 1, suppose that a VM exists on host a. As illustrated in Figure 1(a), when there is no IDPSInst enabled (the service-less case), the internet traffic travels the shortest path from the core switch r to the host a with an intermediate switch m. Let $d(r,a)$ represents the cost of the shortest path between r and a. In the service-less case, the cost of traffic traversal is $d(r,a) = d(r,m) + d(m,a)$. On the other hand, as shown in Figure 1(b), when the IDPSInst is installed on a host b (the IDPSInst enabled case), extra costs are paid. Certain amount of b's resources are allocated to the IDPSInst and certain controlling messages from the controller are exchanged with the host b (the installation cost). This installation cost is independent of the *where* IDPSInsts are located, and only depends on the *number* of IDPSInsts. Moreover, IDPSInst adds certain processing delay time t, and the traffic travels a longer path (the operational cost). Delay time t is independent of where the IDPSInst is placed and related to how much traffic is *assigned* to. Additionally, the traffic is steered from core switch r to host b, and from host b to the host a. In this case, the cost of the traffic steering is $d(r,b) + d(b,a) = d(r,m) + 2d(m,b) + d(m,a)$ (We assume that the shortest path cost is symetric). By deducting the cost of service-less case, the extra cost in the IDPSInst enabled case is $2d(m,b)$. Because a and b are in the same level (host level) $d(m,b) = d(m,a)$, and therefore the extra cost is $2d(m,b) = d(m,b) + d(m,a) = d(a,b)$, which is the cost of shortest path between host a containing IDPSInst and host b containing the VM.

There is another complexity dimension that makes the problem even more complicated. Assignments of demands to the IDPSInsts are not irrevocable decisions, and demands can be reassigned to other IDPSInsts. However, these reassignments are not free of charge and associated with certain costs related to the routing reconfiguration and transferring source IDPSInst's internal state to the destination IDPSInst [22]. Furthermore, after assigning more demands to an IDPSInst during the time, this IDPSInst can migrate to another location in order to minimize its distance to the VMs and subsequently reduce the operational costs; however, migrations are not free and are associated with certain cost.



(a) Service-Less Case (b) IDPSInst Enabled Case

Fig. 1. The comparison of the traffic path

Any model describing this problem must consider the dynamic nature of the problem, optimizing the installation and operational cost of the IDPSInsts, and possibility of assignments switches and IDPS migrations.

B. Increamental Mobile Facility Location Model

In this section, we introduce a new model of facility location problem called **Incremental Mobile Facility Location Problem (IMFLP)** to study the IDPSP problem. Before describing our model, we briefly describe why a new model of this problem is needed to be formulated. The details of other existing models will be discussed in the section V-B.

The offline model of facility location problem has been studied comprehensively in the literature [15, 9, 40, 16]. Unfortunately, it cannot describe IDPSP, becuase this model requires demands and their locations to be known in advanced, but in IDPSP, VMs are installed at any moment and subsequently their demands are not known beforehand. In other words, assignments of demands to IDPSInsts are done without knowledge about the future demands. Hence, the online model of this problem should be used. However, the existing online models in the literature (as will be discussed in section V-B) are not representative for our problem, thus we design a new model of this problem. Our IMFLP model relaxes certain constraints of the these models and resolve their limitations in describing IDPSP problem to model migrations and assignments switches.

We describe our model of facility location problem by defining the *space and metrics*, *facilities*, *demands*, and *allowed actions*.

Space and metrics. Given a connected weighted graph $M = (V, E)$ representing the architecture of the data center network, where V denotes the set of nodes (switches or hosts), and $E : V \times V \to \mathcal{R}^+$ represents the set of network links. $V_{hosts} \subset V$ represents host nodes in which demands and facilities can reside. The shortest path between two nodes $p, q \in V$ is denoted by $d(p, q)$. We also use the notation of $d(V', p)$ to denote the shortest path between the closest node in a subset $V' \subset V$ to a node $p \in V$. Moreover, let $B(p, r) = \{q \in V_{hosts}, r \in \mathcal{R}^+ | d(p, q) \leq r\}$ indicates the nodes within distance r to the node p (the points that lie inside or on the ball with center p and radius r). We assume that the distance metric is symetric and satisfies triangle inequality.

Facility. In IMFLP, a facility represent a VNF instance and is uncapacitated. The location of a facility z in the space is identified by the $\gamma(z) \in V_{hosts}$. We use term *open* or install interchangeably for the installation of a facility. Besides, the notation $C(z)$ represents a set of demands that are assigned to a facility z (z's cluster).

Demand. A demand u denotes a unit of service workload of a VM. Similar to a facility, the location of a demand is given by $\gamma(u) \in V_{hosts}$, which is equal to the node that VM resides. We use term *arrive* to denote that a new demand from a VM should be served. We also assume that each VM has a correct number of demands.

Allowed Actions. In IMFLP following actions are allowed:

- A *facility* can be *opened* in any node $p \in V_{hosts}$ at any time by paying the installation cost $f \in \mathcal{R}^+$. A facility also can *migrate* to another location with the migration cost $k \in \mathcal{R}^+$. we assume that $k < f$. Moreover, a facility can be closed at any time, and its installation cost is refunded. However, if any demand is assigned to that facility, they should switched to a new facility and for each switch, the certain amount of cost as described next is payed.
- A *demand* is allowed to *arrive* and *leave* at any time in any node $p \in V_{hosts}$. The migration of a VM can be modeled by leaving of its demands and their arrivals in the destination node. Furthermore, a demand assignment can be *switched* to another facility by paying the switch cost $h \in \mathcal{R}^+$. We assume that $h \leq k < f$.

Additional notation. Please note, for a demand u and a facility z, instead of $d(\gamma(z), \gamma(u))$ we simply use $d(z, u)$ to represent their distance. In addition, we define $(x - y)_+ = \max(0, x - y)$ for $x, y \in \mathcal{R}^+$.

The model is described as follow. Upon arrival or departure of a demand u_t at time t (the input of our model), a new facility ω or a set of facilities can be opened, closed, or migrated. Likewise, a subset of demands can be switched to other facilities. Therefore, the following costs are defined at time t:

1) **Total installation cost** ($\mathcal{C}_{ins}$) is the cost of installation of a set of facilities F_t at time t.

$$\mathcal{C}_{ins} = |F_t| f \tag{1}$$

Here, $|F_t|$ denotes the number of facilities.

2) **Total operational cost** ($\mathcal{C}_{op}$) represents the operational cost of a set of facilities F.

$$\mathcal{C}_{op} = g \sum_{z \in F_t} \sum_{u \in C(z)} d(u, z) \tag{2}$$

As shown in equation 2, this cost is defined based on the shortest paths between facilities and their assigned demands.

3) **Total migration cost** ($\mathcal{C}_{mig}$) is the cost of migration of a set of facilities since start time until time t.

$$\mathcal{C}_{mig} = k \sum_{i=2}^{t} \sum_{z \in F_i} |\gamma_{i-1}(z) \neq \gamma_i(z)| \tag{3}$$

In equation 3, $\gamma_i(z)$ represents the location of facility z at time i. Please note that term $|\gamma_{i-1}(z) \neq \gamma_i(z)|$ is 1 if $\gamma_{i-1}(z) \neq \gamma_i(z)$, otherwise it is 0.

4) **Total switch cost** ($\mathcal{C}_{sw}$) denotes the switch cost of a set of demands L_t at time t.

$$\mathcal{C}_{sw} = h \sum_{i=2}^{t} \sum_{u \in L_i} |\phi_{i-1}(u) \neq \phi_i(u)| \tag{4}$$

Here, $\phi_i(u)$ represents the facility that demand u is assigned to at time i. Note that $|\phi_{i-1}(u) \neq \phi_i(u)|$ is equal to 1 if $\phi_{i-1}(u) \neq \phi_i(u)$, otherwise it is 0.

The objective of the optimization problem in the IMFLP formulation is to minimize the overall cost ($\mathcal{C}_{overal}$) as defined in equation 5.

$$\mathcal{C}_{overal} = \mathcal{C}_{ins} + \mathcal{C}_{op} + \mathcal{C}_{mig} + \mathcal{C}_{sw} \tag{5}$$

The IDPSP problem can be reduced to the optimization problem defined in the IMFLP model. This optimization problem is NP-hard (facility location problem is NP-hard, and our online model is even more complicated than original problem). Motivated by this observation, we developed an online algorithm for IMFLP model.

III. Adaptive Facility Location (AFL)

In this section, we propose our solution, *Adaptive Facility Location (AFL)*, for the optimization problem introduced in the IMFLP model. We introduce two novel algorithms that use the simple idea of profit and loss for handling a demand arrival and a demand departure.

However, before describing our model, we justify our selection over other candidate approaches. In the area of SDN, some of ubiquitous approaches for modeling the optimization problems are the linear programming [28, 49], simulated annealing [48, 36], and Markov approximation [30, 41]. The linear programming approach solves an offline problem, and is not descriptive enough to model the dynamicity and online nature of these kind of problems. In addition, the linear programming is known that is slow. To deal with the dynamic nature of these optimization problems, simulated annealing and markov approximation are used. In the simulated annealing techniques, at each step again an offline problem is defined, and known to be trapped in the local minimums, and might suffer from the bad initial state. Finally, Markov chain techniques might also affected from bad initial state and slow convergence to the steady state.

A. Demand Arrival

Two functions namely, *migration potential* and *installation potential* are defined to represent how far facilities and assignments of demands are from the *optimal* or *stable* configuration, and how much *profit* is gained by the installation or migration of a facility, respectively. Then by comparing with the cost of certain *actions* (the loss), AFL decides which action is applied.

Installation potential function (Pot_{ins}) is defined as equation 6. This function represents how much of the current cost can be reduced by installation of a facility at a node $p \in V_{hosts}$. In this equation, u_t denotes a new arrived demand at time t. F_{t-1} and L_{t-1} represent a set of opened facilities and demands at time $t-1$ just before arrival of u_t. The first term computes the profit of the situation where u_t is assigned to a facility that can be installed at node p against when u_t is assigned to the closest facility in F_{t-1}. The second term shows that if some demands are switched to a facility that can be installed at node p, how much the operational cost of related to these demands will reduce (recall that each switch incurs switch cost h).

$$\begin{aligned} Pot_{ins}(p) = g.\Big(& \big(d(F_{t-1}, u_t) - d(p, u_t)\big)_+ \\ & + \sum_{v \in L_{t-1}} \big(d(\phi(v), v) - d(p, v) - h\big)_+ \Big) \end{aligned} \tag{6}$$

Migration potential function (Pot_{mig}) is defined in equation 7. This function describes how much the migration of a facility z from its current location to a node p reduces its total operational cost when a new arrived demand u_t is assigned to z as well. This function can be interpreted in another sense as well. Each demand $v \in C_{t-1}(z)$ attempts to reduce its cost by pulling facility z toward its location $\gamma(u_t)$. If a new arrival demand u_t will be assigned to z, u_t also tries to pull facility z toward itself. The potential function $Pot_{mig}(z, p)$ represents how much z becomes more stable by migration form $\gamma(z)$ to p. In other words, z is close enough to each demand $v \in C_{t-1}$ and more closer to u_t in comparison to facilities F_{t-1} including z itself.

$$\begin{aligned} Pot_{mig}(z, p) = g.\Big(& d(F_{t-1}, u_t) - d(p, u_t) \\ & + \sum_{v \in C_{t-1}(z)} \big(d(z, v) - d(p, v)\big)\Big) \end{aligned} \tag{7}$$

Algorithm 1 shows AFL algorithm (for the sake of simplicity, we drop t subscript, but we insist that the presented algorithm is run at time t). By exploiting the aforementioned functions, AFL attempts to improve the current placement of facilities and current assignment of demands. Upon arrival of a new demand u, AFL considers three actions:

1) **Installation action**: Installation of one new facility in the best place with the best possible switches.
2) **Migration action**: Migrating one of the existing facilities (the best one) without any demand switches and assigning u to this facility.
3) **Assignment action**: Assigning u to the nearest existing facility.

As shown in algorithm 1, AFL computes the installation and migration potentials. By comparing the the computed values, AFL applies the best action. For the installation action, AFL calculates the installation potential Pot_{ins} for every point p in the distance of f from u ($B(u, f)$). AFL selects the best point ω_{ins}, which maximizes the Pot_{ins}. If AFL decided to apply this action, it switch the neighbor demands to ω_{ins}, if this switches reduce the service cost and the deducted service cost is bigger than switch cost h.

For the migration action, AFL computes the migration potential Pot_{mig} for each facility z and for each point p in the space V_{hosts}. Eventually, AFL chooses the best facility ω_{mig} to migrate to point ρ that maximizes Pot_{mig}.

Ultimately, AFL decides which action is applied. The installation action is considered first. If it is beneficial ($Pot_{ins}(\omega_{ins}) - f > 0$), and its profit is greater than best migration action ($Pot_{mig}(\omega_{mig}, \rho) - k$), AFL applies the installation action. Otherwise, the best migration action is considered. If this migration is beneficial ($Pot_{mig}(\omega_{mig}, \rho) - k > 0$),

Algorithm 1 AFL-Demand Arrival

```
$F \leftarrow \emptyset; L \leftarrow \emptyset$;
for all new demand $u$ do
    $L \leftarrow L \cup \{u\}$;
    $\rho_{ins} \leftarrow \arg\max_{p \in B(u,f)}\{Pot_{ins}(p)\}$;
    $p_{ins} \leftarrow Pot_{ins}(\rho_{ins})$;
    $\omega_{mig}, \rho_{mig} \leftarrow \arg\max_{z \in F, p \in V_{hosts}/\{\gamma(z)\}}\{Pot_{mig}(z,p)\}$;
    $p_{mig} \leftarrow Pot_{mig}(\omega_{mig}, \rho_{mig}, u)$;
    if $(p_{ins} - f > 0) \wedge (p_{ins} - f \geq p_{mig} - k)$ then
        $\omega_{ins} \leftarrow$ open a facility at $\rho_{ins}$
        $F \leftarrow F \cup \{\omega_{ins}\}$
        Switch facility of each demand $v \in L/\{u\}$ if $d(\phi(v), v) > d(\rho_{ins}, v) + h$;
        Assign $u$ to the nearest facility;
    else if $p_{mig} - k > 0$ then
        Assign $u$ to $\omega_{mig}$;
        Migrate $\omega_{mig}$ to point $\rho_{mig}$;
    else
        Assign $u$ to the nearest facility;
    end if
end for
```

AFL applies this action. Otherwise, it assigns u to the nearest facility.

B. Demand Departure

Similar to the case of demand arrival, AFL defines closing potential and migration potential functions to represent how far the current configuration of a facility whose demand departures is from the stable configuration. Let u'_t denotes a demand departuring at time t, and $z = \phi(u'_t)$ represents the facility to which u'_t was connected at time $t-1$ just before departure.

Closing potential function (Pot_{cls}) is defined in equation 8. This function denotes the profit of *closing* a facility and switching its demands to the closest facilities.

$$Pot_{cls}(z) = g \sum_{v \in C_{t-1}(z)/\{u'_t\}} \big(d(z,v) - d(F_{t-1}, v) - h\big) \quad (8)$$

Migration potential fucntion (Pot'_{mig}) for the departure of a demand is defined by equation 9. It can be interpreted exactly same as the migration potential for a demand arrival.

$$Pot'_{mig}(z,p) = g \sum_{v \in C_{t-1}(z)/\{u'_t\}} \big(d(z,v) - d(p,v)\big) \quad (9)$$

AFL for the departure considers two actions:

1) **Closing action**: Closing facility z and assigning each of its demands to the closest facility in F_{t-1}/z.
2) **Migration action**: Migration of facility z to another location to serve $C_{t-1}(z)/u'_t$ more efficiently.

Algorithm 2 represents AFL's algorithm for handling a demand departure. For the sake of simplicity, we omit subscript t from the notation. AFL computes the closing potential of facility z serving u' and the best migration potential. First, AFL considers the closing action. If closing z is beneficial ($p_{cls} + f > 0$) and is more profitable than migration action ($p_{cls} + f \geq p'_{mig} - k$), AFL applies this action. Otherwise, the migration of z is considered, and if this action is profitable ($p'_{mig} - k > 0$), AFL migrates z to ρ_{mig}. If none of closing and migration actions are beneficial, AFL only remove demand u' from the list of demands.

Algorithm 2 AFL-Demand Departure

```
$z \leftarrow \phi(u')$;
$p_{cls} \leftarrow Pot_{cls}(z)$;
$\rho_{mig} \leftarrow \arg\max_{p \in V_{hosts}}\{Pot'_{mig}(z,p)\}$;
$p'_{mig} \leftarrow Pot'_{mig}(z, \rho_{mig})$;
if $(p_{cls} + f > 0) \wedge (p_{cls} + f > p'_{mig} - k)$ then
    Switch facility of each demand $v \in C(z)/\{u'\}$ to the closest facility in $F/\{z\}$;
    Close facility $z$;
    $F \leftarrow F/\{z\}$;
else if $p'_{mig} - k > 0$ then
    Migrate $z$ to point $\rho_{mig}$;
end if
$L \leftarrow L/\{u'\}$
```

IV. Experiments

We evaluated the effectiveness of our placement algorithm in several simulation studies. We implemented our AFL algorithm in a discrete event simulator and compared it to other five popular algorithms namely: FFL [19], AFL [18], OPTFL [17], RFL [38], and SNFL [20]. The details of these algorithms will be discussed in section V. The OPTFL, AFL, and FFL algorithms have certain input parameters. We ran these algorithms for miscellaneous values of parameters and did not observe substantial difference. Ultimately, their input parameters were set to the values suggested by their authors, specifically for OPTFL $\alpha = 10$ [17], for AFL $\alpha = 18, \beta = 8.0, \psi = 4.0$ [18], and finally for FFL $x = \frac{19}{8}$ [19].

In the last decade a tremendous research has been done to search for an efficient and inexpensive data center networks (DCN) architecture. Several architectures like fat-trees [3], VL2 [24], Portland [39], BCube [25] and DCell [26] have been proposed to address different challenges of current DCN architectures such as scalability, agility, and reconfigurability. For the experiment, we select Al-Fares et al. fat-tree [3] architecture. This architecture is one of the well known DCN architectures [27] [37] [7]. Fat-trees are more scaleable and reliable than conventional tree-based architectures. This topology allows us to leverage identical cheap commodity switches in the all communication layers. In the theory, the over-subscription ratio of this rearrangeable architecture is $1 : 1$, which means that this architecture is non-blocking; however, in the practice preventing packet reordering might make it difficult to guaranty non-blocking network. The fat-tree topology proposed by [3] is a k-ary tree in which k denotes number of ports and number of pods. This topology

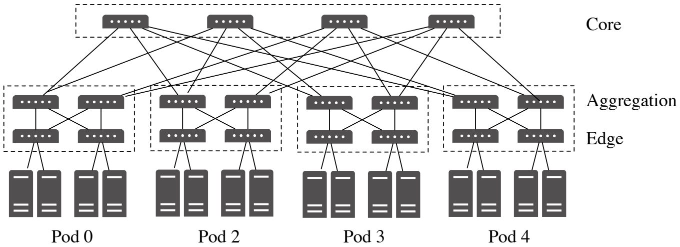


Fig. 2. The fat-tree architecture for 4 pods ($k = 4$)

connects homogeneous switches with the same number of k ports. As depicted in Figure 2, the Al-Fares's fat-tree consists of three switch layers. At the highest level, there are $(\frac{k}{2})^2$ core-switches. Each core-switch is connected to all k pods (i-th port of a core-switch is connected to the i-pod). A pod contains k switches ($\frac{k}{2}$ aggregation-switches and $\frac{k}{2}$ edge-switches). At the second level, aggregation switches are connected to $\frac{k}{2}$ of core-switches upward and $\frac{k}{2}$ edge-switches downward. Furthermore, each aggregation-switch is only connected to edge-switches that are in the same pod. At the third level, edge-switches are linked to the $\frac{k}{2}$ hosts dipping and $\frac{k}{2}$ aggregation-switches mounting. There are $\frac{k^3}{4}$ hosts which are located in the leaves of this architecture. For all experiments, the oversubscribing ratio was set to $1 : 1$, which means that this architecture is non-blocking.

The demands are generated randomly (only in the leaves) from the uniform and normal distributions. The mean and standard deviation parameters of the normal distribution were set to 0.5 and 0.1, respectively, and each generated value was multiplied by the number of leaves and a demand was generated at the position of the result. Moreover, in all experiments, the value of parameter g was set to 1. All algorithms receive one demand at a time and reconfigure the placement of the facilities to serve this demand upon its arrival. The costs of installation, migration, and switch for all algorithms are collected. For each configuration, the average of 10 tests has been reported as the final result. We have conducted three experiments to evaluate the the behavior of AFL under different circumstances.

A. Impact of Number of Demands

In this experiment, the impact of number of demands on the behavior of AFL is examined. As depicted in Figure 3, five tests for 1024, 2048, 3072, 4096, and 5120 number of demands for uniform (Figure 3(a)) and normal distribution (Figure 3(b)) are conducted. We assign 6, 2, 1 for f, k and h, respectively. The idea behind choosing these values is that we assume that the cost of installation of a facility f is always greater than the cost of migration k and switching h, and the cost of migration is equal or greater than the cost of switching. Moreover, The space is the fat-tree with 1024 hosts. For each test, each algorithm receives one demand at a time and returns a placement of facilities.

TABLE I: AFL's Costs

Demands	Facility	Service	Switch	Migration	Total
64	97.30%	2.70%	0.00%	0.00%	584.8
128	95.51%	4.43%	0.04%	0.02%	1667.7
256	92.40%	7.37%	0.15%	0.08%	3551.9
512	87.54%	11.66%	0.42%	0.38%	6525.8
1024	72.19%	24.36%	2.15%	1.30%	12798.7

Note: We omit word *cost* from the headers. For instance, by the *Facility* we mean *Facility Cost*

As shown in Figure 3, the total cost of all algorithms in the uniform distribution are considerably greater than normal case. The reason is that in the uniform case the demands spread in more hosts in comparison to the normal distribution that demands tend to arrive in the middle hosts. As depicted, AFL outperforms all other algorithms in all cases for both distributions. The average of overall costs of AFL is 11.82% and 14.46% lower than the second best algorithm in the case of uniform and normal distributions, respectively.

B. Impact of Number of Hosts

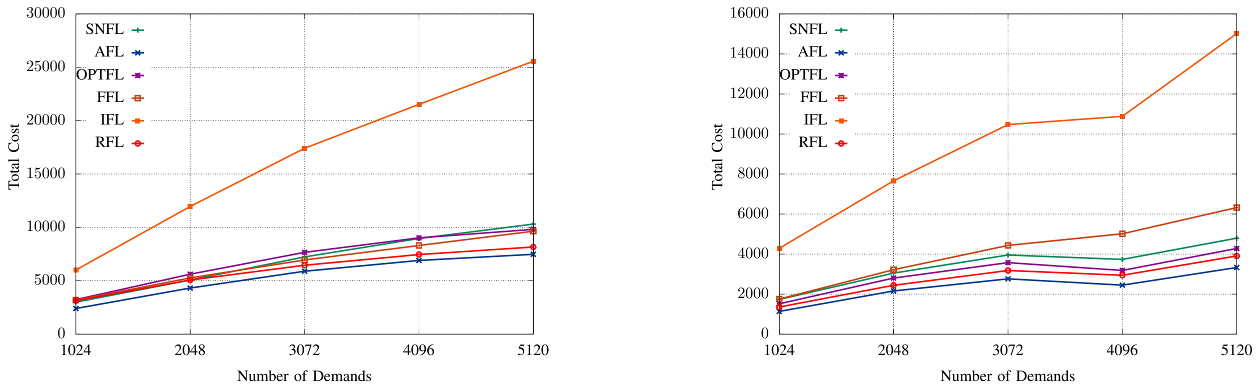
In this experiment, the impact of number of hosts is studied. Fat-trees with 64 ($k = 8$), 250 ($k = 10$), 432 ($k = 12$), 686 ($k = 14$), and 1024 ($k = 16$) hosts are generated and employed as the space for each test. In each test, 1024 demands are generated from uniform and normal distributions. Similar to the previous experiment, values of f, k, and h are set to 6, 2, and 1, respectively.

Figure 4 depicts the results of this experiment. Figures 4(c) and 4(d) represent the results for the uniform and normal distribution, respectively. Similar to the previous experiment, the total costs in the uniform case is noticeably greater than the normal case. As shown, AFL outperforms the other algorithms in both distributions and in all cases. The total cost of AFL is lower than the second best algorithm by the average of 14.16% and 15.66% for the uniform and normal distributions, respectively.

The different costs of AFL for the uniform generated demands are shown by table I. Facility and service costs are the most significant part of the overall cost. By increasing the number of demands, the migration and switching costs increase. In the case of 64 demands, AFL does not migrate or switch, however, when the number of demands increase, AFL migrates certain facilities and switch some of demands in order to reduce the total cost. For instance, the switch and migration costs is 3.45% of the total cost for the 1024 demands. It means that AFL by paying small amount of migration and switch cost saves a significant amount of the facility and service cost. In the case of 1024 points, AFL pays 2914.2 lower than the best second algorithm by paying extra 441.5 migration and switch cost in the average.

C. Impact of Cost Parameters

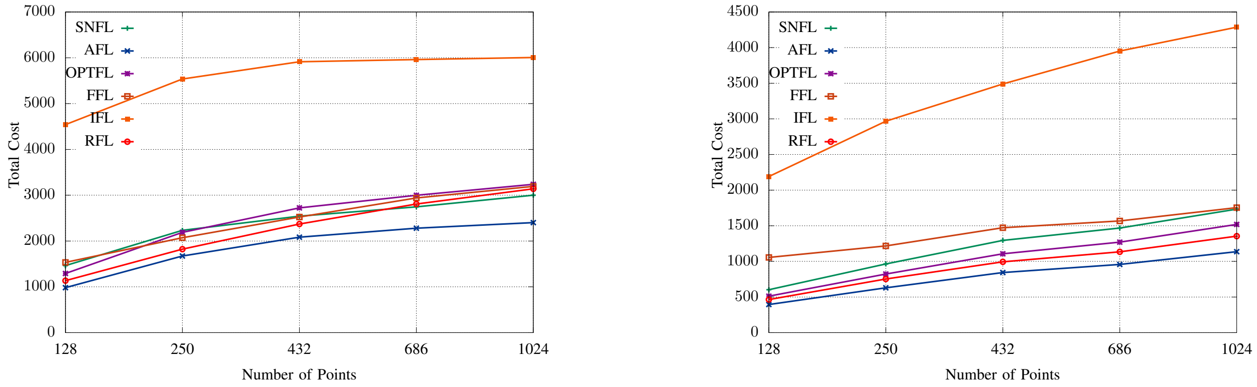
In this experiment, the impact of the costs parameters (f, k, and h) is investigated. For all tests, the space is fixed to the fat-tree with 1024 hosts (16-ary tree). Note that in this k-ary tree, the maximum distance between two points is 6 (Please note that we fixed the value of g to 1 in all



(a) Uniform Distribution

(b) Normal Distribution

Fig. 3. Impact of Number of Demands



(c) Uniform Distribution

(d) Normal Distribution

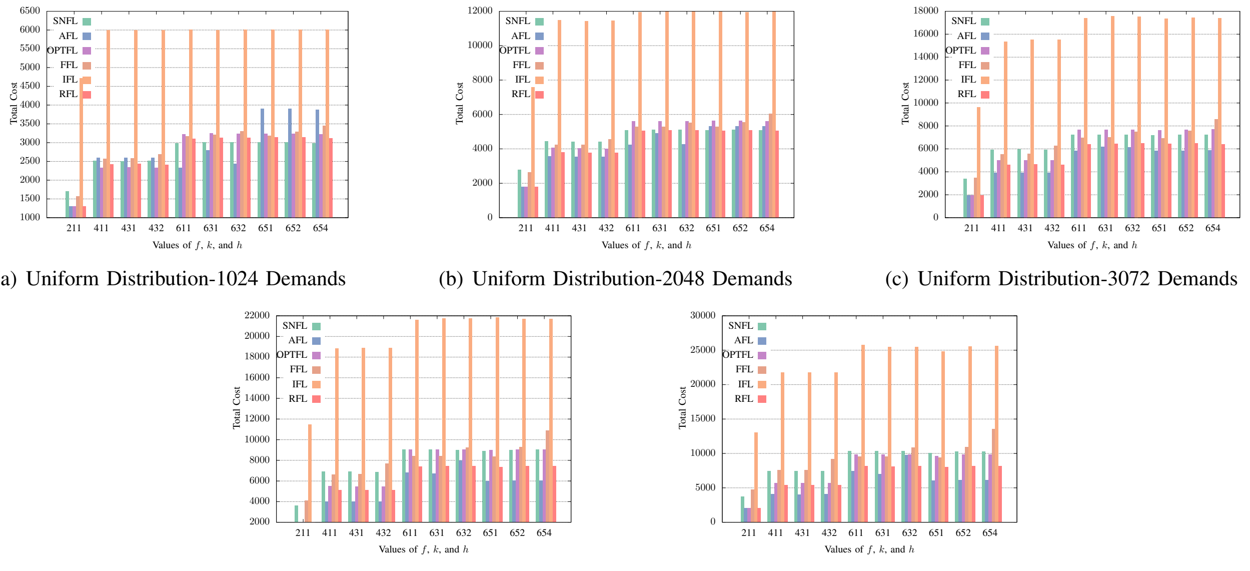
Fig. 4. Impact of Number of Hosts

experiments). Similar to the previous experiment, demands and facilities are located in the hosts. In particular, we vary the cost of installation, switch, and migration to investigate the impact of these parameters on the performance of our algorithm compared to others. We strictly specify that the cost of switching h to be always less or equal than the cost of migration k but cannot exceed the facility installation cost f (i.e, $f < k \leq h$). We run several tests for low, medium, and high values of f, k and h. Specifically for the facility cost f, $2, 4, 6$ are considered as low, medium and high values, respectively. For the migration cost k, the values $1, 3, 5$ and for the switch cost h, the values $1, 2, 4$ are selected as the low, medium and high values, respectively. Ultimately, 10 different configuration of values for the cost parameters are examined. Furthermore, we select the number of demands from 1024, 2048, 3072, 4096, 5120 and randomly place them on leaves based on normal and uniform distribution. Figure 5 shows the overall cost of our algorithm compare to others when changing the installation, switch, and migration cost parameters.

Figures 5(a), 5(b), 5(c), 5(d), and 5(e) represent the results of tests for 1024, 2048, 3072, 4096, and 5120 demands, respectively. As shown, for all configuration of cost values in all tests AFL outperforms the other algorithms except for the three configurations of cost values, 651, 652, and 654 in Figures 5(a), 5(b). However, as can be seen in figures 5(c), 5(d), and 5(e), by increasing the number of demands, AFL again outperforms the other algorithms in these configurations as well. It seems that by increasing the number of demands, AFL converges to the more stable configuration and performs more efficient.

D. Evaluation of Demand Departure

In this experiment, we examine the behavior of our solution for demand departures. Because we did not find any online algorithm in the literature considering demand departures, we compared our algorithm with a famous offline facility location greedy algorithm with approximation ratio of 1.61 [32]. The same configuration as experiments for the impact of number of demands and number of hosts (sections IV-A, IV-B) are used. A fat-tree with 1024 hosts, and values of 6, 2, 1, and 1 for parameters f, k, h, and g, respectively.



(a) Uniform Distribution-1024 Demands

(b) Uniform Distribution-2048 Demands

(c) Uniform Distribution-3072 Demands

(d) Uniform Distribution-4096 Demands

(e) Uniform Distribution-5120 Demands

Note: Please note that x-axis in the figures represent different values for f, k, h as the first, second, and third digit, respectively. For instance, 652 represents 6, 5, and 2 for f, k, and h, respectively.

Fig. 5. The total costs of the algorithms for the different values of f, k, and h for 1024 leaves

TABLE II: Algorithms for OnFLP and IncFLP models

Algorithm	Competitive Ratio		Time complexity [a]
	Adversarial	Random	
RFL	$O(\log n)$[b]	8	$\Omega(\log m)$[c]
OPTFL	$O(\frac{\log n}{\log\log n})$	-	$O(m^2 + \log d_{max})$[d]
SNFL	$4\log n + 1 + 2$	-	$O(m\lvert M\rvert\lvert F\rvert)$
IFL	$O(1)$	$O(1)$	$O(mm'\lvert F_{max}\rvert)$[e]
FFL	14	$O(1)$	$O(m\lvert F_{max}\rvert)$

[a] The complexity of processing m^{th} demand

[b] n is the number of all demands

[c] The number of demands at time t

[d] d_{max} is maximum distance in the space

[e] F_{max} denotes the maximum number of facilities opened by the algorithm

V. Related Works

A. Existing Systems

Network Function management solutions in the existing literature can be classified in to two separate groups. 1) Systems that deal with NFs that are deployed on pre-designated static hardware. These include systems such as CoMB [45], SIMPLE [42], xOMB [5] and PLayer [33]. 2) Systems that deal with VM based NF deployments, such as Stratos [23].

Static NF deployments are a step up from the traditional NFs, and introducing software based NFs within pre-placed commodity or specialized hardware. This gives such NFs the ability to use the best of both software and hardware world: multiple NFs can co-exist on the same high speed hardware and work in a coordinated manner to provide superior performance [45]. Since the NF itself is in software, it is easy to update and maintain. The hardware can also evolve independently of the software as the hardware and servers on which the NFs are hosted can be upgraded and replaced. This comes at price - the location of the NF, due to its rigid placement, might not always be ideal. The demand for NFs is not always uniformly distributed with in the data center [23].

B. Facility Location Problem

Facility Location Problem (FLP) is one of the well-known problems in the location theory. This classical optimization problem is concerned with optimal locations of certain facilities to minimize the cost of providing service to demands [46] with the offline settings. This problem is known to be NP-Hard, and several approximation algorithms has been developed for this problem [46]. The best known algorithm is proposed by Li et al. [35] and achieves 1.448 approximation ratio. In addition, several models of this problem with offline settings have been defined in the litereture. Farahani and Hekmatfar [15] provided extensive review of different models of offline FLP, and Boloori Arabani and Farahani [10] overviewed the dynamic models.

Unfortunately, none of the above models are not applicable for our problem, becuase they do not consider the online nature of the problem. Hence, we focus on online models of FLP. Fotakis [21] overviewed the online models of FLP and identified two major models of online version of FLP, namely *Online Facility Location Problem (OnFLP)* and *Incremental Facility Location Problem (IncFLP)*. In addition, some of the other relaxed version are discussed here. Table II represents the well-known algorithms, and their competitive ratios for all models.

1) Online Facility Location Problem: Meyerson et. al. [38] for the first time designed the *Online Facility Location*

Problem (OnFLP). In OnFLP, demands come one at a time and each demand is irrevocablely assigned to a facility upon its arrival; however the location of demands and facilities are not change during the time. Meyerson also proved that there is no algorithm that can be constant competitive against an adversary.

a) Random Facility Location Problem (RFL): Meyerson also proposed the first algorithm for OnFLP [38] called *Random Facility Location (RFL)*. RFL is pretty straightforward for the uniform facility cost. Upon arrival of each new demand u_t, RFL opens a facility with probability $\min\{1, \frac{d(F_{t-1}, u_t)}{f}\}$ in location $\gamma(u_t)$.

b) Optimal Facility Location Problem (OPTFL): The first deterministic algorithm for OnFLP is *Optimum Facility Location (OPTFL)* proposed by Fotakis [17], which achieves to the optimum competitive ratio for OnFLP. OPTFL defines the unsatisfied demands L that contains demands no having contributed in opening a new facility. Each $v \in L$ at time t contributes in opening a facility by $d(F_{t-1}, v)$. OPTFL marks each new arrived demand u_t as unsatisfied and appends u_t to L. $S = B(u_t, \frac{d(F_{t-1}, u_t)}{\alpha} \cap L)$ is a set of unsatisfied demands that are close to u_t. The potential function of S is defined by $P(S) = \sum_{v \in S} d(F_{t-1}, v)$. If $P(S) < f$, OPTFL opens no facility and assign u_t to the nearest facility. If $P(S) \geq f$, then the algorithm opens a new facility in a location $\omega \in S$ and removes S from L ($L = L/S$). The location ω is the center of the smallest radius ball $S' \subseteq S$ whose potential $P(S') \geq \frac{1}{2}P(S)$.

c) Simple Non-Uniform Facility Location (SNFL): Simple Non-Uniform Facility Location (SNFL) [20] uses the same idea of OPTFL [17], but this algorithm defines the potential function in a different way. There are no unsatisfied demands, and for each point z (either demand or facility) at time step t, the potential function is $p(z) = \sum_{v \in L}(d(F_{t-1}, v) - d(z, v))_+$ in which L denotes the set of previous demands and also the new arrived demand u_t. Upon arrival of demand u_t, SNFL adds u_t to L, computes potential $p(z)$ for all $z \in M$, and finds the point ω maximizing $p(\omega) - f_\omega$. If $p(\omega) > f_\omega$, SNFL opens a new facility at ω ($F_t = F_{t-1} \cup \{\omega\}$) and assigns u_t to ω. If $p(\omega) \leq f_\omega$ then SNFL does not open any new facility and assigns u_t to the nearest existing facility.

2) Increamental Facility Location Problem (IncFLP): Motivated by the framework of incremental clustering [11] and incremental k-median [12], *Incremental Facility Location Problem (IncFLP)* is developed. In contrast to ONFLP, two existing facilities and corresponding demands clusters can be merged in this model. A merge rule procedure determines whether two facilities will be merged. When a new demand u_t arrives, the algorithm applies a facility-opening rule and a merge rule to determine whether a new facility opens, and two existing facilities will be merged.

a) Incremental Facility Location (IFL): Incremental Facility Location (IFL) [18] is the first algorithm proposed for the IncFLP model. IFL introduced a new concept merge ball $B(\omega, m(\omega))$ for each facility ω, in which $m(\omega)$ is the merge-radius. IFL also defines $C(\omega)$ and $Init(\omega)$. In fact, $Init(\omega)$ is a set of demands initially assigned to a just opened facility ω, and $C(\omega)/Init(\omega)$ are demands that initially are assigned to other facilities different from ω, and gradually are assigned to ω by the merge rule ($m(\omega) \subseteq C(\omega)$). Moreover, IFL makes sure that no merge operation increase the total service cost of the demands in $Init(\omega)$ intensely. Hence, it keeps $m(\omega)$ decreasing by maintaining invariant $|Init(\omega) \cup B(\omega, \frac{m(\omega)}{\psi})|.m(\omega)$ lower than or equal to βf, in which ψ and β are appropriate positive constant integers.

3) Relaxed incremental Facility Location Problem (RFLP): Fotakis [19] suggested another model for FLP that is similar to IncFLP, though the demands can be reassigned to a nearest facility. We call this model Relaxed incremental Facility Location Problem (RFLP).

a) Fast Facility Location (FFL): Fast Facility Location (FFL) [19] introduces the final distance function used for the merge rule. The final distance for each facility ω and a point p is defined as $g(\omega, p) = d(\omega, p) + 2m(\omega)$. In this definition $m(\omega)$ denotes ω's replacement distance. For a point p and a facility set F, the replacement distance is $g(F, p) = \min_{\omega \in F}\{g(\omega, p)\}$. FFL works as follows. When a new demand u_t arrives, the algorithm computes $\delta = g(F_{t-1}, u_t)$ and opens a new facility ω in the location of u_t with probability of $\min\{1, \frac{\delta}{xf}\}$. If ω opens, the replacement radius $m(\omega)$ is set to $\frac{\min\{xf, g(F_{t-1}, u_t)\}}{6}$. Then, FFL considers every facility z which ω is inside $B(z, m(z))$, and merges z with ω. Notice that the demands assigned to these facilities will be reassigned to the nearest facility, not necessarily to ω.

4) Comparison of IMFLP with the other works: It is worth noting that our formulation of IMFLP is different from OnFLP, because assignment decisions and locations of facilities can be changed. Furthermore, the difference the incremental version [19] is that switch a demand to any other facility is allowed in the IMFLP model. Regarding the RFLP, switches are allowed but is free of charge. Finally, [14, 4] are the most similar model to ours, however; they assumed that the number of demands is fixed and known in advance, and they just change their locations.

VI. Conclusion and Future Works

We formulated the Intrusion Detection and Prevention Systems Placement (IDPSP) problem for cost-effective support as services in Cloud-based environment. Incremental Mobile Facility Location Problem (IMFLP) model was proposed to study this problem, and Adaptive Facility Location (AFL) solution was presented and evaluated for solving the optimization problem in this model. For the future works, we plan to implement this system as a real cloud service. Moreover, we intend to improve upon the proposed placement model. Certain constraints have been ignored in the formulation of the optimization problem in the IMFLP model for the sake of simplicity. We plan to add these constraints in the future model to make IMFLP model more applicable.

References

[1] ETSI Network Functions Virtualisation Introductory White Paper. https://portal.etsi.org/nfv/nfv$_{w}hite_{p}aper.pdf$.

[2] OpNFV - White Paper.

[3] Mohammad Al-Fares, Alexander Loukissas, and Amin Vahdat. A scalable, commodity data center network architecture, 2008.

[4] Hyung-Chan An, Ashkan Norouzi-Fard, and Ola Svensson. Dynamic facility location via exponential clocks. *ACM Transactions on Algorithms (TALG)*, 13(2):21, 2017.

[5] James W Anderson, Ryan Braud, Rishi Kapoor, George Porter, and Amin Vahdat. xomb: Extensible open middleboxes with commodity servers. In *Proceedings of the eighth ACM/IEEE symposium on Architectures for networking and communications systems*, pages 49–60. ACM, 2012.

[6] Junaid Arshad, Paul Townend, and Jie Xu. An automatic intrusion diagnosis approach for clouds. *International Journal of Automation and Computing*, 8(3):286–296, 2011.

[7] Md Faizul Bari, Raouf Boutaba, Rafael Esteves, Lisandro Zambenedetti Granville, Maxim Podlesny, Md Golam Rabbani, Qi Zhang, and Mohamed Faten Zhani. Data center network virtualization: A survey. *Communications Surveys & Tutorials, IEEE*, 15(2):909–928, 2013.

[8] Md Faizul Bari, Arup Raton Roy, Shihabur Rahman Chowdhury, Qi Zhang, Mohamed Faten Zhani, Reaz Ahmed, and Raouf Boutaba. Dynamic controller provisioning in software defined networks. In *CNSM*, pages 18–25, 2013.

[9] Alireza Boloori Arabani and Reza Zanjirani Farahani. Facility location dynamics: An overview of classifications and applications. *Computers & Industrial Engineering*, 62(1):408–420, February 2012.

[10] Alireza Boloori Arabani and Reza Zanjirani Farahani. Facility location dynamics: An overview of classifications and applications. *Computers & Industrial Engineering*, 62(1):408–420, 2012.

[11] Moses Charikar, Chandra Chekuri, Tomás Feder, and Rajeev Motwani. Incremental clustering and dynamic information retrieval. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 626–635. ACM, 1997.

[12] Moses Charikar and Rina Panigrahy. Clustering to minimize the sum of cluster diameters. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 1–10. ACM, 2001.

[13] András Császár, Wolfgang John, Mario Kind, Catalin Meirosu, Gergely Pongrácz, Dimitri Staessens, Attila Takács, and J Westphal. Unifying cloud and carrier network. *Proceedings of. DCC, Dresden, Germany, to appear Dec*, 2013.

[14] David Eisenstat, Claire Mathieu, and N Schabanel. Facility location in evolving metrics. *arXiv preprint arXiv:1403.6758*, pages 1–12, 2014.

[15] Reza Zanjirani Farahani and Masoud Hekmatfar. *Facility location: concepts, models, algorithms and case studies*. 2009.

[16] Björn Feldkord and Friedhelm Meyer auf der Heide. The mobile server problem. In *Proceedings of the 29th ACM Symposium on Parallelism in Algorithms and Architectures*. ACM, 2017.

[17] Dimitris Fotakis. On the competitive ratio for online facility location. *Algorithmica*, 14186:637–652, 2003.

[18] Dimitris Fotakis. Incremental algorithms for Facility Location and k-Median. *Theoretical Computer Science*, 361(2-3):275–313, 2006.

[19] Dimitris Fotakis. Memoryless facility location in one pass. *STACS 2006*, pages 608–620, 2006.

[20] Dimitris Fotakis. A primal-dual algorithm for online non-uniform facility location. *Journal of Discrete Algorithms*, 5(1), 2007.

[21] Dimitris Fotakis. Online and incremental algorithms for facility location. *ACM SIGACT News*, 42(1):97–131, 2011.

[22] A Gember, R Viswanathan, C Prakash, R Grandl, J Khalid, S Das, and A Akella. Opennf: Enabling innovation in network function control. Technical report, University of Wisconsin-Madison, 2014.

[23] Aaron Gember, Robert Grandl, Ashok Anand, Theophilus Benson, and Aditya Akella. Stratos: Virtual middleboxes as first-class entities. *UW-Madison TR1771*, 2012.

[24] By Albert Greenberg, James R Hamilton, Srikanth Kandula, Changhoon Kim, Parantap Lahiri, A Maltz, Parveen Patel, Sudipta Sengupta, Albert Greenberg, Navendu Jain, and David A. Maltz. VL2: a scalable and flexible data center network. In *Proc. ACM SIGCOMM 2009 Conf. Data Commun.*, volume 09, pages 51–62, 2009.

[25] Chuanxiong Guo, G Lu, Dan Li, Haitao Wu, and X Zhang. BCube: a high performance, server-centric network architecture for modular data centers. In *ACM SIGCOMM 2009 Conf. Data Commun.*

[26] Chuanxiong Guo, Haitao Wu, Kun Tan, Lei Shi, Yongguang Zhang, and Songwu Lu. Dcell: a scalable and fault-tolerant network structure for data centers, 2008.

[27] Ali Hammadi and Lotfi Mhamdi. A survey on architectures and energy efficiency in Data Center Networks, 2014.

[28] Qichao He, Ying Wang, Wenjing Li, and Xuesong Qiu. Traffic steering of middlebox policy chain based on sdn. In *Integrated Network and Service Management (IM), IFIP/IEEE Symposium on*. IEEE, 2017.

[29] Brandon Heller, Rob Sherwood, and Nick McKeown. The controller placement problem. In *Proceedings of the first workshop on Hot topics in software defined networks*, pages 7–12. ACM, 2012.

[30] Huawei Huang, Song Guo, Jinsong Wu, and Jie Li. Service chaining for hybrid network function. *IEEE Trans. on Cloud Computing*, 2017.

[31] Jinho Hwang, KK Ramakrishnan, and Timothy Wood. Netvm: high performance and flexible networking using virtualization on commodity platforms. In *11th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2014.

[32] Kamal Jain, Mohammad Mahdian, and Amin Saberi. A new greedy approach for facility location problems. In *Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing*, STOC '02, pages 731–740, New York, NY, USA, 2002. ACM.

[33] Dilip A Joseph, Arsalan Tavakoli, and Ion Stoica. A policy-aware switching layer for data centers. In *ACM SIGCOMM Computer Communication Review*, volume 38, pages 51–62. ACM, 2008.

[34] Md Tanzim Khorshed, ABM Ali, and Saleh A Wasimi. A survey on gaps, threat remediation challenges and some thoughts for proactive attack detection in cloud computing. *Future Generation Computer Systems*, 28(6):833–851, 2012.

[35] Shi Li. A 1.488 approximation algorithm for the uncapacitated facility location problem. *Information and Computation*, 222:45–58, 2013.

[36] Jiaqiang Liu, Yong Li, Ying Zhang, Li Su, and Depeng Jin. Improve service chaining performance with optimized middlebox placement. *IEEE Transactions on Services Computing*, 10(4):560–573, 2017.

[37] Yang Liu, Jogesh K Muppala, and Malathi Veeraraghavan. A Survey of Data Center Network Architectures.

[38] Adam Meyerson. Online facility location. *. . . of Computer Science, 2001. Proceedings. 42nd . . .*, pages 0–5, 2001.

[39] Radhika Niranjan Mysore, Andreas Pamboris, Nathan Farrington, Nelson Huang, Pardis Miri, Sivasankar Radhakrishnan, Vikram Subramanya, Amin Vahdat, and Radhika Niranjan Mysore. PortLand: a scalable fault-tolerant layer 2 data center network fabric. In *ACM SIGCOMM 2009 Conf. Data Commun.*

[40] VT Paschos. *Paradigms of Combinatorial Optimization: Problems and New Approaches*. 2013.

[41] Chuan Pham, Nguyen H Tran, Shaolei Ren, Walid Saad, and Choong Seon Hong. Traffic-aware and energy-efficient vnf placement for service chaining: Joint sampling and matching approach. *IEEE Transactions on Services Computing*, 2017.

[42] Zafar Ayyub Qazi, Cheng-Chun Tu, Luis Chiang, Rui Miao, Vyas Sekar, and Minlan Yu. Simple-fying middlebox policy enforcement using sdn. In *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*, pages 27–38. ACM, 2013.

[43] Sebastian Roschke, Feng Cheng, and Christoph Meinel. Intrusion detection in the cloud. In *Dependable, Autonomic and Secure Computing, 2009. DASC'09. Eighth IEEE International Conference on*, pages 729–734. IEEE, 2009.

[44] Karen Scarfone and Peter Mell. Guide to intrusion detection and prevention systems (idps). *NIST special publication*, 800, 2007.

[45] Vyas Sekar, Norbert Egi, Sylvia Ratnasamy, Michael K Reiter, and Guangyu Shi. Design and implementation of a consolidated middlebox architecture. In *NSDI*, pages 323–336, 2012.

[46] David B Shmoys, Éva Tardos, and Karen Aardal. Approximation algorithms for facility location problems. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 265–274. ACM, 1997.

[47] Anna Sperotto, Gregor Schaffrath, Ramin Sadre, Cristian Morariu, Aiko Pras, and Burkhard Stiller. An overview of ip flow-based intrusion detection. *IEEE Communications Surveys and Tutorials*, 12(3):343–356, 2010.

[48] Wentao Wang, Lingxia Wang, and Fang Zheng. An improved adaptive scheduling strategy utilizing simulated annealing genetic algorithm for data center networks. *KSII Transactions on Internet and Information Systems (TIIS)*, 11(11):5243–5263, 2017.

[49] Jie Zhang, Deze Zeng, Lin Gu, Hong Yao, and Muzhou Xiong. Joint optimization of virtual function migration and rule update in software defined nfv networks. In *GLOBECOM IEEE Global Communications Conference*, pages 1–5. IEEE, 2017.

Insights on Car Relocation Operations in One-Way Carsharing Systems

Rabih Zakaria
Université de Technologie et de Sciences Appliquées Libano-Française
Tripoli, Lebanon

Laurent Moalic
Université de Haute-Alsace
Mulhouse, France

Mohammad Dib
Engie
Paris, France

Alexandre Caminada
Université Bourgogne Franche-Comté Belfort, France

***Abstract*—One-way carsharing system is a mobility service that offers short-time car rental service for its users in an urban area. This kind of service is attractive since users can pick up a car from a station and return it to any other station unlike round-trip carsharing systems where users have to return the car to the same station of departure. Nevertheless, uneven users' demands for cars and for parking places throughout the day poses a challenge on the carsharing operator to rebalance the cars in stations to satisfy the maximum number of users' requests. We refer to a rebalancing operation by car relocation. These operations increase the cost of operating the carsharing system. As a result, optimizing these operations is crucial in order to reduce the cost of the operator. In this paper, the problem is modeled as an Integer Linear Programming model (ILP). Then we present three different car relocation policies that we implement in a greedy search algorithm. The comparison between the three policies shows that car relocation operations that do not consider future demands do not effectively decrease rejected demands. On the contrary, they can generate more rejected demands. Results prove that solutions provided by our greedy algorithm when using a good policy, are competitive with CPLEX solutions. Furthermore, adding stochastic modification on the input data proves that the results of the two presented approaches are highly affected by the input demand even after adding threshold values constraints.**

***Keywords*—Carsharing; car relocation; ILP; greedy algorithm; CPLEX; green city**

I. Introduction

It is straightforward that convenient transportation systems are crucial for supporting the economic development of cities [1], [2]. Generally, people in urban areas commute using different modes of transportation, such as public transport buses, trains, taxis, private cars, bikes, etc. Private cars are more attractive to users for their high flexibility and comfort. However, the increasing number of private cars has serious consequences related to environment issues, traffic and parking congestion [3]. Then, numerous efforts have been made to motivate people to use more sustainable modes of transportation like biking, walking or the use of public transportation facilities when possible. In June 2007, Vélib was successfully launched in Paris. 20,000 self-service bikes were deployed over 1500 stations [4]. Within the first year, the number of subscribed members exceeded 200,000 members and the bikes have been used 26 million times. The success of this system has motivated cities all over the world to adopt this idea of sharing vehicles, which includes carsharing, bikesharing and other vehicle sharing concepts. Carsharing is one of the innovative solutions that can contributes in promoting sustainable car use. Many studies stated that private cars spend most of their time parked, since many car owners use their cars occasionally. Thus, in this case, one shared car can replace many private owned ones. Carsharing offers on demand access for cars distributed in a defined urban area. Therefore, carsharing systems offer the benefits of owning a private car without actually having to buy it. Carsharing is based on the model of Pay As You Go service, so users do not have to afford all the fees of owning a car like insurance and maintenance, they just pay during the time they access the service as an alternative of ownership in a market shift as predicted in "the age of access" [5]. Usually, users of this kind of systems rent cars for short periods of time. It is a complementary solution for the existing public transportation facilities. It offers the comfort and flexibility of private cars and the reduced costs of public transportation. According to Navigant Research, global carsharing service revenues will grow up to US$6.2 billion by 2020 [6]. This kind of system has been implemented since the end of forties in Europe [7]. However, they were not successful since it was not easy to monitor the system and protect it from vandalism. Thanks to the advances in Information and Communications Technology (ICT), better facilitation, monitoring and management of reservations and payment operations of these systems have become available [8]. System operators and users are able to locate the stations and check the availability of vehicles in real time. In our study, we are dealing with one-way carsharing systems. Unlike round-trip carsharing systems, one-way carsharing systems allow users to take a car from a station and to drop it off in any other one. Although the one-way option makes the system more attractive to users, carsharing operators encounter difficulties in maintaining enough numbers of vehicles in stations to satisfy user demands. If stations are full, users who want to drop off their cars at the destination station cannot find a free parking place. On the other hand, user demands to take a car from empty stations will not be satisfied. If this imbalance problem occurs frequently, system clients lose their enthusiasm for using the service since it is not reliable and available when they need it. Recently, vehicle-sharing systems have generated a great interest of research in its different majors to solve the problems that arise upon operating these systems. We will focus in this paper on the problem of car relocation in order to meet user demands. In one-way carsharing systems, the relocation problem is technically more difficult than the relocation problem in bikesharing systems. In the latter, we can use a truck to move several bikes at the same time, while we cannot do this in carsharing system because of the size of cars and the difficulty of loading and unloading cars. This paper presents an exact approach for the relocation problem in one-way carsharing system, followed by a heuristic approach using a greedy search algorithm. The focus is on providing different analysis and results to highlight the particular aspects related to this problem. In the literature, papers do not emphasize on the workload and cost of employees recruited to locate cars between the stations. The objective in this paper is to bring the attention to the complexity of these operations and to provide different analysis for this problem. This paper is structured as follows. The next section presents a practical example for the relocation problem. This is followed by the formulation of an Integer Linear Programming model for the problem. Then, the platform and mobility data used for this study are described. After that, a greedy algorithm and three relocation policies are explained. Different results and analysis are presented in Section VI. Finally, conclusion and future works are provided.

II. Literature Overview

In the literature, we find many papers dealing with the relocation problem in one-way carsharing system. In 1999, Barth et al. developed a simulation model performance analysis of a multiple stations shared-use vehicles [9]. They found that the carsharing system is most sensitive to vehicle to trip ratio, to the relocation algorithm, and to the charging policy used in case of electrical vehicles. Other papers proposed that carsharing operators can involve users in the system to relocate cars [10], [11]; despite the fact that this technique was fruitful in alleviating the imbalance problem, it highly depends on clients participation, which is obviously not always guaranteed. Mitchel et al. proposed dynamic pricing for mobility-on-demand systems which include carsharing [12]. A price incentive strategy is used to motivate users to change their origin or destination stations to other stations near to them based on system needs. In a different study, a decision support system is presented by [13] to help carsharing operators to decide the values of operating parameters in a near-optimal way. Tuning these parameters reduces, between 37.1% and 41.1%, the number of relocation operations, it decreases the staff cost of 50%. It also reduces the zero vehicle-time between 4.6% and 13.0%. The author in [14] presented a multistage stochastic linear integer model with recourse for dynamic decision-making problem of vehicle allocation. They optimize trip selection in a way that the operator accepts or refuses trips reservations that maximize the profit of the carsharing system. Results showed profit increase but the model was not applied on real network and under real conditions. In [15], the author proposed stochastic mixed-integer programming model to minimize the cost of cars relocation operations in a way that satisfy p-proportion of all near-term demand scenarios. The study used historical data originated from the Intelligent Community Vehicle System (ICVS), which is no longer operational. Authors proved the robustness of the solutions through simulations that consider stochasticity in generating redistribution plans.

III. Relocation Problem in a Practical Example

This paper is dealing with one-way carsharing systems, which consists of many stations scattered in an urban area. A station has a predefined number of parking spots for its users. System users can take a car from a station and return it to any other station. When a user arrives at an empty station to drive a car, his request will be rejected. On the other side when a user wants to return a car to a station that is full, his request will be rejected as well. Users expect that cars are always available in stations when they need it, and they expect to find a free parking place at the destination station when they want to return the rented car as well. However, maintaining this level of service is not an easy task. This should be done by employees recruited to redistribute cars between the stations; in the following, we refer to these employees by “jockeys”. However, when the operator fails to solve this imbalance problem, users tends to abandon the system, which leads to potential system failure. We modeled our one-way carsharing system by a simple time-space network. To simplify the idea, an example of a simple carsharing system is provided below. Table I shows how many vehicles are available av_{i_t} in each station i for each time step t. We have three stations S_1, S_2 and S_3. At time t = 0, we have the initial number of available vehicles in each station. Table II shows the number of cars out_{i_t} that users would like to rent by station and by time step. Table III shows the number of cars in_{i_t} that users would like to return to each station at each time step. In Table IV, we see the number of rejected user demands to take a car because a station is empty $out^r_{i_t}$ while Table V shows the number of rejected demands to take a car because a station is full $in^r_{i_t}$ for each time step. In this example, we consider that a station can host at most five cars. It is obvious that the values in these five tables must be non-negative. The input for the system consists of the initial number of available vehicles at t = 0 and the values in Tables II and III. While all other values are calculated based on the aforementioned input. To get the number of available cars, we use this equation:

$$av_{i_t} = av_{i_{t-1}} + (in_{i_t} - in^r_{i_t}) - (out_{i_t} - out^r_{i_t}) \qquad (1)$$

In (1), available vehicles in station i at time t is equal to the number of available vehicles at the same station in the previous time step added to the number of arriving cars to the same station at time t minus the number of cars that could not be returned because the station is full. Then, we subtract the number of cars that go out of the station minus the number of rejected requests to take a car out of the station because there is a lack of cars.

TABLE I. Number of Available Cars

t	0	1	2	3	4	5	6	...	T
S_1	2	2	3	3	3	4	5	...	...
S_2	4	2	2	1	0	0	0	...	...
S_3	3	3	4	4	5	3	3	...	...

As we can see in Table IV, we have one rejected demand in station S_2 at time t = 6. This rejected demand occurs because station S_2 does not have any vehicle at t = 5 and there is one request for vehicle at t = 6. On the other side, we see in Table

TABLE II. Number of Requests for Departing Cars

t	0	1	2	3	4	5	6	...	T
S_1	0	0	0	0	0	0	0	...	...
S_2	0	2	0	1	1	0	1	...	...
S_3	0	0	0	0	0	2	0	...	...

TABLE III. Number of Requests for Arriving Cars

t	0	1	2	3	4	5	6	...	T
S_1	0	0	1	0	0	1	2	...	...
S_2	0	0	0	0	0	0	0	...	...
S_3	0	0	1	0	1	0	0	...	...

TABLE IV. Number of Rejected Demands because of an Empty Station

t	0	1	2	3	4	5	6	...	T
S_1	0	0	0	0	0	0	0	...	...
S_2	0	0	0	0	0	0	1	...	...
S_3	0	0	0	0	0	0	0	...	...

TABLE V. Number of Rejected Demands because of a Full Station

t	0	1	2	3	4	5	6	...	T
S_1	0	0	0	0	0	0	1	...	...
S_2	0	0	0	0	0	0	0	...	...
S_3	0	0	0	0	0	0	0	...	...

V that we have one rejected demand in station S_1 at time t = 6 since this station has four vehicles at t = 5 and there are two requests to return cars at t = 6.

IV. ILP for the Relocation Problem

We adapted the car relocation model presented in [13] to our study. Thus, a two dimensional time-space matrix $N \times T$ is used to model the relocation problem, N stands for the total number of stations $S = \{1, 2, .., N\}$ and T is the number of time steps during a day starting from 1 to T. Each element of the matrix represents a station S_i at time t. For each station $s \in S$ we generate T nodes to represent that station at each time t. Then we put all the $S \times T$ nodes in one row vector $V = (1_1, ..., 1_T, ..., N_1, ..., N_T)$. An employee has three sorts of activities:

1) Relocating: is the action taken by the jockey to move a car from a station i to another station j.
2) Moving: is the action taken by the jockey to move himself from his current station to another station in order to begin a relocation activity.
3) Waiting: when the jockey is not involved in relocating or moving activities we say that the jockey is waiting.

Therefore, to represent these activities three sets of arcs are generated in the time-space network. An arc a_1 is constructed for each node $i_t \in V$, to represent a waiting activity between i_t and i_{t+1}; this set is called $A_1 = \{..., a_1(i_t, i_{t+1}), ...\}$. Also, $N-1$ arcs a_2 are constructed for each node i_t in V, to represent the move activity from station i and station $j, \forall\, i, j \in S, i \neq j$, from time t to time $t + t_{ij}$ where t_{ij} stands for the number of time steps required to move from station i to station j; this set is named $A_2 = \{..., a_2(i_t, j_{t+t_{ij}}), ...\}$. likewise, $N-1$ arcs a_3 are built to represent relocation activities, this set is denoted $A_3 = \{..., a_3(i_t, j_{t+t_{ij}}), ...\}$. The staff that is responsible for these activities is denoted by a set $E = \{1, ..., e, ..., W\}$, W represents the number of recruited employees. An ILP Model is formulated for the relocation problem. Six different decision variables are declared:

- u^e: When an employee e is used at least once during the day, its associated binary variable takes the value 1, while it takes the value 0 otherwise.
- $wait^e_{i_t i_{t+1}}$: When an employee e is involved in a waiting activity at station i from time t to $t+1$, the associated binary variable is assigned the value 1, while it remains 0 otherwise.
- $move^e_{i_t j_{t+t_{ij}}}$: When an employee e is involved in a moving activity from the set A_2, the associated binary variable is assigned with the value 1, while it remains 0 otherwise.
- $rel^e_{i_t j_{t+t_{ij}}}$: When an employee e is involved in one of the relocation activities within the set A_3, the associated binary variable is assigned the value 1, while it remains 0 otherwise.
- $out^r_{i_t}$: This variable can be assigned with integer values. It represents the number of rejected user demands to rent a car from station i at time t.
- $in^r_{i_t}$: This variable can be assigned with integer values. It represents the number of rejected user demands to give back the rented car to a station i at time t.

On other side, the input parameters for the ILP are listed below:

- av_{i_0}: Represents the number of available cars in station i at time 0.
- out_{i_t}: Represents the number of requests to rent a car at time t from station i.
- in_{i_t}: Represents the number of requests to give back a car at time t to a station i.
- p_i: Represents the number of parking spots in a station i.
- c_{ij}: Denotes the estimated cost of a relocation or moving activity from a station i to station j.
- c_e: Denotes the estimated cost of an employee during a day.
- c_{in}: Stands for the estimated cost of the rejection of a demand to give back a vehicle to a station.
- c_{out}: Stands for the estimated cost of the rejection of a demand to rent a vehicle from a station.

Also, one dependent variable is used:

- av_{it}: Denotes the remaining available cars at station i at time t.

The relocation problem can be modeled by the ILP model below:

$$\begin{aligned}&Min \quad Z = c_{ij}(\sum_{(i_t,j_{t+t_{ij}})\in A_3} \sum_{e\in E} move^e_{i_t j_{t+t_{ij}}} \\ &+ \sum_{(i_t,j_{t+t_{ij}})\in A_4} \sum_{e\in E} rel^e_{i_t j_{t+t_{ij}}})+ c_{out} \sum_{i_t\in V} out^r_{i_t} \\ &+c_{in} \sum_{i_t\in V} in^r_{i_t} + c_e \sum_{e\in E} u^e\end{aligned} \tag{2}$$

Subject to:

$$\begin{aligned}&\sum_{i\in S} wait^e_{i_1 i_2} + \sum_{\substack{i,j\in S\\ i\neq j}} move^e_{i_1 j_{1+t_{ij}}} + \sum_{\substack{i,j\in S\\ i\neq j}} rel^e_{i_1 j_{1+t_{ij}}} \\ &= u^e \qquad\qquad\qquad\qquad \forall e \in E\end{aligned} \tag{3}$$

$$\begin{aligned}&wait^e_{i_{t-1} i_t} + \sum_{(j_{t-t_{ij}},i_t)\in A_3} move^e_{j_{t-t_{ij}} i_t} \\ &+ \sum_{(j_{t-t_{ij}},i_t)\in A_4} rel^e_{j_{t-t_{ij}} i_t} - wait^e_{i_t i_{t+1}} \\ &- \sum_{(i_t,j_{t+t_{ij}})\in A_3} move^e_{i_t j_{t+t_{ij}}} - \sum_{(i_t,j_{t+t_{ij}})\in A_4} rel^e_{i_t j_{t+t_{ij}}} \\ &= 0 \qquad\qquad \forall\; i_t \in V,\; e \in E, t > 1\end{aligned} \tag{4}$$

$$\begin{aligned}&av_{i_t} = av_{i_{t-1}} + (in_{i_t} - in^r_{i_t}) - (out_{i_t} - out^r_{i_t}) \\ &+ \sum_{(j_{t-t_{ij}},i_t)\in A_4} \sum_{e\in E} rel^e_{j_{t-t_{ij}} i_t} - \\ &\sum_{(i_t,j_{t+t_{ij}})\in A_4} \sum_{e\in E} rel^e_{i_t j_{t+t_{ij}}} \qquad \forall\; i_t \in V\end{aligned} \tag{5}$$

$$av_{i_t} \le p_i \;\; \forall\; i_t \in V \tag{6}$$

$$in^r_{i_t} \;\le\; in_{i_t} \;\forall\; i_t \in V \tag{7}$$

$$out^r_{i_t} \;\le\; out_{i_t} \;\forall\; i_t \in V \tag{8}$$

$$u^e = (0, 1) \;\; \forall \;\; e \in E \tag{9}$$

$$wait^e_{i_t i_{t+1}} \in \{0, 1\} \;\; \forall\; (i_t, i_{t+1}) \in A_1,\; e \in E \tag{10}$$

$$move^e_{i_t j_{t+t_{ij}}} \in \{0, 1\} \;\; \forall\; (i_t, j_{t+t_{ij}}) \in A_2,\; e \in E \tag{11}$$

$$rel^e_{i_t j_{t+t_{ij}}} \in \{0, 1\} \;\; \forall\; (i_t, j_{t+t_{ij}}) \in A_3,\; e \in E \tag{12}$$

$$in^r_{i_t} \ge 0 \;\;\; \forall\; i_t \in V \tag{13}$$

$$out^r_{i_t} \ge 0 \;\; \forall\; i_t \in V \tag{14}$$

$$av_{i_t} \ge 0 \;\; \forall\; i_t \in V \tag{15}$$

Equation (2) represents the objective function. It minimizes the weighted aggregation of the number of rejected requests to rent or to give back a car, the number of needed staff and the required moving and relocating operations needed to decrease the number of rejected demands. Constraint (3) serves to ensure that an employee cannot perform more than one task at a time and to assign the value 1 to the variable u^e when the associated employee e is engaged in an activity at $t = 1$. Constraint (4) ensures that an employee cannot be engaged with a new activity before he completed the last one and to assure the continuity of activities for an employee if he is engaged at t = 1. Constraint (5) is used to get the remaining available cars at each station at each time step. It is calculated based on the number of remaining cars in the previous time step, the number of cars entering and leaving the station by the users and the number of cars moved in/out of the station by the staff. Constraint (6) ensures that the number of available cars at a station will not exceed its capacity. Constraints (7) and (8) are used to make sure that the number of rejected demands will not be greater than the number of demands. Constraints (9)-(12) are used to impose binary values to the associated variables, and constraints (13)-(15) ensure that the associated variables are non-negative.

V. Greedy Algorithm to Solve the Car Relocation Problem

A. Motivation Behind the Greedy Algorithm

As described earlier in Section IV, the car relocation problem is modeled as an ILP model. The model is solved using CPLEX. The model was tested using different configurations, it was evident that the running time gets significantly bigger when the number of jockeys increased. It was also noted that stations number, the average number of trips per car and the maximum number of parking spots in each station highly affect the running time. These parameters highly influence number of rejected demands. After running the model with CPLEX with some complex configurations, CPLEX could not give a solution before two days of execution. While the solver could not give any solution for other complex configurations. Fig. 1 shows the variation of CPLEX's running time with respect to the number of jockeys used. This experiment uses a configuration of a simple carsharing system which has 10 parking spaces per station, 18 stations and 83 cars. The average trips per car for this experiment was 12. It was clear that the running time of CPLEX grows significantly when the number of jockeys is increased. This observation was a good motivation to look for another approach to get a result more quickly. For this reason, a simple greedy algorithm is developed to minimize the number of rejected demands while reducing the number of required relocation operations. A relocation operation is performed through two steps: first, a station is selected to take a car from it, then the destination station is chosen to move the car for it by the jockey to rebalance the system.

In this paper, different relocation policies are proposed which are implemented later using a greedy algorithm. This algorithm uses a policy pattern to assess the impact of each policy on resulting rejected demands. When the greedy algorithm is executed, one second of running time was enough to build a good non-optimal solution for configurations regardless of their complexity, and the used policy.

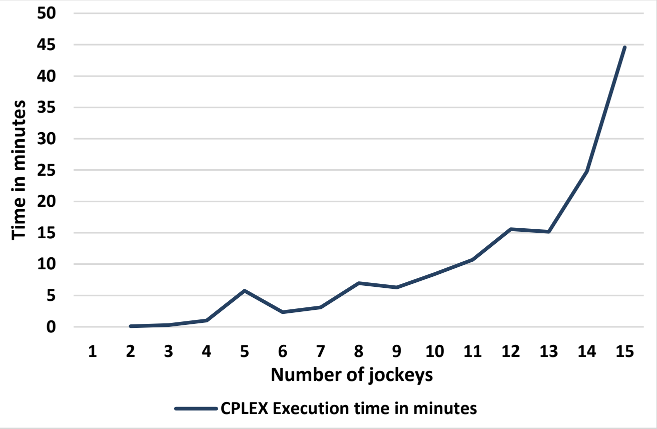


Fig. 1. CPLEX execution time when solving the relocation problem (10 parking places per station, 18 stations, 12 trips per car, 83 cars).

B. Relocation Policies

To increase client satisfaction, it is important to choose a good relocation policy in order to reduce the total number of rejected demands. Different approaches were tested using the greedy algorithm:

- Policy 1: The jockey moves one car from the nearest station to his current station, and if several, he chooses the one that has the highest number of available vehicles, to the nearest station, and if there are several ones, to the station that has the lowest number of available vehicles.
- Policy 2: The jockey moves one car from the station that has the biggest number of available vehicles, and if several, the nearest station, to the station that has the lowest number of available vehicles, and if several, at the closest station.
- Policy 3: The jockey moves one car from the station having the soonest rejected demand because it is full to the station having the soonest rejected demand because it is empty.

NB: In our examples (Fig. 2 and 3), we consider that the only car movements are done by jockeys for relocation purposes.

1) Policy 1: In this policy, the priority is given to the time needed to move between the stations. For each operation decision, the jockey chooses the operation that takes the shortest possible time with the objective of having enough time to do the maximum number of relocation operations that can be done during the day. During each relocation operation, the jockey starts by determining the closest station to his present location in order to take a car from this station. If more than one station found on same distance, he selects randomly a station having the biggest number of available vehicles. Then, the jockey chooses the closest station again and if he finds many, he selects the one that has the minimum number of available vehicles among them. For example, Fig. 2 shows a representation of this policy with four stations. The circles represent the stations. The name of the station and the number of available cars at a specified time are displayed in each circle.

In our example, the jockey starts the relocation operation by going to the nearest station from his location which is S_3. This moving activity is done without a car from the system. Then, in a second step, he drives a car from the selected station S_3 to the station S_1, because, first, it is the closest station and, second, because it presents the minimum number of available cars.

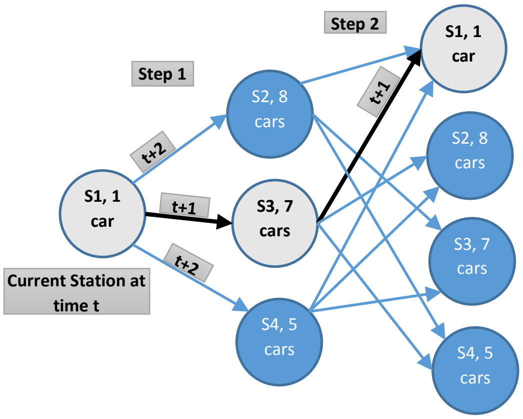


Fig. 2. Simple relocation operation using Policy 1.

2) Policy 2: Policy 2 prioritizes the balancing of the cars over the station, aiming to rebalance the number of available cars in each station. Policy 1 has some similarities with Policy 2, but here, the order of selecting stations is reversed. During each relocation operation, the jockey starts by looking for stations having the highest number of available vehicles, and he selects the closest station in the list. After that, stations having the minimum number of cars are selected, and the jockey chooses the closest. As we can see in Fig. 3, the jockey chooses station S_2 in the first step since it presents the highest number of cars, while the choice in the later step remains the same since station S_1 has the lowest number of cars.

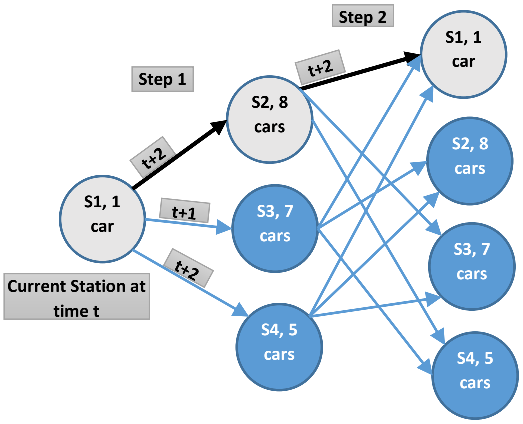


Fig. 3. Simple relocation operation using Policy 2.

3) Policy 3: Policy 3 considers that an estimation of what will happen in the future is known by the jockey, so he can foresee the rejected demands even if they occur after several time steps. In addition, in this policy the jockey can see the effect of each relocation operation on the overall system, so the jockey will not remove or add cars when this may cause a station to be empty or full respectively. Here, the aim is to get rid of the maximum number of rejected demands during a single relocation operation. The relocation operation starts at step 1 by looking for the list of stations where the soonest rejected demands will occur as a result of stations filling up ($in^{r}_{i_t} > 0$) and the list of stations that have cars that can be delivered to other stations that will need cars for future demands. In step 2, the jockey tries to find the list of stations where the soonest rejected demands will occur because empty stations ($out^{r}_{i_t} > 0$) and the list of stations that may be in shortage for future demands. From these lists, the jockey chooses the target station for the relocation operation in order to reduce the maximum number of rejected demands while preventing to generate future rejected demands. When choosing the best relocation operation, if we have many possibilities with the same effect on rejected demands, we privilege the operation that reduces rejected demands in the closest stations and the soonest possible. In Fig. 4 we propose the flow chart to implement Policy 3 in a greedy algorithm. A greedy algorithm makes the optimal choice at each iteration up to the local optimum.

VI. Experimentation and Results

A. Mobility Data

The mobility data used for this study consists of socio-economical information and survey data that are collected by professional for the objective of regional planning. This data describes people mobility flows in a region of 20 km x 10 km in Paris. The region of the study is plotted into a grid of cells having the same size. A cell has two characteristics:

- The type of the terrain: it describes structure types that are dominant in the area associated to the cell (commercial center, business center, buildings, roads, houses, etc.).
- Attraction weight: based on the terrain type and survey data, this information attributes a dynamic attraction weight to each cell for each 15 minutes of the day.

A 3D matrix $F = (f_{i,j,t})$ represents the people mobility between different cells, where $f_{i,j,t}$ stands for the number of persons who want to move from cell i to cell j at time t. We consider t to be a period of 15 minutes during the day, which makes 96 time periods. Then the flow mobility data is plotted on a map using GIS shapefiles. As a result, 400 cells have been detected as a potential origin or destination point knowing that some cells are eliminated because of their geographical nature e.g. lakes, plains, etc. The final flow mobility data consists of 400 x 400 x 96 elements, which makes 15,360,000 records to represent how people move during the day.

B. Platform for Locating Stations

In order to locate station for a carsharing system in the region on the study, [16] developed a dedicated platform. The platform uses the mobility data that are described earlier in this study. To locate the stations, a multiobjective memetic algorithm has been implemented in the platform. The algorithm optimizes three objective functions:

- Objective 1: The location of the stations should maximize the mobility flow between the cells.

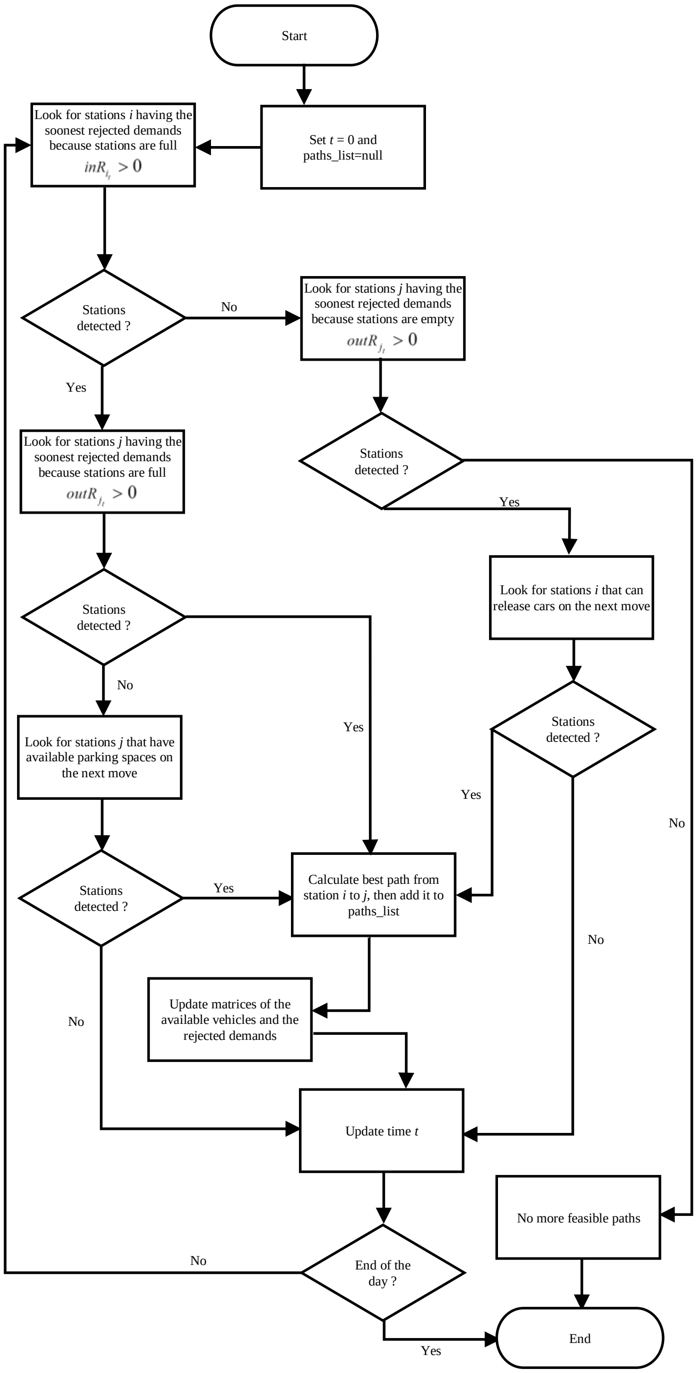

Start
Set t = 0 and paths_list=null
Look for stations i having the soonest rejected demands because stations are full $inR_{i_t} > 0$
Stations detected ?
No
Look for stations j having the soonest rejected demands because stations are empty $outR_{j_t} > 0$
Yes
Look for stations j having the soonest rejected demands because stations are full $outR_{j_t} > 0$
Stations detected ?
Yes
Stations detected ?
No
Look for stations i that can release cars on the next move
Look for stations j that have available parking spaces on the next move
Yes
Stations detected ?
Yes
No
Stations detected ?
Yes
Calculate best path from station i to j, then add it to paths_list
No
No
Update matrices of the available vehicles and the rejected demands
Update time t
No more feasible paths
No
End of the day ?
Yes
End

Fig. 4. Flowchart of the Relocation Algorithm using Policy 3.

- Objective 2: The location of the stations should maximize the balance between the ingoing and outgoing flows in each station.
- Objective 3: The location of the stations should minimize the standard deviation of the flows in order to obtain a uniform flow during the day.

Each cell is considered to cover the demand in a radius of 300 meters. Special filters and probability distribution are applied on the mobility data to forecast the potential users for the service. A study has been carried out with the carsharing operator to set the desired system parameters. We used this platform to generate the data for this study. The generated dataset consists of the four matrices described in Section III. For each generated dataset, we use four parameters:

1) Total number of cars in the system.
2) Total number of stations in the system.
3) Average trips number by car.
4) Parking spots number for each station.

C. Relocation Policies Comparison

Fig. 5 shows a comparison of three policies described earlier, these results concern a generated dataset for a car-sharing system that consists of 20 stations having 10 parking places each and 150 cars that have an 9 average trips. As we can see, the performance of policy 1 and policy 2 is rather similar in the beginning. After that, when the number of jockeys is increased to be more than 19, policy 1 and policy 2 start to generate new rejected demands rather than reducing them. Policy 1 is worse than policy 2 in minimizing the rejected demands number. This result is logical, since policy 2 gives the priority to the relocation operations which aims to redistribute the cars in order to rebalance the system. Bad relocation operations may lead to an augmentation in the number of remaining rejected demands in the future. On the other hand, policy 3 performs much better than the other two policies. This can be explained by the fact that the jockey has an estimated knowledge of the future rejected demands. With this knowledge, the jockey is able to take better relocation decisions in order to reduce the maximum number of rejected demands, keeping in mind not to generate future rejected demands. This policy is better from the other two policies since relocation decisions are taken only when needed. When relocation operations are not advantageous to the system, the jockey does not relocate cars but he waits until the appropriate moment for better relocation operations. Fig. 6 clearly shows that the number of relocation operations is somehow constant using policy 1 and policy 2. While in policy 3, the number of relocation operations is decreasing, likewise the number of remaining rejected demands which also decreases.

D. Comparison of CPLEX and Greedy Algorithm

As we can see in the subsection VI-C, the comparison of the performance of the three proposed policies shows that Policy 3 is the best approach for the relocation problem. In the remaining part of this paper, our greedy algorithm implements the policy 3, exclusively. In order to assess the performance of our greedy algorithm we solved the same problem with the same data with CPLEX. Fig. 7 shows that the results of the greedy algorithm are competent with the results obtained by CPLEX; especially the greedy algorithm takes less than one second to deliver a solution while CPLEX may take a long time before delivering a solution as shown in Section V.

E. Stochastic Data Results

After solving the relocation problem using our greedy algorithm and CPLEX, each jockey is affected to a path that should be followed in order to cut down rejected demands. The path is constituted of a series of relocation operations to be done during the day. A relocation operation tells the jockey from which station and when, a car should be moved, and to which station and when, it should be dropped off. When the number of jockeys is increased in Policy 3, the number of remaining rejected demands decreases as well as the number of needed relocation operations as shown in Fig. 5 and 6. In order to measure the robustness of the resulted relocation operations, we used a special Gaussian method to add stochastic noise to the input data for the incoming and outgoing cars; knowing that the added stochastic noise does not exceed 10% of the original data. In Fig. 8 we see an example of stochastic input data modification on the number of incoming cars at a station in our carsharing system. After that, using the original data, the resulted relocation operation plan is applied, then these operations were evaluated on the input data that was modified in a stochastic manner regardless of the number of available cars; we call this step blind relocation. As shown in Fig. 9, when the number of jockeys is increased the number

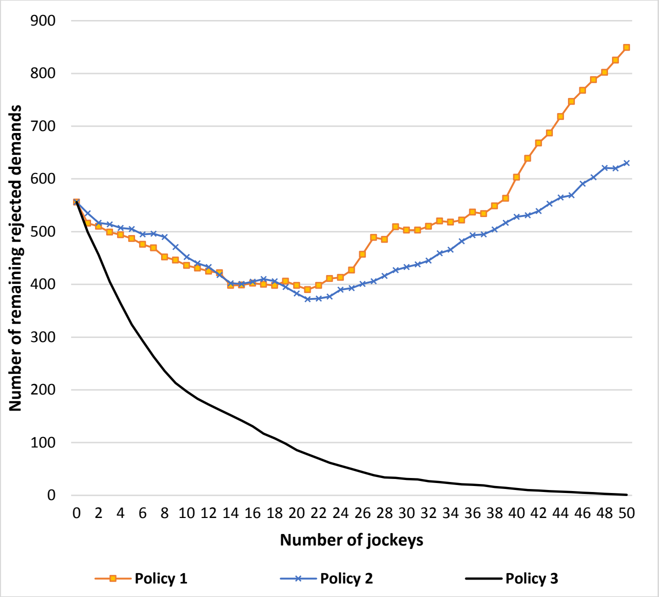


Fig. 5. Comparison of the performance of the three relocation policies.

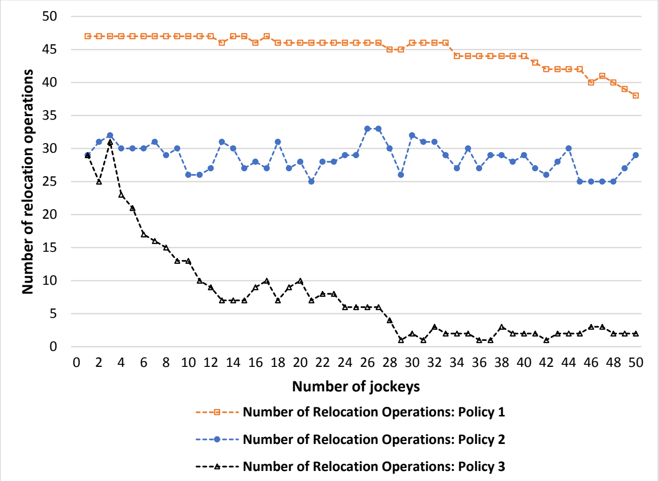


Fig. 6. Number of relocation operations with the three relocation policies.

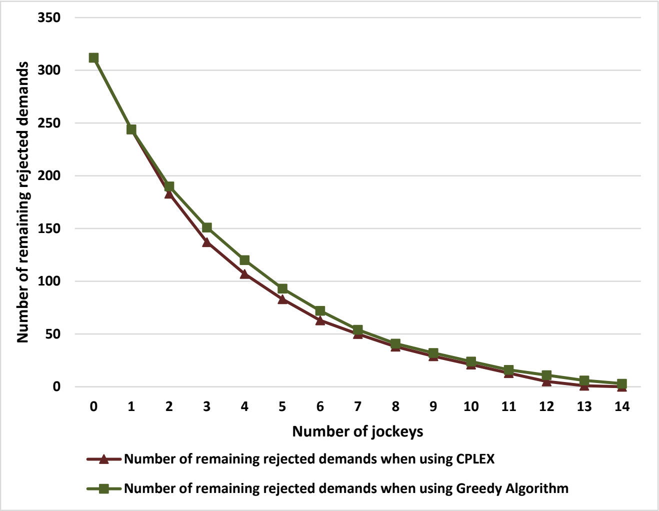


Fig. 7. Comparison results between the greedy algorithm and CPLEX (18 stations, 10 parking spots by station, 88 cars with the average of 12 trips by car).

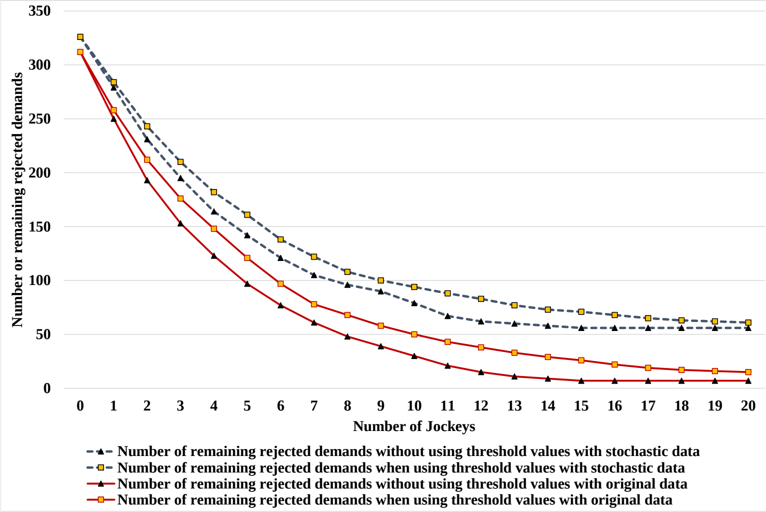


Fig. 9. Effects of threshold values and stochastic data on the total number of remaining rejected demands.

of remaining rejected demands is decreased with stochastic input data. However, as much as the number of jockeys is increased, the difference between remaining rejected demands increases when using the original data as well as when using stochastic modified data (see both curves with triangles). This is due to the fact that each stochastic modification on the input data, in any station at any time step, will be aggregated and propagated to all the following time steps that concern this station. Thus, since the number of available vehicles in each station at each time step is used to make the decision of the relocation operation, the resulted relocation operations can lose its efficiency drastically when the input data of user demands is changing.

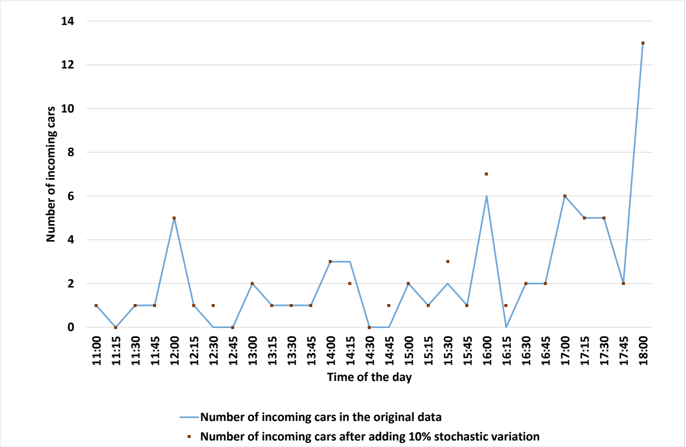


Fig. 8. Stochastic data variation on the number of incoming cars from 11:00 to 18:00 in one station.

F. Integrating Threshold Values in our Greedy Algorithm

In another step, the greedy algorithm is changed by integrating lower and upper threshold values in order to measure the effect of threshold values on the relocation operations when using stochastic modified data. The lower threshold is used to avoid relocating a car from a station when the number of available vehicles in the station before the relocation operation is less than or equal to the lower critical threshold. In this example, this value is set to one. That is if the station has only one car before the relocation operation, then the algorithm does not apply the relocation in that station. The upper threshold value is used to avoid relocating a car to a station when the number of available cars in the destination station is greater than or equal to the upper critical threshold. In this example, it is set to the maximum number of places in the station minus one. As shown in the chart below, the performance of the threshold strategy in terms of reducing the number of rejected demands (curves with squares) is worse than without threshold constraints. In addition, shown in Fig. 9 the threshold values does not bring improvement for the blind relocation on stochastic data compared to blind relocation without using threshold values. In both cases, the difference in the number of reduced rejected demands starts to be small when the number of jockeys is small, but gets bigger as the number of jockeys increases. On the other side, it is clear that the number of relocation operations when using threshold values, is less than the number of relocation operations without using threshold values since threshold adds a constraint on the decision of a relocation operation until all rejected demands problems are solved as shown in Fig. 10.

G. Identification of Mobility Patterns During the Day

As described earlier, when the number of jockeys is increased, the number of rejected demands decreases as well. However, the cost of relocation operations increases as well. In this study, we consider that we are using each jockey for the whole day, which is impractical. However, in real life there will be staff shifts that depend directly on the demands and needs of relocation operations. In the literature, relocation operations that are carried out at night are called static relocation since user demands for cars is considered negligible during this period. Static relocation is necessary to provide the stations with the appropriate number of cars for the next morning. In static relocation, there are no time window constraints to

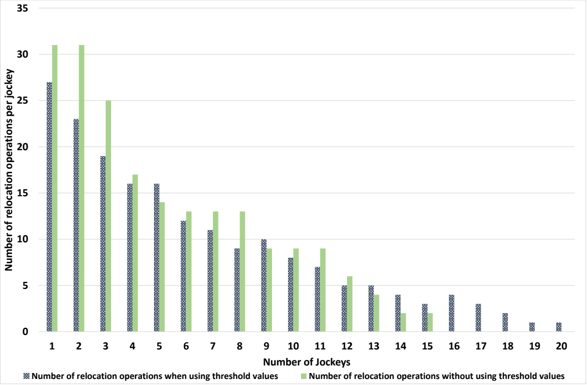


Fig. 10. Number of relocation operations when using threshold values and without using them.

deliver cars to stations at specific times, unlike cars relocation during the day where some stations have urgent needs for cars to satisfy user demands on time. In our approach, relocation operations are carried out during the day. When analyzing the time at which the rejected demands are solved using our greedy algorithm with policy 3, we get the histogram in Fig. 11. This histogram compares the total number of reduced rejected demands per hour of the day when using 15 jockeys for the whole day and when using them from 7:00 to 19:00. When analyzing the histogram in Fig. 11 we can detect some relocation patterns during the day. There are some periods of high activity such as the period from 8:00 to 10:00 and from 17:00 to 19:00. There are also periods of low activity such as period from 11:00 to 16:00. These patterns can be explained by the fact that these intervals correspond to periods of high mobility of customers in the morning when they go to work and in the evening when they come back home. On the other side, we notice that when we limit the working time until 19:00, the number of reduced rejected demands in the late hours (17:00 to 19:00) increases. This can be explained by the fact that the jockey can anticipate rejected demands and reduce them even before their occurrences. Knowing that the number of reduced rejected demands at any time t of the day, does not only represent the number of reduced rejected demands that occur at time t, but it also includes the anticipated rejected demands that occur in the future but reduced by relocation operations performed at time t.

Thereby, the required effort for the jockeying operations, will not be the same during the day; likewise, the number of jockeys should vary as well. Thus choosing the appropriate number of jockeys per time interval is a key factor to reduce the cost of jockeying operations.

In another experience, we divided the working time of jockeys into three periods with an interruption of work between them:

1) From 7:00 to 9:00
2) From 11:00 to 13:00
3) From 17:00 to 19:00

Then we compared the performance of the jockeys in this case with their performance when they work from 7:00 to 19:00. As we can see in the chart below in Fig. 12, even when we divide the working time into three periods, the number

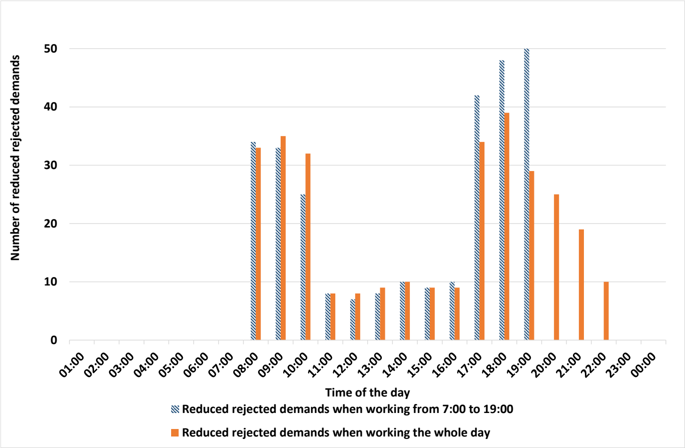


Fig. 11. Number of solved rejected demands in each hour of the day using 15 jockeys.

of reduced rejected demands decreases. However, the slope is smaller since the number of working hours is smaller. We conclude that the company must evaluate the cost of rejected demands in regard to the cost of the jockeying hour.

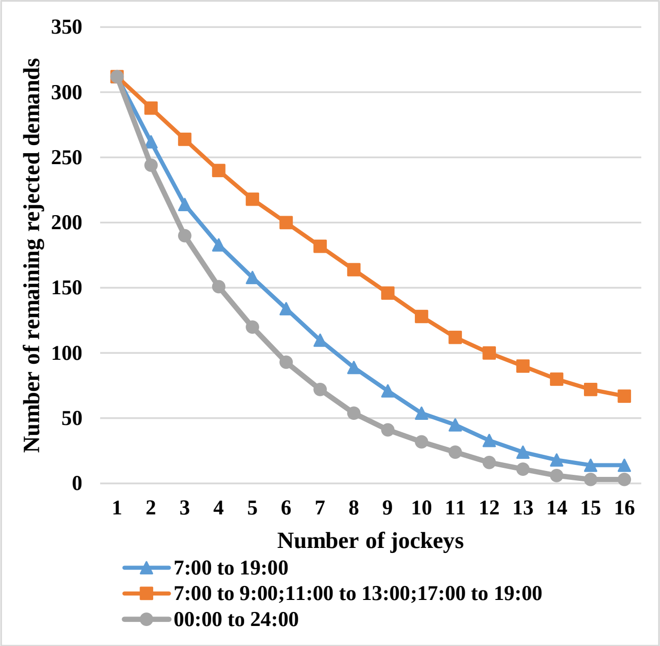


Fig. 12. Number of remaining rejected demands when varying working hours.

VII. Conclusion and Perspective Works

The one-way car sharing service is appealing to users since they are not required to return the car of departure station and for its flexibility. Nevertheless, this flexibility leads to an imbalance in cars distribution. The imbalance problem affects the image of the service and makes it less attractive to users. To cope with problem, relocation operations are vital to increase the satisfaction of the clients. In this study, three different policies of car relocation are compared. The performance of Policy 3, where the jockey has information on the future state of the system based on historical data and predictions, is much better than the two other policies that do not consider any future information. It was proven that the implementation of intuitive policies that are based on basic decisions such as the total distance covered by jockeys and the available cars at stations without considering the propagation of the effect of these relocation operations on the future, which may influence the overall service, will not have a great impact in minimizing the number of rejected demands. On the contrary, applying these policies may require more relocation operations, which will eventually rise the total cost of the system. Taking into consideration the historical data to make future estimation is crucial in order to minimize the number of rejected demands. From another side, we see that jockeys pass by inactivity periods when there is no need for relocation operations. Analyzing these periods, suggests that working hours of each jockey can be reduced and so, we can decrease the car sharing operation cost. In addition, we found that the effectiveness of the resulted relocation operations is highly dependent on the input data even when we use threshold values for the relocation operations. As perspective, it is possible to implement a new heuristic approach based on stochastic model using historical data, in the aim of solving the relocation problem. This can be modeled in a simulation environment that considers real life parameters.

References

[1] A. Reno and G. Weisbrod, “Economic impact of public transportation investment,” 2009.

[2] J.-P. Rodrigue, C. Comtois, and B. Slack, *The geography of transport systems*. Routledge, 2013.

[3] R. Katzev, “Car sharing: A new approach to urban transportation problems,” *Analyses of Social Issues and Public Policy*, vol. 3, no. 1, pp. 65–86, 2003.

[4] F. Meunier, “Véhicules partagés: des défis pour la ro.”

[5] J. Rifkin, *The Age of Access. How the Shift from Ownership to Access is Transforming Capitalims*. Putnam, New York et al, 2000.

[6] [Online]. Available: http://www.navigantresearch.com/research/carsharing-programs

[7] S. A. Shaheen, D. Sperling, and C. Wagner, “A short history of carsharing in the 90’s,” *Institute of Transportation Studies*, 1999.

[8] D. Jorge and G. Correia, “Carsharing systems demand estimation and defined operations: a literature review,” *EJTIR*, vol. 13, no. 3, pp. 201–220, 2013.

[9] M. Barth and M. Todd, “Simulation model performance analysis of a multiple station shared vehicle system,” *Transportation Research Part C: Emerging Technologies*, vol. 7, no. 4, pp. 237–259, 1999.

[10] M. Barth, M. Todd, and L. Xue, “User-based vehicle relocation techniques for multiple-station shared-use vehicle systems,” *Transportation Research Record*, vol. 1887, pp. 137–144, 2004.

[11] K. Uesugi, N. Mukai, and T. Watanabe, “Optimization of vehicle assignment for car sharing system,” in *Knowledge-Based Intelligent Information and Engineering Systems*. Springer, 2007, pp. 1105–1111.

[12] D. Jorge, G. H. Correia, and C. Barnhart, “Comparing optimal relocation operations with simulated relocation policies in one-way carsharing systems,” 2014.

[13] A. G. Kek, R. L. Cheu, Q. Meng, and C. H. Fung, “A decision support system for vehicle relocation operations in carsharing systems,” *Transportation Research Part E: Logistics and Transportation Review*, vol. 45, no. 1, pp. 149–158, 2009.

[14] W. D. Fan, R. B. Machemehl, and N. E. Lownes, “Carsharing: Dynamic decision-making problem for vehicle allocation,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2063, no. 1, pp. 97–104, 2008.

[15] R. Nair and E. Miller-Hooks, “Fleet management for vehicle sharing operations,” *Transportation Science*, vol. 45, no. 4, pp. 524–540, 2011.

[16] L. Moalic, S. Lamrous, and A. Caminada, “A multiobjective memetic algorithm for solving the carsharing problem,” in *WORLDCOMP’13-The 2013 World Congress in Computer Science, Computer Engineering, and Applied Computing*, vol. 1, 2013, pp. 877–883.

Load Balancing in Cloud Complex Systems using Adaptive Fuzzy Neural Systems

Mohammad Taghi Sadeghi
Bachelor of Computer, Faculty of Engineering
Birjand University
Jajarm, North Khorasan, Iran

***Abstract*—Load balancing, reliability, and traffic are among the service-oriented issues in software engineering, and cloud computing is no exception to this rule and has put many challenges ahead of experts in this field. Considering the importance of the load balancing process in cloud computing, the purpose of this paper is to provide an appropriate solution for load balancing load in complex cloud systems using an adaptive fuzzy neural system. This system consists of four layers, and a particular operation is performed on each layer. The results of the experiments show that the system has better performance in the criteria mentioned above (balancing, traffic and reliability).**

Keywords—Load balancing; cloud computing; adaptive fuzzy neural system

I. Introduction

Cloud computing has caused many changes in the world of information and communication technology because of its efficiency. It is gaining more and more popularity due to its easy access, low costs, and comfortable usage. Small companies can conduct business in the field of technology and information processing, and bigger companies can significantly cut their expenses in maintenance, manpower, and processing equipment by using cloud computing [1], [2]. The majority of the physical servers in cloud computing work with the virtualization technique. Each virtual machine needs a specific amount of materials such as central processing unit, memory, bandwidth, sending and receiving data, etc. to maintain the application Function Segmentation and security. Furthermore, virtualization technology enables a number of virtual servers to work on the same physical machine used to improve the efficiency of Physical server resources and reduce energy consumption. Therefore, virtualization can help managers achieve an efficient solution for the flexible management of resources by using virtualization Technique, we can enable more than one virtual machine to work on one physical machine, and each of them will provide their required resources by cutting a section of the machine Physical resources. Data centers of virtual machines in cloud computing have broadened to such an extent that the quality of positioning the virtual machines has become an important issue for producers of cloud computing in order to prevent interference in load balance between virtual machines and losses due to non-compliance with the quality of the virtual machine based on the Service Level Agreements (SLAs) in cloud computing [5]-[8].

II. Literature Review

In [9], the authors have claimed that I/O virtualization is a big challenge, and there is no ideal solution. When virtual machines want to access these non-sliceable resources, an interference in their performance occurs, and the SLA is violated. Hence, one of the main problems of cloud computing is the interference between the virtual machines and their load imbalance of the SLA, and the effective placement of virtual machines greatly reduces or increases the profitability of the cloud computing of virtual machines. In [10], researchers have only considered the strategy of placing virtual machines but have not considered the quality requirements of user applications in load balancing of virtual machines [11]. They have provided a framework for load balancing strategies in the cloud computing environment and have proposed a method to evaluate the resource allocation strategies in the cloud computing environment, seeking to focus on optimizing the awareness and compatibility of network load balancing strategies. The framework for network load balancing in cloud computing is based on active criteria. Network topology, taking into account traffics, and the optimal change of the criteria corresponding to the user's dynamic needs, which plays a major role in determining the Internet architectures and protocols and shaping load balancing management strategies in cloud computing, are the most important results of the research [12].

It is possible to provide a linear scheduling strategy for load balancing in the cloud environment. The separate scheduling of resources and tasks includes the waiting time and the response time. In this research, a linear scheduling algorithm for scheduling tasks and resources called LSTR is designed which schedules tasks and resources, respectively. Here is a combination of Nimbus and Cumulus services to create a server-provider. IaaS cloud environment, KVM/Xen virtualization, and LSTR scheduler have been used to balance load and maximize operational capacity and resource utilization [13]. In [14], dynamic load balancing is done using virtual machines for the cloud computing environment. This research adopts a system which uses virtualization technique to balance the dynamic load of the data center based on applied demands and service resource optimization. In this research, this concept is introduced as a rough criterion for exploiting multidimensional resources in the service. In addition, this research effectively develops a series of innovations to prevent overload in the system and use it in energy storage. The experimental results show that the proposed algorithm is desirable.

In the distributed type, dynamic load balancing algorithms are implemented by all the nodes in the system, among which the task of load balancing is divided. The interaction between nodes to achieve load balancing can be in two forms of cooperation and without cooperation. In the first form, the nodes work along with each other to achieve a common goal (for example, to improve the overall response time). In the second form, each node works independently towards a local goal, such as improving the response time of a local work. Dynamic load balancing algorithms with distributed feature generate more messages than the non-distributed type. An advantage of this method is that if one or more nodes fail within this system, this does not cause the entire load balancing process to stop. To create resources in cloud computing, an efficient model provides two interactive performance evolutionary classes to determine the minimum number of servers required for SLAs. For both classes, the probability of a response time smaller than x is considered to be y. Two server allocation strategies are used; the first one is the shared allocation and the other is exclusive allocation. The FCFS scheduler determined an allocation strategy for distributing response time to develop an innovative algorithm that required the least number of servers. This algorithm was used in operational conditions and yielded favorable results [15].

In [6], the researcher presented a game theory approach to load balancing for cloud computing services. The purpose of this research is to solve the problem of the service quality by limited load balance. Service seekers require that their complex parallel computing problem be provided by requesting the resources they need. In this research, the game theory has been used to solve the load balancing problem and a suitable two-step solution has been proposed. First, each customer independently solves his/her problem without considering the load balancing multiplex. An appropriate binary programming method is proposed for the independent optimal solution. Second, an evolutionary mechanism is designed that changes the multiplex strategy of the initial solutions of different customers. The overall result is that an appropriate solution can always be found [16]. Moreover, in another study, the dynamic load balancing used in cloud computing was studied by using multi-criteria distributed analysis. In this research, a two-step process method is proposed for managing dynamic autonomous resources in cloud computing in two stages. First, a distributed architecture divides resource management into independent tasks, each of which is run by agent nodes that are physical machines of firm connection at the data center. Second, automated agent nodes are configured through a number of decision-making criteria using the fuzzy configuration method. Simulation results show that the proposed method is flexible [17].

In different studies, load balancing was done to provide services in cloud computing environments. These load balancing algorithms are provided for Software-as-a-Service (SaaS) to minimize the cost of infrastructure. The proposed algorithms designed are a safe way enabling SaaS providers to manage client changes, map client requests to infrastructure layer parameters, and set up virtual machines [18]. This research analyzes and validates the algorithms for minimizing the costs of SaaS providers in the cloud computing environment. The optimal allocation algorithms are investigated in cloud computing clusters and a possible model for tasks (requests such as CPU, memory, and storage) in cloud computing is presented. The proposed model can split the source allocation problem into two balancing and scheduling problems and consider the join-the-shortest-queue and power-of-two-choice algorithms with the maxweight scheduling algorithm. This research shows that algorithms optimize operational power and limit the optimal queue length in heavy traffic [19]. In [20] a tool is presented for modeling and simulating the real-time assignment of virtual machines in the cloud data center. The innovative method was applied to schedule resources dynamically in a cloud data center that has runtime constraints on RMs and PMs. This method focuses on timing simulation in the Infrastructure-as-a-Service (IaaS) layer. Simulations indicate that the multidimensional information source designed and implemented in real time shows that the results have improved compared to previous approaches. In [21], load balancing and automatic load balancing for cloud services are surveyed with fuzzy logic and ant colony, and a method is developed to support cloud agents for an optimal configuration in using dependencies on cloud-based applications that require high security. This method uses the model-driven principles of UML and the Bayesian networks to store, analyze, and optimize the configuration in the cloud space.

Therefore, the load balancing of the whole system is performed by the central nodes of each cluster. Centralized dynamic load balancing receives fewer messages. Hence, the total number of interactions within the system is reduced compared to that of the distributed type. However, centralized algorithms can cause a limiting operation (bottlenecks) on the central node, and the load balancing process fails when the central node collapses. Therefore, this algorithm is more suitable for networks with a smaller size.

III. Proposed Approach

For the virtual machine load balancing problem, the virtual machines are embedded in the physical machines in such a way that the cloud provider’s profit is maximized. Each physical machine limits sources for hosting a number of virtual machines. Each virtual machine has its own resource requirement and rental rates. In the virtual machine load balancing problem, each physical and virtual machine can be considered as a backpack and an item in the multi-backpack problem, respectively. The physical machine limits sources to host a number of virtual machines [9]. A virtual machine has its own demand for resources and rental price. The purpose of the virtual machine load balancing problem is to balance the load of virtual machines in the physical machine as much as possible, while not exceeding the resources of each physical machine. In addition, the cloud provider can maximize profits by load balancing of the virtual machines in the physical machines. With the above mappings, the problem of the virtual machine load balancing can be turned into a multi-backpack problem.

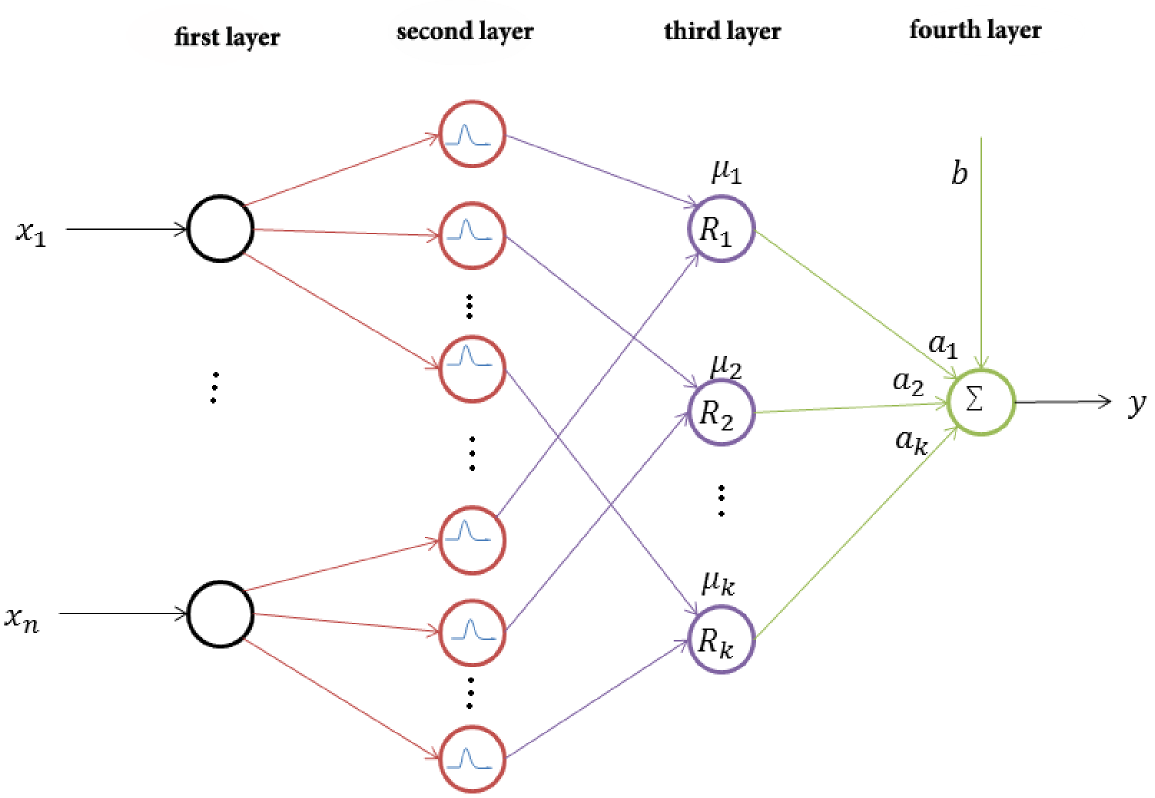


Fig. 1. Proposed system.

Without considering the interference of the virtual machine in the virtual machine load balancing, the requirements of the virtual machine's quality of running applications in the virtual machine may be violated. Different from the problem of the usual placement of the virtual machine, the load balancing problem aware of the quality of the virtual machine considers the following three factors in the virtual machine load balancing: 1) demands for resource from a virtual machine, 2) requirements for the virtual machine's quality of applications, and 3) interference in virtual machines. To integrate these three factors, we consider the problem of virtual machine load balancing aware of the quality of service to the virtual machine load balancing based on the prediction of the cloud provider's profit for these factors. Then, the load balancing problem aware of the quality of service can be formulated as an integer linear programming model to obtain an optimal solution. However, solving the integer linear programming requires noticeable computing time. In this section, a load balancing algorithm is proposed in a complex cloud system using an adaptive fuzzy neural system.

Fig. 1 shows the structure of the proposed fuzzy system with four layers. Each layer has its own operations, each of which is described below.

A. Definition of the Rules

In this fuzzy network, each rule in the proposed method is defined as follows:

The Kth rule: If x_1 is equal to A_{k1} and x_n is equal to A_{kn}, then $y' = a_k$

Where a_k refers to a single fuzzy rule and A_{kn} refers to a set of fuzzy rules. Here are four network layers.

B. The First Layer (Input Layer)

In the first layer, the values for each node (input variable) are transmitted directly to the next layer. Therefore, no calculations are required. In other words, each node is associated with an input variable, and as a result of this layer, each input is transmitted exactly to the next layer. The trained dataset tag (S) will be described in relation (1):

$$= \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_N, y_N)\} \tag{1}$$

C. The Second Layer (Calculation Of Membership Function)

In the second layer, each node is calculated corresponding to a fuzzy set and membership value. In this layer, the fuzzy set A_{kj} is obtained by the Gaussian membership function and from equation (2):

$$M_{kj}(x_j) = \exp\left(-\frac{(x_j - m_{kj})^2}{\sigma_k^2}\right)) \tag{2}$$

In this equation, m_{kj} shows the center of the fuzzy set and σ_k^2 shows the width of the fuzzy set. This Gaussian membership function is widely used in neural fuzzy systems. In other words, in the second layer, each node corresponds to a fuzzy rule and the value of its membership function is calculated, and the Gaussian function is used for this network.

D. The Third Layer (Fuzzy Rules)

In the third layer, each node represents a fuzzy logic rule, and the *AND* operation of their set is done by multiplying the output of the membership functions.

$$\mu_k(\vec{x}) = \prod_{j=1}^{n} M_{kj}(x_j) = \exp\{-\sum_{j=1}^{n}[\frac{(x_j-m_{kj})^2}{\sigma_k^2}]\} = \exp\{-\frac{\|\vec{x}-\vec{m}_k\|^2}{\sigma_k^2}\} \quad (3)$$

$\mu_k(\vec{x})$ is used as a criterion for deciding whether to produce a new fuzzy rule. In this equation, $\vec{x}$ is obtained from equation (4):

$$\vec{x} = [x_1, \dots, x_n]^T, \vec{m}_k = [m_{k1}, \dots, x_{kn}]^T \quad (4)$$

E. Fourth Layer (Output Layer)

In the fourth layer, each node corresponds to an input variable, and the defuzzification operation is performed in this section. In this layer, the total weight of the outputs of the previous layer with a bias value is considered for calculating the output of the layer. The output of the fourth layer is obtained using equation (5):

$$y' = \sum_{k=1}^{r} a_k . exp\left\{-\frac{\|\vec{x}-\vec{m}_k\|^2}{\sigma_k^2}\right\} + b = \sum_{k=1}^{r} a_k . \mu_k(\vec{x}) + b \quad (5)$$

In the layer *b*, the bias is related to the values computed by the system in this layer.

The purpose of training the network structure is to partition the input space generated under the influence of a number of fuzzy rules. As mentioned above, $\mu_k(\vec{x})$ is used as a criterion for deciding whether to produce a new fuzzy rule. For the first input data $\vec{x}(0)$, a new fuzzy rule is generated to which the membership function with the Gaussian center and Gaussian width are allocated using equation (6):

$$m_{1j} = x_j(0), j = 1, \dots, n, \sigma_1 = \sigma_{init} \quad (6)$$

In this equation, σ_{init} is a predetermined value that defines the initial Gaussian width in the first cluster. For successful input data $\vec{x}(t)$, the value of K is obtained using equation (7):

$$K = arg \max_{1 \le k \le r(t)} \mu_k(\vec{x}(t)) \quad (7)$$

Where r(t) is the number of rules available at time t, and as long as $\mu_k < \mu_{th}$, the new rules are created.

F. Parameters of the Proposed Method

In the experiments conducted in this paper, it is assumed that the cloud computing system has 250 physical machines randomly distributed in 10 clusters. Each of the clusters is located in a local area on a 100 × 100 square plate. Of the 10 clusters, one as the central cluster organizes all of the 10 clusters of the physical machine and is determined as a tree topology. In each cluster, the physical machines are randomly located in a local area in the cluster. In addition, there is a switch in each cluster of the physical machine to provide intra-cluster and inter-cluster communications between physical machines.

Based on the architecture of the cloud computing system, simulation experiments have been performed on the following parameters:

1) In each physical machine, there are a number of virtual machines available. The number of virtual machines available is randomly assigned between 0 and 10.

2) The amount of resources available (the available central processing in GHz, the available memory space in gigabytes, the available storage space in gigabytes) is triple. The source interval [(12,129,200), (96, 3000, 9600)] is randomly used to decide the resources available to each physical machine.

3) The bandwidth is assumed to be in the range of 10 gigabytes/second to 40 gigabytes/second.

4) The number of virtual machines in each simulation run is 200 to 1000.

5) To simulate the profit metric, the cost of a virtual machine (revenue) and the payment of a fine resulting from a service quality violation in a virtual machine are allocated.

IV. Experiments

Some parameters are considered for load balancing oudtechniques in the cl environment. Some algorithms create and improve these parameters the optimal load balance parameter over other algorithms. Thes of traffic, ,reliability efficiency, etc. are used to calculate the number of tasks executed. These factors need to be optimized to improve These .system performance factorneed to be improved at a s The .reasonable cost factor of resource efficiency and usefulness of resources used to control the is usefulness of for This factor should be optimized .resources useful load balance. Scalability demonstrates the ability of an algorithm to perform load balancing on a system with any limited number of nodes. This parameter should be improved. time Response is the eamount of tim spent by an algorithm to answer specific balancing load in distributed systems.This parameter should be minimized. Failover Tolerance Parameters show the ability of an algorithm to perform load balancing when a node loses its connection . This section indicates results related to criteria such as the overall the traffic of the adaptive fuzzy neural system, computational benefits, points sensitive to delays, and running time. Each of these criteria a discussed in is graph.

A. Overall Traffic

Fig. 2 shows the outcomes of overall traffic in three different methods. As the diagram shows, the adaptive fuzzy neural system has a better effect on the overall traffic, the ant colony algorithm is somehow similar to clustering. Also, the overall traffic of the topology of the tree is less than that of the FAT tree.

Moreover, Fig. 3 shows the comparison of efficiency using the proposed method, the clustering method, and the ant colony base algorithm in different topologies.

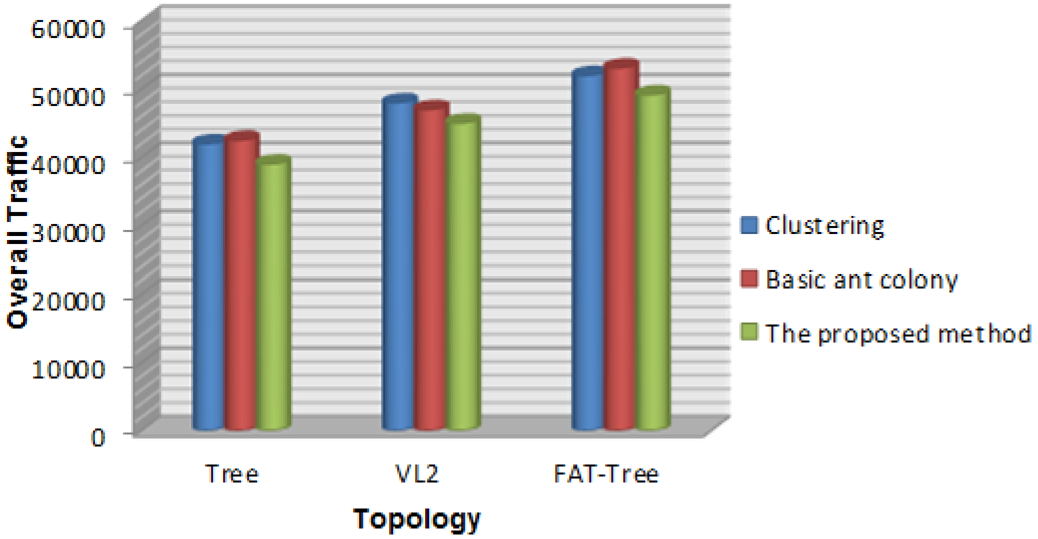


Fig. 2. Change in overall traffic generated by the proposed method.

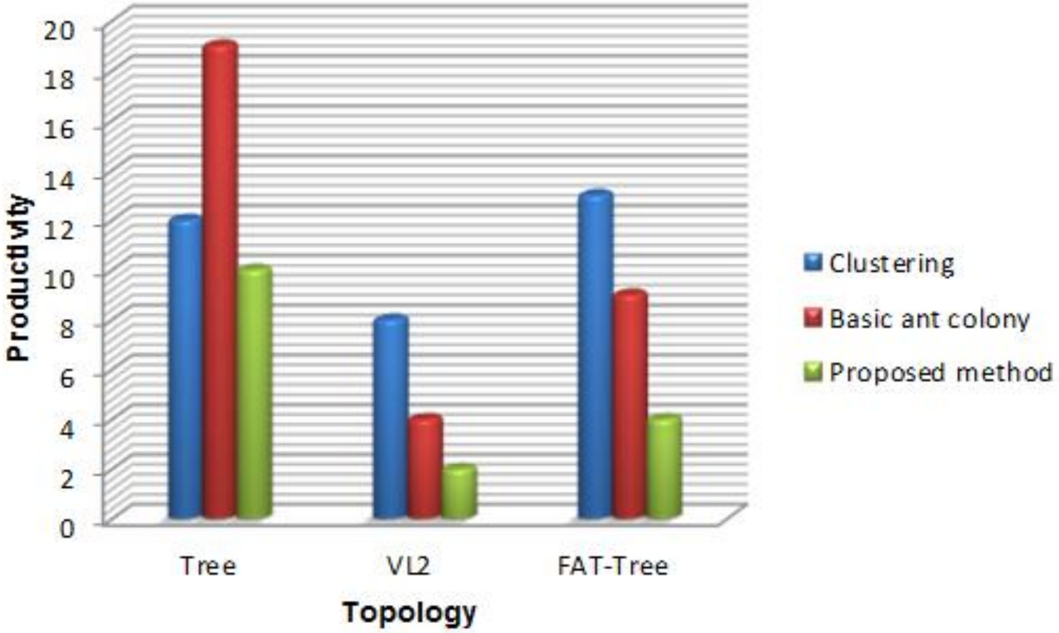


Fig. 3. Minimum productivity.

The proposed algorithm not only leads to the optimization of overall network traffic, but it is also for network productivity. The adaptive fuzzy neural system refers to a situation in which the method of overall traffic and productivity is optimized. By using the data in this paper, traffic has been used among the same virtual machines for different network topologies. As seen in Fig. 2 and 3, although the proposed method slowly increases overall traffic, it further reduces overall traffic and improves network performance. Also, when the FAT tree and VL2 use multiple paths, their effects are more significant.

B. *Patterns of Critical Points*

When the traffic matrix of the virtual machine is created, if the system wants to overlap some traffic, the network path generates Critical points, leading to aggregation. Fig. 4 shows the change in the number of Critical points created in the three methods, including the adaptive fuzzy neural system, clustering, and ant colony in different topologies. Compared to the two methods, in the proposed method, the number of paths of Critical points has decreased significantly, and the number of paths of the Critical points in the FAT tree has been to 8%. Although the proposed method cannot completely avoid network aggregation, it greatly improves network performance.

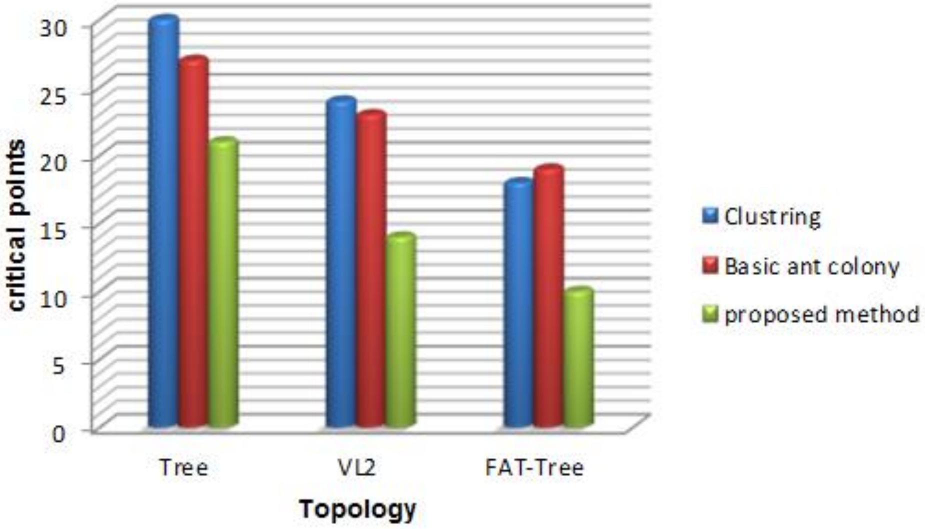


Fig. 4. Change in the number of critical points created.

C. Changes in Profit

The following simulation results are shown for 50 simulation runs on average. The benefits obtained are shown by creating different numbers of the virtual machines from 200 to 1000. As seen in the figure, the profit from all algorithms increases along with the number of virtual machines. However, random fit algorithms, the first fit, do not consider the minimum fit of virtual machine interference. By creating more virtual machines, the virtual machine interference increases. The virtual machine interference influences the quality requirements of the virtual machine of the applications. If the effect of virtual machine interference is not controlled, violating the quality of the virtual machine will result in fines being imposed to reduce the profitability of the cloud provider. In an adaptive fuzzy neural system, the virtual machine interference has dropped as much as possible. There is a linear process in profit growth by increasing the number of virtual machines created. Compared to the two algorithms, the proposed algorithm can increase their profits by approximately 12%, 9%, and 21%.

D. Reliability and Load Balance

Reliability should be calculated by a timetable that calculates the time associated with different activities such as sending and receiving packet. The reliability of the initial phase (when the node enters the network) is ignored because the lifespan of the network is very high and is considered about tens or even hundreds of days, and can be ignored due to the very low initial phase (Fig. 5 & 6).

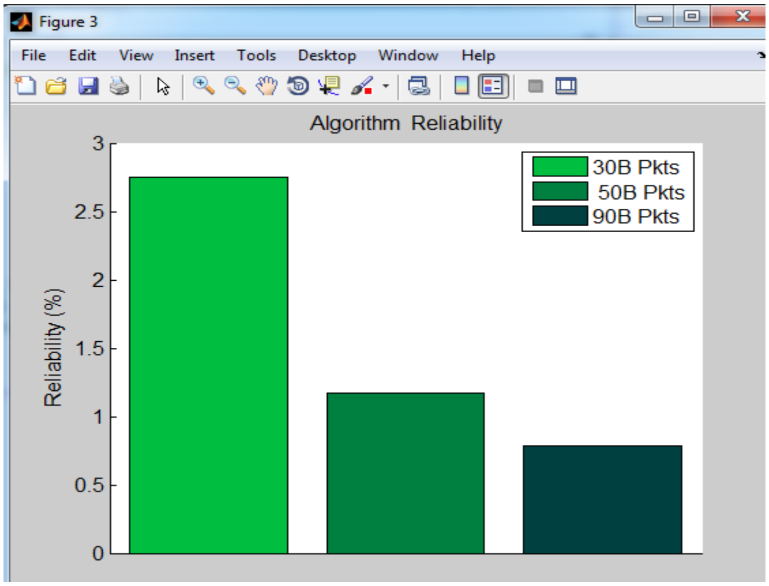


Fig. 5. Comparison of the reliability based on the size of packages.

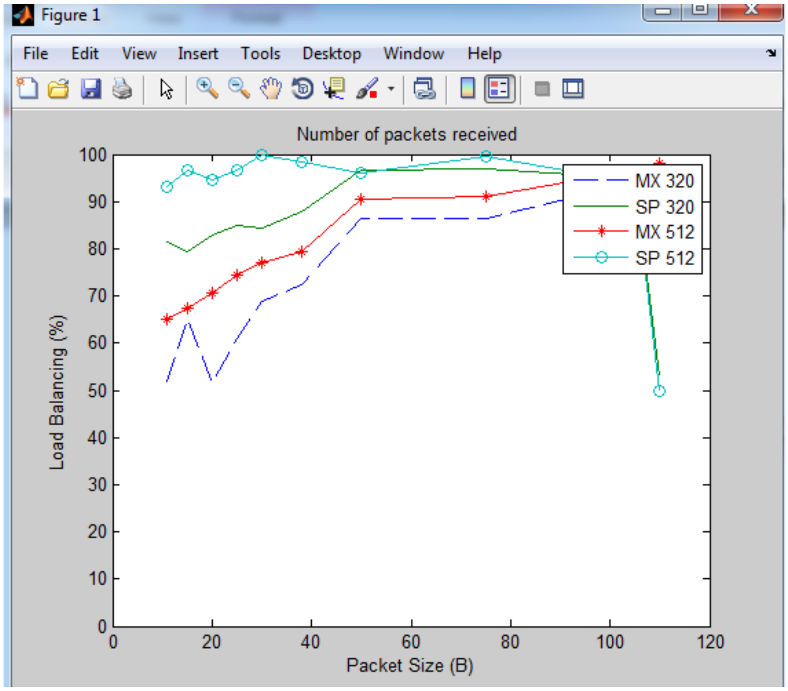


Fig. 6. Comparison of load balances with packages received.

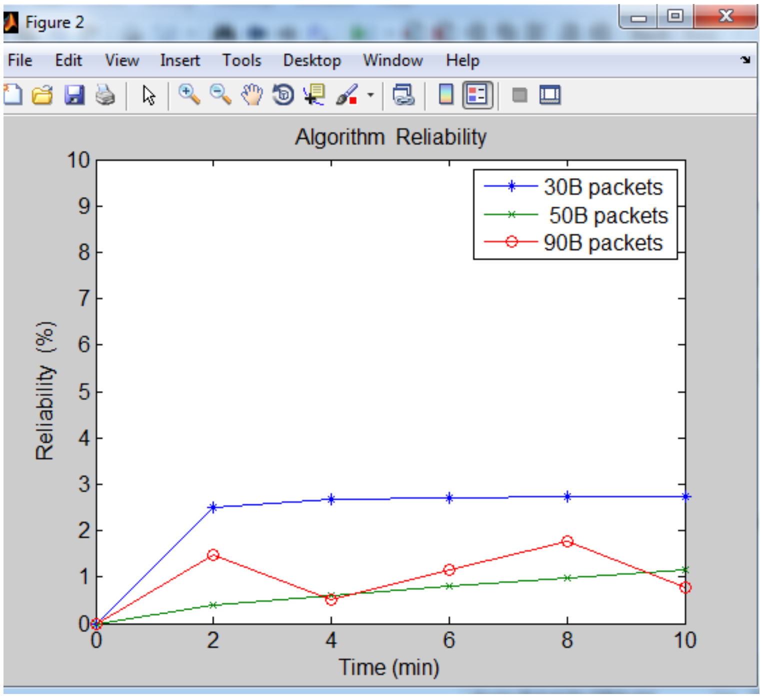


Fig. 7. Comparison of reliability towards time.

In Fig. 7, the most important comparison has been made in terms of reliability which is calculated based on the number of packages received. As can be seen, the reliability of the proposed method for the number of packages received is higher than that of other methods.

E. *Execution Time*

The execution time of the proposed method is measured in seconds. Due to the parallel characteristics of the adaptive fuzzy neural system, the algorithm can be applied to memory machines and parallel calculations can be used to reduce time and improve performance. The execution time in seconds is acceptable for analysis in the data center, especially for solving the NP-Hard problem. Although the time performance of the adaptive fuzzy neural system is slightly weak, the algorithm shows more accurate results. Experiments confirm that the use of the adaptive fuzzy neural system is suitable for analyzing problems.

V. Conclusion

In general, load balancing can be effective by dividing the flow of performance between all nodes. This paper investigated the load balancing in complex cloud systems using the adaptive fuzzy neural system. As can be seen in the figures, tree topology has a better overall optimization and also contains the largest amount. It can be concluded that the tree topology is more likely to generate aggregation among similar virtual machines. Although the FAT tree has a large overall traffic, it can smooth traffic due to its multi-path connections. There is a great difference between the minimum path efficiency and the optimal amount for VL2. Therefore, the distribution of traffic is non-uniform. According to the results of the graphs, the efficiency of the proposed method is greater than that of the clustering and base ant colony method. The delays in the Critical points of the proposed method are approximately 7% better than that of the other two methods.

References

[1] Einollah Jafarnejad Ghomia, Amir Masoud Rahmania, Nooruldeen Nasih Qaderb, Load-balancing algorithms in cloud computing: A survey, Journal of Network and Computer Applications, Volume 88, 15 June 2017, Pages 50-71.

[2] Tang Linlin, Li Zuohua, Ren Pingfei, Pan Jengshyang, Lu Zheming, Jingyong Su, Meng Zhenyu, Online- and offline-based load balance algorithm in cloud computing, Knowledge-Based Systems, Volume 138, 15 December 2017, Pages 91-104.

[3] AvnishcThakur, Major Singh Goraya, A taxonomic survey on load balancing in cloud, journal of Network and Computer Applications, Volume 98, 15 November 2017, Pages 43-57

[4] Mohit Kumar, S.C. Sharma, Dynamic load balancing algorithm for balancing the workload among virtual machine in cloud computing, Computer Science, Volume 115, 2017, Pages 322-329.

[5] Ren et al., “The load balancing algorithm in cloud computing environment,” in International Conference on Computer Science and Network Technique, Changchun, China, 2012.

[6] J. Bhatia et al., “HTV Dynamic Load Balancing Algorithm for Virtual Machine Instances in Cloud,” in International Symposium on Cloud and Services Computing, Mangalore, KA, 2012.

[7] Aarti Singha, Dimple Junejab, Manisha Malhotra, Autonomous Agent-Based Load Balancing Algorithm in Cloud Computing, Computer Science, Volume 45, 2015, Pages 832-841.

[8] B. Shaoo et al., “Analyzing the Impact of Heterogeneity with Greedy Resource Allocation Algorithms for Dynamic Load Balancing in Heterogeneous Distributed Computing System,” Int J Comput Appl, Jan. 2013.

[9] P. Samal and P. Mishra, “Analysis of variants in Round Robin Algorithms for load balancing in Cloud Computing,” International Journal of Computer Science and Information Technologies, 2013.

[10] A. Lakra, and D. Yadav, "Multi-Objective Tasks Scheduling Algorithm for Cloud Computing Throughput Optimization." Procedia Computer Science, 2015.

[11] A. Thomas et. al., "Credit-Based Scheduling Algorithm in Cloud Computing Environment." Procedia Computer Science, 2015.

[12] X. Ren et al., “A Dynamic Load Balancing Strategy for Cloud Computing platform based on Exponential Smoothing Forecast,” in International Conference on Cloud Computing and Intelligence Systems., Beijing., China, 2011.

[13] H. Chen et al., “Towards energy-efficient scheduling for real-time tasks under uncertain cloud computing environment,” J.Syst. Software, Jan. 2015.

[14] K. Navdeep and K. Kaur, "Improved Max-Min Scheduling Algorithm.," IOSR Journal of Computer Engineering, vol. 17, May 2015.

[15] E. Pacini et al., “Balancing throughput and response time in online scientific Clouds via Ant Colony Optimization (SP2013/2013/00006)”, Advances in Engineering Software, June 2015.

[16] F. Ramezani and F. Khadeer Hussain, “Task-based System Load Balancing in cloud computing using Particle Swarm Optimization” Int JParallel Prog, Oct. 2013.

[17] D. Babu and P. Venkata, “Honeybee behavior inspired load balancing of tasks in cloud computing environments,” Appl Soft Comput, May 2013.

[18] N. Tziritas et al., "On minimizing the resource consumption of cloud applications using process migrations." Journal of Parallel and Distributed Computing, 2013.

[19] Himani and H. Sindhu, "Comparative analysis of scheduling algorithms of Clouds in cloud computing." *International Journal of Computer Applications,* 2014.

[20] Ranesh Kumar, Naha Mohamed Othman, Cost-aware service brokering and performance sentient load balancing algorithms in the cloud, Journal of Network and Computer Applications, Volume 75, November 2016, Pages 47-57.

[21] A. Saffar, R. Hooshmand, A. Khodabakhshian, A new fuzzy optimal reconfiguration of distribution systems for loss reduction and load balancing using the ant colony search-based algorithm, Applied Soft Computing, Volume 11, Issue 5, July 2011, Pages 4021-4028

Context Aware SmartHealth Cloud Platform for Medical Diagnostics

Using Standardized Data Model for Healthcare Analytics

Sarah Shafqat, Almas Abbasi
Department of basic & Applied Sciences,
International Islamic University (IIU),
Islamabad, Pakistan

Muhammad Ahsan Qureshi, Tehmina Amjad
Department of basic & Applied Sciences,
International Islamic University (IIU),
Islamabad, Pakistan

Muhammad Naeem Ahmad Khan
Department of Computer Science,
Shaheed Zulfiqar Ali Bhutto Institute of Science and
Technology (SZABIST),
Islamabad, Pakistan

Hafiz Farooq Ahmad
Department of Computer Science, College of Computer
Sciences and Information Technology (CCSIT)
King Faisal University (KFU)
Alahsa 31982, Kingdom of Saudi Arabia

***Abstract*—Healthcare has seen a great evolution in current era in terms of new computer technologies. Intensive medical data is generated that opens up research in healthcare analytics. Coping with this intensive data along with making it meaningful to deliver knowledge and be able to make decisions are the most important tasks. To deduce the authenticity of the data on basis of precision, correction, associations and true meaning is important to validate the understanding of correct semantics. In case of medical diagnosis to form accurate understanding of associations while removing ambiguity and forming a correct picture of the case is of utmost importance. To come up with the right metrics for the diagnostic solution we have explored the known criteria to validate healthcare analytics techniques involved in formation of diagnosis that results in betterment and safety of patients under observations and heading towards possible treatments. In this work, we have proposed a thematic taxonomy for the comparison of existing healthcare analytics techniques with emphasis on diabetes and its underlying diseases. This analysis lead us to propose a data model for hybrid distributed simulation model for future Context Aware SmartHealth cloud platform for diagnostics. This platform is designed to inherit smartness of unsupervised learning which in turn would keep updating itself under supervised learning by qualified experts. Finally, the accuracy would be determined using HUM approach with biomarkers or a better accuracy model than AUC. The recommended action plan is also presented.**

Keywords*—*Healthcare analytics; medical diagnostics; HL7; cloud platform; SmartHealth; big data

I. Background

Going through some early histories of medical informatics [1] we get to know of the major research that was carried out in 1977 in Bosnia and Herzegovina. In it the periodic data analysis of healthcare services and performance available in Bosnia and Herzegovina was done and later in 1982, the first Local Health Information System (LHIS) with health databases with supervision of 6000 citizens was tested. Izet Masic and Arif Agovic, an electronics engineer in Energoinvest Ltd. in Sarajevo, were known as the creator and pioneer of Medical Informatics in Bosnia and Herzegovina. A Healthcare Informatics Society BiH was established in 1987. Realization of ‘Development of Information System of healthcare B&H in circumstances of electronic data manipulation’ project started in 1986 after approval by the Executive Board of Association of healthcare communities B&H. It was planned to be in three phases: (i) first phase (1985-1990), (ii) second phase (1991-1995), and (iii) third phase (1996-2000). Several trials and projects failed during their inception or due to war.

Only, biggest progress was seen in pharmaceutical industry where 43 pharmacies in Sarajevo centrally connected for receipt collection and analysis. It took three years of testing and Izet Masic finally defended it in his Master’s thesis [1] but the later planned activities got interrupted during and after war (1992-1995) and due to lack of funds. A need was there to introduce health informatics in medical profession to train medical specialists and physicians to be able to give quality services in healthcare using technologies that have quantitative and qualitative growth in diagnostics and therapy. In 1985, Professor Izet Masic launched a separate course “Informatics and Economics in Health” of 30 hours’ duration and some of these postgraduate students became MSc and PhDs in this subject who were then able to offer their services in B&H universities and abroad. After the war (1992-1995), the B&H went through a very tough time with lack of electricity, gas, water supply and food. And, during these circumstances Society for Medical Informatics carried out eight scientific and professional events where 500 papers were presented and published in the proceedings. That was a miracle.

From then onwards several proceedings and congress participations were done. Finally in 1993, SMI BiH launched the professional and scientific journal Acta Informatica Medica (AIM) and it is being published continually since then [1]. In Bosnia and Herzegovina, there is also a library system BiH-

SBMNI been established during 1984 to 1990, by a group of medical librarians and informatics professionals in BiH. Currently, the interest is in establishing higher education in the field of Medical Informatics and the question revolves around Cathedra for Medical Informatics, University of Sarajevo. Several surveys have been carried out to analyze the level of higher education in medical informatics and future development required.

From then onwards, the healthcare informatics is rapidly adopting current trends of cloud computing to store big data coming in from miscellaneous platforms; hospitals, social networks, IoT and other wearable devices, etc. The study in [2] led researchers community to form consensus for adopting cloud computing in hospitals in a strategical manner as part of smart city project led by ICT [3] in Bandung city of Indonesia [4].

Currently, the use of ICT in hospitals [3] for playing strategic role is a viewpoint that is being worked out for rightfully implementing healthcare analytics in the field. Exploration through exploitation is carried out extensively to enhance hospital performance in administration as well as services sector; prediction, diagnosis, treatment, etc.

The need is realized to conquer these problems through computation [5]. The use of computation specially in medical diagnostics is a complex task. Till now to expect a complete diagnostic system is understood as unrealistic. But, no matter how much complex it is advances are being made using artificial intelligence (AI) techniques. Computers have advantage over humans as they do not get fatigued or bored. Computers update themselves within seconds and are rather economical. If automated diagnostic system is made such that it would take care of routine clinical tasks in which patients are not too sick, then doctors would be free to focus on serious patients and complex cases. For analysing complex medical data, the use of artificial intelligence is known to be more capable [6]. The way artificial intelligence exploit and relate the complex datasets giving it a meaning to predict, diagnose and treat the particular disease in a clinical setup. Several artificial intelligence techniques with significant clinical applications are reviewed. Its usefulness is explored in almost all fields of medicine. It was found that artificial neural networks were the most used technique while other analytical tools were also used and those included fuzzy expert systems, hybrid intelligent systems, and evolutionary computation to support healthcare workers in their duties, and assisting in tasks of manipulation with data and knowledge. It was concluded that artificial intelligent techniques have solution to almost all fields of medicine but much more trials are required in a carefully designed clinical setup before these techniques are utilized in real world healthcare scenarios. Artificial intelligence (AI) is termed to be part of science and engineering known to exhibit computational understanding to be said as behaving intelligently through the creation of such artifacts that form the stimulus in it. Alan Turing (1950) explained intelligent behaviour [6] of a computer that could act as any human for cognitive tasks applying logics (right thinking) and this theory was later termed as ‘Turing Test’. The application of AI techniques in the field of surgery was first experimented in 1976 by Gunn, exploring the diagnosing of abdominal pain with computer analysis. The challenge lies in collecting, analysing and applying all the medical knowledge for solving complex medical problems. Medical AI relates to developing AI programs such that they help in forming the diagnosis, making therapeutic decisions and predicting the outcome. Artificial Neural Networks (ANN) on the basis of their recognition for classifying and pattern identification have been known to be widely used for solving several clinical problems.

This paper is to highlight important factors and criteria required for establishing successful and trusted means of automated medical diagnostics through healthcare analytics. The challenge lies in development of automated medical diagnostics to reap its long term benefits to human kind is enough of motivation to dig deep into this domain.

II. Literature Review

Previous research confirms [3], [7], the successful application of analytics is found to be effective after significant degree of digitization accomplished. It was also recognized that the potential of analytics in clinical domain is stronger than in the administrative domain. Specific to automate medical diagnosis, the artificial neural networks (ANN) approach is greatly studied in combination with fuzzy approach [5]. There are some most common risks and precautions associated when forming a medical diagnosis [5]:

- To reach a well-established diagnosis a physician is required to be well versed with some very experienced cases and that experience does not come through completing academics but after lots of experience in a specialized field or disease.
- In case of new or rare disease even the experienced physicians feel to be at the same level as an entry level doctor.
- Humans are good at observing patterns or objects but fail when need to find the associations or probabilities for their observations. Here computer statistics helps.
- Quality in diagnosis relates to physician’s talent and years of experience.
- Doctor’s performance gets affected with the emotions or fatigue.
- To train the doctors in specialties is expensive as well as a long procedure.
- Medical field is always evolving with new treatments and new diseases are coming up with time. It is not easy for a physician to keep abreast with so much change and new trends in medicine.

Currently, 14 hospitals in Italy were observed for their activities and users’ involvement in the use, adoption, and improvement of ICT infrastructure and services based solutions [3]. The prime data source was 107 semi-structured interviews of level C and other hospital informants in a time period of 3 years. There are three paths taken for exploration and exploitation paradox management for enhancing hospital performance: (i) digitization of the assets utilized within hospitals, (ii) ICT-based integration among healthcare stakeholders, and (iii) the disruption of clinical and administrative decision making through the use of analytics, coping the conflicting demands attached to medical sector. Analysing various case studies, the possible wastage of energies is saved in ICT-based innovation and prioritization of possible paths is achieved when moving towards the management of exploration and exploitation paradox. This research [3] divides its analysis activities in three domains: (i) ICT introduction process, (ii) hospital, and (iii) healthcare system within a hospital. The analysis through different case studies produces evidences to demonstrate the theory in Fig. 1 [3]. As shown in Fig. 1 [3], the three complimentary ICT-based paths maintaining a balanced exploratory and exploitative stimuli act in a given approach:

i. Digitization of assets utilized within hospitals ((1)—(2))

ii. ICT-based integration among healthcare stakeholders ((2)—(3)—(5))

iii. The use of analytics for disruption of clinical and administrative decision making ((2)—(4)—(5))

The three paths are successful in managing the exploration-exploitation paradox for short as well as long term. It is found that their overall effectiveness is felt stronger in the clinical domain rather the administration domain. The other limitation observed is that it is not easy to distinguish and separate these interlinked forces. The whole operationalization has to keep into consideration all the pros and cons and there are many factors attached to its success.

Disease outbreaks are happening all over and at all times. Computational prediction, identification, confirmation and responsiveness to these diseases is important as well. Therefore, Predictive Analytical Decision Support System (PADSS) [8] integrates in a cloud based healthcare platform that is Message Oriented Middleware (MOM). It connects healthcare organizations to share patients' data using a customized Health Level Seven (HL7) platform having Fast Healthcare Interoperability Resources (FHIR) specification.

Considering chronic kidney disease (CKD) [9], accurate prediction with time is important for lowering cost and mortality rate. Adaptive Neurofuzzy Inference System (ANFIS) is proposed that uses real clinical data of 10 years for newly diagnosed patients with CKD to predict renal failure time frame. It is deemed as highly successful measure to predict GFR variations in long future periods within uncertain body condition and dynamic nature of CKD progression. The limitation of this experiment was only that urine protein was missing as variable input for predicting GFR in 6 to 18 months' evaluation time.

For detection in medical imaging [10] is also required in healthcare practice and Automated Computer-Aided Detection (CADe) tool is there. There is a high rate of false positives (FP) as well that would add to cost for achieving success (high true positive rate) in highly sensitive cases per patient using state-of-the-art methodologies. But for high false positive or false negative rate that would add to the low sensitivity the application of CADe is not seen in the clinical practice. An updated multi-tiered hierarchical CADe system improved in performance and was named as deep convolutional neural networks (ConvNets). ConvNets is open for advancements as it has been commercially tried on natural images as well as biomedical applications for detection of mitosis in digital pathology. It learns from supervised training to detect features from images through two cascaded layers for filtration using convolutional filters. It can be applied to detect from two and two and a half dimensional observations.

Usage of ConvNets through multiple 2D, 2.5D and 3D representations on CADe problems proves its success by avoiding overfitting analogous data. ConvNets in applications build better accurate classifiers for CADe systems pruning FPs maintaining high sensitivity recalls. Even with the threshold value of 95% sensitivity with 1FP per patient it is not found optimal for colonic polyp CADe System for polyp sizes between 6mm to 9mm. Better results are observed with the polyp sizes larger than 10mm.

Diagnostic accuracy is determined by applying statistical measures in medicine [11] in the context of multi-category classification.

To determine the success of data fuzzification an experimental setup was created to analyse the medical diagnosis done by physicians and automate it with machine implementable format. Eight different diseases were selected for extracting symptoms from few hundreds of cases and MLP Neural Networks was applied [5]. Later results were discussed to conclude that effective symptoms selection for data fuzzification using neural networks could lead to automated medical diagnosis system. A diagnostic procedure is work of an art of specialized doctors and physicians. It starts from patient's complaints and discussion with doctor that leads to perform some tests and examinations and on basis of results the patient's status is judged and diagnosis is formed. Then the possible treatment is prescribed. Patient remains under observation for some time where the whole diagnostic procedure is repeated and refined or even rejected if needed. We all are aware of the complexity of forming a medical diagnosis as even the profession requires twice the study than other professions. There are diversified symptoms history that is caused for diversified reasons. All these causes have to be included in the patient history.

To propose [5] a medical diagnostic system several interviews were conducted of expert doctors getting some diagnostic flow diagrams of various diseases and associated list of symptoms. The dataset was created such that hundreds of patients were tested against 11 symptoms (features) and 9 diseases (classes). First eight classes were associated with specific illnesses and the ninth one was defined for normal/healthy person. Multilayer perception (MLP) neural network was used with application of back propagation GDR training algorithms. The simulation was developed on MATLAB with NETLAB toolbox. A three-layer feed forward perceptron was kept to keep the structure simple and focus on the hidden nodes with the training iterations as variable parameters. The performance of classifier was tested while changing the parameters. Also, feature fuzzification rule [5] on the accuracy of diagnosis was investigated. With k-folding scheme, the lack of dataset was overcome to give better accuracy for validation. So, the training procedure was repeated k=5 times with 80% as training dataset and 20% as testing. The mean was taken for all the outcomes of 5 tests. 88.5% best accuracy was achieved with 30 nodes at the hidden layer. Then, membership-based fuzzification scheme [5] was applied to the dataset converting it to fuzzified set of symptoms. A linear membership function was selected for each symptom with experts' consultation. From three to five linguistic variables were linked to each symptom and classification tests were repeated. Maximum performance was achieved with diagnostic accuracy of 97.5%.

ANNs an exploration of Baxt is also found a successful technique in clinical domain [6] for diagnostics. He came up with a neural network model to diagnose acute myocardial infarction accurately validating his work later with similar accuracy. ProstAsure Index is a classification algorithm extracted from ANN that classifies prostates as benign or malignant. This model gave the diagnostic accuracy of 90%, sensitivity of 81% and 92% specificity. Other relevant surgical diagnostics with application of ANNs are appendicitis and abdominal pain, glaucoma, retained common bile stones and back pain. PAPNET [6] is another computerized screening application based on ANN that assists cytologist in cervical screening and this application was commercially available. Thyroid, breast, oral epithelial, gastric, urothelial cells, peritoneal effusion cytology and pleural enjoyed varied level of diagnostic accuracy with application of ANNs. In the field of radiology, the inputs to ANNs are both human observations and digitized images. ANNs are good at analysing plain radiographs, CT, ultrasound, radioisotope and MRI scans. Various wave forms like ECGs, EEG, EMG, and Doppler ultrasound as well as hemodynamic patterns are well interpreted using ANNs pattern recognition ability. Correct prognosis is also very important to carry out appropriate treatment strategy and follow up. ANNs exploiting non-linear relations between variables are well suited to analyze complex cancer data and it can predict survival of breast and colorectal cancer patients better than colorectal surgeons. ANNs have also been applied to predict outcomes of lung and prostate cancer.

Then there is another AI domain 'fuzzy logic' [6]. It is the science of thinking, reasoning, and inference been applied in real world phenomenon of varying degrees. It sees beyond black and white into the varying shades of grey. Medicine is known to be a continuous field and the data is imprecise and thus fuzzy logic applies to it. Fuzzy expert systems reach to the conclusion through its 'if-then' structure modelling. Fuzzy logic was found to be better than multiple logistic regression analysis for diagnosing the patients with lung cancer having tumour. Then, fuzzy logic has been applied to acute leukaemia, and breast and pancreatic cancer diagnosis. It has also been applied to capture ultrasound images of breast and ultrasound, and CT scan images of liver lesions and MRI images of brain tumours. It is also being used to predict survival of patients with breast cancer. Fuzzy controllers are there for administration of vasodilators that controls blood pressure, and anesthetics in the operating room.

Evolutionary computation is yet another AI domain that mimic the natural selection and survival of the fittest mechanism in natural world to solve the problems and its most widely used form in medical field is 'genetic algorithms' [6]. Genetic Algorithms are applied to reach diagnosis, prognosis, processing signals and analyse medical images, and plan and schedule. They are used to predict outcomes in critically ill patients, melanoma, lung cancer and response to warfarin. They are also used to analyse mammographic micro calcification, MRI segmentation of brain tumours for measuring efficacy of treatment strategies, 2-D images to diagnose malignant melanomas.

Then finally there is hybrid intelligent systems [6] combining the strengths of ANNs, fuzzy logic, and evolutionary computation.

In spite of many different AI techniques [6] that can be readily used for solving several clinical problems there is hindrance in its acceptance by clinicians who would mostly reject the biochemical results produced by autoanalyzer or images resulted from magnetic resonance imaging. So the obligation lies at the researcher's end to authenticate and validate its successful application in real clinical setup. There is no doubt that this future technology resulting in 'medical intelligence' would add to the ability of future clinician. It is concluded that for managing the exploration-exploitation paradox for short as well as long term the overall effectiveness is felt stronger in the clinical domain rather the administration domain. The other limitation observed is that it is not easy to distinguish and separate the three interlinked forces identified in Fig. 1 [3]. The whole operationalization has to keep into consideration all the pros and cons and there are many factors attached to its success. It is also observed that there is no universal solution till now for healthcare diagnostics meeting a standard protocol.

III. Analysis of Previous Diagnostics Systems

For managing the exploration-exploitation paradox for short as well as long term, it is found the overall effectiveness is felt stronger in the clinical domain rather the administration domain. The other limitation observed is that it is not easy to distinguish and separate the three interlinked forces identified in Fig. 1 [3]. The whole operationalization has to keep into consideration all the pros and cons and there are many factors attached to its success. The successful application of PADSS [8] is a strong motivation to take it to the next step in healthcare towards medical diagnostics in compliance with HL7 standards. It also demonstrates the successful utilization of Google Cloud Platform with Google BigQuery where there is need to make SQL queries more efficient by tuning. While evaluating of various healthcare analytics platforms it is visualized that mostly artificial neural networks or fuzzy logic algorithms form the basis of several proposed techniques and recent research has included deep learning algorithms as deep convolutional neural networks ConvNets on CADe problems proved its success.

Further it is observed that by selecting symptoms of eight different diseases a dataset containing few hundreds of cases was created and Multilayer Perception (MLP) Neural Networks was applied [5]. Later results were discussed to conclude that effective symptoms selection for data fuzzification using neural networks could lead to automated medical diagnosis system. Simulation was developed on MATLAB with NETLAB toolbox. A three-layer feed forward perceptron was kept to keep the structure simple and focus on the hidden nodes with the training iterations as variable parameters.

Recently, evolutionary computation is another AI domain exploited that mimic the natural selection and survival of the fittest mechanism in natural world to solve the problems and its most widely used form in medical field is ‘genetic algorithms’ [6]. Genetic Algorithms are applied to reach diagnosis, prognosis, processing signals and analyze medical images, and plan and schedule. Then finally there is hybrid intelligent systems [6] combining the strengths of ANNs, fuzzy logic, and evolutionary computation.

These previously proposed healthcare analytics systems and techniques have been validated against the accuracy achieved. There are different metrics for evaluating accuracy level. It is seen that sensitivity of CADe [10] is improved from 57% to 70%, from 43% to 77%, and 58% to 75%. Statistical model based on HUM analysis [11] over biomarkers showed 65% accurate classification. ANFIS [9] is evaluated by accessing Normalized Mean Absolute Error that is lower than 5%. PAPNET [6] is said to have diagnostic accuracy of 90%, sensitivity of 81%, and 92% specificity. In MLP neural network [5] 88.5% best accuracy was achieved with k-modelling and maximum performance was achieved with diagnostic accuracy of 97.5% with membership-based fuzzification scheme. Further accuracy results are depicted in Table I.

TABLE I. TAXONOMY MATRIX FOR EVALUATION OF HEALTHCARE ANALYTICS

Authors	Diseases	Proposed Technique/System	Algorithms/Platforms	Tested	Accuracy Level
Gastaldi and Corso, 2015	X	Management of exploration and exploitation paradox in 14 Italy Hospitals	X	Case study using 107 semi-structured interviews	X
Ramesh et al., 2004	Acute myocardial infarction and Surgical Diagnostics including appendicitis and abdominal pain, glaucoma, retained common bile stones and back pain	Neural network model	Artificial Neural Networks (ANN), ProstAsure Index	Simulation	Diagnostic accuracy of 90%, sensitivity of 81%, and 92% specificity
Ramesh et al., 2004	Thyroid, breast, oral epithelial, gastric, urothelial cells, peritoneal effusion cytology and pleural	Neural network model	ANN	Simulation	Diagnostic accuracy of 90%, sensitivity of 81%, and 92% specificity
Ramesh et al., 2004	Cervical screening	PAPNET	ANN	Experimented	Diagnostic accuracy of 90%, sensitivity of 81%, and 92% specificity
Alan Turing, 1950	X	Turing Test	Cognitive behavior	Theory	X
Gunn, 1976	Diagnosing of abdominal pain	Computational analysis	AI techniques	Applied in field of surgery	X
Neto & Ferraz, 2016	X	PADSS integrated in cloud based healthcare platform	Google Cloud Data Store	Prototype	Compliant to HL7 standard
Boukenze et al., 2016	Chronic kidney diseases (CKD)	Predicts the chronic kidney diseases (CKD)	C4.5 learning algorithm using classification model with ORANGE or TANAGRA	Simulation	63%

Norouzi J. et al., 2016	Chronic kidney diseases (CKD)	Adaptive Neurofuzzy Inference System (ANFIS)	Neurofuzzy	Simulation	Normalized Mean Absolute Error lower than 5%
Roth et al., 2015	Lymph nodes, sclerotic metastases and colonic polyps	Computer-Aided Detection (CADe)	Deep convolutional neural networks (ConvNets)	Experimentation on three datasets	Improves the sensitivity of CADe from 57% to 70%, from 43% to 77%, and 58% to 75%
Li J., 2013	Synovitis and leukemia	Statistical Accuracy Model	(HUM) based analysis	1000 simulations using Monte Carlo	65% accurately classified
Moein et al., 2009	Eight diseases classified with ninth one a healthy person	Medical diagnostic system	Multilayer perception (MLP) neural network	Simulation was developed on MATLAB with NETLAB toolbox	88.5% best accuracy was achieved with k-modeling , maximum performance was achieved with diagnostic accuracy of 97.5% with membership-based fuzzification scheme
Izet Masic, 1977	X	First Local Health Information System (LHIS) in Bosnia and Herzegovina	X	It took 3 years to process and test health data by a specially designed software package ARCHIVE on a Personal Sinclair QL computer that was later redesigned in Clipper, DBASE III, and FoxBASE	Realization of Healthcare Informatics was formed
Sordo, 2002	Screening of Pap (cervical) smears	PAPNET	ANN	In cardiovascular domain, neural networks were trained using ECG measurements of 1120 patients suffering from heart disease and 10,452 normal people without any history of heart attack	15.5% more sensitive than the interpretation program and 10.5% more sensitive than the cardiologist for diagnosing abnormalities
Sordo, 2002	**Predict breast cancer**	**Entropy Maximization Network (EMN)**	**ANN**	**Experimented**	**Most accurate in predicting five-year survival of 25 cases**
Sordo, 2002	Cancer	Drug Development Solution	ANN	Simulation	Correctly classified 91.5% from anticancer agents (drugs)

A. Comparison of Analytical Tools and Techniques used in HealthCare

In Fig. 1, the taxonomy, clearly demonstrates that there is an exemplary work already been done in developing and using various healthcare analytics techniques for prediction, detection and diagnosis of various chronic diseases.

There are various artificial intelligence and machine learning algorithms been employed that contributed in emergence of different clinical systems. There are some specific tools and platforms used that are mostly believed to be state-of-the-art technologies including google cloud platform, MATLAB, Weka, and design methodology like DFDs. But in most cases these systems have been proposed and only limited number of systems have been commercialized yet. Major example of commercially used systems is Papnet that is for diagnostics of cervical cancer. This is due to the limitation posed by the level of accuracy achieved and in acceptance from clinical domain.

If we carefully elaborate the picture we see that there have been systems developed for focused areas like; prediction, detection and diagnosis with respect to the particular disease mostly chronic. Narrowing our focus, we see that PADSS is a prediction system proposed for large scale patient data that was distinguishable by the standard nomenclature understood by HL7 using FHIR specification. It became part of the Global Public Health Intelligence Network (GPHIN) project that was supported by WHO and Centers for Disease Control. This is the most unique and large scale project that underwent in medical field. The platform used was Google cloud platform and Google bigquery. Its basis was previous system PREVENT that used XML-based syntax and object-oriented approach for which it was criticized.

Then for detection there is hierarchical two-tiered CADe system been discussed here that uses ConvNets a deep convolutional neural networks algorithm to detect lymph nodes and colonic polyps. Analogous data from three datasets is represented in 2D, 2.5D and 3D representation.

For diagnosis there is system that is prominent is Papnet which is commercially available. For other diagnostic measures there are fuzzy expert systems and hybrid intelligent systems that use various state-of-the-art algorithms and tools.

In 2009, there was conducted an experiment for proposing a diagnostic system that included eight different diseases. The algorithm used was Multilayer Perception (MLP) neural networks within MATLAB integrated with NETLAB. The accuracy achieved was 97%.

Accuracy is determined in all these systems through measuring sensitivity or using statistical accuracy tools like HUM or biomarkers. The most authentic accuracy measure would be to compare the system with a known standard or the system should be in compliance with a standard as in PADSS.

B. Application of Healthcare Analytics on Data of Diabetes and Underlying Diseases using Various Learning Algorithms

Taking the comparison done in Table II of various healthcare analytics systems we aim to provide universal diagnostic solution for diabetics and patients forming risk or suffering from underlying other diseases of liver cirrhosis, kidney disease, cardiovascular complications, etc. [12] already refers to a detailed systematic review on application of data mining and machine learning techniques to perform analytics on diabetes mellitus (DM) for prediction and diagnosis, finding complications associated to it and its linkage with the genetic background and environment to assist betterment in healthcare management.

It is known that the patient diagnosed with diabetes has to be very careful in keeping the blood sugar controlled otherwise there are chances that long-term diabetes may develop certain complications in form of some known chronic diseases (mayoclinic.org) mainly; cardiovascular disease, brain stroke, nerve damage (neuropathy), kidney damage (nephropathy), eye damage (retinopathy), foot damage, worse skin conditions, hearing impairment, alzheimer’s disease, etc.

IV. Aims and Objectives

LHS [17] is in its development phase with a lot to compute and found extendable. Based on the thematic taxonomy in Tables I, II and III the optimal diagnostics system for diabetics patients those are threatened by underlying diseases is yet a problem to be achieved with many challenges.

1) Data and Simulation Model for a universal diagnostic system primarily for diabetes leading to other chronic diseases with the perspective of efficiency and reliability with maximum accuracy is required.

2) Providing a test bed with automated real time patient data to input into diagnostics system is an important concern for validation through simulation.

3) diagnostics system for diabetic and its underlying diseases is not present in current scenario and has our attention.

4) Finding an optimal big data healthcare analytics technique for diagnostics in compliance with HL7 standard still remains a challenge.

TABLE II. Comparison of Various Healthcare Analytics Systems

	Platform	Model	Algorithm	Tool	Standard
PADSS	**Google Cloud**	**OLAP**	**X**	**BigQuery**	**HL7**
Predictive Analytics	**X**	**Classification Model**	**C4.5**	**Weka**	**X**
ANFIS	**X**	**ANFIS prediction Model**	**Genfis3**	**MATLAB**	**X**
Hierarchical two-tiered CADe	**CADe**	**X**	**Deep ConvNets**	**X**	**X**
Medical Diagnostic System	**X**	**X**	**MLP neural networks**	**MATLAB with NETLAB**	**X**
PAPNET	**X**	**Neural network model**	**ANN**	**X**	**X**

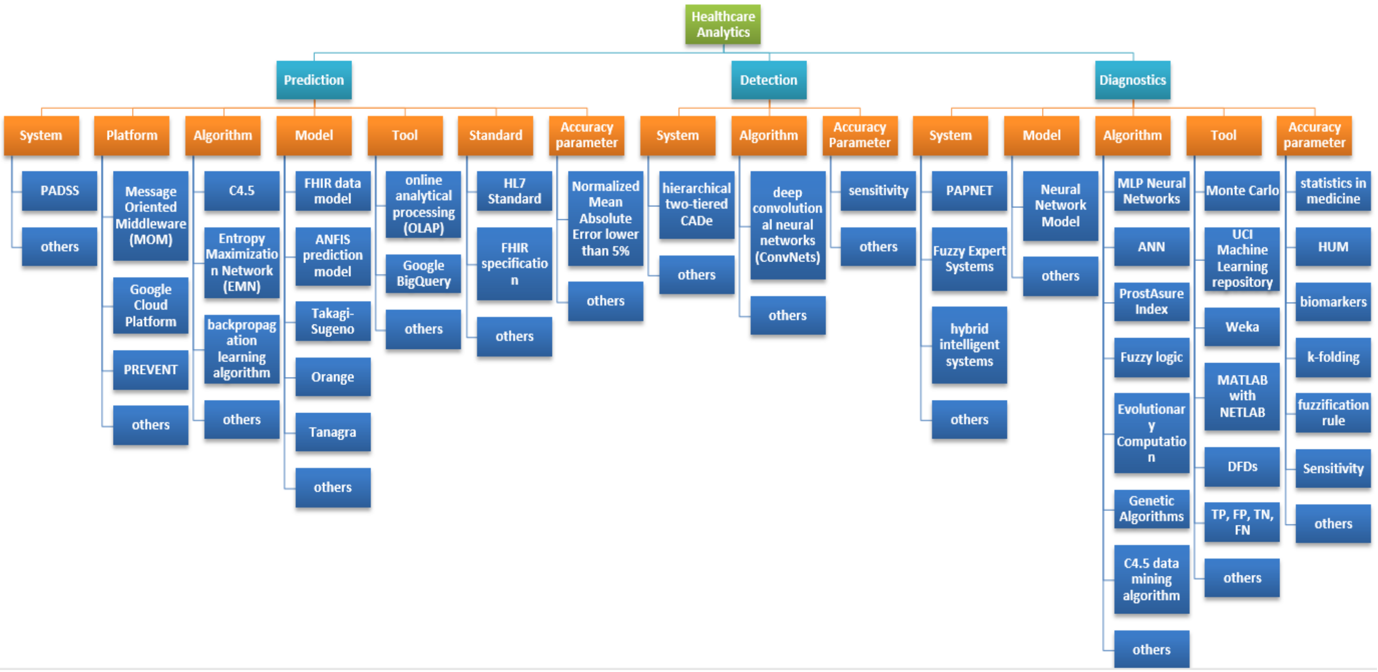


Fig. 1. Detailed taxonomy of healthcare analytics.

TABLE III. HEALTHCARE ANALYTICS APPLIED FOR DIAGNOSTICS OF DIABETES AND LINKED DISEASES

Citation	Disease	Application	Algorithm/System	Dataset	Tool/s	Accuracy
[13]	Diabetes	Diagnosis	Hybrid System (ANN & FNN)	Pima Indians Diabetes (taken from UCI machine learning repository)	Matlab 7.0.0	84.2%
[13]	Heart Disease	Diagnosis	Hybrid System (ANN & FNN)	Cleveland Heart Disease	Matlab 7.0.0	86.8% but IncNet outperformed it with 90% accuracy
[14]	Diabetes, Liver, Breast Cancer, Hepatitis, etc.	Diagnosis	Hybrid Intelligent System (optimizing SVM & MLP using GSA, PSO & FA evolutionary algorithms)	11 datasets of various diseases obtained from UCI (that included Pima Indian Diabetics & BUPA Liver Disease)	Weka	Optimized MLP technique with Firefly Algorithm (FMLP) in proposed Hybrid System was found better
[15]	Diabetes, heart disease, breast cancer, iris, skin cancer etc.	Compression of Training Set	Comparison of instance selection algorithms	Wisconsin breast cancer, Cleveland heart disease, appendicitis, Iris, wine, Pima Indians diabetes, Skin Cancer	Tested for classification algorithms kNN, SVM, SSV, NRBF, FSM & IncNet	Prototype selection algorithms Explore, RMHC, MC1, LVQ, DROP2-4 and DEL, are the most effective. RMHC and LVQ algorithms also finished tests with high level of accuracy on the unseen data. ENN algorithm came at the top, especially for KNN & NRBF and ENRBF. These algorithm may stabilize the learning process, except the SSV or SVM models.
[16]	Clinical Medicine (Psychological)	Approximation and classification (Diagnosis)	IncNet is integrated with Extended Kalman Filter learning algorithm and bi-central transfer functions while rotating constant values in multidimensional spaces	Minnesota Multiphasic Personality Inventory (MMPI) test	Not mentioned	IncNet outperforms C4.5 & FSM Rule Opt. with accuracy of 99.23%

V. Proposed Solution

This paper has started off with proposing a solution to our first problem of defining a data model and validate a proposed simulation model in [18] based on it that would give us a testbed in future to propose detailed solution.

A. Universal Data Model

As mentioned previously researchers are interested to propose Healthcare Analytics Platform for Diagnostics of Diabetic and underlying other chronic diseases like; Liver Cirrhosis. Comprehensible Knowledge Model for Cancer Treatment (CKM-CT) proposed in [19] greatly contributes to visualize the limitations posed when defining a universal data model. In [19] researchers acknowledge the close interaction and support of clinicians with technologists. Experts in knowledge domain are highly recognized to assist in retrieving raw patient data in form of EHRs and transforming it into structured form using CRT algorithm integrated in CKM-CT model. CKM-CT is said to be generalized with some limitations of availability of latest technological facilities and infrastructure. Our aim is to use machine learning techniques that would assist to apply deep learning over the universal diagnostic data model that would become part of context aware SmartHealth cloud platform derived for our problem. Our confidence in proposed model is based on study of various healthcare analytics platforms earlier been used in cloud context as in PADSS mentioned in Table II.

B. Proposed Data Modelling Methodology

Our research would be an extended version to connect with LHS in future that would be in compliance with latest HL7 coding standards for coming up with medical diagnostics as in PADSS [8] that is for prediction of diseases integrated in cloud based healthcare platform. Further when setting up a benchmark for our system Adaptive Neurofuzzy Inference System (ANFIS) is seen as one of the successful approach for diagnosis of chronic kidney diseases [9]. It is focused over a single disease where Multilayer perception (MLP) neural network [5] is tested over eight diseases with 97% accuracy level achieved. Our measure of accuracy would be based on HUM using biomarkers or a better approach than AUC. The major simulation platform for SmartHealth used would be AnyLogic as modelled in [18] and Google Cloud as in PADSS [8]. The system may further be standardized if the simulation platform is HIPAA enabled or the dataset is following or is convertible to HIPAA HL7 coding standard.

By standardized we mean the specification given by Fast Healthcare Interoperability Resources (FHIR) regulating HIPAA HL7 standard to be followed. But to converge the heterogeneous healthcare big data coming in from different platforms it has to be structured and modeled to fit FHIR data modeling guidelines. After being transformed in structured diagnostic data model if still it fails to map with FHIR specified data model fields then it may be modified to fit our built model (shown in Fig. 2). Later these modeled patients' diagnostic profiles would be clustered to train diagnostic learning model using analytics.

EHR is relatively structured form of health data. Therefore, it may be used to train the diagnostic learning model as in [20]. We would use it to cluster patient's data in three categories: (i) diabetic, (ii) prediabetic, and (iii) having family history (as shown in Fig. 3). Further nested clusters would be found for various other variables like; age, and diabetes type, etc.

Thus, learning model (Fig. 3) would be trained to find different sequences in symptoms and lab test results through finding correlations in one diagnosis or multiple diagnosis characteristics [21]-[23] of a patient in multiple visits. This diagnostics and history of diagnosis would also reach prognosis of co-occurrence [24], [25] of other chronic underlying diseases.

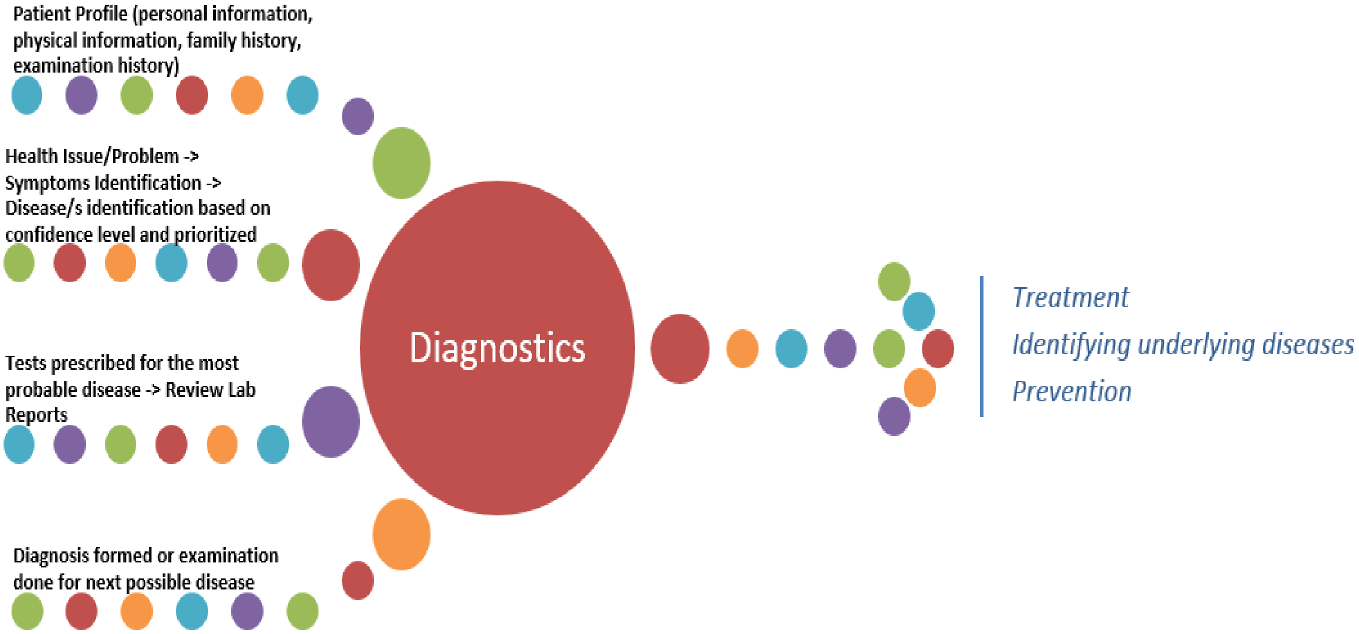

Patient Profile (personal information, physical information, family history, examination history)
Health Issue/Problem -> Symptoms Identification -> Disease/s identification based on confidence level and prioritized
Tests prescribed for the most probable disease -> Review Lab Reports
Diagnosis formed or examination done for next possible disease
Diagnostics
Treatment
Identifying underlying diseases
Prevention

Fig. 2. Universal data model for diagnostics to apply on any disease.

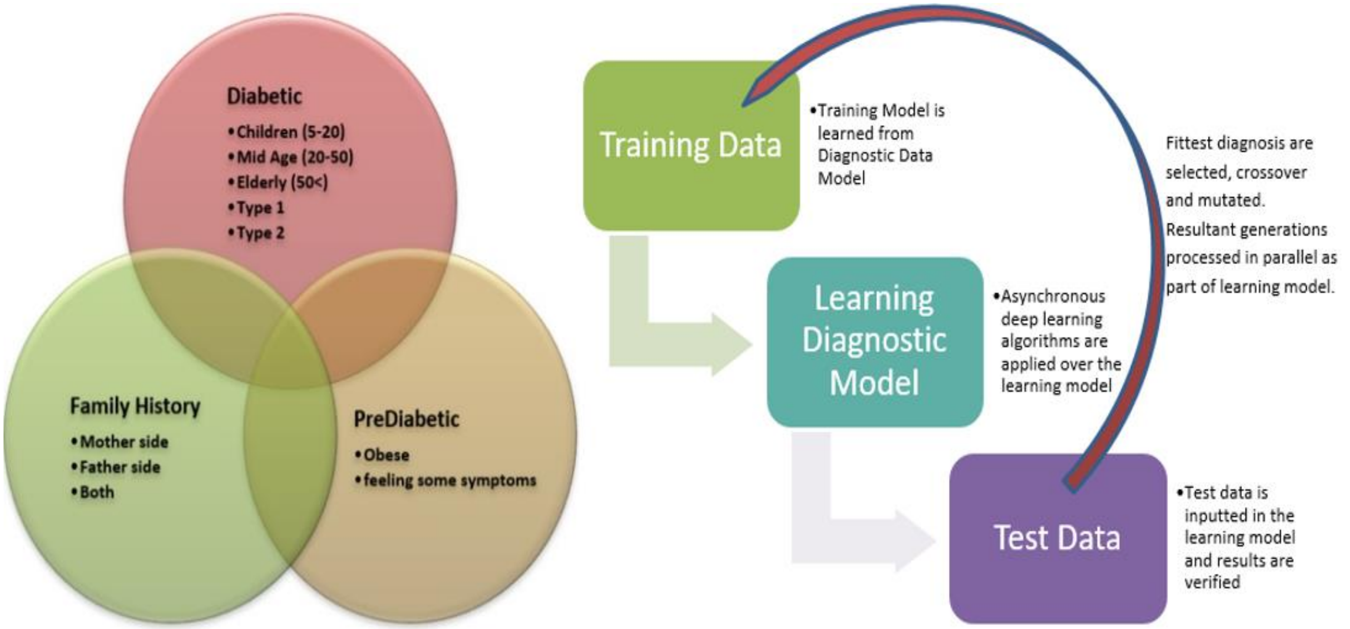

Diabetic
•Children (5-20)
•Mid Age (20-50)
•Elderly (50<)
•Type 1
•Type 2
Family History
•Mother side
•Father side
•Both
PreDiabetic
•Obese
•feeling some symptoms
Training Data
•Training Model is learned from Diagnostic Data Model
Fittest diagnosis are selected, crossover and mutated. Resultant generations processed in parallel as part of learning model.
Learning Diagnostic Model
•Asynchronous deep learning algorithms are applied over the learning model
Test Data
•Test data is inputted in the learning model and results are verified

Fig. 3. Nested cluster model to develop accurate learning diagnostics model for diabetes.

VI. Evaluating Proposed Solution

Based on the extensive analysis of healthcare analytics platforms (Table I) detailed taxonomy (Fig. 1) is extracted for prediction, detection and diagnostics systems spanning various diseases. Moving forward we focused particularly on systems involving our domain of diabetes and linked diseases (Table III). This analysis lead us to compose our problem statement to construct a universal diagnostic system for diabetes and its underlying diseases that is context aware based on its ability to integrate with the latest technology of IoT embedded with cloud to give personalized service to patients wherever they go. To start with solvation of our problem we were able to devise a generic data model (Fig. 2) for any particular disease and in our case that would take patients data related to diabetes or other interlinked diseases. This data would be heterogeneous in nature coming from biosensors and other apps to become part of socially context aware SmartHealth cloud platform. The data would be structured with respect to its nature based on patient profile and examination history, current health issue or symptoms identified, the list of most probable diseases that are associated to a particular patient, any tests underwent and reports, etc. If the diagnosis is reached then patient gets the feedback for the next immediate action otherwise he/she is examined for next possible disease. The detailed scenario of how this data is evaluated at different stages using healthcare analytics techniques (as shown in Fig. 3) as part of hybrid distributed simulation environment within AnyLogic is demonstrated in Fig. 3 [18].

Detailed data model abstraction enabled us to build the simulation in detail in [18] and not compromising on abstract level to miss out important details. Our cloud platform considers to be integrated with societal role in future but at the moment is assumed to take input from biosensors and computing the in-formal information gathered by patients forming a patient profile graph as mentioned in [26] and shown in Fig. 1 of [8]. We have already selected our data formation clustering method to be Random Dynamic Subsequent method proposed in [27]. Further when performing machine learning heuristics we consider the proposed deep patient representation using unsupervised deep feature learning method shown in Fig. 2 of [28] to form diagnostics of diabetes and the underlying disease of liver cirrhosis. Still it is kept in supervision of human doctors not to let patients suffer any risk. The metrics to validate the accuracy of our model would be determined using HUM approach with biomarkers or a better accuracy model than AUC.

VII. Conclusion and Future Directions

Based on our study of previous healthcare analytics techniques been utilized for prediction, diagnosis, treatments, and prognosis it is determined that researchers have focused over best approaches with respect to the maximum accuracy level achieved. These approaches properly integrated with the latest techniques would be used for coming up with future standardized medical diagnostics. We are working towards proposing a diagnostic system keeping in view for its commercial universal use by clinicians for various chronic diseases like; diabetes mellitus and its underline diseases particularly liver cirrhosis. The system should be such that it would be visualized for coming up with optimal healthcare analytics technique. In future the system would be enhanced exclusively for treatments or cover the diagnosis of diseases that have been left out due to limited time or limitation of proposed healthcare analytics technique.

The simulation in [18] embedded with proposed data model for SmartHealth cloud platform is found to be technically intelligent and sound in its process flow when compared in [18] to SelfServ platform and Societal Information System for Healthcare simulations [29], [30]. It acquires the smartness of unsupervised learning while at the same time it keeps it under supervision of qualified doctors that would assist in complicated diagnostic case and continuously update the knowledgebase not compromising on the risk associated. This SmartHealth cloud platform is a very comprehensive and complete visualization of our future vision of LHS [17]. Currently, it lacks actual implementation due to limited resources and time. Still we have done considerable effort in paving the way with clear detail of the scenario for making a hybrid distributed cloud in future that is intelligent enough to assist in Complex Event Processing (CEP) in service oriented community (SOC) for health. It is clearly understandable as we explored the challenges of building DSM [31] integrated with big data that forms the bases of our hybrid distributed cloud that actual simulation and implementation of SmartHealth cloud is a complex task. It is estimated to take huge span of time and expert skills giving us a novel machine learning framework to support our SmartHealth cloud platform.

VIII. Recommended Action Plan

Our detailed data model would be in compliance with latest HL7 coding standards for coming up with self-learning medical diagnostics as in PADSS [8] that is for prediction of diseases integrated in cloud based healthcare platform. Further when setting up a benchmark for our system Adaptive Neurofuzzy Inference System (ANFIS) is seen as one of the successful approach for diagnosis of chronic kidney diseases [9]. But, it is focused over a single disease where Multilayer perception (MLP) neural network [5] is tested over eight diseases with 97% accuracy level achieved whereas we would structure the data using sequential and hierarchical clustering to become part of learning diagnostic model in Fig. 3. Our measure of accuracy would be based on HUM using biomarkers or a better approach than AUC. The simulation platform mostly used is MATLAB with NETLAB tool box. Other good platform for real scenario comprising of large scale patients' data would be Google Cloud Platform [8] used for PADSS or AnyLogic for hybrid cloud simulation as evaluated in [18] against NetLogo and other distributed simulation tools discussed.

References

[1] Masic I., & Novo A., "History of Medical Informatics in Bosnia and Herzegovina", MedArh 2006, [Medical Faculty, University of Sarajevo]

[2] Shafqat S., Awan M. D., & Javaid Q., "Evaluating Cloud Computing for Futuristic Development", (2013), International Journal of Computer Applications, Volume 61-No.6, [SZABIST, Pakistan]

[3] Gastaldi L., & Corso M., "Managing ICT to Solve the Exploration-Exploitation Paradox in Healthcare", (2015), [Italy]

[4] Kusnandar T., & Surendro K., "Adoption model of hospital information system based on cloud computing: Case study on hospitals in Bandung city", (2013), International Conference on ICT for Smart Society, [STMIK Mardira Bandung, Indonesia]

[5] Moein S., Moallem P., & Monadjemi A., "A novel fuzzy-neural based medical diagnosis system", (2009), [University of Isfahan, Isfahan, Iran]

[6] Ramesh A., Kambhampati C., Monson J., & Drew P., "Artificial Intelligence in Medicine", (2004), [The Royal College of Surgeons of England]

[7] Shafqat S., Kishwer S., Rasool R. U., Qadir J., Amjad T., and Ahmad H. F., "Big data analytics enhanced healthcare systems: a review," The Journal of Supercomputing, pp. 1–46, 2018.

[8] Neto S., & Ferraz F., "Disease Surveillance Big Data Platform for Large Scale Event Processing", (2016), ANCS'16, March 17–18, ACM, [Recife Center for Advanced Studies and Systems, Brazil]

[9] Norouzi J. et al., "Predicting Renal Failure Progression in Chronic Kidney Disease Using Integrated Intelligent Fuzzy Expert System", (2016), Computational and Mathematical Methods in Medicine Volume 2016, Hindawi Publication, [University of Technology and University of Medical Sciences, Iran]

[10] Roth H. R., Lu L., Liu J., Yao J., Seff A., Cherry K., Kim L., & Summers R. M., "Improving Computer-aided Detection using Convolutional Neural Networks and Random View Aggregation", (2015), [National Institutes of Health Clinical Center, USA]

[11] Li Jialiang, "Multicategory reclassification statistics for assessing improvements in diagnostic accuracy", (2013), [National University of Singapore]

[12] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I., "Machine Learning and Data Mining Methods in Diabetes Research", (2017), Computational and Structural Biotechnology Journal

[13] Kahramanli H. & Allahverdi N., "Design of a Hybrid System for the Diabetes and Heart Diseases", Expert Systems with Applications 35 (2008) 82–89, [Selcuk University, Konya, Turkey]

[14] Nalluri et al., "Hybrid Disease Diagnosis Using Multiobjective Optimization with Evolutionary Parameter Optimization", Journal of Healthcare Engineering Volume 2017, [SASTRA University, Thanjavur, Tamil Nadu, India]

[15] Grochowski M. & Jankowski N., "Comparison of Instances Selection Algorithms", International Conference on Artificial Intelligence and Soft Computing 2004, LNAI 3070, pp. 580–585, 2004. [Nicholaus Copernicus University, Poland]

[16] Jankowski N., "Approximation and Classification in Medicine with IncNet Neural Networks", Workshop on Machine Learning in Medical Applications, [Norbert Jankowski Department]

[17] Kaggal et al., "Toward a Learning Healthcare System – Knowledge Delivery at the Point of Care Empowered by Big Data and NLP", (2016), Innovations in Clinical Informatics, [Division of Information Management and Analytics, Mayo Clinic, Rochester, MN, USA]

[18] Shafqat, S., Abbasi, A., Amjad, T., & Ahmad, H. F. (2018). SmartHealth simulation representing a hybrid architecture over cloud integrated with IoT: a modular approach. In Future of Information and Communications Conference (FICC).

[19] Afzal et al., "CKM-CT: Comprehensible knowledge model creation for cancer treatment decision making", (2017), Computers in Biology and Medicine, [Kyung Hee University, South Korea and King Faisal University, KSA, Saudi Arabia]

[20] Syahputra M. F., Felicia V., Rahmat R. F., and Budiarto R.. "Scheduling Diet for Diabetes Mellitus Patients using Genetic Algorithm," International Conference on Computing and Applied Informatics 2016. Conf. Series: Journal of Physics: Conf. Series 801 (2017). [IOP Publishing]

[21] Cynthia M. Boyd, Bruce Leff, Jennifer L. Wolff, Qilu Yu, Jing Zhou, Cynthia Rand, Carlos O. Weiss. "Informing Clinical Practice Guideline Development and Implementation: Prevalence of Co-existing Conditions Among Adults with Coronary Heart Disease." Journal of the American Geriartrics Society, 2011.

[22] Heather A. Close, Li-Ching Lee, Christopher N. Kaufmann, Andrew W. Zimmerman. "Co-occurring Conditions and Change in Diagnosis in Autism Spectrum Disorder." American Academy of Pediatrics, 2012.

[23] Konstadina Griva, Nandakumar Mooppil, Eric Khoo, Vanessa Yin Woan Lee, Augustine Wee Cheng Kang, Stanton P Newman. "Improving outcomes in patients with coexisting multimorbid conditions—the development and evaluation of the combined diabetes and renal control trial (C-DIRECT): study protocol." BMJ Open, 2015.

[24] Stephen K Gruschkus, Joseph M Darragh, Michael A Kolodziej, Roy A Beveridge, Michael Forsyth and Carolina Reyes. "Impact of Disease Progression On Healthcare Cost and Resource Utilization Among Follicular NHL Patients Treated within the US Oncology Network." Blood Journal - Vol 114, October 5, 2015.

[25] Zhou Zhou, Paresh Chaudhari, Hongbo Yang, Anna P. Fang, Jing Zhao, Ernest H. Law, Eric Q. Wu, Ruixuan Jiang, Raafat Seifeldin. "Healthcare Resource Use, Costs, and Disease Progression Associated with Diabetic Nephropathy in Adults with Type 2 Diabetes: A Retrospective Observational Study." Diabetes Therapy, 2017: 555-571.

[26] Kulkarni, S. M., & Babu, B. S., “Cloud-Based Patient Profile Analytics System for Monitoring Diabetes Melli-tus”, (2015), IJITR, 228-231

[27] Zhao et al., “Learning from heterogeneous temporal da-ta in electronic health records”, (2017), Journal of Bio-medical Informatics 65 (2017) 105–119, [Stockholm University, Sweden]

[28] Miotto et al., “Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Elec-tronic Health Records”, (2016), Scientific Reports, [Icahn School of Medicine at Mount Sinai, New York, NY, USA]

[29] Florio et al., “Towards a Smarter organization for a Self-Servicing Society”, (2016), ACM DSAI 2016, [Mo-rocco & Belgium]

[30] Du H., Taveter K., and Huhns M. N., “Simulating a So-cietal Information System for Healthcare”, (2012), Pro-ceedings of the Federated Conference on Computer Sci-ence and Information Systems pp. 1239–1246, [Univer-sity of South California, USA]

[31] Marshall et al., “Transforming Healthcare Delivery: In-tegrating Dynamic Simulation Modelling and Big Data in Health Economics and Outcomes Research”, (2015), PharmacoEconomics Springer, [Clinical Engineering Learning Lab, Mayo Clinic Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Rochester, MN, USA]

Detection of Parkinson’s Disease through Acoustic Parameters

[1]Imen Daly, [1]Zied Hajaiej, [2]Ali Gharsallah

[1]Laboratory of Systems and signal Processing (LRSITI) National Engineering School of Tunis (ENIT)
[2]Research Units (CSEHF), Department of Physics, Faculty of Sciences of Tunis, Tunis El Manar University

***Abstract*—Parkinson’s disease is a neurological disorder. It is the second most common disease after Alzheimer’s. Incidence rates for this disease are increasing rapidly with increasing life expectancy. The search for measures to diagnose the disease and monitor symptoms is an important step, despite the fact that it presents a number of challenges. Among the symptoms related to this disease is the disturbances of the voice which particularly occur in a remarkable way called hypokinetic dysarthria which is presented by the poverty of the gesture in all the characteristics of the speech (phonatory, prosodic, articulatory and rhythm). Our goal is to do a study based on voice analysis at the level of the glottis to examine some early parameters measured using the LF model and clinical manifestations to help diagnosis of the disease.**

***Keywords*—Parkinson’s disease; LF model**

I. Introduction

The work presented in this paper focuses on variations and analysis of the parameters of the glottal source in Parkinson's speech. This research can be relevant for different speech signal domains; it announces the characteristics of the speech source of normal speakers and those with Parkinson's disease (PD). It is a chronic neurodegenerative disease with a variety of motor and non-motor symptoms. The second most commonly diagnosed after Alzheimer's disease [4]. It results in the slow and progressive death of a set of neurons of the human brain that have an important role in the control of movement [1]. It is a loss of muscular control and cognitive impairments, known by the trembling of the body and the voice called dysarthria [2]. Age is the simplest factor in this disease, so that the prevalence rate is 0.5% to 1% for the age of 65 to 69 years and 3% for the oldest 80 years [5], [3]. It is therefore preferable to detect it in advance before it is well developed. Although there are other symptoms of this disease, no final biomarker exists or specific measures for diagnosis [10]. The latest research findings indicate that approximately 70-90% of patients with this disease have a form of vocal disability [6] and this index may be one of the best symptoms of this disease. The idea of studying the speech signal and its tremors that either occur in instantaneous frequencies or in a variation in amplitudes is a good index for the diagnosis and monitoring of the progression of this disease. Ideally, such a measure would be non-invasive and could be acquired outside the clinical setting without the need for expert assistance. In this context, several studies have been carried out on acoustic measurements based on jitter and Shimmer and NHR ratios, fundamental frequency [7], intensity and formants [8] or joint model [9]. In our work, we are interested in measuring the spectrum of the glottal source, estimated from the speech signal to identify parameters that behave differently in Parkinson’s and Healthy speech. This study was motivated by all the research that used glottal behavior and characteristics [10]-[13].

The organization of this document and as follows. In the first section, we present the glottal model also the voice data used in the analysis of the sources. The next section describes the experience and the results obtained and finally a conclusion.

II. Methodology

A. Acoustic Modeling of Speech Production

The acoustic theory of speech production presented by [15] is based on the analysis of glottal sources. It is a theory that allows the functional separation of speech production in two parts (source and filter) to improve the understanding of this phenomenon. The filter is assumed linear time invariant (LTI), which means that each short-term speech segment contains constant parameters without any interaction with the glottal source. These hypotheses are clearly simplifications, since the exact physical description of the generation and propagation of sound in the vocal organs leads to a complex set of differential equations [16] in order to facilitate and separate the components of the model and understand their functioning. The acoustic speech signal S (n) is the result of the convolution of the waveform of the glottis source g (n), which presents the volume of the airflow exiting the lung to the vocal cords with an impulse response of the filter leaving the voice path representing the frequencies of resonant formants and the radiation of the lip r (n).

$$S(n) = g(n).v(n).r(n) \tag{1}$$

In the z domain, the model can be defined as:

$$\mathrm{S(Z) = \ G(Z).V(Z).R(Z)} \tag{2}$$

Where R(Z) represents the discrete time radiation impedance and V(Z) the discrete time speech signal of the volume velocity at the glottis and E(Z) the discrete signal enters the time domain. Any loss in the system occurs by radiation on the lips through which has a high-pass filtering effect modeled with a single Zero such that R(Z) = 1-$\boldsymbol{Z}^{-\mathbf{1}}$

1) Glottal source: The production of the speech signal requires the use of several organs, a flow of air coming out of the lungs passes through the larynx where the vibration of the area is produced by the voice piles. The opening between these two stacks is called the Glottis slit or glottis [13], see Fig. 1.

The air pressure exerted on the cells is sufficiently high to separate them and allows the flow of air to flow so that the glottis is released to close. Several factors contribute to movement. The first is the nature of the elastic tissue which forces the vocal folds to return to its initial position, the second is described by the Bernoulli effect [18] (the retro-aspiration effect of the cordial mucosa) and the third by the action of the Decreased Pressure [14].

This cycle is repeated several times periodically, its duration is called the fundamental period noted T0 and its rhythm determines the fundamental frequency of the voice (number of vibrations per second denoted F0 expressed in Hz).

The flow waveform is the result of vibration of the vocal chords from an open position and a closed position that form the two most important phases of the glottal flow impulse during each cycle.

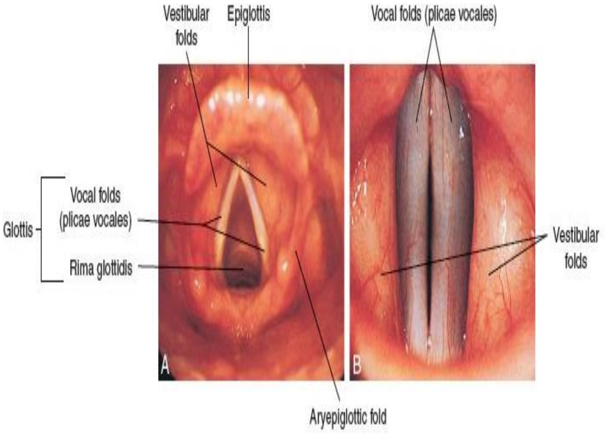


Fig. 1. The Positions of the Open Vocal Cords (A) and Closed (B).

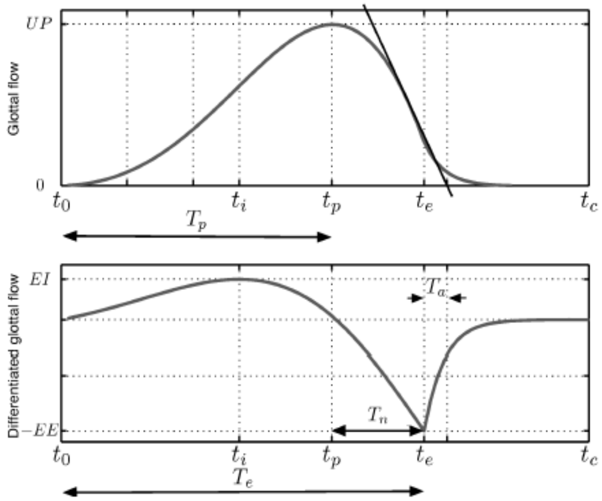


Fig. 2. The different instants and periods of the LF Model.

a) LF Model

Several models exist to determine the characteristics of the glottal flow. We chose to use the Liljencrants fant (LF) model [19] in this document because it gives a good fit in the waveform.

This model can be presented with parameters such as tp, te and ta (see Fig. 2). It is composed in two parts opening and closing and calculated as follows:

$$E(t) = \begin{cases} E_0\, e^{\alpha t}\ \sin w_g \mathrm{t} & 0\ \leq t \leq t_e \\ -\frac{\mathrm{EE}}{\varepsilon t_a}\left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}\right], & t_e < t \leq t_c \end{cases} \tag{3}$$

Where ta is related to the return phase of the model and tp is the positive peak of the flux also defined by the zero point before the derivative of the flux and α, ε the parameters controlling the shape of the model [20].

The pulse form of this model can also be presented with other parameters.

- The R parameters

In the section, a set of parameters was determined and are expressed in quotients for to describe the shape of the glottal source signal E (t) and characterize the shape of the pulse of the model LF.

The Rk parameter is defined by (4), it is the measure of the asymmetry in the cycle, Rg represents the normalizer of the fundamental frequency presented in (5) and Rd in (6) to capture all the variation of the model LF It is related to the fundamental frequency and the other two parameters Rg and Rk.

$$R_k = \frac{t_e - t_p}{t_p} \tag{4}$$

$$R_g = \frac{T}{2T_p} \tag{5}$$

$$R_d = 1000\left(\frac{A}{E_e}\frac{f0}{110}\right) \tag{6}$$

The interval between the beginning of the glottal impulse and the instant where tp reaches the maximum is called the opening phase, see (2). At this moment, the vocal folds begin to close and the amplitude of the flow begins to decrease until the sudden closure of the globe. The time that corresponds to the duration of passage of the area flux through the globe when the folds are open is known by the open phase is measured by the following formula te –Ta. With Ta the point where the tangent to the exponential in t = te, it also serves to determine the phase of transition between the open phase and the closed phase known by the return phase is calculated by the following formula [21]:

$$T_a\ = t_a - t_e \tag{7}$$

Finally, the last phase of the cycle when the folds are completely closed at time tc. The model can also be presented by other parameters, which are correlated with the quality of the voice and its spectral properties as in [22].

- The parameters NAQ and QOQ

Several temporal characteristics can be determined as a function of time derived from the glottal form [23], using the model LF and its parameters such as the open quotient Qq, the asymmetry coefficient αm and the speed quotient Sq. These moments are always difficult to detect correctly. To solve this problem a new parameter has been proposed by Hacki [24] describing the open time of the glottis named quasi-open quotient (QOQ). It is defined by the ratio between the time of opening and closing of the glottis and corresponds to the period during which the flow of the glottis is greater than 50% of the difference of two maximum and minimum flows. Laukkanen [25] has to experience that this parameter can be used to study the variation of the glottal source in case of stress and emotion as well as Airas and alku which have announce that this quotient and the best parameter to use to test the degrees of reflection of the changes of phonation [26].

The temporal characteristics of the glottal source can also be presented by measuring the amplitude of the glottis and its derivative.

$$NAQ = \frac{f_{ac}}{d_{peak}.T0} \tag{8}$$

It is the normalized amplitude quotient (NAQ) proposed by Alku et al [27] using two amplitudes f_ac the amplitude of flow and d_peak the negative peak amplitude (see Fig. 3).

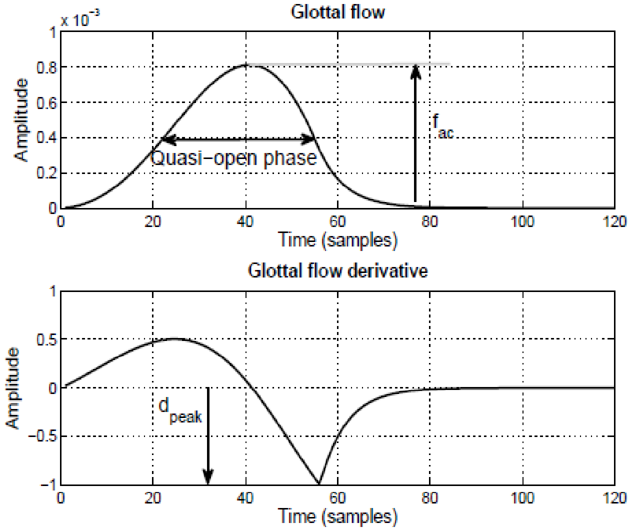


Fig. 3. Parameters needed for NAQ and QOQ measurement.

III. Experimental Details

The aim of this article is to determine whether the parameters of the glottal signal have remarkable characteristics between the two databases PD and HC. To evaluate this, we used Matlab to measure its various parameters and anova for statistical measurements.

A. Database

In this study we used, recording of 36 native Czech speakers, the PD age group 64.21 ± 9.46 SD (extreme 41-82). Before the experience each patient did a neurological examination to rank them according to the scale of which varies from 1 to 5 according to the degree of unilateral motor disorder as well as the classification scale (UPDRS III) also The H&Y score was 2.1 ± 0.4(1–3),. The healthy control group (HC) with a mean age of 64.21 ± 9.22 (42-80) years. None of the HC group participants had a history of neurological disorders or speech. The study was approved by the Ethics Committee of the General University Hospital in Prague, Czech Republic, and all participants provided informed and written consent.

The description of some records used in this document is displayed in the following Table I:

TABLE I. Some Characteristic of Parkinson's Patients

Participant code	Gender	Age (years)	Age of disease onset (years)	Duration of disease from first symptoms (years)	UPDRS III total	UPDRS III 18. Speech
PN301	M	74	64	10	22	2
PN302	M	77	70	7	14	1
PN304	M	76	70	6	23	1
PN307	M	66	61	5	-	-
PN314	F	62	57	5	32	0
PN316	F	61	57	4	38	2
PN401	M	56	55	1	8	0
PN403	M	68	59	9	12	0
PN405	M	76	65	11	27	2
PN407	M	55	44	11	16	0
PN409	M	79	70	9	9	0

B. Recording

The recordings of the speakers are recorded in a quiet room with a professional microphone (Beyer-Dynamics Opus 55, Heilbronn, Germany) placed 5 cm from the mouth of each patient. All of his recordings were digitized at a sampling rate of 48 Khz with 16-bit quantization. All speakers pronounced the vowel a under the control of a speech specialist without any time limit was imposed during the recording.

IV. Results

Glottic pulse parameters related to time and frequency provides quantitative information for the examiner about their importance in biomedical applications. To prove it, we used both types of recording of PD and HC speakers, regardless of age or gender. The results for each glottal parameter are shown in the following Table II. This table presents the mean value and standard deviation calculated for all patients in two PD and HC groups. A remarkable difference can be observed in the different parameters.

The analysis of group differences between PD and controls performed using a rank sum of Wilcoxon. It is a non-parametric statistical test that uses the data distribution assumption the same with the Spearman correlation reliability test that is used if the static correlation variables do not have affine relationships, see Table III.

These parameters were also tested following the clinical manifestations with a significant correlation between the NAQ parameter and the UPDRS discourse 18 (Z = 0.36, p <0.04) (Table IV).

TABLE II. THE RESULTS OF MEASURING THE MEAN AND THE STANDARD DEVIATION OF GLOTTAL PARAMETERS FOR PD AND HEALTHY SPEECH

	Results			
	PD speech		**HC speech**	
	Average	**Standard Deviation**	**Average**	**Standard Deviation**
Ee	0,0566	0,0926	-0,1447	0,1393
Te	0,0027	0,0029	0,0043	0,0149
Tp	0,0025	0,0027	-0,0256	0,0999
Rd	0,0140	0,0179	0,0391	0,0667
Rg	0,1730	0,1383	0,2971	0,3193
Rk	1,1767	1,3183	1,2827	3,3212

TABLE III. THE RESULTS STATISTICAL DIFFERENCES BETWEEN ALL GROUPS

Variable	**Spearman correlation**			**Wilcoxon rank sum test**	
	PD vs HC			PD vs HC	
	Z	**p**		**Z**	**p**
NAQ	0,81	p< 0.001		-2,19	p < 0,05
QOQ	0,78	p< 0.001		-1,84	p = 0,06

TABLE IV. THE CLINICAL MANIFESTATIONS WITH NAQ ET QOQ

Variable						
	UPDRS		UPDRS speech		UPDRS rigidity	
	Z	**p**	**Z**	**p**	**Z**	**p**
NAQ	0,24	0,17	0;36	0,04	-2,19	0,96
QOQ	-0,17	0,93	0,96	0,98	0,078	0,66

V. DISCUSSION AND CONCLUSION

This work, we presented a glottal analysis based on the estimation of flux and the extraction of some parameters in the time domain in order to examine their relations in the two bases PD and HC. The results found are based on several direct correlations to demonstrate voice quality and glottal velocity. A significant effect was found with the NAQ parameter through a positive correlation describing the amplitude variation during pronunciation and reflecting the distinctive effect of Bradykinesia in Parkinson's disease.

Another presented part of our paper is the analysis of glottal source local relations and the clinical manifestations described by the UPDRS speech scale. Our results show that the NAQ parameter confirms the intelligibility of the manifestation.

There is a limit in this analysis, in particular that our results are based on a relatively small number of PD patients who are heterogeneous in terms of age, duration of illness and gender of the patient.

REFERENCES

[1] Kostikis, N., et al. "A smartphone-based tool for assessing parkinsonian hand tremor." IEEE journal of biomedical and health informatics 19.6 (2015): 1835-1842.

[2] Rhouma, A. Ben, and S. Ben Jebara. "Features based on quasi-sinudoidal modeling for tremor detection in Parkinsonian voice." Advanced Technologies for Signal and Image Processing (ATSIP), 2014 1st International Conference on. IEEE, 2014.

[3] Nussbaum, Robert L., and Christopher E. Ellis. "Alzheimer's disease and Parkinson's disease." New england journal of medicine 348.14 (2003): 1356-1364.

[4] Liu, Yao, et al. "The analysis of regularity and synchrony of Parkinsonian tremor using Approximate Entropy and Cross-Approximate Entropy." Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), International Congress on. IEEE, 2016.

[5] Liu, Chen, et al. "Closed-Loop Control of Tremor-Predominant Parkinsonian State Based on Parameter Estimation." IEEE Transactions on Neural Systems and Rehabilitation Engineering 24.10 (2016): 1109-1121.

[6] Ho, Aileen K., et al. "Speech impairment in a large sample of patients with Parkinson's disease." Behavioural neurology 11.3 (1999): 131-137.

[7] Jones, Harrison N., et al. "Speech motor program maintenance, but not switching, is enhanced by left-hemispheric deep brain stimulation in Parkinson's disease." International journal of speech-language pathology 12.5 (2010): 385-398.

[8] Rusz, J., et al. "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease." The journal of the Acoustical Society of America 129.1 (2011): 350-367.

[9] Bocklet, Tobias, et al. "Detection of persons with Parkinson's disease by acoustic, vocal, and prosodic analysis." Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on. IEEE, 2011.

[10] Hanson, Helen M., and Erika S. Chuang. "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data." The Journal of the Acoustical Society of America 106.2 (1999): 1064-1077.

[11] Kreiman, Jody, Bruce R. Gerratt, and Norma Antonanzas-Barroso. "Measures of the glottal source spectrum." Journal of speech, language, and hearing research 50.3 (2007): 595-610.

[12] Li, Haoxuan, Ronan Scaife, and Darragh O'Brien. "LF model based glottal source parameter estimation by extended Kalman filtering." Proceedings of the 22nd IET Irish Signals and Systems Conference. 2011.

[13] Drugman, Thomas. "Advances in glottal analysis and its applications." University of Mons, Belgium (2011).

[14] Drugman, Thomas, et al. "Glottal source processing: From analysis to applications." Computer Speech & Language 28.5 (2014): 1117-1138.

[15] Fant, G. (1970). Accoustic theory of speech production: with calculations based on X-ray studies of Russian articulations. Mouton & Company.

[16] Rabiner, L. R., & Schafer, R. W. (1978). Digital processing of speech signals (Vol. 100, p. 17). Englewood Cliffs, NJ: Prentice-hall.

[17] Hamlet, S. M. (2012). Comparing Acoustic Glottal Feature Extraction Methods with Simultaneously Recorded High-speed Video Features for Clinically Obtained Data.

[18] Pelorson, X., Hirschberg, A., Van Hassel, R. R., Wijnands, A. P. J., & Auregan, Y. (1994). Theoretical and experimental study of quasi steady flow separation within the glottis during phonation. Application to a modified two mass model. The Journal of the Acoustical Society of America, 96(6), 3416-3431.

[19] Fant, G., Liljencrants, J., & Lin, Q. G. (1985). A four-parameter model of glottal flow. STL-QPSR, 4(1985), 1-13.

[20] Doval, Boris, Christophe d'Alessandro, and Nathalie Henrich. "The spectrum of glottal flow models." Acta acustica united with acustica 92.6 (2006): 1026-1046.

[21] Childers, Donald G. "Glottal source modeling for voice conversion." Speech communication 16.2 (1995): 127-138.

[22] Alku, Paavo. "An automatic method to estimate the time-based parameters of the glottal pulseform." Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on. Vol. 2. IEEE, 1992.

[23] Alku, P. (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech communication, 11(2-3), 109-118.

[24] Hacki, T. (1989). Klassifizierung von glottisdysfunktionen mit hilfe der elektroglottographie. Folia Phoniatrica et Logopaedica, 41(1), 43-48.

[25] Laukkanen, A. M., Vilkman, E., Alku, P., & Oksanen, H. (1996). Physical variations related to stress and emotional state: a preliminary study. Journal of Phonetics, 24(3), 313-335.

[26] Airas, M., & Alku, P. (2007). Comparison of multiple voice source parameters in different phonation types. In Eighth Annual Conference of the International Speech Communication Association.

[27] Alku, P., Bäckström, T., & Vilkman, E. (2002). Normalized amplitude quotient for parametrization of the glottal flow. the Journal of the Acoustical Society of America, 112(2), 701-710.